Chapter 4: Summarizing & Exploring Data (Descriptive Statistics)

Graphics! Graphics! Graphics! (and some numbers)

Slides prepared by Elizabeth Newton (MIT) with some slides by Jacqueline Telford (Johns Hopkins University) and Roy Welsch (MIT).

Graphical Excellence

"Complex ideas communicated with clarity, precision, and efficiency"

- Shows the data
- Makes you think about substance rather than method, graphic design, or something else
- Many numbers in a small space
- Makes large data sets coherent
- Encourages the eye to compare different pieces of the data

Charles Joseph Minard

Graphic Depicting Exports of Wine
from France (1864)

Available at
http://www.math.yorku.ca/SCS/Gallery/

Source: Minard, C. J. *Carte figurative et approximative des quantités de vin français exportés par mer en 1864*. 1865. ENPC (École Nationale des Ponts et Chaussées), 1865.
Also available in: Tufte, Edward R. *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press, 2001.

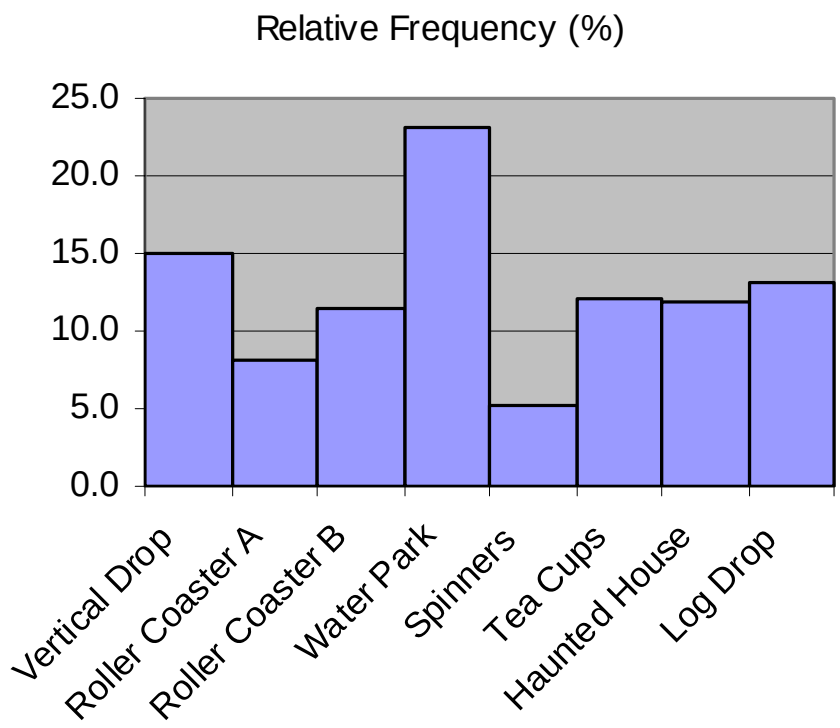
Summarizing Categorical Data

A frequency table shows the number of occurrences of each category. Relative frequency is the proportion of the total in each category.

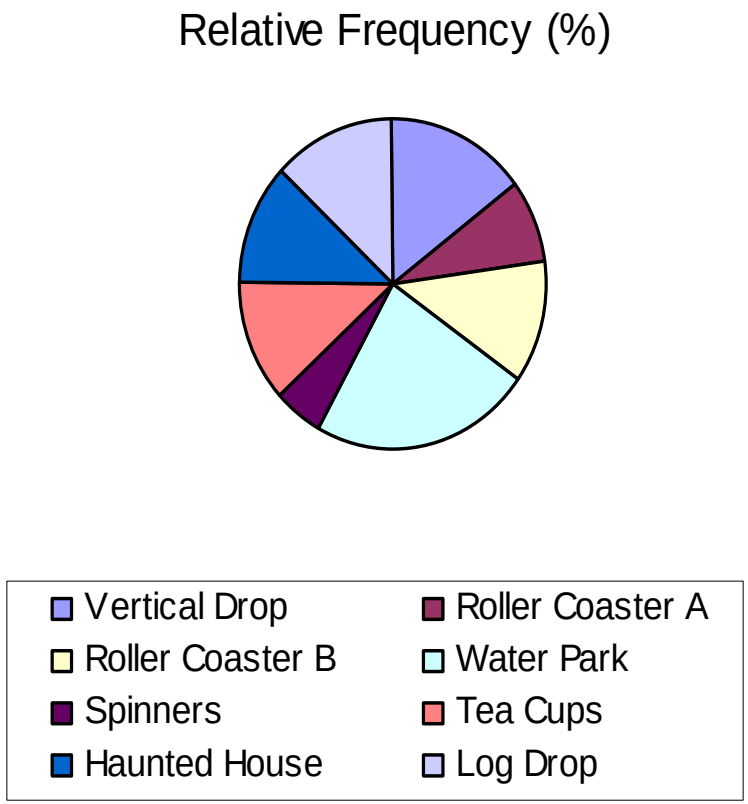
Bar charts and Pie Charts are used to graph categorical data. A Pareto chart is a bar chart with categories arranged from the highest to lowest (QC: "vital few from the trivial many").

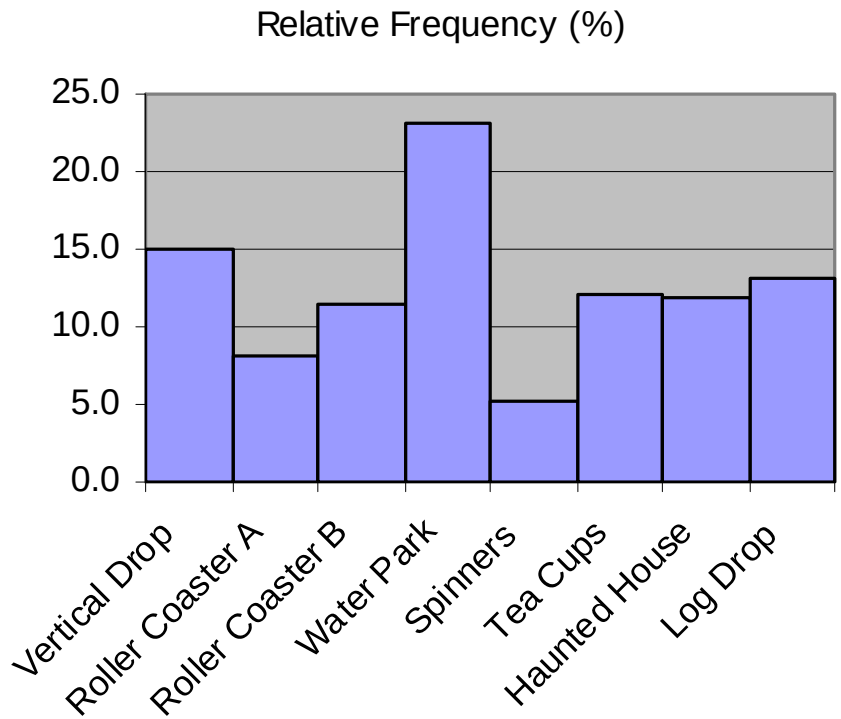
Attraction	Frequency	Relative Frequency (%)
Vertical Drop	101	15.1
Roller Coaster A	54	8.1
Roller Coaster B	77	11.5
Water Park	155	23.1
Spinners	35	5.2
Tea Cups	81	12.1
Haunted House	79	11.8
Log Drop	88	13.1
Total	670	100.0

Popularity of attractions at an amusement park



Pie Chart and Bar Chart of Attraction Popularity at an Amusement Park





Charles Joseph Minard

Graph showing quantities of meat sent from various regions of France to Paris using pie charts overlaid a map of France (1864)

Available at
http://www.math.yorku.ca/SCS/Gallery/

Source: Minard, C. J. *Carte figurative et approximative des quantités de viande de boucherie envoyées sur pied par les départments et consommées à Paris*. ENPC (École Nationale des Ponts et Chaussées),1858, pp. 44.

Plots for Numerical Univariate Data

Scatter plot (vs. observation number)

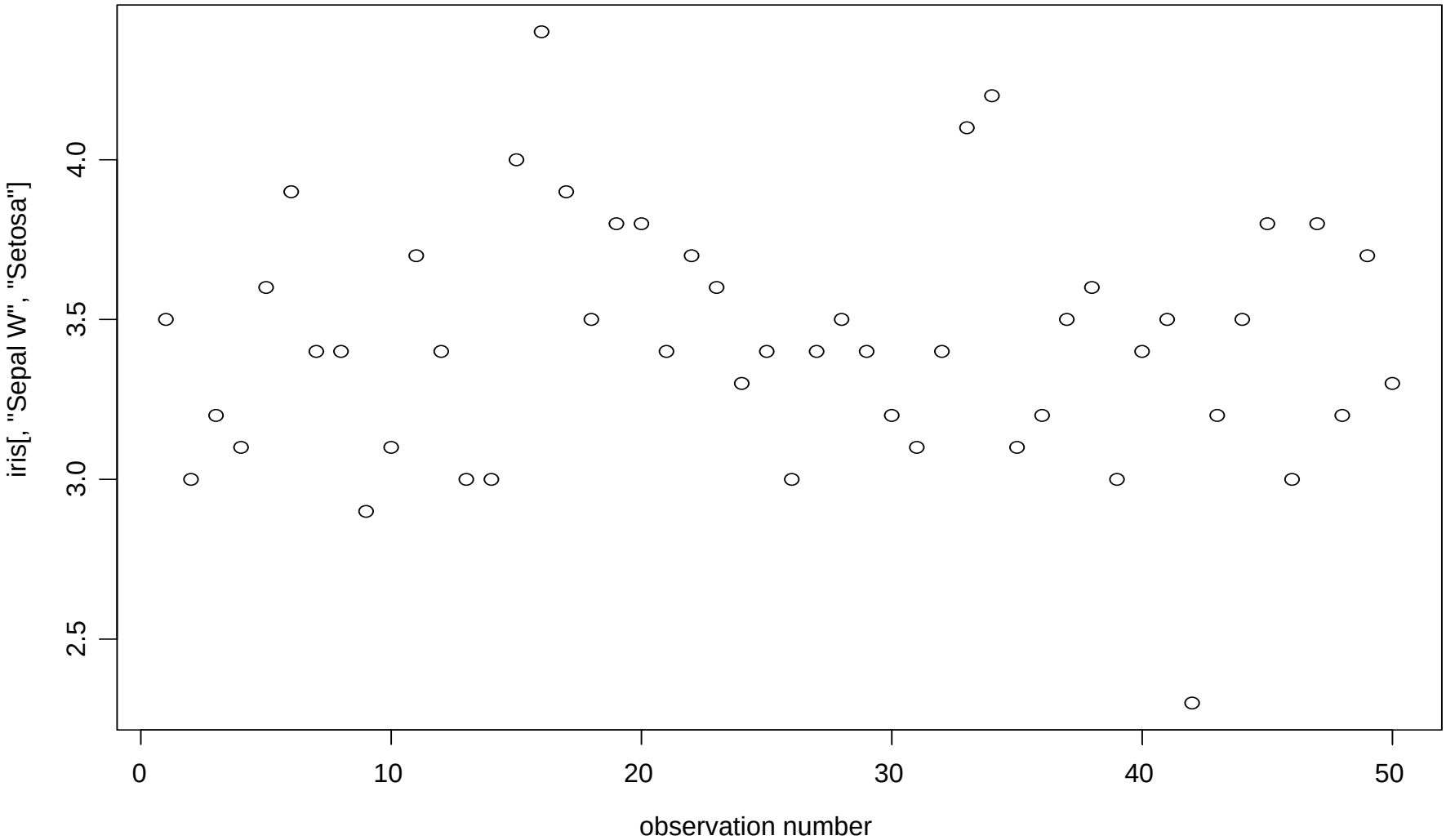
Histogram

Stem and Leaf

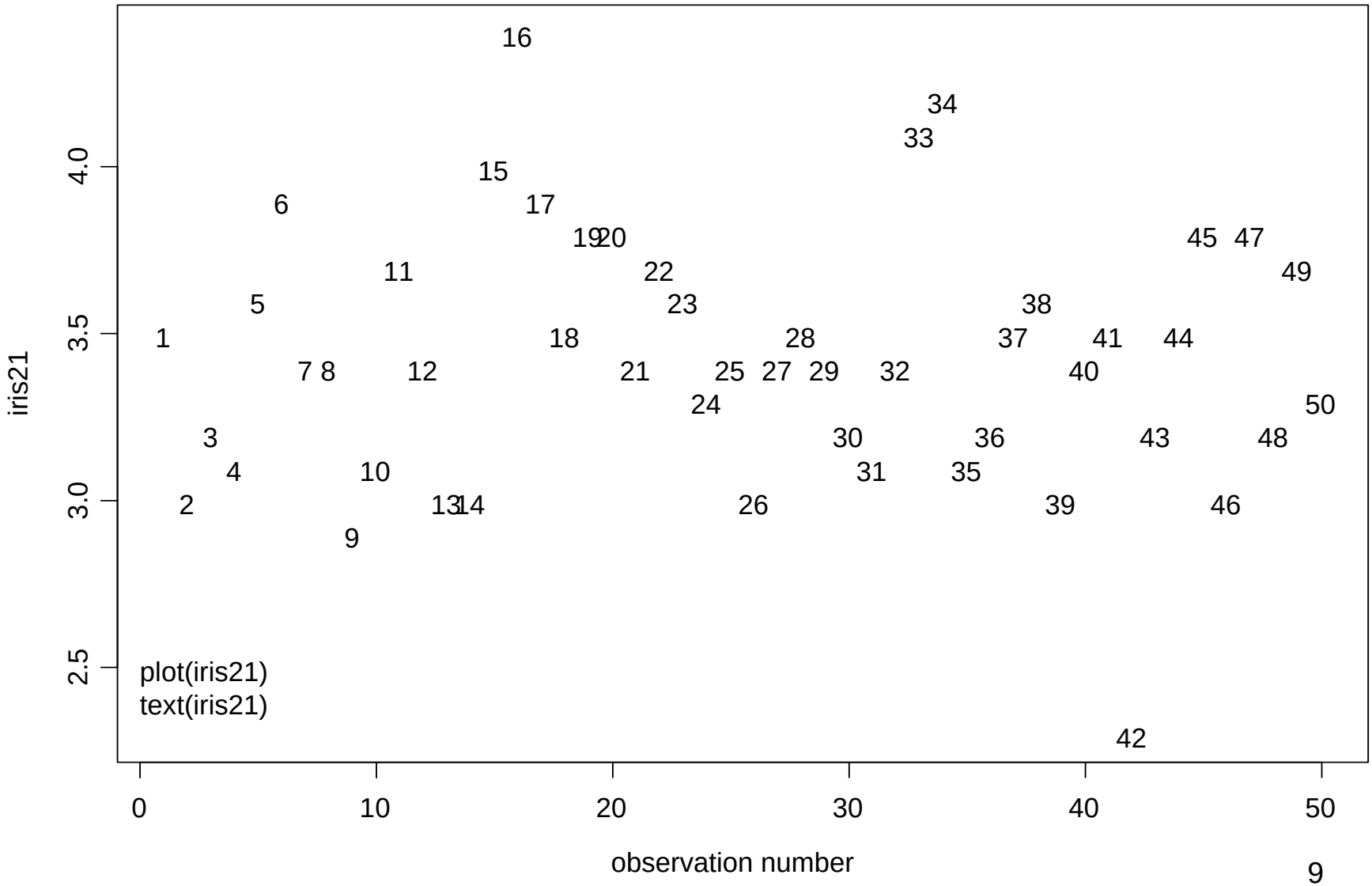
Box Plot (Box and Whiskers)

QQ Plot (Normal probability plot)

