

DESCRIPTIVE STATISTICS

BIOS 662

Michael G. Hudgens, Ph.D.

mhudgens@bios.unc.edu

<http://www.bios.unc.edu/~mjudgens>

2008-08-19 08:51

Descriptive Statistics

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

Types of Variables

- A *variable* is a quantity that may vary from object to object
- A *sample* or *data set* is a collection of values of one or more variables.
- Types of variables
 - Quantitative variable intrinsically numerical
e.g. age, height, counts
 - Qualitative (categorical) - intrinsically nonnumerical
e.g. gender, province, country

Types of Variables

- Qualitative (categorical) - intrinsically nonnumerical
 - 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, numerical values may be assigned

Descriptive Statistics

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

Measures of Location

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

Arithmetic mean

- Data:

$$x_1, x_2, \dots, x_n$$

- Mean:

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

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 Mean

- Let c be any constant

- If

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

then

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

- If

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

then

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

Properties of 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 ozs. is

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

Order statistics

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

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

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

$$x_{(1)} = \min\{x_1, x_2, \dots, x_n\}$$

$$x_{(n)} = \max\{x_1, x_2, \dots, x_n\}$$

Example

- Duration of hospital stay in days:

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

- Order statistics:

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

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 $(p \times 100)^{th}$ percentile of a sample

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

for $0 < p < 1$

- Note: $\lfloor y \rfloor$ is the greatest integer $\leq y$; i.e., the *floor* function

$\lceil y \rceil$ is the smallest integer $\geq y$; i.e., the *ceiling* function

- Cf Def 3.11 of text

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 $m = p$ and

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

then $j = \lfloor pn + p \rfloor$ and we recover text definition

Percentiles: Software

- SAS Proc Univariate: 5 definitions of percentile
- R: 9 definitions
- Claim: none of these match the book definition

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

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

for $0 < p < 1$

- Definition 2 of R/Hyndman and Fan: $m = 0$ 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

$$np = 278 \times .75 = 208.5$$

such that

$$\hat{\zeta}_{.75} = x_{(209)}$$

- R

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

Example: SAS

```
data;  
  infile "H:/WWW/bios/662/2007fall/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

$$\hat{\zeta}_{.5} = \begin{cases} y_{(\frac{n+1}{2})} & \text{if } n \text{ is odd} \\ \{y_{(n/2)} + y_{(n/2+1)}\}/2 & \text{if } n \text{ is even} \end{cases}$$

for $0 < p < 1$

Example

- Duration of hospital stay in days:

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

- Median:

$$\hat{\zeta}_{.5} = \{x_{(2)} + x_{(3)}\}/2 = (6 + 10)/2 = 8$$

Mode

- The mode is the most frequently occurring value in the data set
- E.g., if

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

then mode is 11

Geometric Mean

- Data: x_1, x_2, \dots, x_n

- The geometric mean of x is

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

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

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

- \bar{x}_g is used when data are of the form c^k

- Eg, suppose $x_1 = 10$ and $x_2 = 0.1$. Then $\bar{x}_g = 1$

Comments

- Mean is most often used measure
- Median is better if there are influential observations (more robust to extreme values)
- Mode rarely used (exception: nominal data)

Example

- Duration of hospital stay in days:

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

$$\hat{\zeta}_{.5} = \bar{x} = 8, \bar{x}_g = 7.6$$

- Alter last observation:

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

$$\hat{\zeta}_{.5} = 8, \bar{x} = 17.7, \bar{x}_g = 11.1$$

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 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