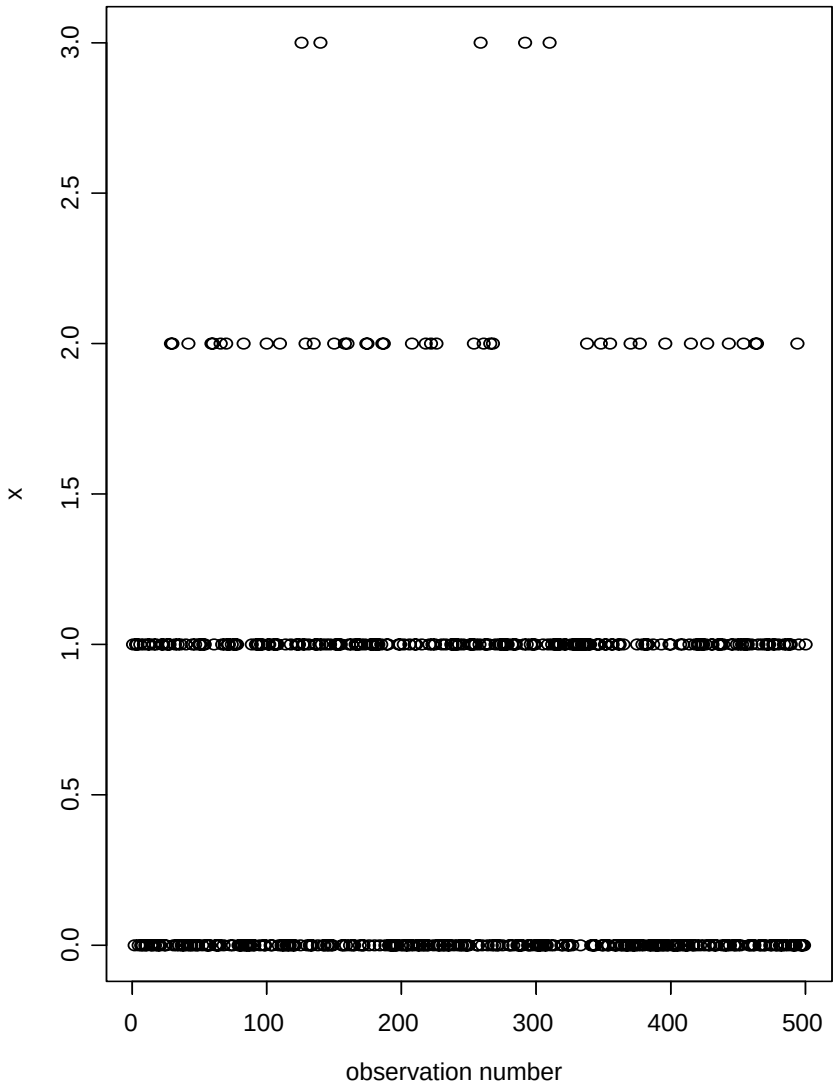
Scatter Plot of Iris Data

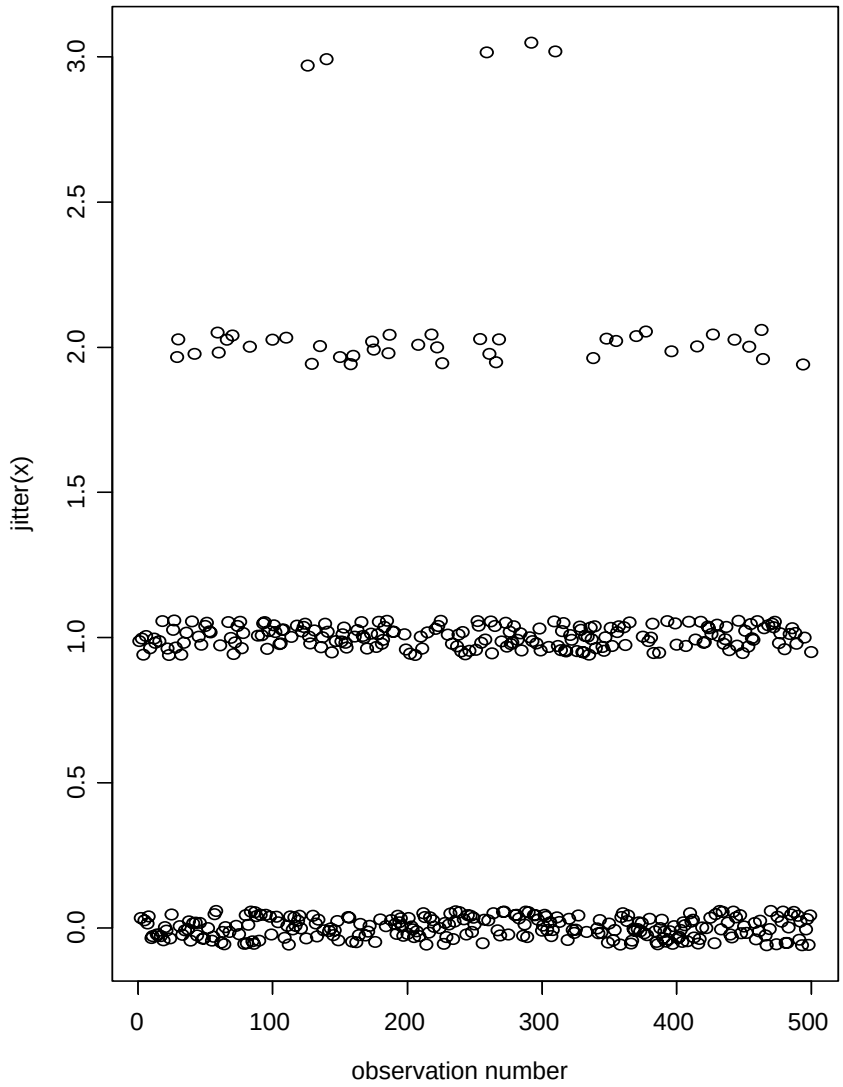


Scatter Plot of Iris Data with Observation Number Indicated



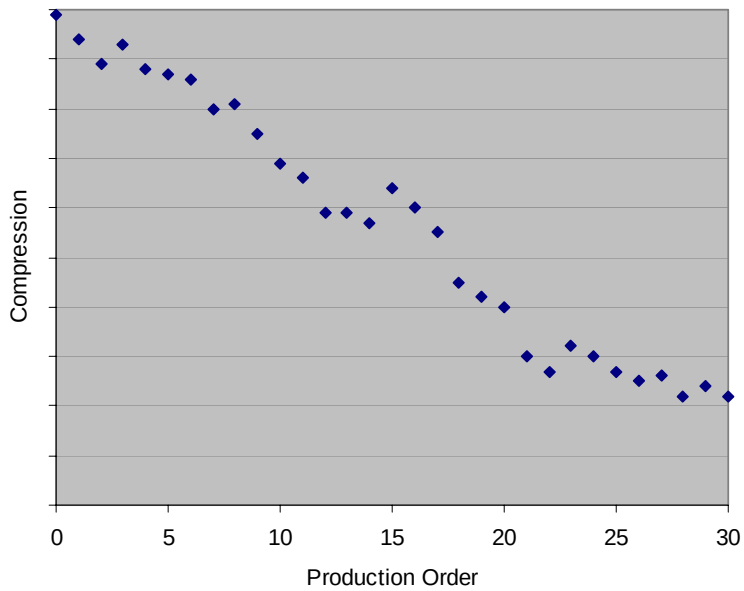
Plot of data using jitter function in S-Plus

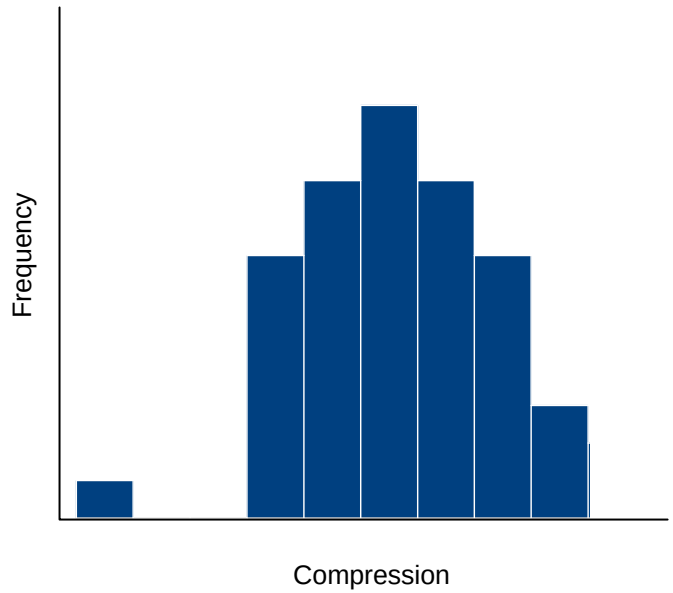




Run Chart

For time series data, it is often useful to plot the data in time sequence. A run chart graphs the data against time.

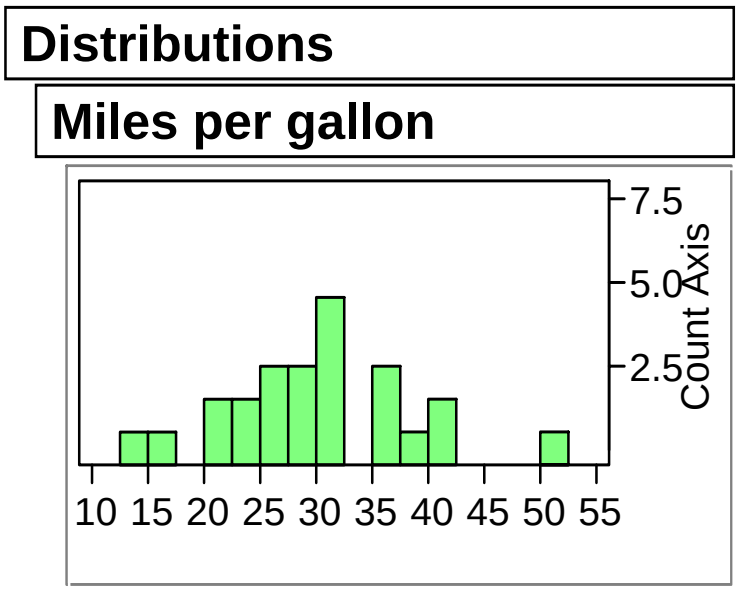




Always Plot Your Data Appropriately - Try Several Ways!

Histogram

Data: n=24 Gas Mileage
{31,13,20,21,24,25,25,27,28,
40,29,30,31,23,31,32,35,28,
36,37,38,40,50,17}

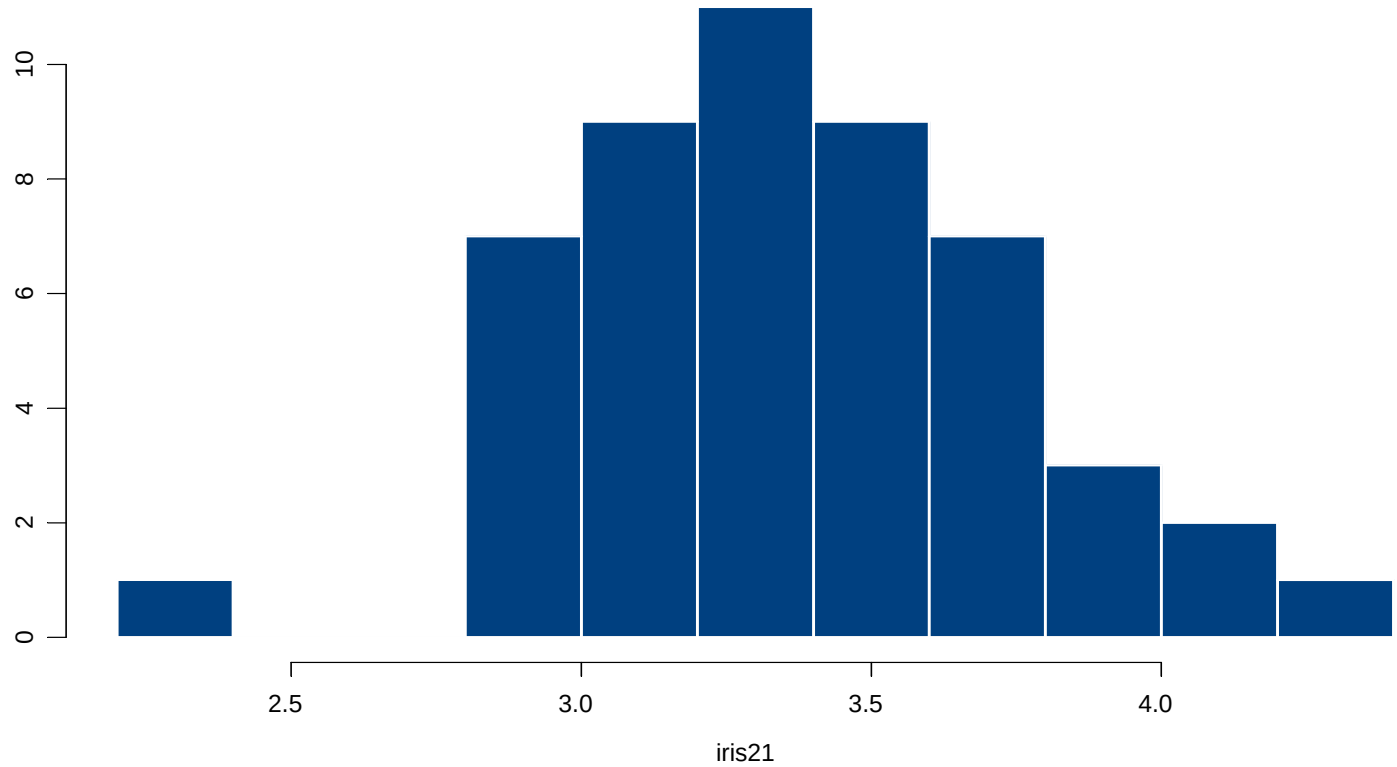


Note: Bars touch for continuous data, but do NOT touch for discrete data.

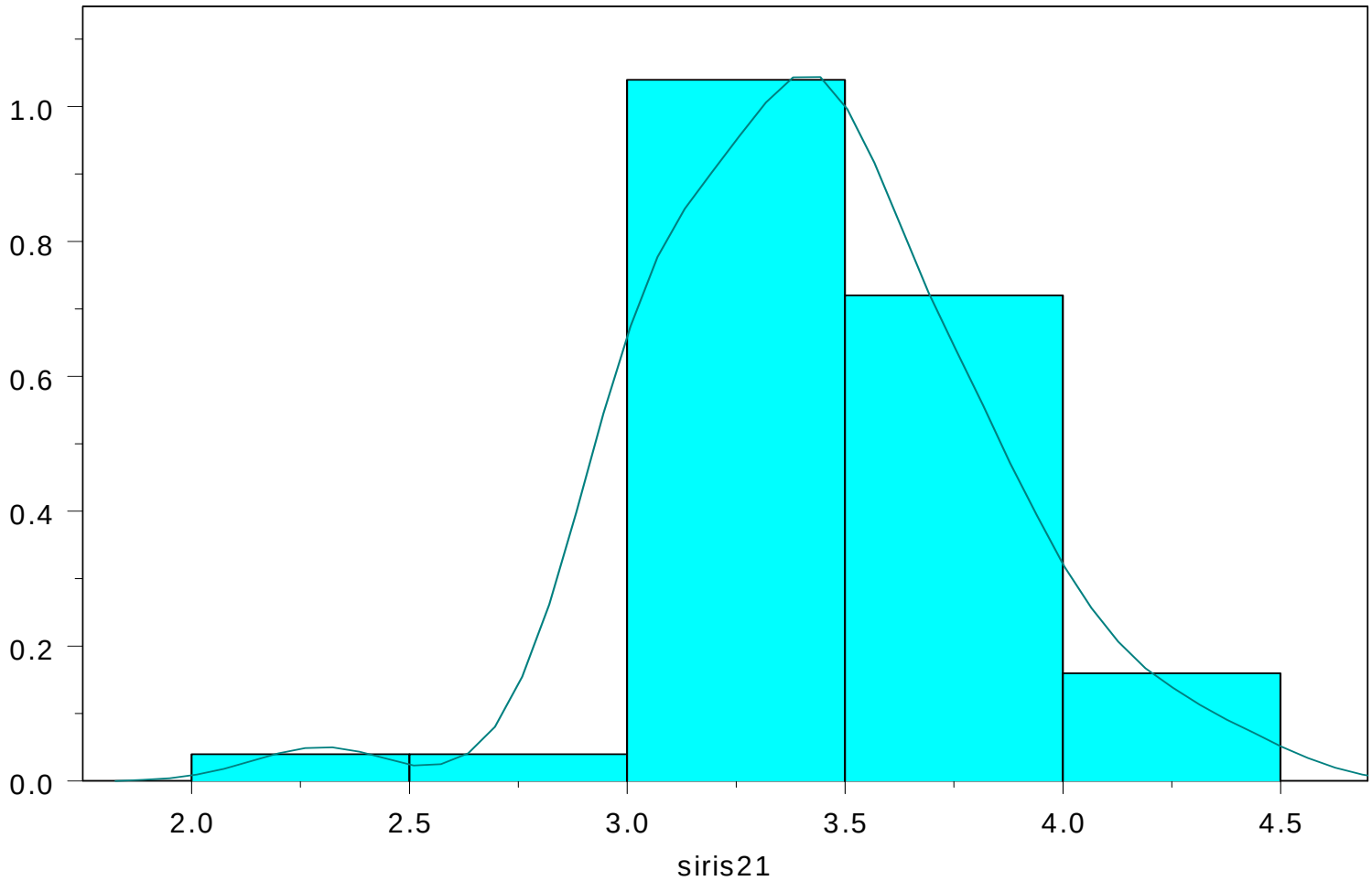
Gives a picture of the distribution of data.