- Report 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 $\bar{x} \pm s$ will contain approx 68% of the observations
- The interval $\bar{x} \pm 2s$ will contain approx 95% of the observations
- Approx s by

$$s \approx \frac{\hat{\zeta}_{.75} - \hat{\zeta}_{.25}}{1.35}$$

- Note

$$\hat{\zeta}_{.75} - \hat{\zeta}_{.25}$$

is called *interquartile range*

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

Skewness

- Define r th sample moment about the mean

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

- 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 page 51
- SAS Proc Univariate VARDEF=N

Interpretation?

- Text:

“skewed to the right if mean is greater than mode”

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

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

$\bar{x} = 2.2$, mode equals 2, and skewness equals -0.37

Alternative Definitions

- Another definition of skewness:

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

- 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 page 51

Kurtosis: SAS

- Proc Univariate VARDEF=N

$$a_4 = \frac{1}{n} \sum \left(\frac{y_i - \bar{y}}{s} \right)^4 - 3$$

i.e.,

$$a_4 = \frac{\sum (y_i - \bar{y})^4 / n}{s^4} - 3$$

i.e.,

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

- Why minus 3?

Descriptive Statistics

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

Data display

- Simplest form is a line listing
- 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	1st	2nd
< 106	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 example of *empirical frequency distribution*
- Difficult to compare blood pressure distributions due to different sample sizes