- Area under the histogram represents sample proportion.
- Use approx. sqrt(n) “bins” - if too many, too jagged; if too few, too smooth (no detail)
- Shows if the distribution is:
  - Symmetric or skewed
  - Unimodal or bimodal
- Gaps in the data may indicate a problem with the measurement process.
- Many quality control applications
  - Are there two processes?
  - Detection of rework or cheating
  - Tells if process meets the specifications

Histogram of Iris Data



Histogram of Iris Data with Density Curve



Stem and Leaf Diagram

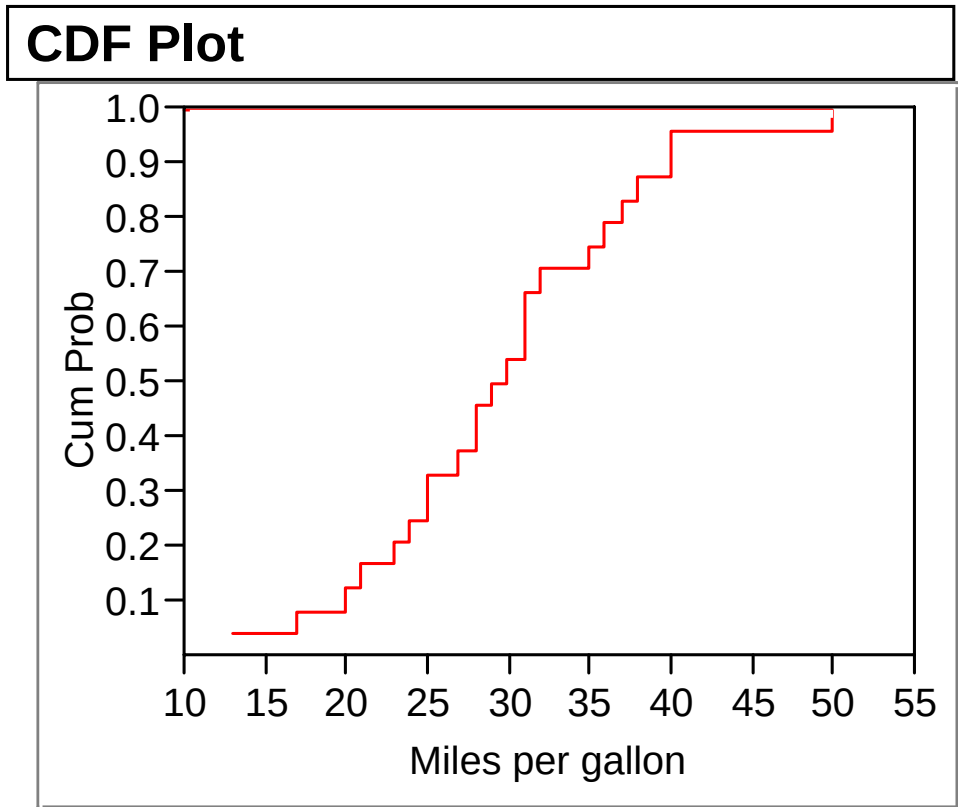
Cum. Dist. Function

Data: Gas Mileage

Stem	Leaf	Count
5	0	1
4		
4	00	2
3	5678	4
3	01112	5
2	557889	6
2	0134	4
1	7	1
1	3	1

Shows distribution of data similar to a histogram but preserves the actual data.

Can see numerical patterns in the data (like 40's and 50).



Step occurs at each data value (higher for more values at the same data point).

Stem and Leaf Diagram for Iris Data

- Decimal point is 1 place to the left of the colon

- 23 : 0
- 24 :
- 25 :
- 26 :
- 27 :
- 28 :
- 29 : 0
- 30 : 000000
- 31 : 0000
- 32 : 00000
- 33 : 00
- 34 : 000000000
- 35 : 000000
- 36 : 000
- 37 : 000
- 38 : 0000
- 39 : 00
- 40 : 0
- 41 : 0
- 42 : 0
- 43 :
- 44 : 0

Summary Statistics for Numerical Data

Measures of Location:

Mean ("average"): $\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$

Median: middle of the ordered sample (like $\theta_{.5}$ for distribution)

$$x_{min} = x_{(1)} \le x_{(2)} \le \ \ldots \le x_{(n)} = x_{max}$$

$$\text{median} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left[x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n+2}{2}\right)} \right] & \text{if } n \text{ is even} \end{cases}$$

Median of {0,1,2} is **1** : n=3 so n+1=4 & (n+1)/2=2 (2nd value)

Median of {0,1,2,3} is **1.5** (assumes data is continuous): n=4

Mode: The most common value

Mean or Median?

Appropriate summary of the center of the data?

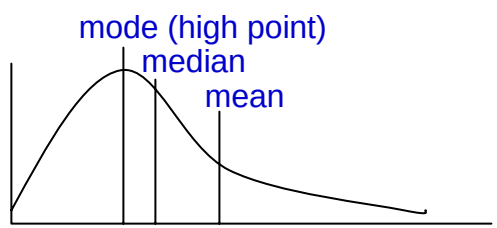
– Mean if the data has a symmetric distribution with light tails (i.e. a relatively small proportion of the observations lie away from the center of the data).

– Median if the distribution has heavy tails or is asymmetric.

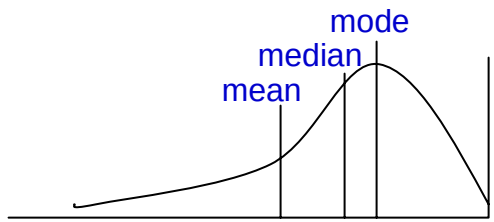
Extreme values that are far removed from the main body of the data are called outliers.

– Large influence on the mean but not on the median.

Right and left skewness (asymmetry)



(reverse alphabetic - RIGHT skewed)



(alphabetic - LEFT skewed)

Quantiles, Fractiles, Percentiles

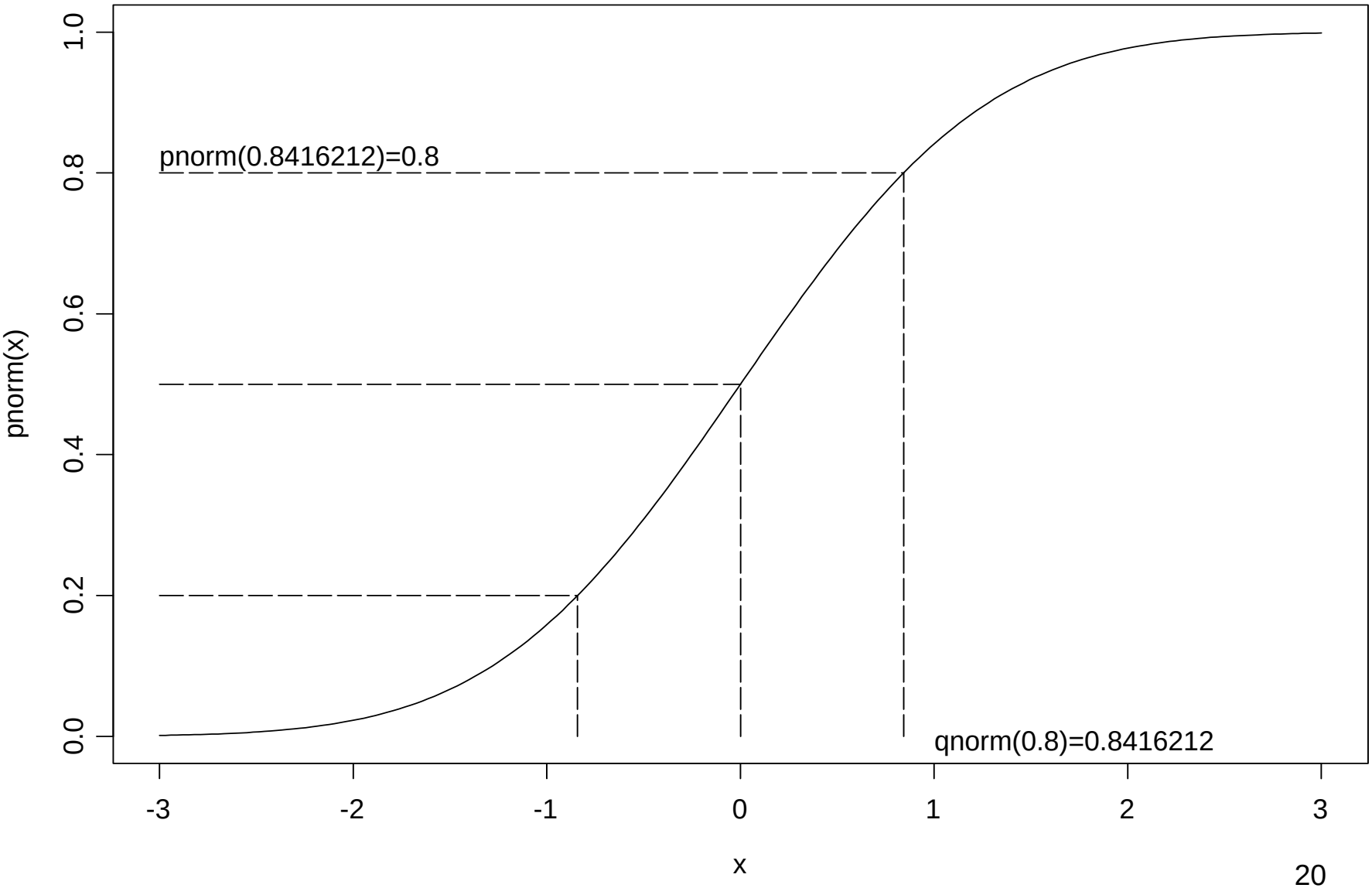
For a theoretical distribution:

- **The pth quantile is the value of a random variable X, x_p, such that $P(X<x_p)=p$. For the normal dist'n:**
- **In S-Plus: qnorm(p), 0<p<1, gives the quantile.**
- **In S-Plus: pnorm(q) gives the probability.**

For a sample:

- **The order statistics are the sample values in ascending order. Denoted $X_{(1)}$,…$X_{(n)}$**
- **The pth quantile is the data value in the sorted sample, such that a fraction p of the data is less than or equal to that value.**

Normal CDF



An algorithm for finding sample quantiles:

1) Arrange observations from smallest to largest.
2) For a given proportion p, compute the sample size $\times$ p = np.
3) If np is NOT an integer, round up to the next integer (ceiling (np)) and set the corresponding observation = x_p.
4) If np IS an integer k, average the kth and (k + 1)st ordered values. This average is then x_p.

– *Text has a different algorithm*

Quantiles, continued
(p^{th} quantile is $100p^{th}$ percentile)

Example:

Data: {0, 1, 2, 3, 4, 5, 6}

$= \{x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)}, x_{(5)}, x_{(6)}, x_{(7)}\}$

n=7

Q1 = ceiling(0.25*7) = 2 $\Rightarrow$ Q1 = $x_{(2)}$ = 1 = 25th percentile

Q2 = ceiling(0.50*7) = 4 $\Rightarrow$ Q2 = $x_{(4)}$ = 3 = median (50th percentile)

Q3 = ceiling(0.75*7) = 6 $\Rightarrow$ Q3 = $x_{(6)}$ = 5 = 75th percentile

S-Plus gives different answers!

Different methods for calculating quantiles.

Measures of Dispersion (Spread, Variability):

Two data sets may have the same center and but quite different dispersions around it.

Two ways to summarize variability:

1. Give the values that divide the data into equal parts.

- **Median is the 50^{th} percentile**
- **The 25^{th}, 50^{th}, and 75^{th} percentiles are called quartiles (Q1,Q2,Q3) and divide the data into four equal parts.**
- **The minimum, maximum, and three quartiles are called the "five number summary" of the data.**

2. Compute a single number, e.g., range, interquartile range, variance, and standard deviation.

Measures of Dispersion, continued

Range = maximum - minimum

Interquartile range (IQR) = Q3 – Q1

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 mean, variance, and standard deviations are sample analogs of the population mean, variance, and standard deviation (μ, σ^2, σ)

Other Measures of Dispersion

Sample Average of Absolute Deviations from the Mean:

$$\frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|$$

Sample Median of Absolute Deviations from the Median

$$\text{Median of } \{|x_i - x_{.5}|,\ i = 1, \ldots, n\}$$

Computations for Measures of Dispersion

Example:

Data: {0, 1, 2, 3, 4, 5, 6}

$= \{x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)}, x_{(5)}, x_{(6)}, x_{(7)}\}$

mean = (0+1+2+3+4+5+6)/ 7 = 21/ 7 = 3

min = 0, max = 6

$Q1 = x_{(2)} = 1 = 25^{th}$ percentile

$Q2 = x_{(4)} = 3 =$ median (50^{th} percentile)

$Q3 = x_{(6)} = 5 = 75^{th}$ percentile

Range = max - min = 6 - 0 = 6

IQR = Q3 - Q1 = 5 - 1 = 4

$s^2 = [(0^2+1^2+2^2+3^2+4^2+5^2+6^2) - 7(3^2)]/(7-1) = [91-63]/6 = 4.67$

s = sqrt(4.67) = 2.16

Sample Variance and Standard Deviation

s^2 and s should only be used to summarize dispersion with symmetric distributions.

For asymmetric distribution, a more detailed breakup of the dispersion must be given in terms of quartiles.

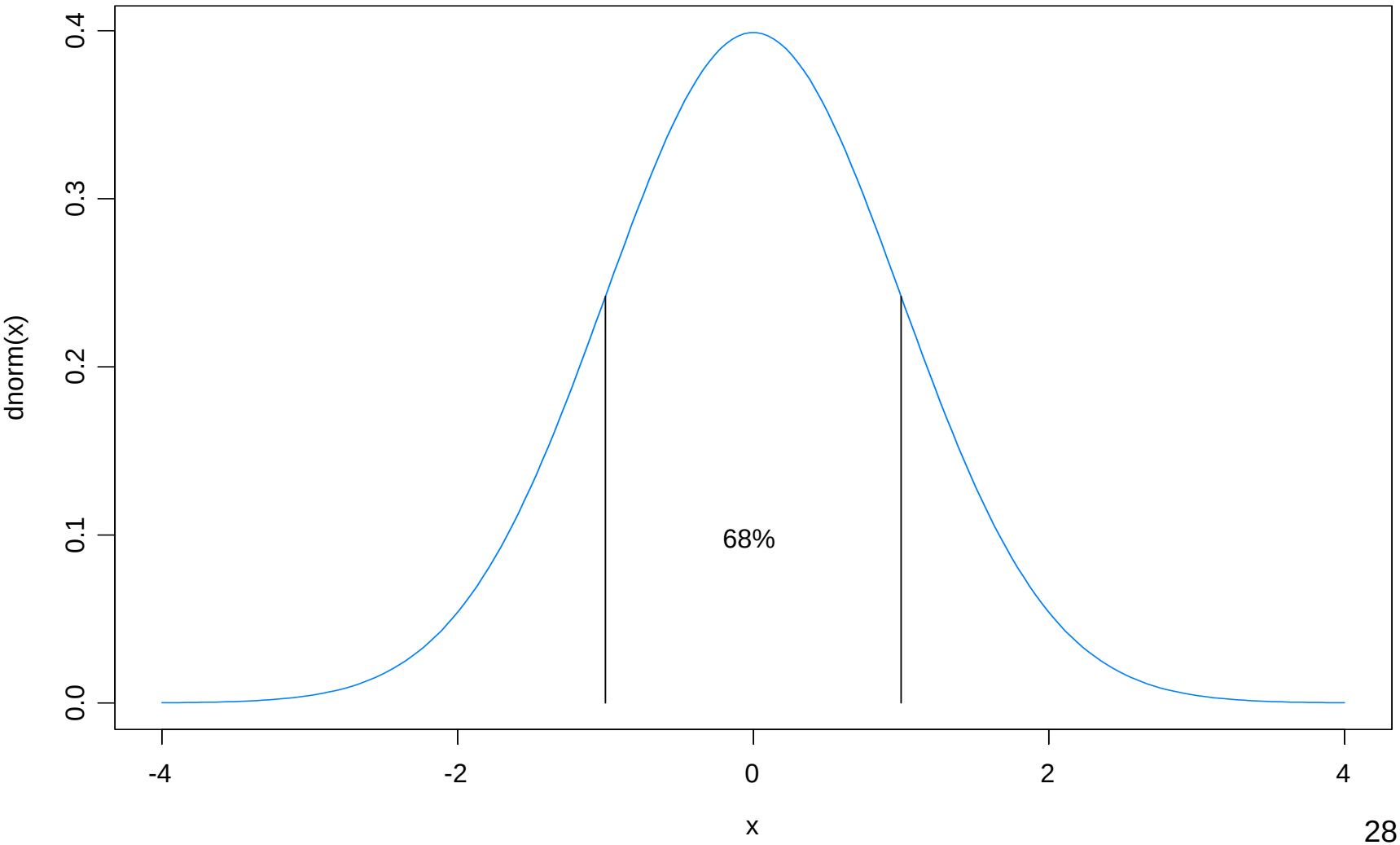
For normal data and large samples:

- 50% of the data values fall between mean ± 0.67s
- 68% of the data values fall between mean ± 1s
- 95% of the data values fall between mean ± 2s
- 99.7% of the data values fall between mean ± 3s

For normally distributed data:

$$\text{IQR}=(\text{mean} + 0.67s) - (\text{mean} - 0.67s) = 1.34s$$

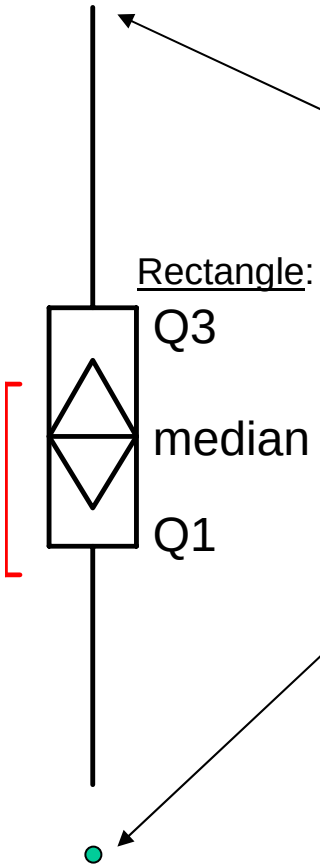
Standard Normal Density



Box (and Whiskers) Plots

Visual display of summary of data (more than five numbers)

Outlier Box Plot Data: Gas Mileage Quantile Box Plot



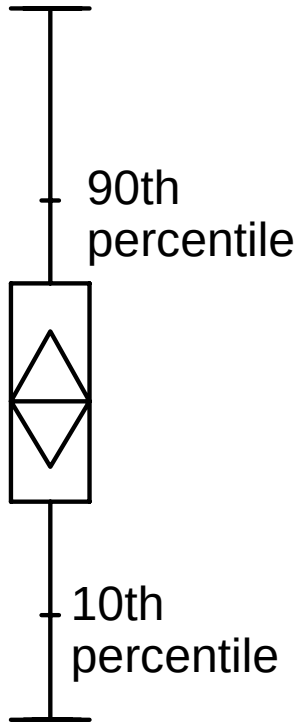
IQR = Q3 - Q1

Upper Fence = Q3 + 1.5 x IQR

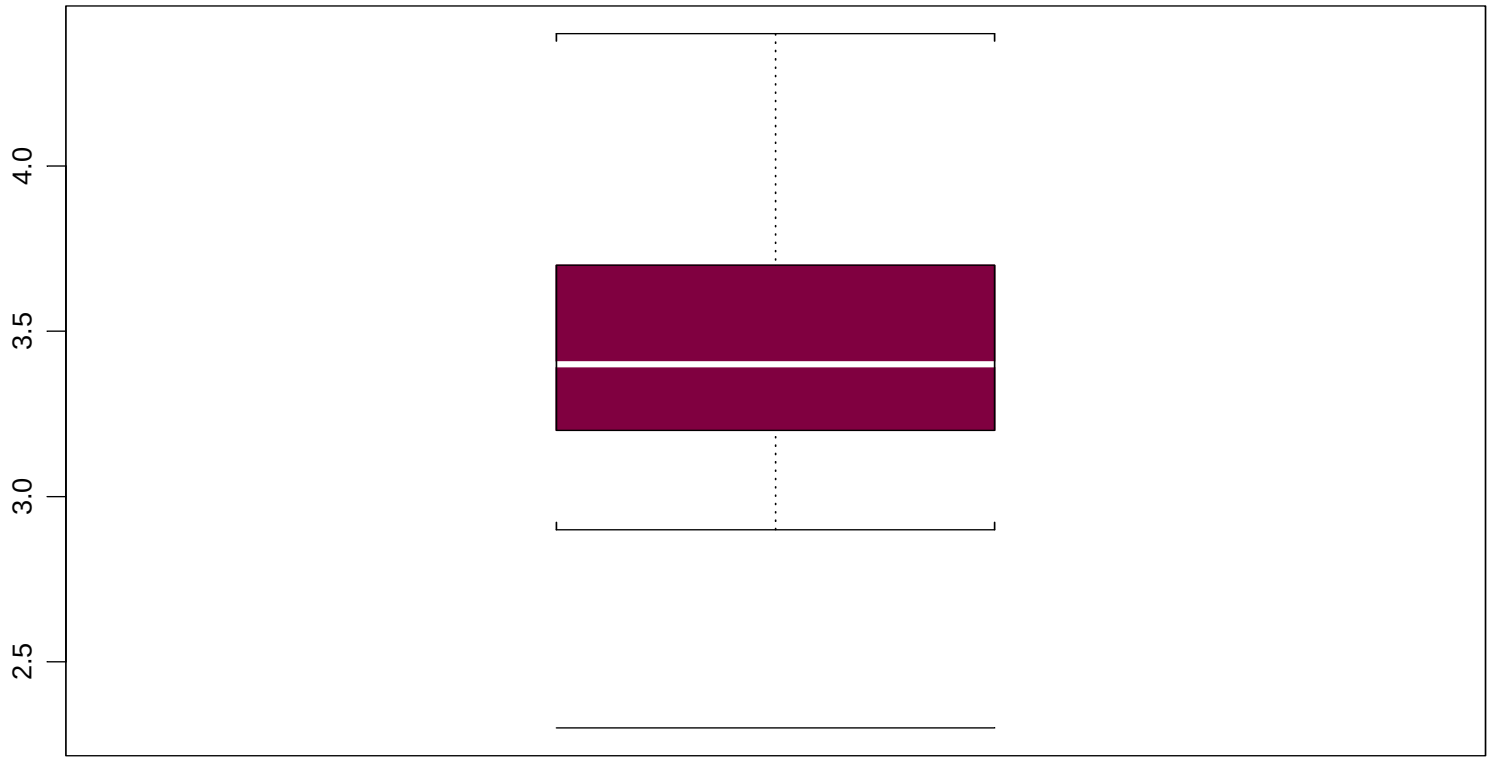
Lower Fence = Q1 – 1.5 x IQR

Two lines are called whiskers and extend to the most extreme data values that are still inside the fences.

Observations outside the fences are regarded as possible outliers and are denoted by dots and circles or asterisks.



Box Plot for Iris Data



iris21

QQ Plots
Compare Sample to Theoretical Distribution

Order the data. The i^{th} ordered data value is the pth quantile, where p = (i - 0.5)/n, 0<p<1.

Text uses i/(n+1).

(Why can't we just say i/n)?

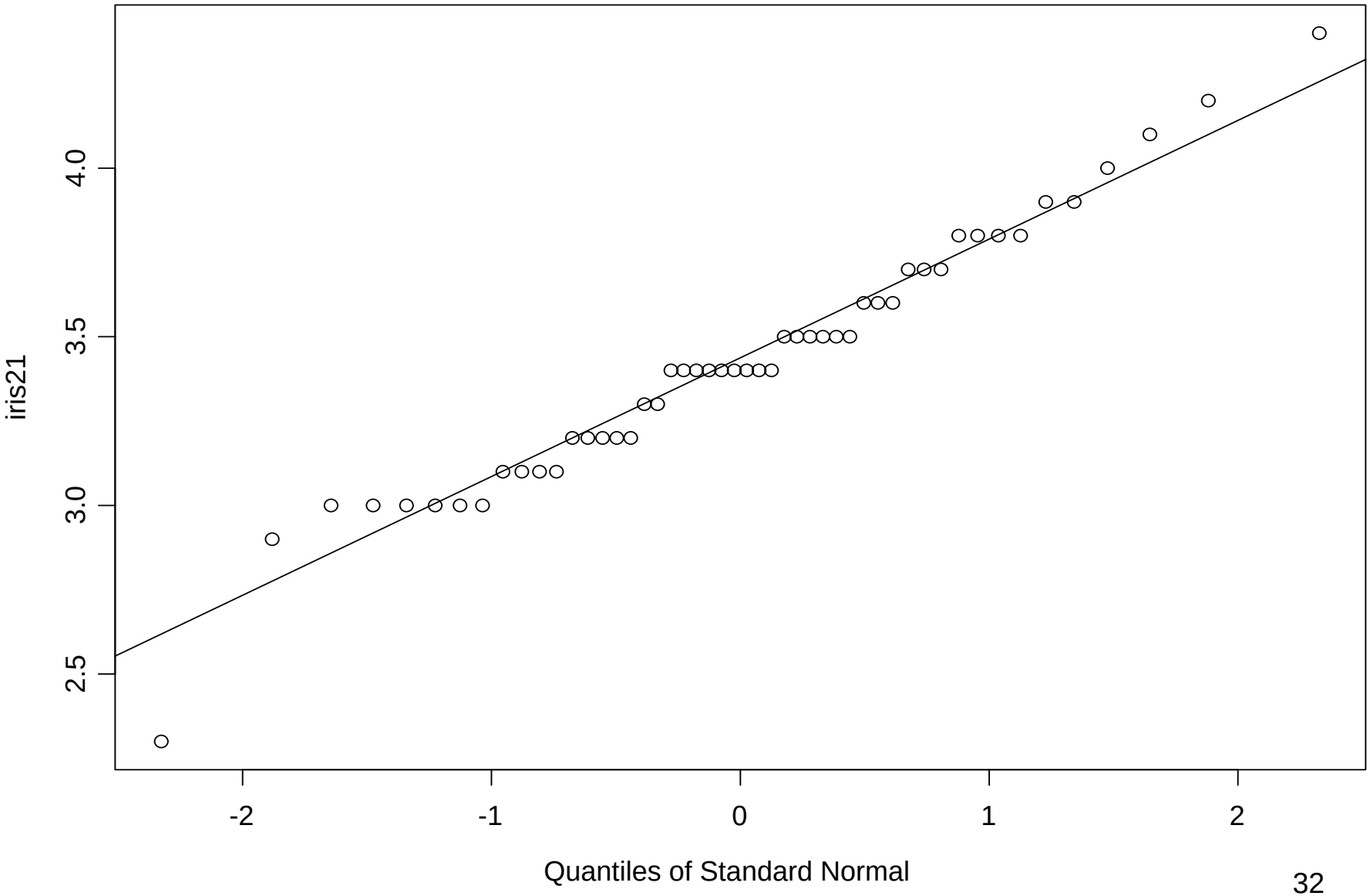
Obtain quantiles from theoretical distribution corresponding to the values for p.

E.g. qnorm(p), in S-Plus for normal distribution.

Plot theoretical quantiles vs. empirical quantiles (sorted data). S-Plus: plot(qnorm((1:length(y)-0.5)/n),sort(y))

Fit line through first and third quartiles of each distribution.

QQ (Normal) Plot for Iris Data



Normalizing Transformations

Data can be non-normal in a number of ways, e.g., the distribution may not be bell shaped or may be heavier tailed than the normal distribution or may not be symmetric.

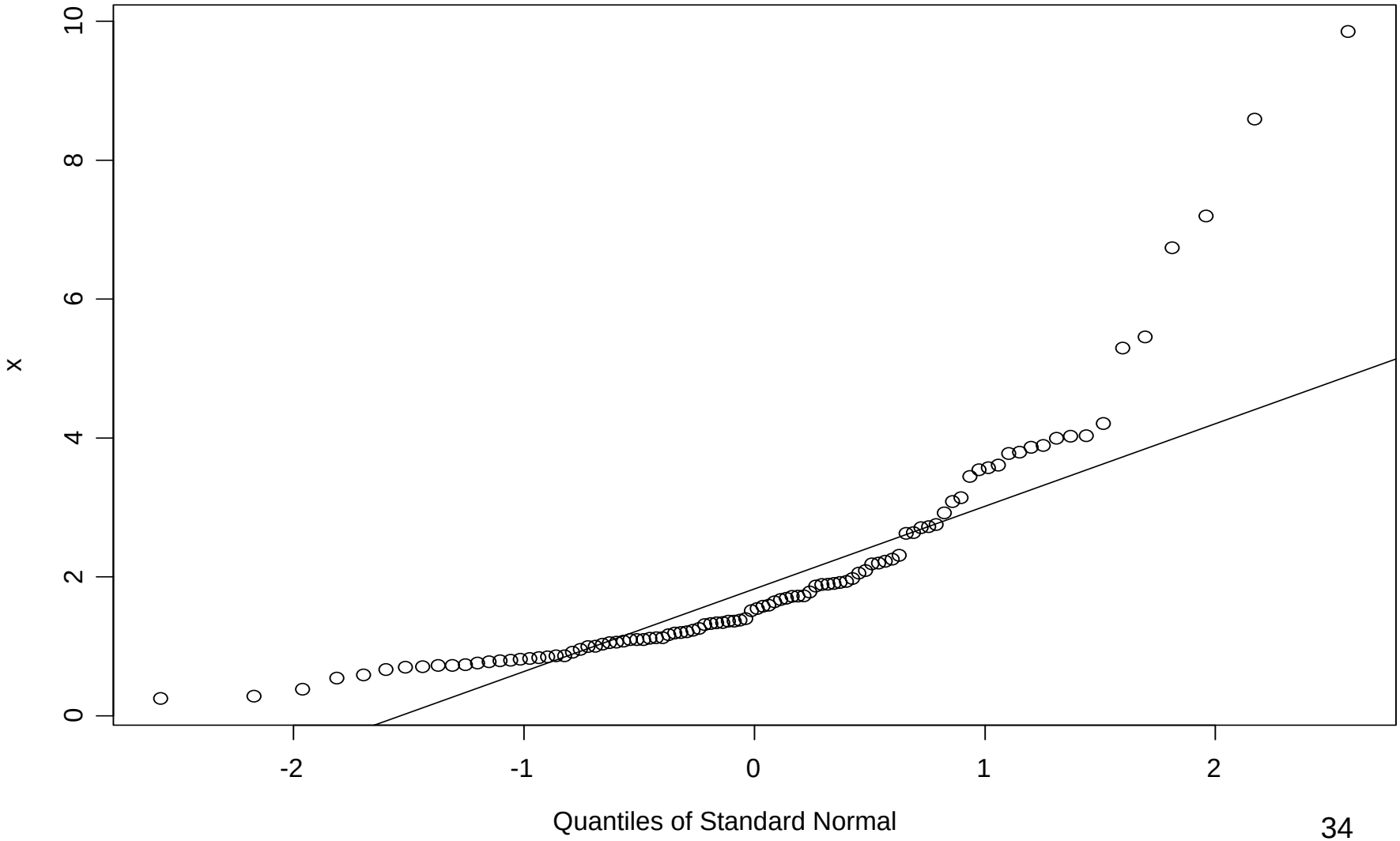
Only the departure from symmetry can be easily corrected by transforming the data.

If the distribution is positively skewed, then the right tail needs to be shrunk inward. The most common transformation used for this purpose is the log transformation: $x \rightarrow \log x$ (e.g., decibels, Richter, and Beaufort (?) scales); see Figure 4.11.