- Divide by sample size to get *empirical relative frequency distribution*

ERFD - Example: Table 3.7

Blood Pressure	Native	1st	2nd
< 106	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

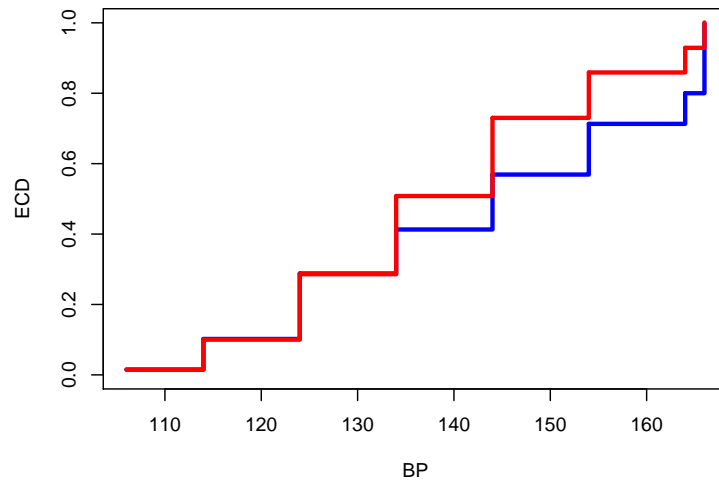
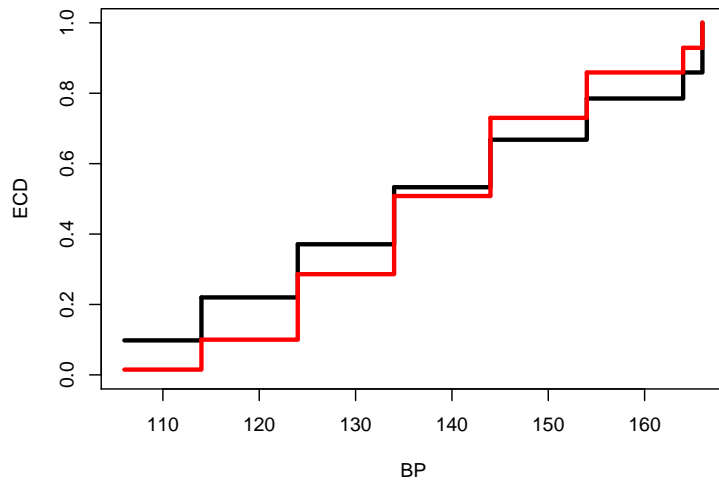
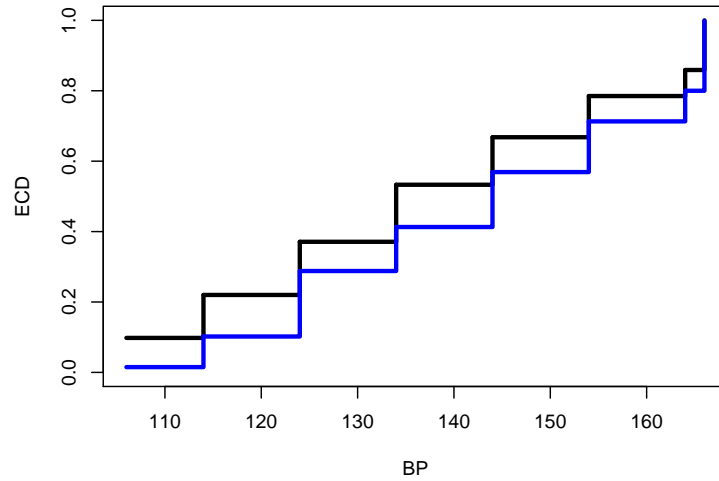
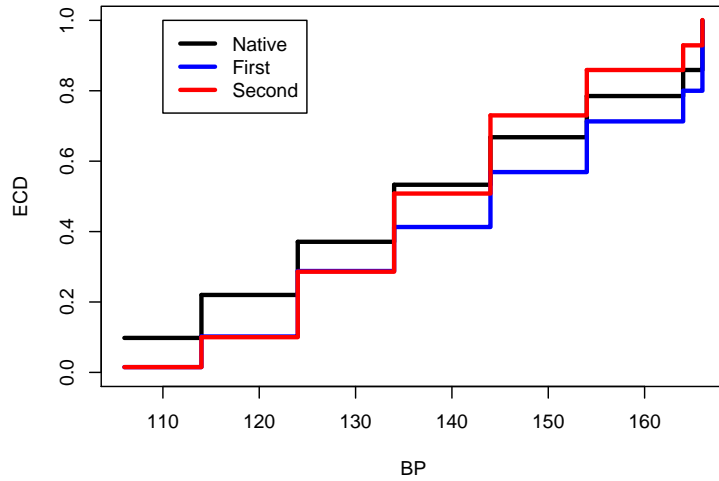
Empirical distribution function

- Def 3.9 The *empirical cumulative distribution* of a variable is a listing of the variable with the proportion of observations less than or equal to that value (cumulative proportion)
- Aka *empirical distribution function* (EDF)
- Does not necessarily entail binning

ECD - Example

Blood Pressure	Native	1st	2nd
< 106	0.098	0.015	0.015
≤ 114	0.220	0.102	0.100
≤ 124	0.371	0.288	0.286
≤ 134	0.533	0.413	0.508
≤ 144	0.668	0.569	0.730
≤ 154	0.785	0.713	0.859
≤ 164	0.859	0.800	0.929
$< \infty$	1.000	0.998	1.001
Total	2232	263	1561

ECD - Example



Graphs

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

Histogram

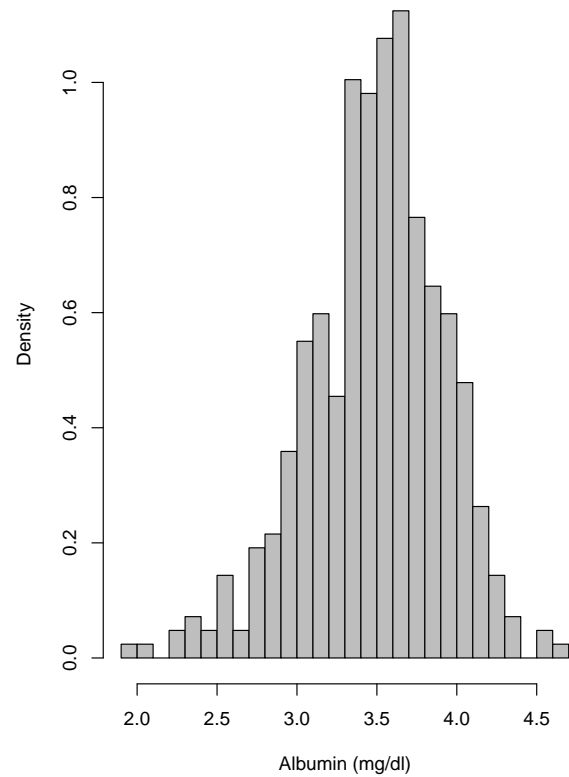
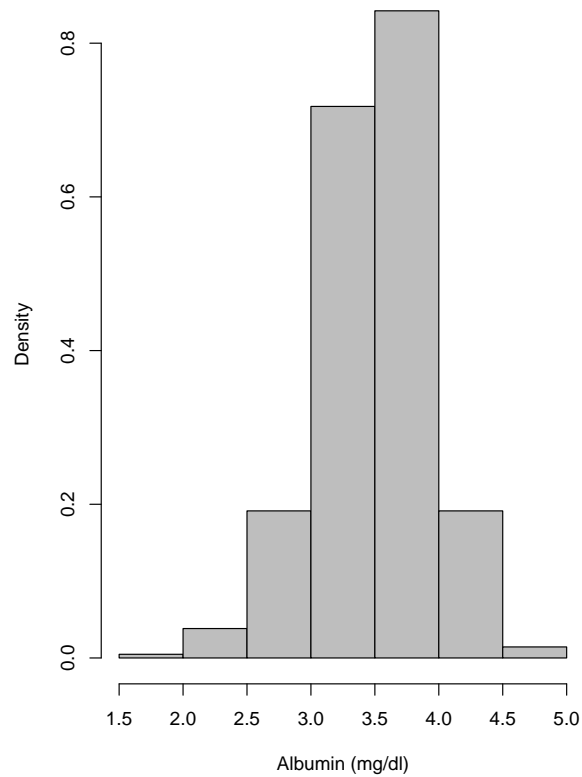
- Data are divided into intervals as in a frequency table
- A histogram is a bar graph with the area of each bar equal to the relative frequency in the interval.
- Can compare histograms from samples of different size
- Intervals need not be the same width
- Beware effect of choice of interval width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

Histogram: Example (Fig 3.1 text)

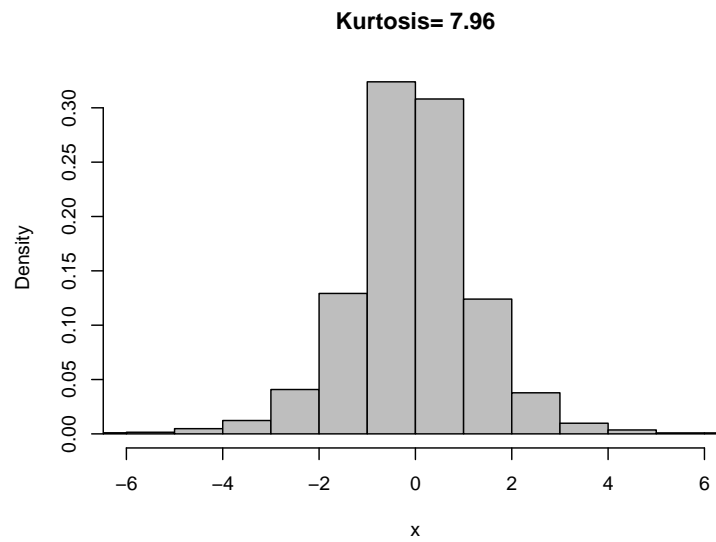
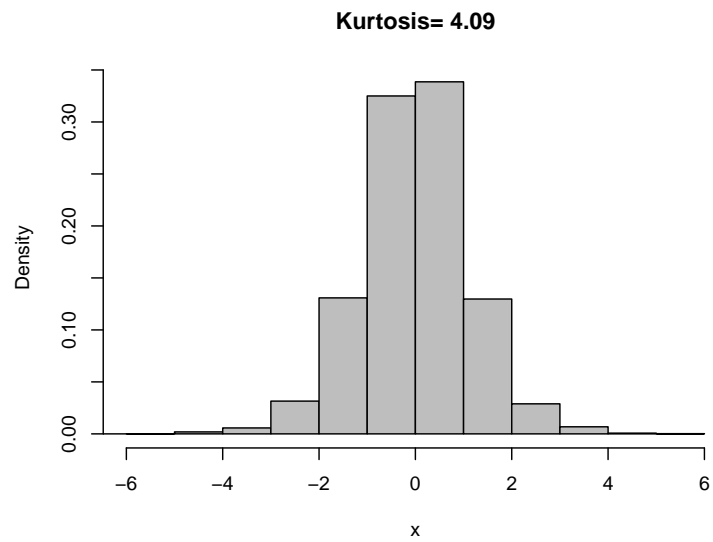
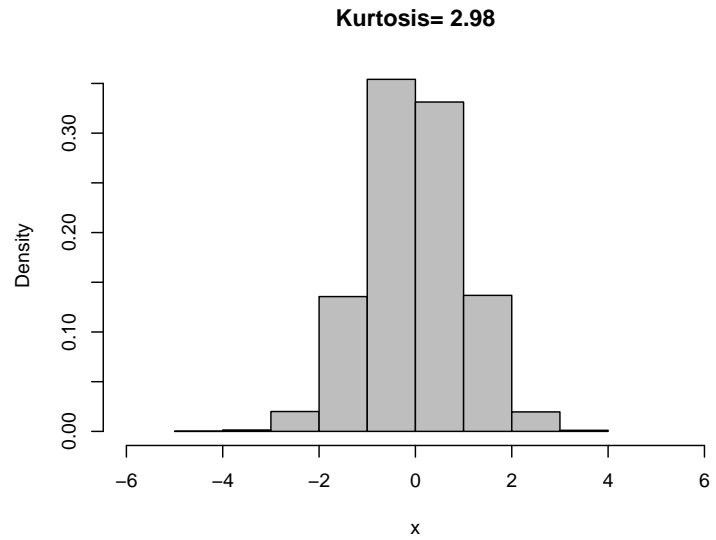
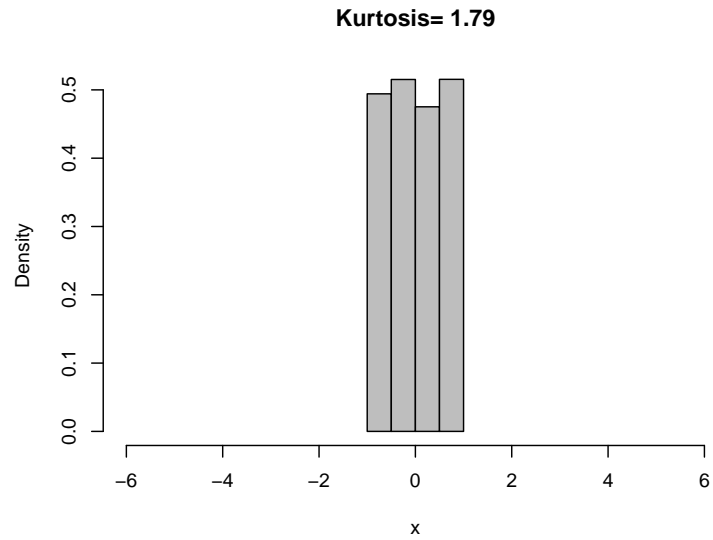
```
> par(mfcol=c(1,2))
```

```
> 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="")
```



Sample Kurtosis



Stem and Leaf Plot

- Stem consists of leading digits
- Leaves consist of last digit
- Example: $x=496$, 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 | 00001111223334456667778880011222234455567889999
32 | 000001223334456666699901111233444445555555556666777789
34 | 000000011122222333334445566667888889990000000011112222333444455555+5
36 | 00000000111122233333344445555555666677778889999900000002233344445556+2
38 | 0000011122333333445555567799900012233334445567788889999
40 | 00111334467888889999003456678999
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 | 00001111223334456667778880011222234455567889999
32 | 0000012233344566666999011112334444455555555566666777789
34 | 000000011122222333334445566666788888999000000011112222333444455556666667777788889
36 | 000000001111222333333444455555566667777888999990000000223334444555666666777778999
38 | 00000111223333334445555567799900012233334445567788889999
40 | 00111334467888889999003456678999
42 | 022340088
44 | 022
46 | 4
```

```
> stem(liver$albumin,scale=2)
```

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 | 0000111122333445666777888
31 | 0011222234455567889999
32 | 00000122333445666666999
33 | 0111123344444455555555566666777789
34 | 0000000111222233333444455666667888888999
35 | 00000000111122233334444555556666667777788889
36 | 000000001111222333333444455555566666777788899999
37 | 0000000223334444555666666777778999
38 | 00000111223333334555555677999
39 | 00012233334445567788889999
40 | 001113344467888889999