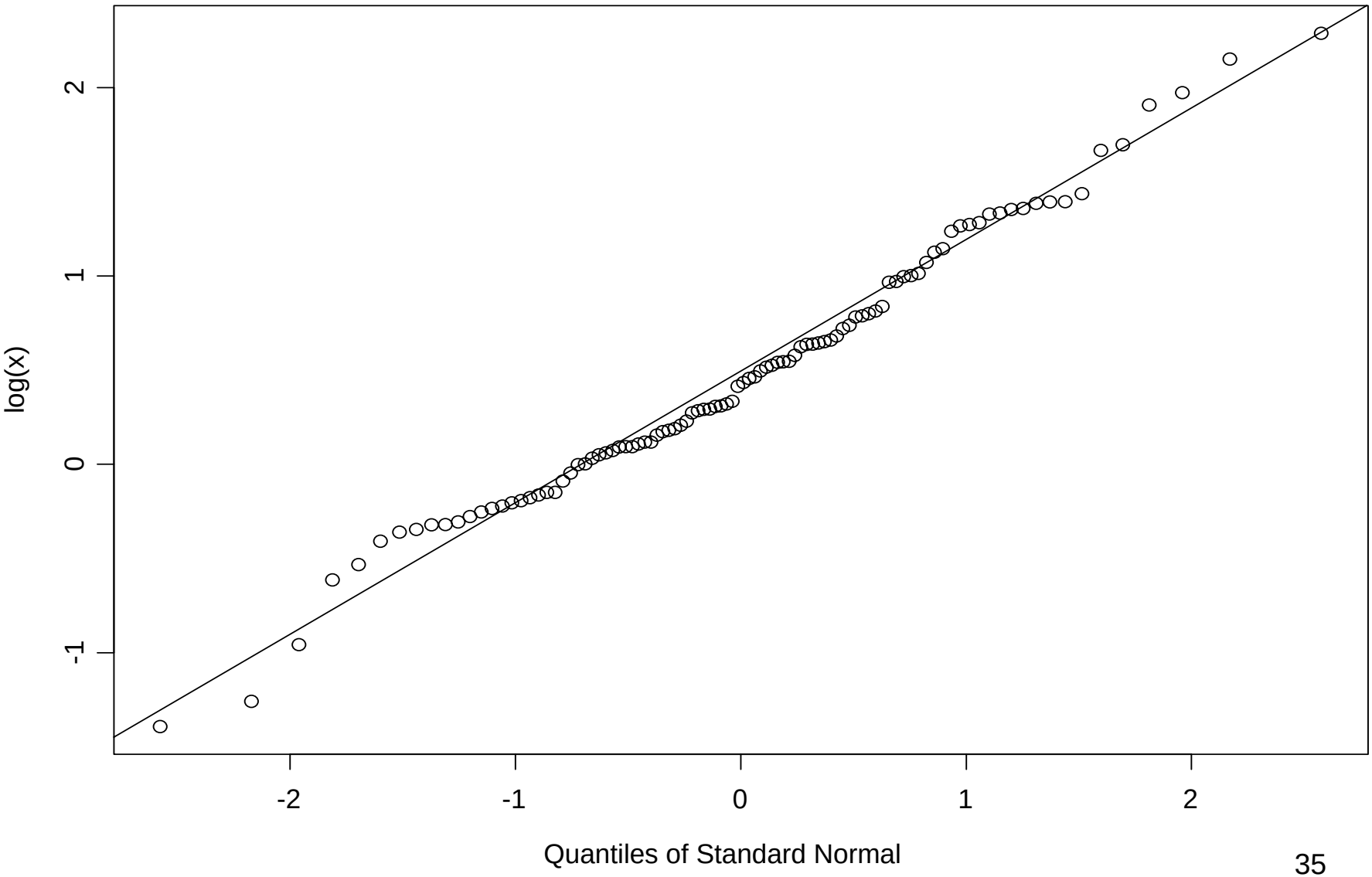
The square-root ($\sqrt{x}$) transformation provides a weaker shrinking effect; it is frequently used for (Poisson) count data.

For negatively skewed data, use the exponential (e^x) or squared (x^2) transformations.

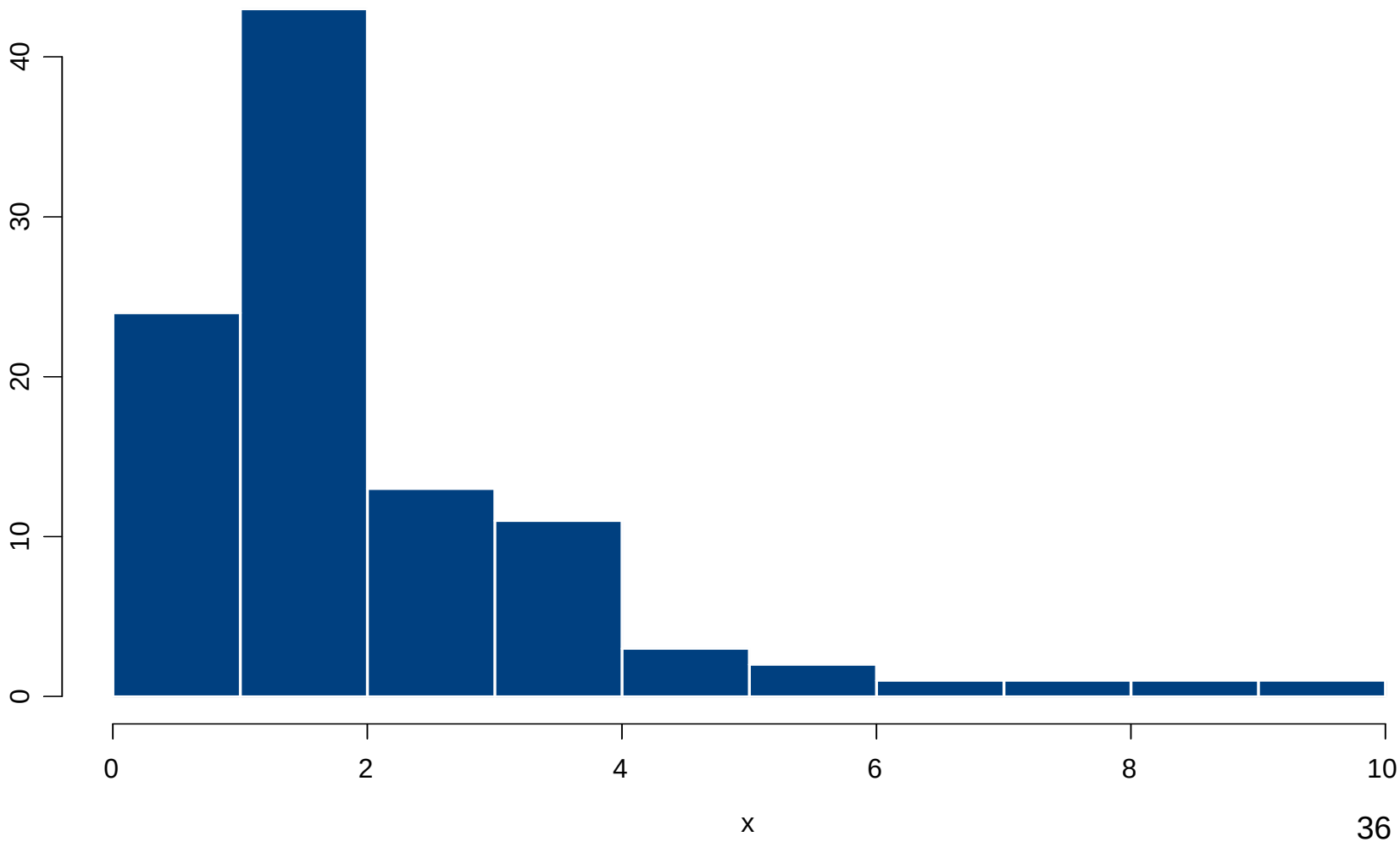
Normal Probability Plot of data generated from a certain distribution



Normal probability plot of log of same data



Histogram of the same data



Summarizing Multivariate Data

When two or more variables are measured on each sampling unit, the result is multivariate data.

If only two variables are measured the result is bivariate data. One variable may be called the *x* variable and the other the *y* variable.

We can analyze the *x* and *y* variable separately with the methods we have learned so far, but these methods would NOT answer questions about the relationship between *x* and *y*.

– What is the nature of the relationship between *x* and y (if any)?

– How strong is the relationship?

– How well can one variable be predicted from the other?

Summarizing Bivariate Categorical Data

Two-way Table

	Overall Job Satisfaction				
Annual Salary	Very Dissatisfied	Slightly Dissatisfied	Slightly Satisfied	Very Satisfied	**Row Sum**
Less than $10,000	81	64	29	10	184
$10,000-25,000	73	79	35	24	211
$25,000-50,000	47	59	75	58	239
More than $50,000	14	23	84	69	190
Column Sum	215	225	223	161	**824**

The numbers in the cells are the frequencies of each possible combination of categories.

Cell, row and column percentages can be computed to assess distribution.

Column Percentages for Income and Job Satisfaction Table

	Overall Job Satisfaction			
Annual Salary	Very Dissatisfied	Slightly Dissatisfied	Slightly Satisfied	Very Satisfied
Less than $10,000	37.7	28.4	13.0	6.2
$10,000-25,000	34.0	35.1	15.7	14.9
$25,000-50,000	21.9	26.2	33.6	36.0
More than $50,000	6.5	10.2	37.7	42.9

Simpson's Paradox

"Lurking variables [excluded from consideration] can change or reverse a relation between two categorical variables!"

Doctors' Salaries

- The interpreter of a survey of doctors' salaries in 1990 and again in 2000 concluded that their average income *actually declined* from $97,000 in 1990 to $91,000 in 2000."
- Income is measured here in nominal (not adjusted for inflation) dollars.

What about the “Rest of the Story”?

- What deductive piece of logic might clarify the real meaning of this particular pair of statistics?
- Look more deeply: Is there a piece missing?
- Here is a very simple breakdown of “the numbers” that may help.

Doctors' Salaries by Age

	1980		1990	
Age	fraction, f1	Income	fraction, f2	Income
<=45	0.5	$60,000	0.7	$70,000
>45	0.5	$120,000	0.3	$130,000
	Mean	$90,000		$88,000

Conclusion

- If MD salaries are broken into two categories by *age:*
  - Doctors younger than 45 constituted 50% of the MD population in 1980 and 70% in 1990
  - Younger doctors tend to earn less than older, more experienced doctors
  - Parsed by age, MD salaries *increased in both age categories!*

Gender Bias in Graduate Admissions

For this example, see Johnson and Wichern, *Business Statistics: Decision Making with Data*. Wiley, First Edition, 1997.

Statistical Ideal

Randomized study

Gender should be randomly assigned to applicants!

This would automatically balance out the departmental factor which is not controlled for in the original plaintiff (observational) study.

Practical reality

Gender cannot be assigned randomly.

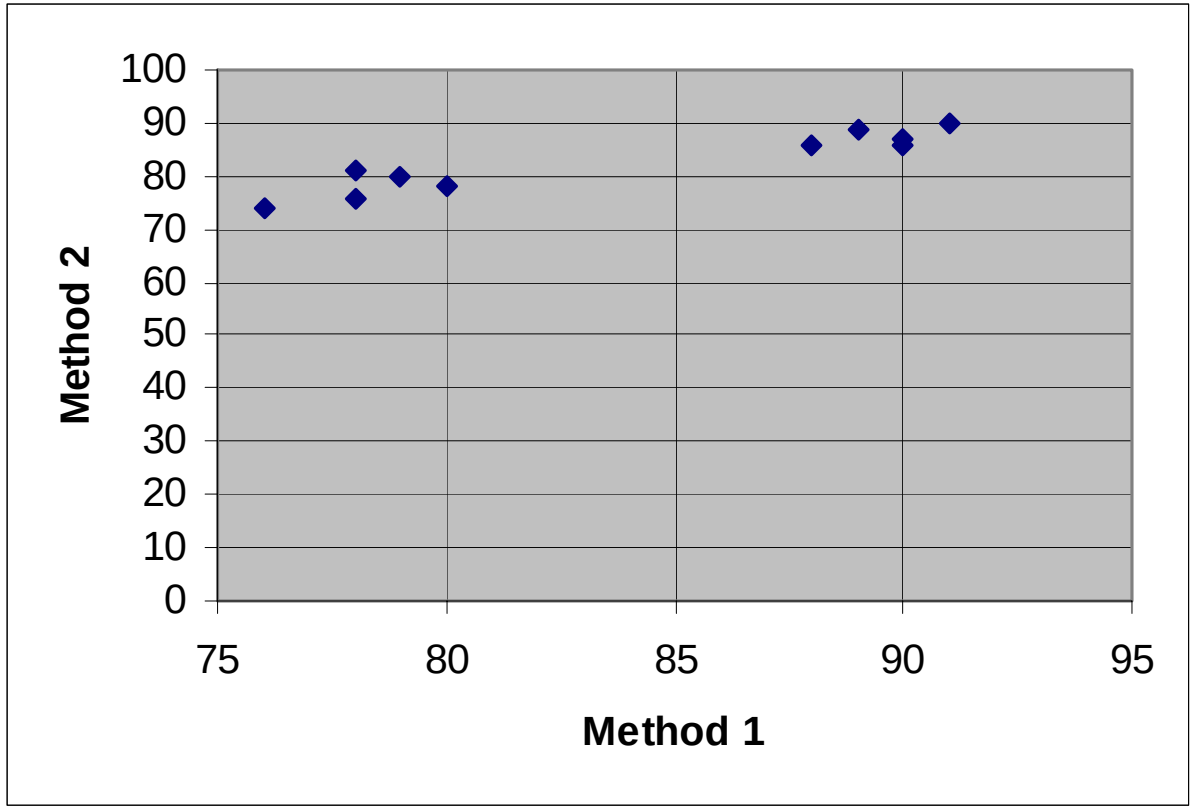
Control for department factor by comparing admission within department, i.e. controlling for the confounding factor after completion of the study.

“There are lies, damn lies and then there are statistics!”

Benjamin Disraeli

Summarizing Bivariate Numerical Data

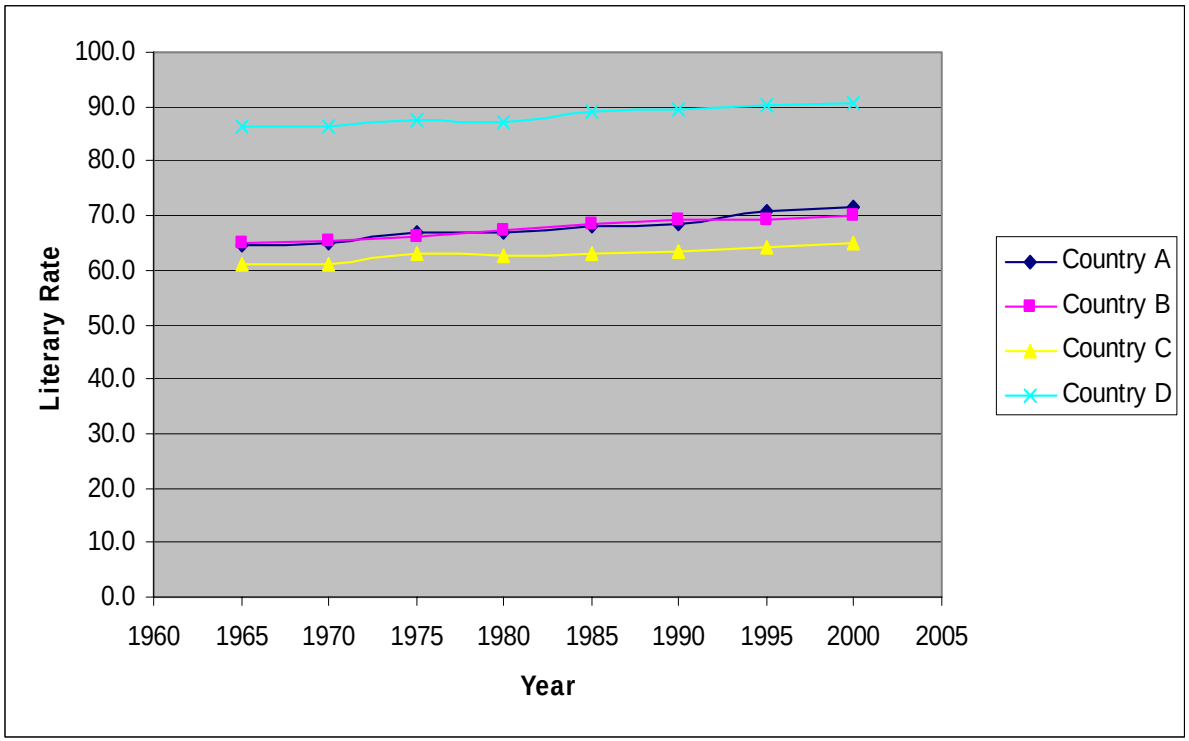
No.	Method 1 (x_i)	Method 2 (y_i)
1	88	86
2	78	81
3	90	87
4	91	90
5	89	89
6	79	80
7	76	74
8	80	78
9	78	76
10	90	86



Is it easier to grasp the relationship in the data between Method A and Method B from the Table or from the Figure (scatter plot)?

Labeled Scatter Plot

Year	Country A	Country B	Country C	Country D
1965	64.7	64.8	61.1	86.2
1970	65.0	65.2	61.2	86.5
1975	66.8	66.3	63.0	87.4
1980	66.9	67.4	62.8	87.0
1985	67.9	68.5	63.1	89.2
1990	68.3	69.1	63.5	89.4
1995	70.8	69.4	64.3	90.1
2000	71.7	70.0	65.1	90.5



Can you see the improvements in the literacy rates for these four countries more easily in the Table or in the Figure?

Sample Correlation Coefficient

A single numerical summary statistic which measures the strength of a linear relationship between *x* and *y*.

$$r = \frac{s_{xy}}{s_x s_y} \text{ where } s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

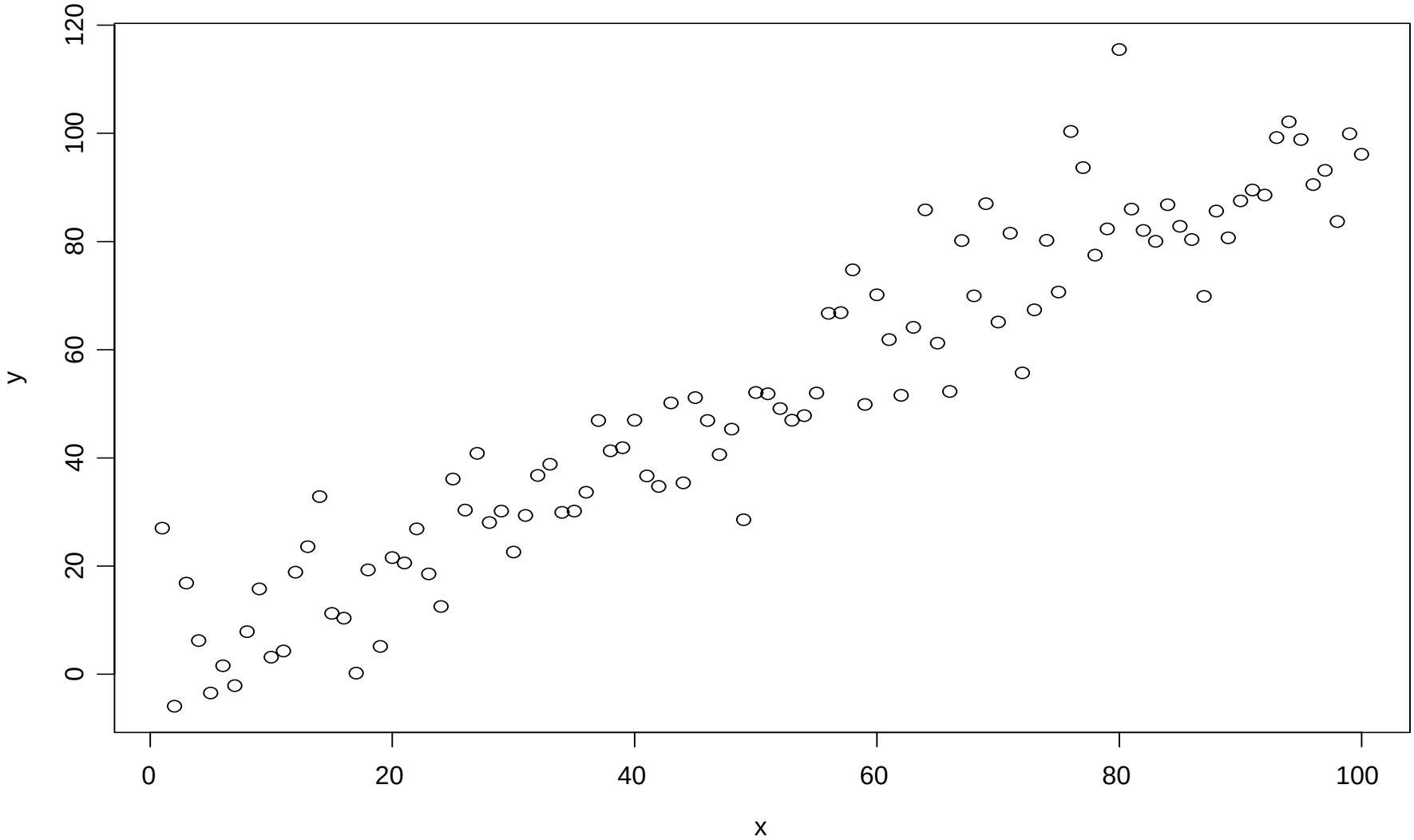
***r* = covar(x,y)/(stddev(x)*stddev(y))**

Properties similar to the population correlation coefficient ρ

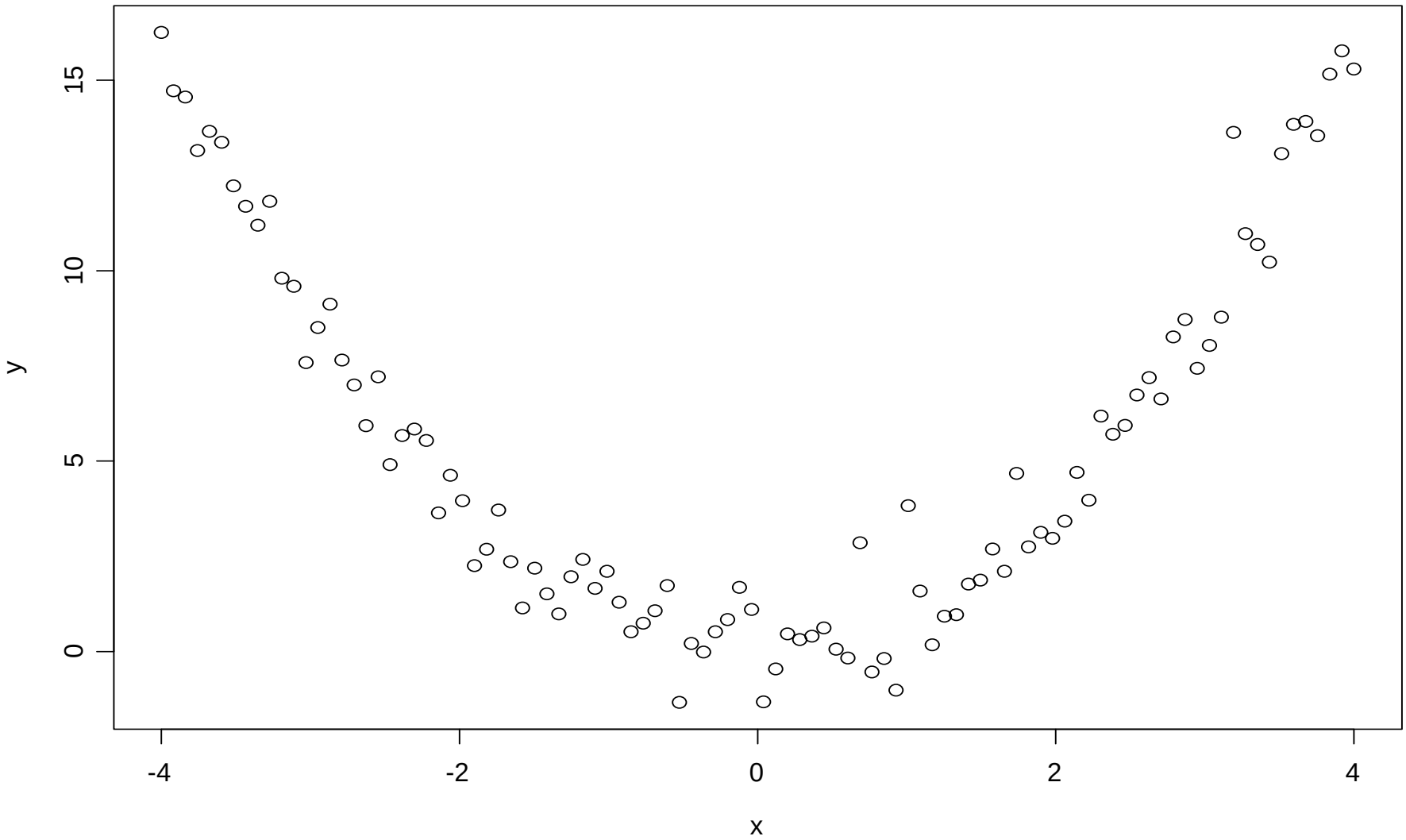
- **– Unitless quantity**
- **– Takes values between –1 and 1**
- **– The extreme values are attained if and only if the points (x_i , y_i) fall exactly on a straight line (*r* = -1 for a line with negative slope and *r* = +1 for a line with positive slope.)**
- **– Takes values close to zero if there is no linear relationship between *x* and *y*.**

- **See Figures 4.15, 4.16, 4.17 (a) and (b)**

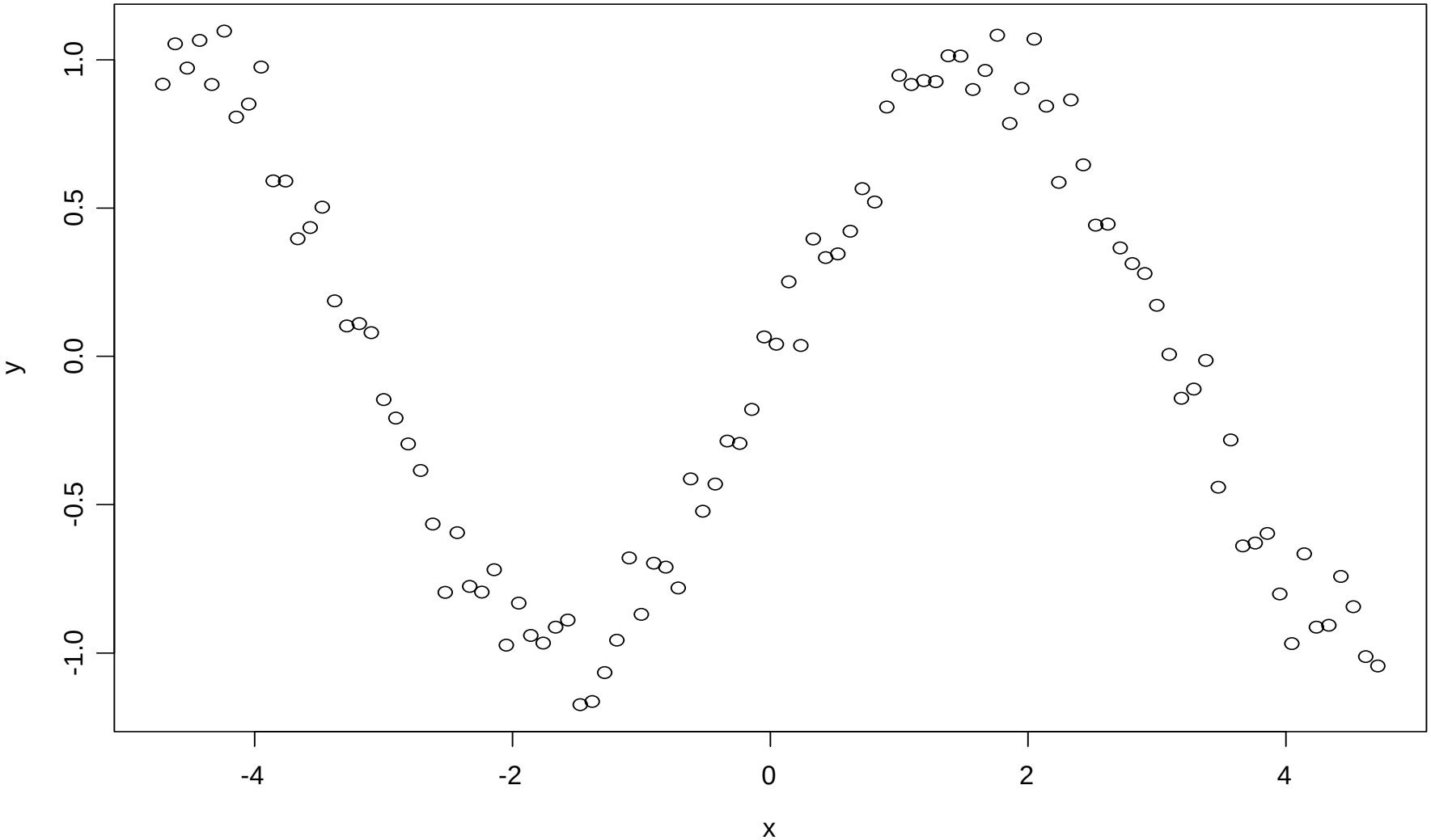
What is the correlation?



What is the correlation?



What is the correlation?



Correlation and Causation

High correlation is frequently mistaken for a cause and effect relationship. Such a conclusion may not be valid in observational studies, where the variables are not controlled.

- A lurking variable may be affecting both variables.
- One can only claim association, not causation.

Countries with high fat diets tend to have higher incidences of cancer. Can we conclude causation?

A common lurking variable in many studies is time order.

- Wealth and health problems go up with age.

  Does wealth cause health problems?

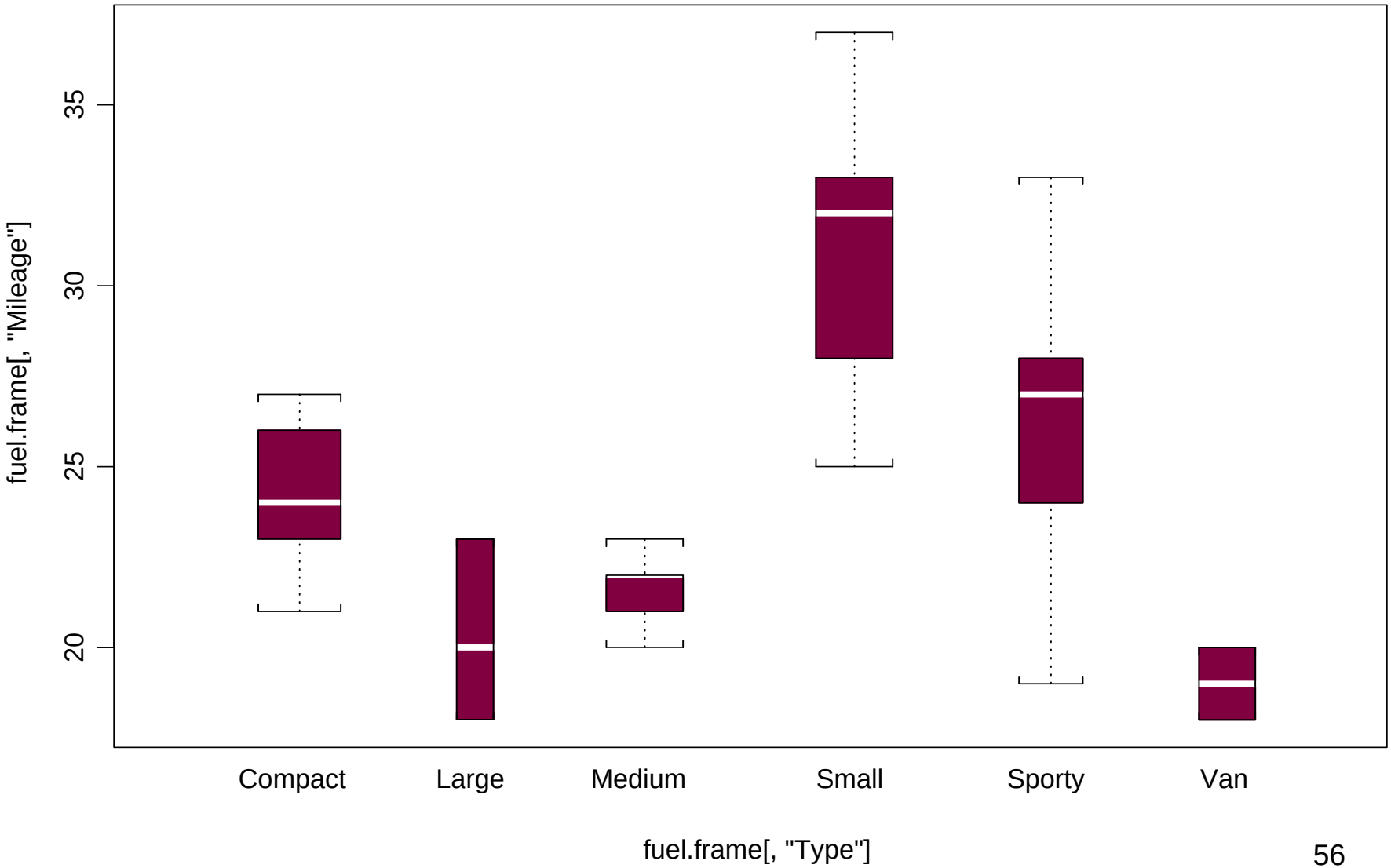
Sometimes correlations can be found without any plausible explanation, e.g., sun spots and economic cycles.