41 | 003456678999
42 | 02234
43 | 0088
44 | 0
45 | 22
46 | 4
```

Stem and Leaf

- Advantage: visualize all (or almost all) of the data
- Disadvantage: loss of ordering of data set

Box plot

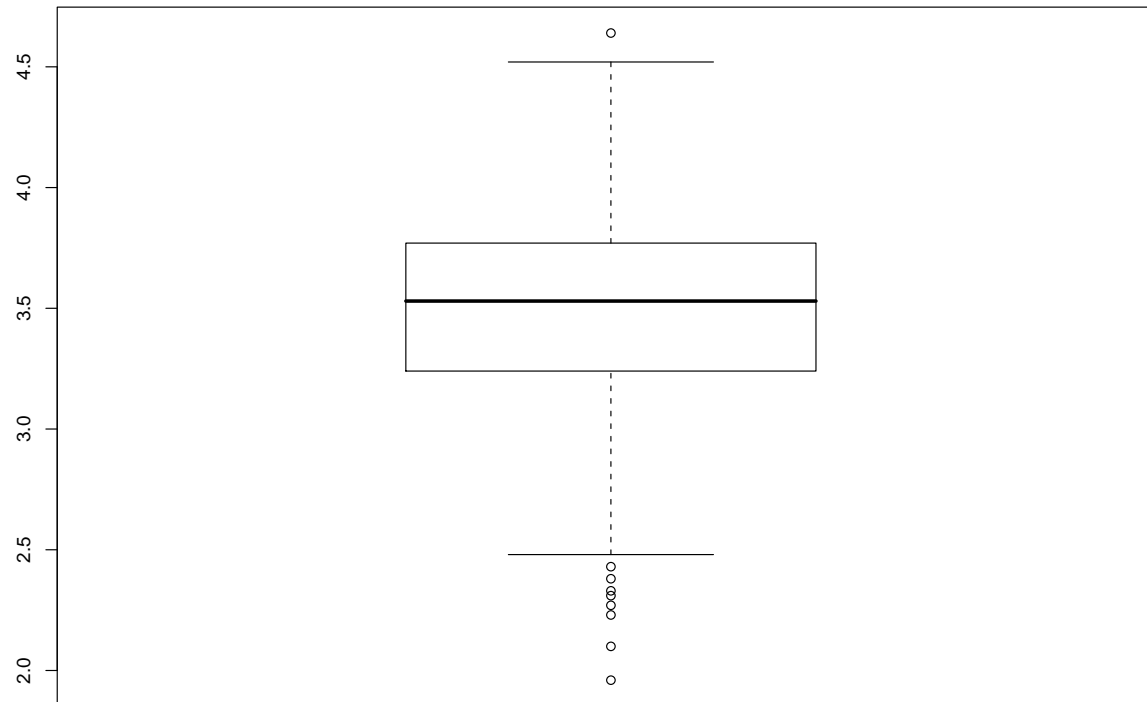
- The top of the box is the 75th percentile ($\hat{\zeta}_{.75}$); the bottom is the 25th percentile ($\hat{\zeta}_{.25}$)
- A line through the box is drawn at the median

Box plot

- The lines extending out of the box (*whiskers*) may extend to
 - the 90th and 10th percentiles
 - the largest and smallest values
 - largest observation $\leq \hat{\zeta}_{.75} + 1.5 \times \text{IQR}$;
smallest observation $\geq \hat{\zeta}_{.25} - 1.5 \times \text{IQR}$
(text is wrong! cf Tukey 1977, Chambers et al 1983)
- Data beyond whiskers may be plotted individually