Plots for Multivariate Data

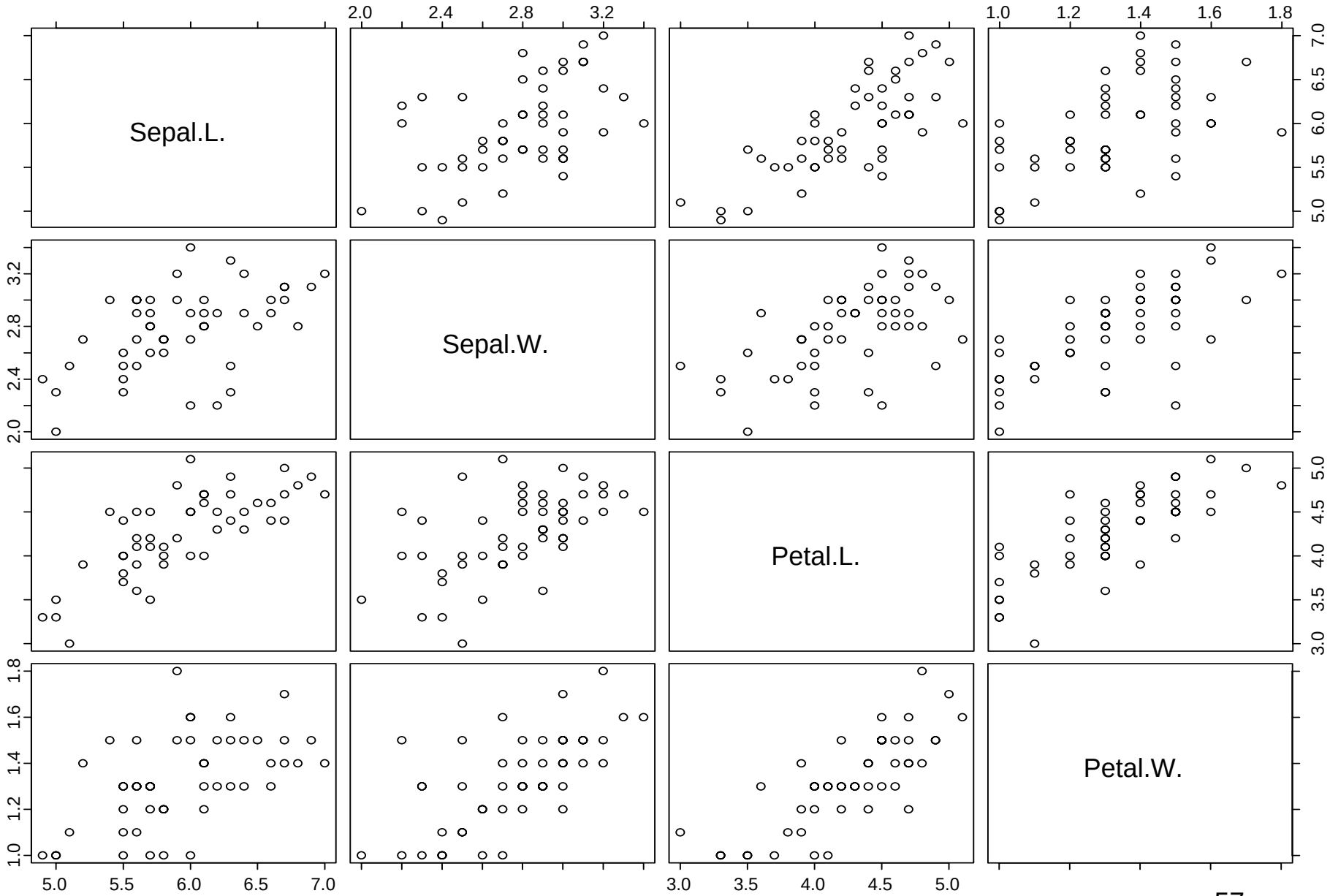
- Side by Side Box Plots
- Scatter plot matrix
- Three dimensional plots
- Brush and Spin plots – add motion
- Maps for spatial data

Box Plots of Auto Data

widths indicate number of each type

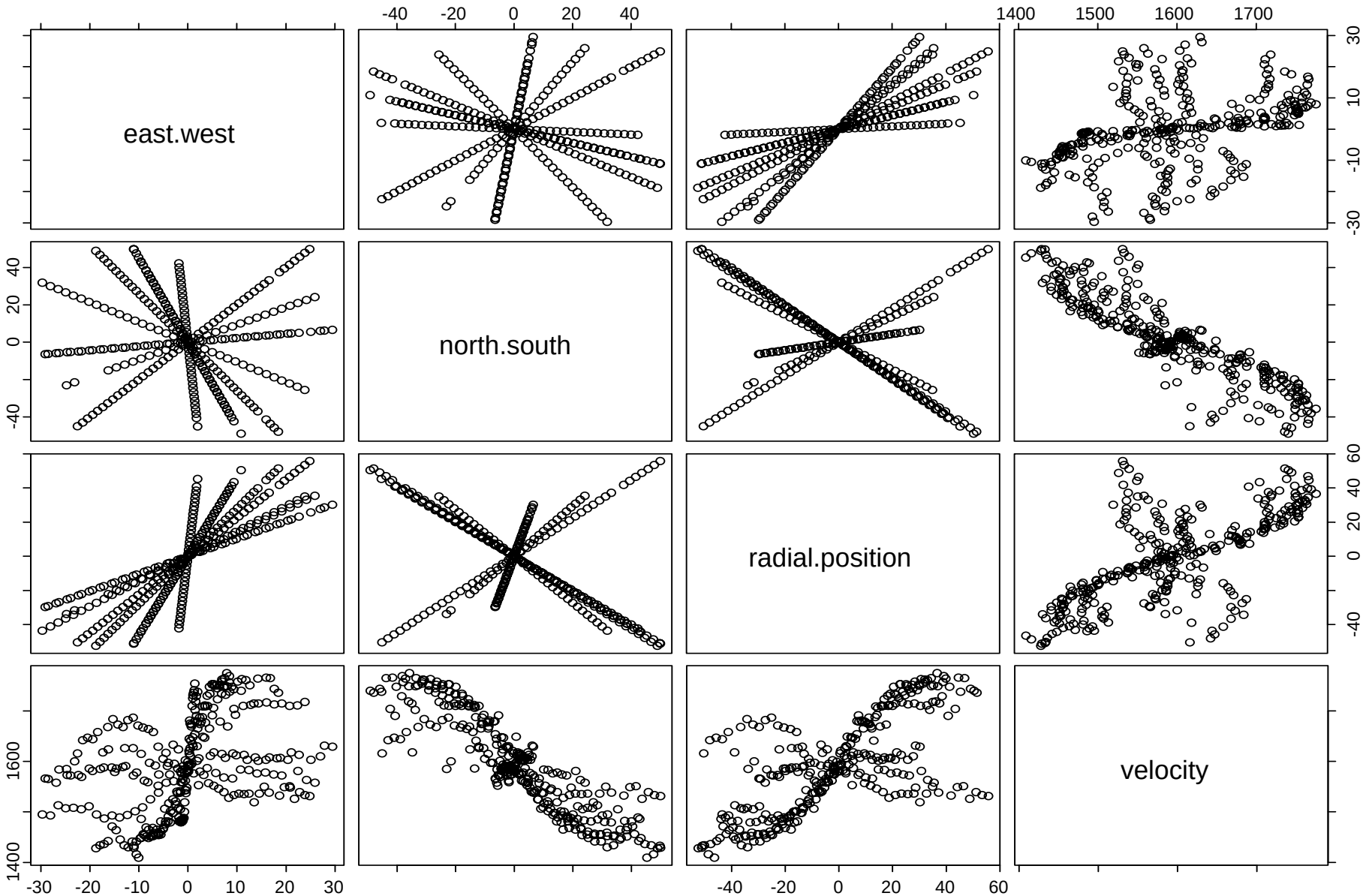


Scatter plot matrix Iris –(Versicolor)

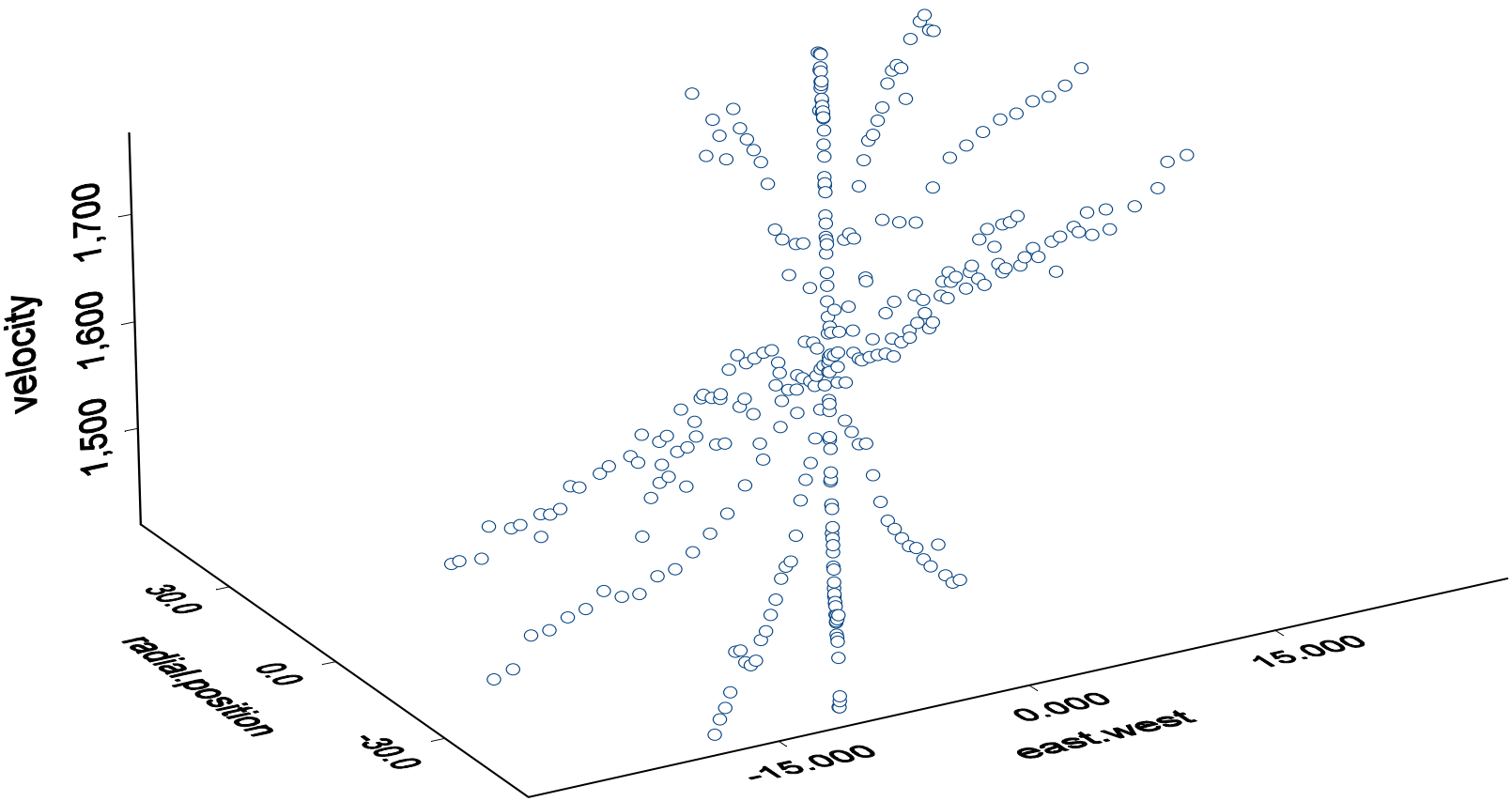


- Galaxy S-PLUS Language Reference
- **Radial Velocity of Galaxy NGC7531**
- **SUMMARY:**
- The galaxy data frame records the radial velocity of a spiral galaxy measured at 323 points in the area of sky which it covers. All the measurements lie within seven slots crossing at the origin. The positions of the measurements given by four variables (columns).
- **ARGUMENTS:**
- **east.west**
  - the east-west coordinate. The origin, (0,0), is near the center of the galaxy, east is negative, west is positive.
- **north.south**
  - the north-south coordinate. The origin, (0,0), is near the center of the galaxy, south is negative, north is positive.
- **angle**
  - degrees of counter-clockwise rotation from the horizontal of the slot within which the observation lies.
- **radial.position**
  - signed distance from origin; negative if east-west coordinate is negative.
- **velocity**
  - radial velocity measured in km/sec. .