Box plot: Example

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



Box plot

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

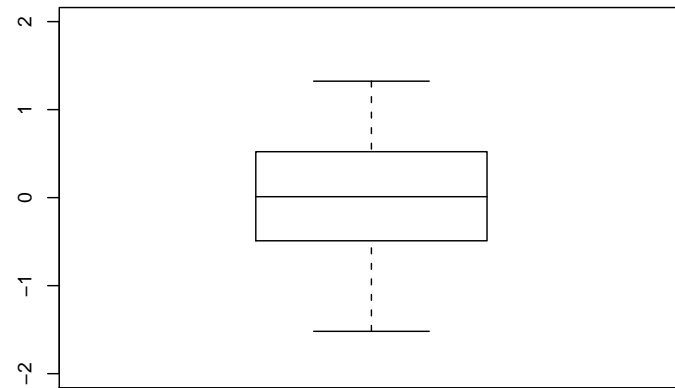
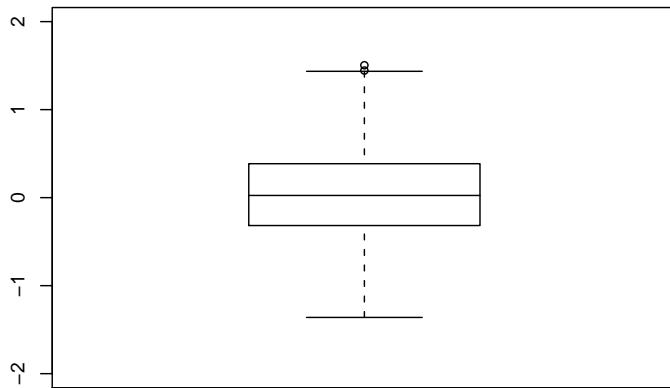
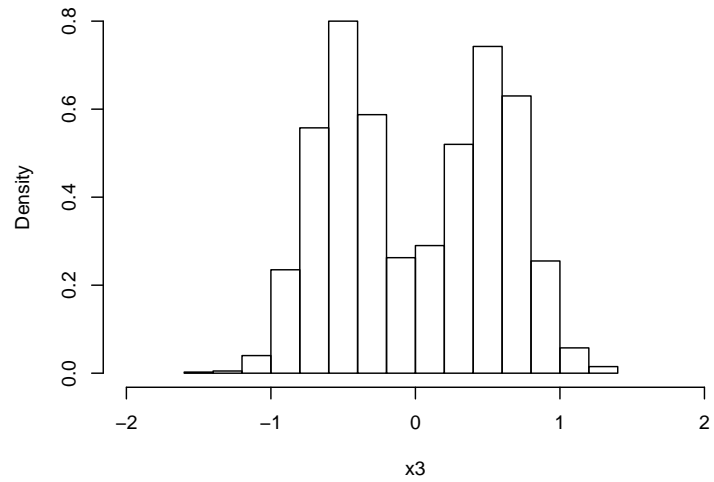
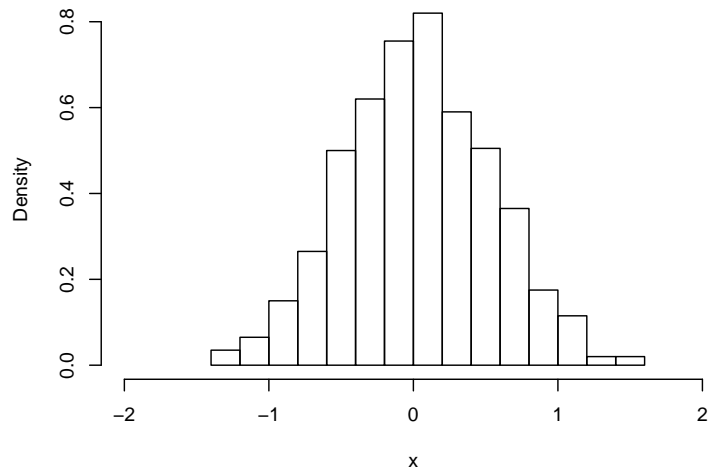
- Note

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

so whiskers cover

$$\approx \hat{\zeta}_{.5} \pm 2.68s$$

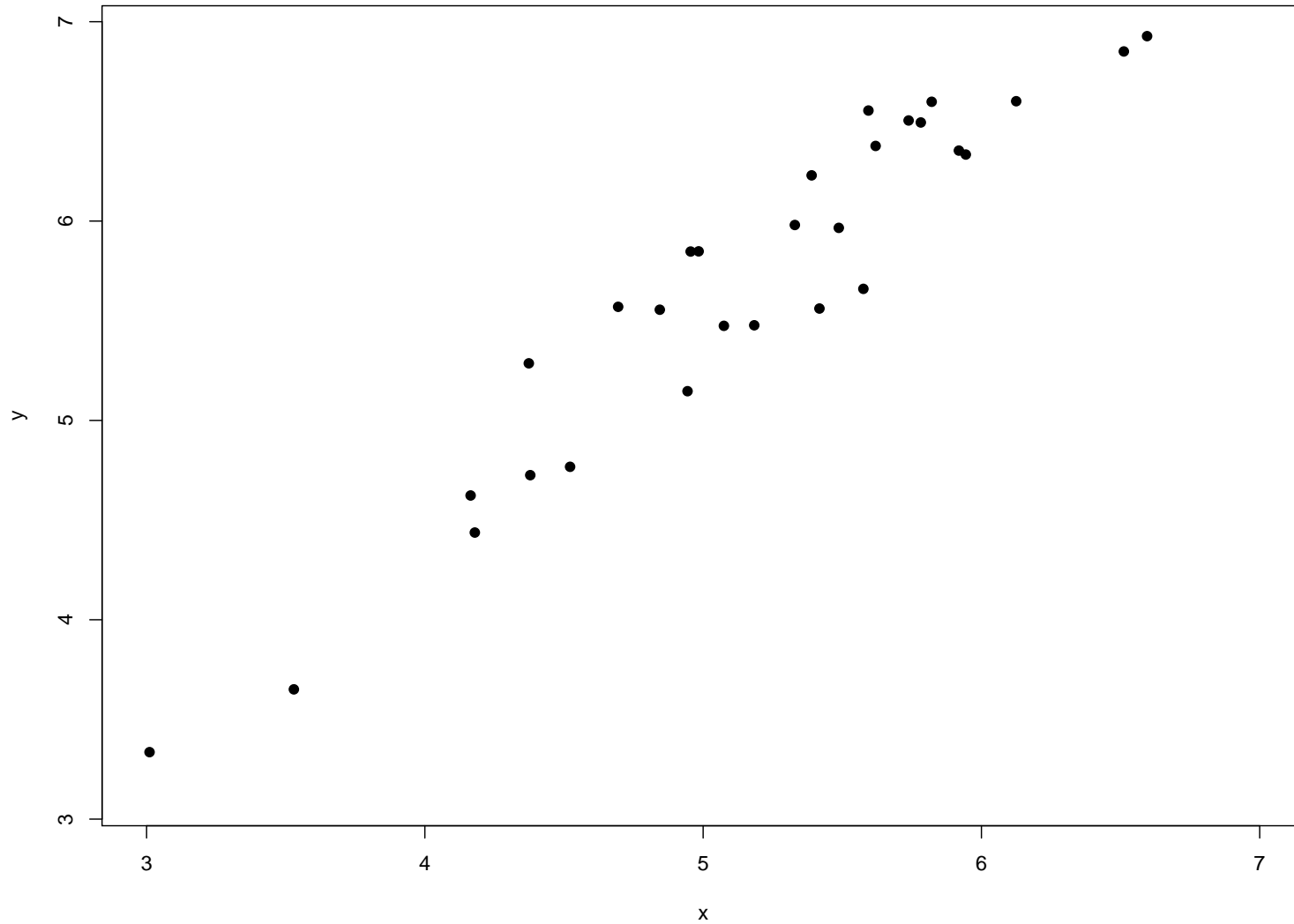
Box plot and Histogram Example



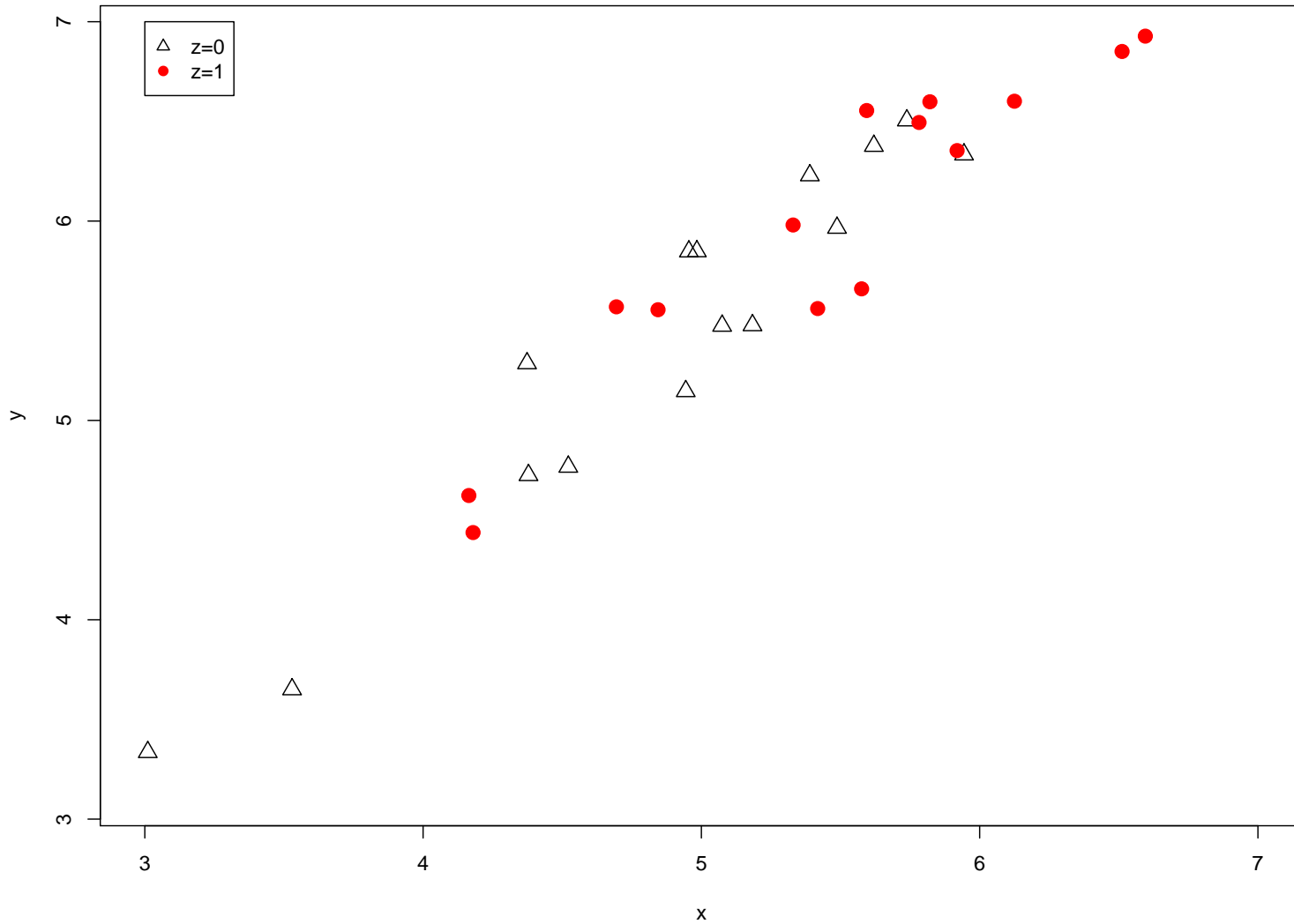
Multivariate plots

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

Scatterplot Example I



Scatterplot Example II



Trellis plots

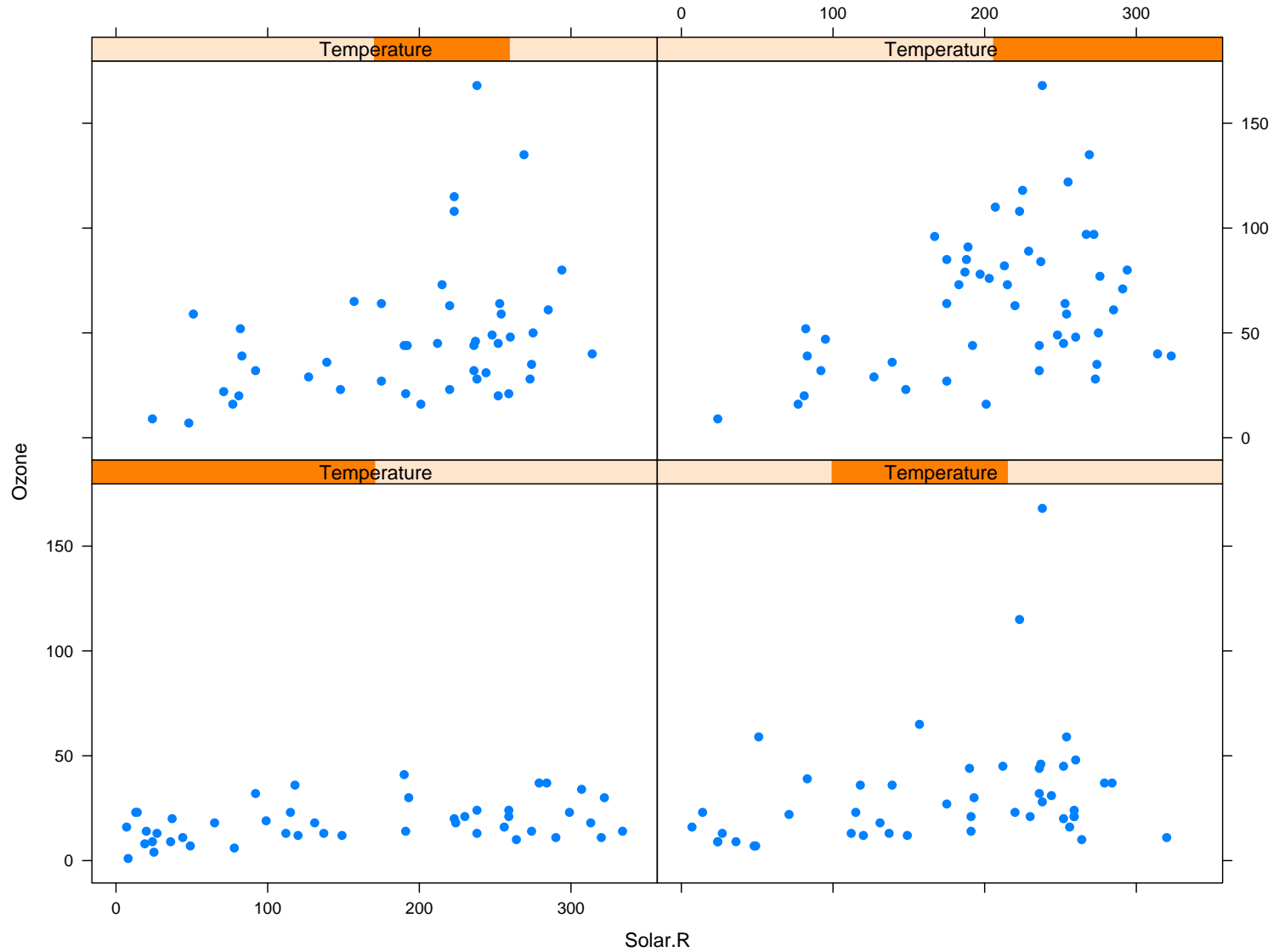


Table or graph?

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