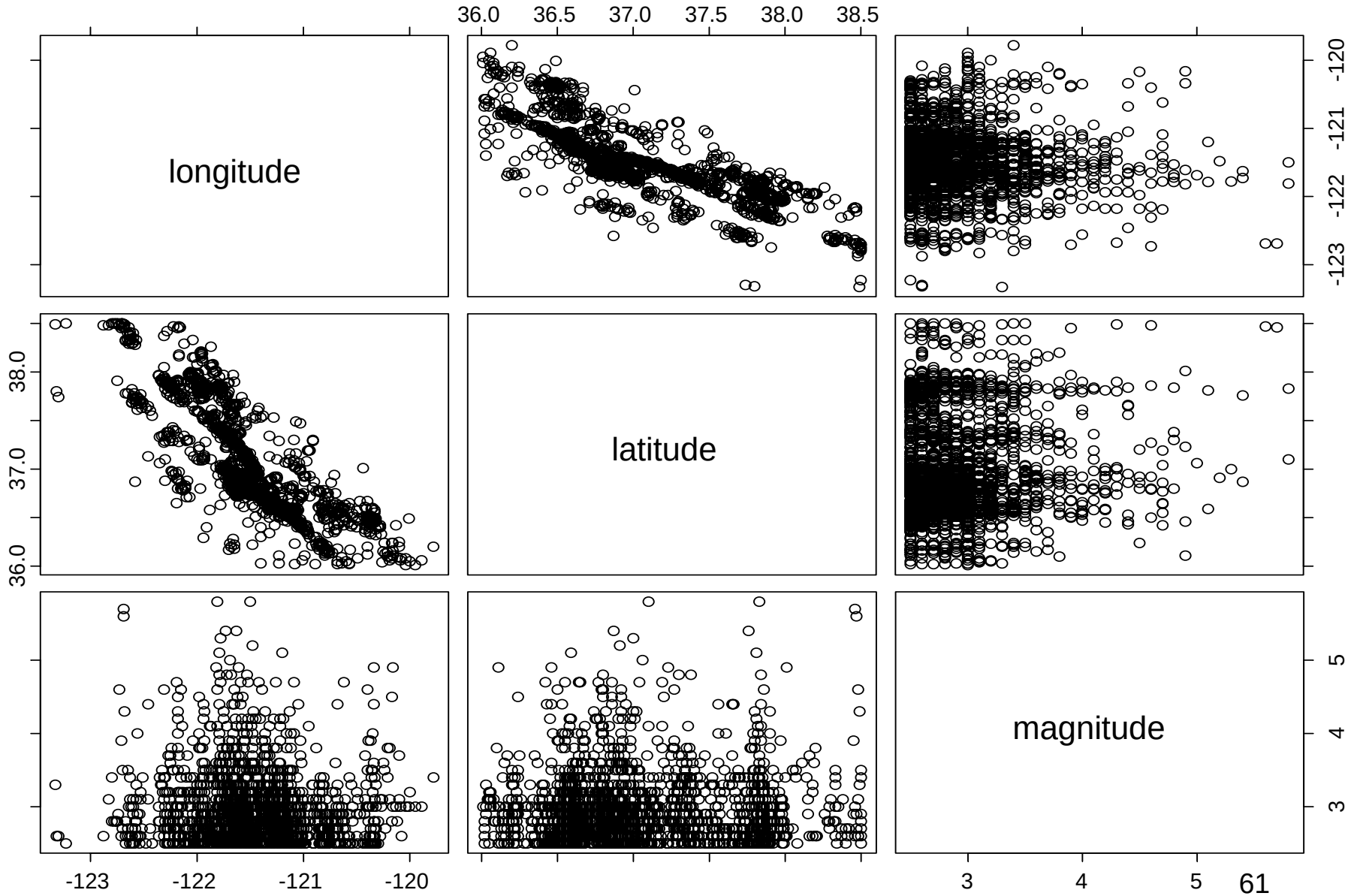
Galaxy Data



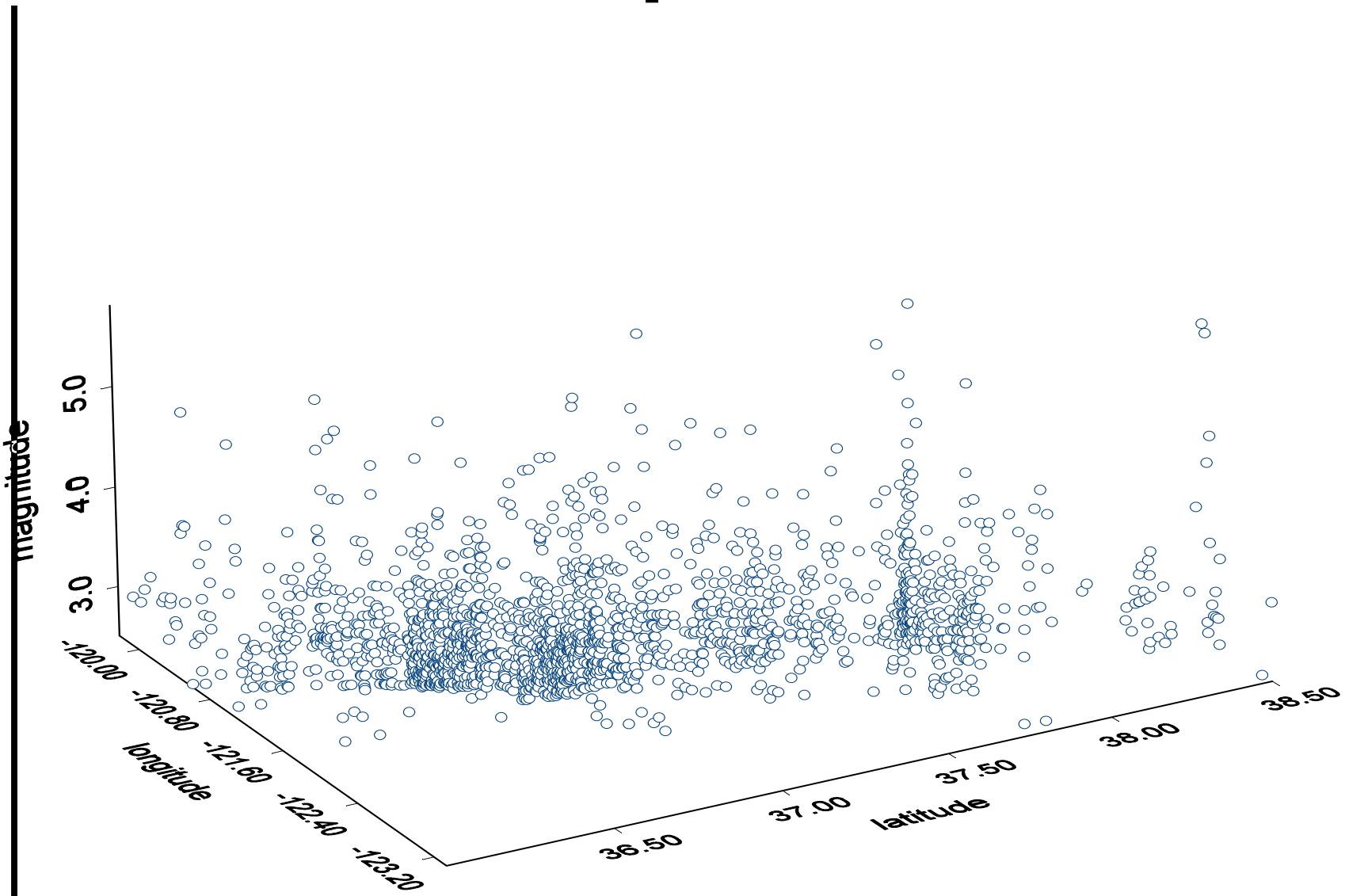
Galaxy 3D



Earthquake Data



Earthquake 3D



Narrative Graphics of Space and Time

- Adding spatial dimensions to a graph so that the data are moving over space and time can enhance the explanatory power of time series displays
- The Classic of Charles Joseph Minard (1781-1870) shows the terrible fate of Napoleon's army during his Russian campaign of 1812. A copy of the map is available at http://www.math.yorku.ca/SCS/Gallery/

Map Source: Minard, C. J. *Carte figurative des pertes successives en hommes de l'armée qu'Annibal conduisit d'Espagne en Italie en traversant les Gaules (selon Polybe). Carte figurative des pertes successives en hommes de l'armée française dans la campagne de Russie, 1812-1813*. École Nationale des Ponts et Chaussées (ENPC), 1869. Also available in: Tufte, Edward R. *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press, 2001.

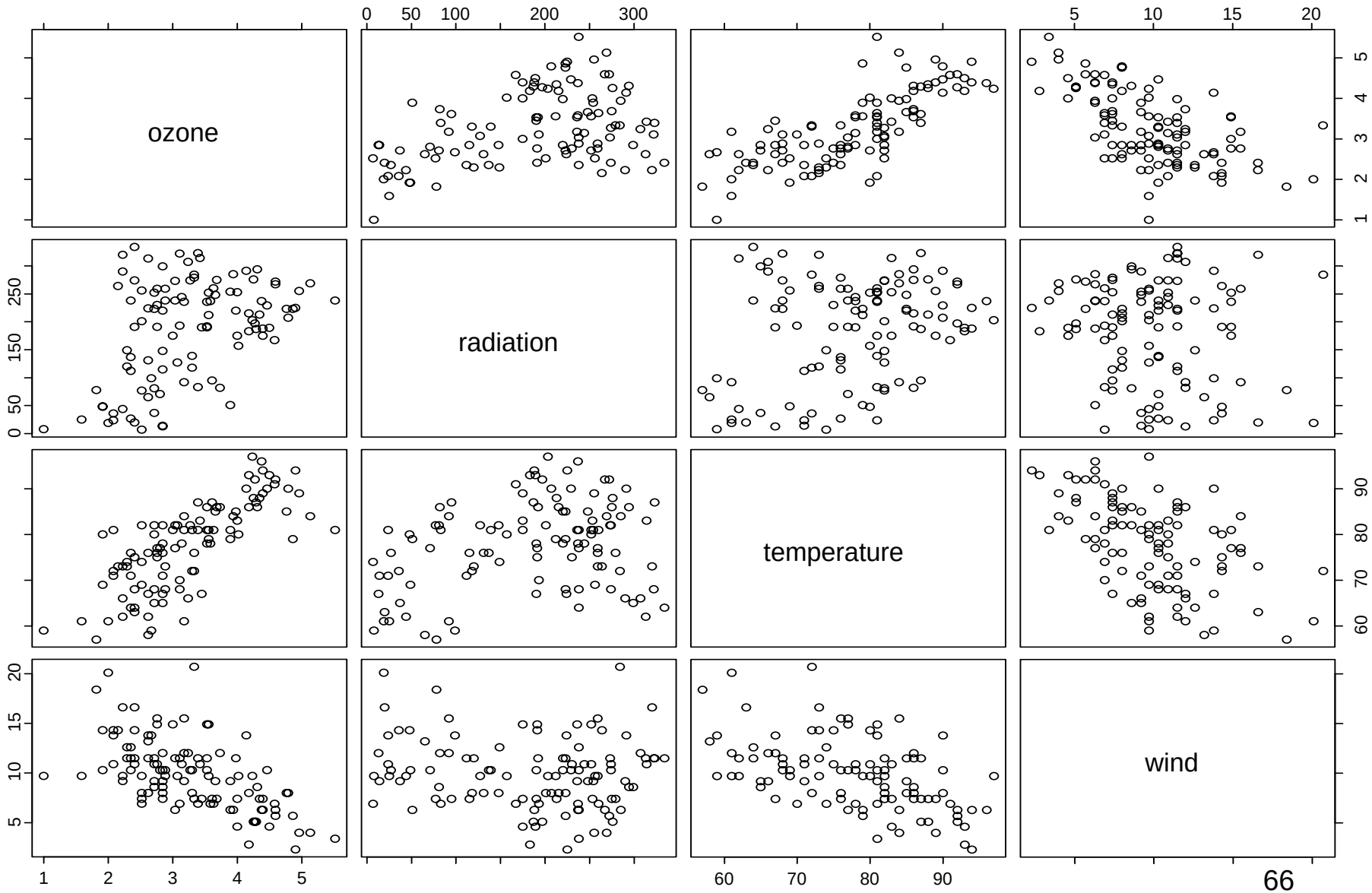
Beginning at the left on the Polish-Russian border near the Niemen River the thick band shows the size of the army (422,000) as it invaded Russia in June 1812.

- The width of the band indicates the size of the army…
- The army reached a sacked and deserted Moscow with 100,000 men
- Napoleon's retreat path from Moscow is depicted by a dark, lower band, linked to a temperature scale and dates at the bottom.
- The men struggled into Poland with only 10,000 troops remaining.

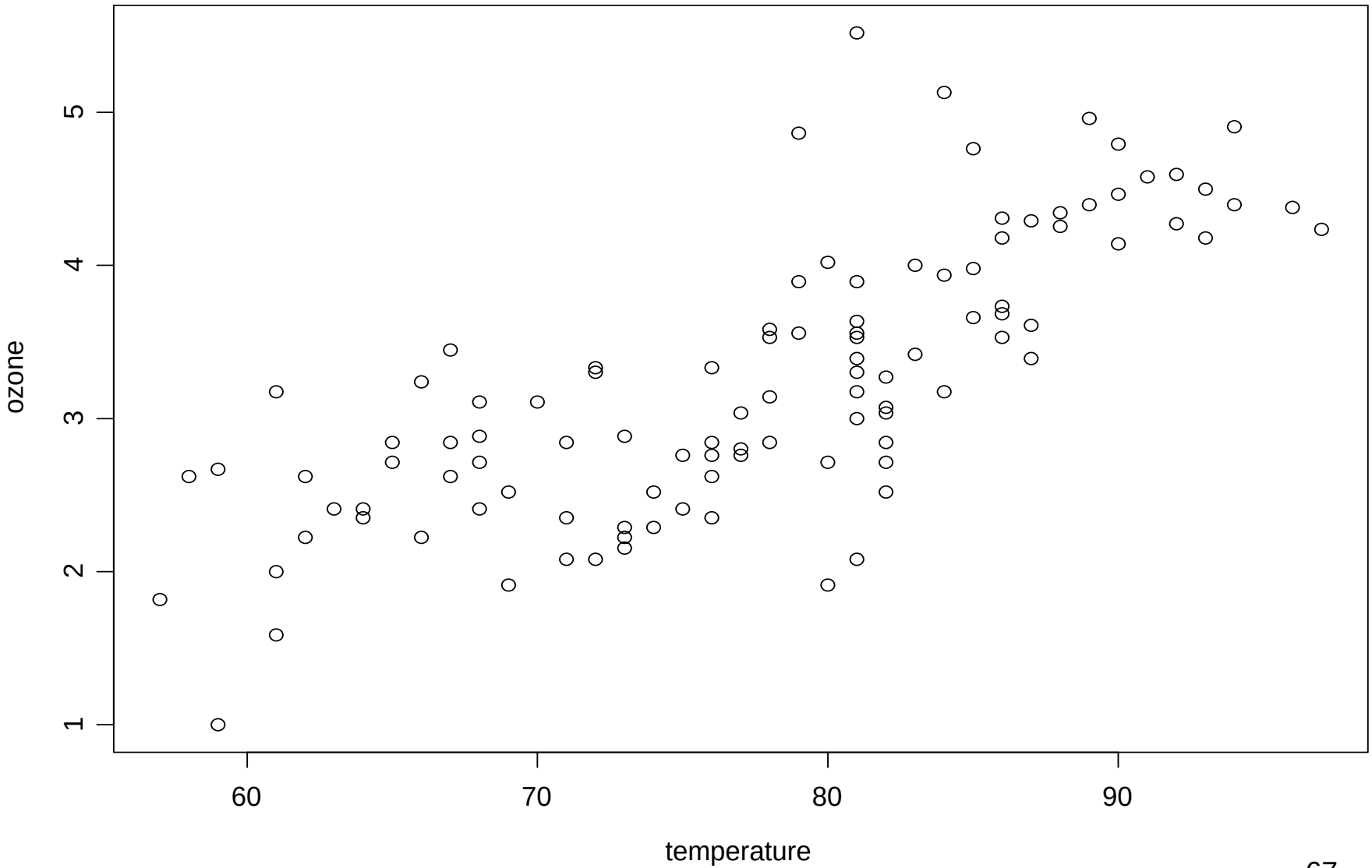
- **Minard's graphic tells a rich, coherent story with its multivariate data, far more enlightening than just a single number**

- ***SIX* variables are plotted:**
  - ***Its location on a two-dimensional surface***
  - ***Direction of army's movement***
  - ***Temperature as a function of time during the retreat***
  - ***The size of the army***
- ***"It may well be the best statistical graphic ever drawn." Edward Tufte*** (*The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press, 2001, pp. 40)

Scatter plot matrix of air data set in S-Plus



plot(temperature,ozone)



Fitting Lines

We often try to fit a straight line to bivariate data as a way to summarize bivariate data:

y = data = fit + residual

fit = $a + bx$

The parameter (coefficients) a and b can be found in many ways. Least-squares is commonly used.

$$\min_{a,b} \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

$$b = S_{xy} / S_x$$

$$a = \bar{y} - b\bar{x}.$$

The fit is often denoted by $\hat{y}_i = a + bx_i$. The residuals are $y_i - \hat{y}_i$.
What about curvature and outliers?

Resistant Line

Divide x data into thirds. Find median of x in each third, and median of the y's that correspond to the x's in each third. Call these three pairs (x_a, y_a), (x_b, y_b), (x_c, y_c). Fit a least-squares line to these three points.

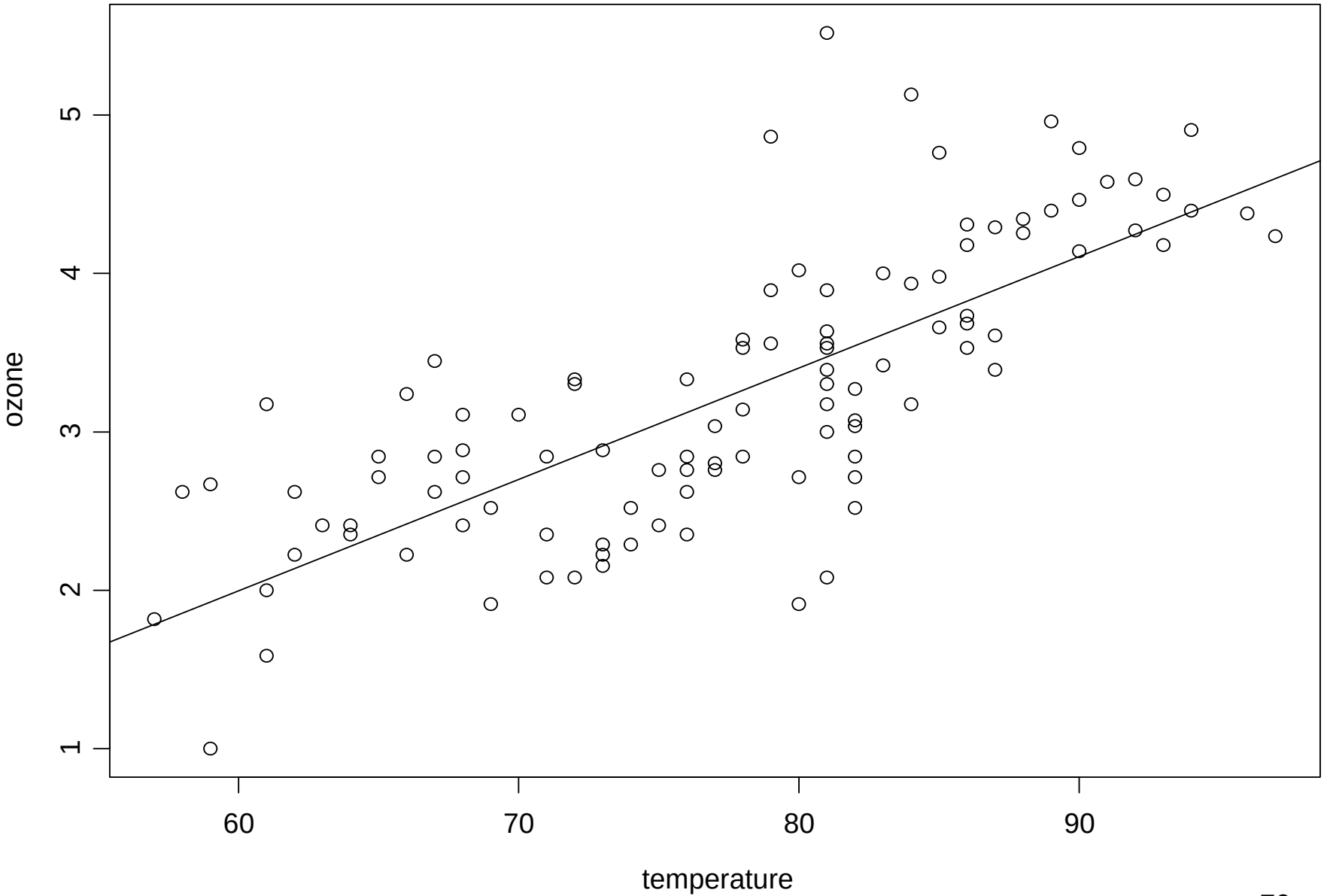
Or consider other metrics

$$\min_{a,b} \sum_{i=1}^{n} |y_i - a - bx_i|$$

$$\min_{a,b} \operatorname{median}_{i} |y_i - a - bx_i|.$$

These are alternatives to least-squares.

abline(lm(ozone~temperature))



Prediction and Residuals

Fitted lines can be used to predict. If we go too far beyond range of *x*-data, we can expect poor results. Consider problems of interpolation and extrapolation.

Examination of residuals help tell us how well our model (a line) fits the data.
We also compute

$$s = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

and call *s* the standard deviation of the residuals. Note use of $n - 2$ because two degrees of freedom are used to find *a* and *b*.

Residual Plots

1. against fitted values $(\hat{y}_i)$
2. against explanatory variable
3. against other possible explanatory variables
4. against time, if applicable.

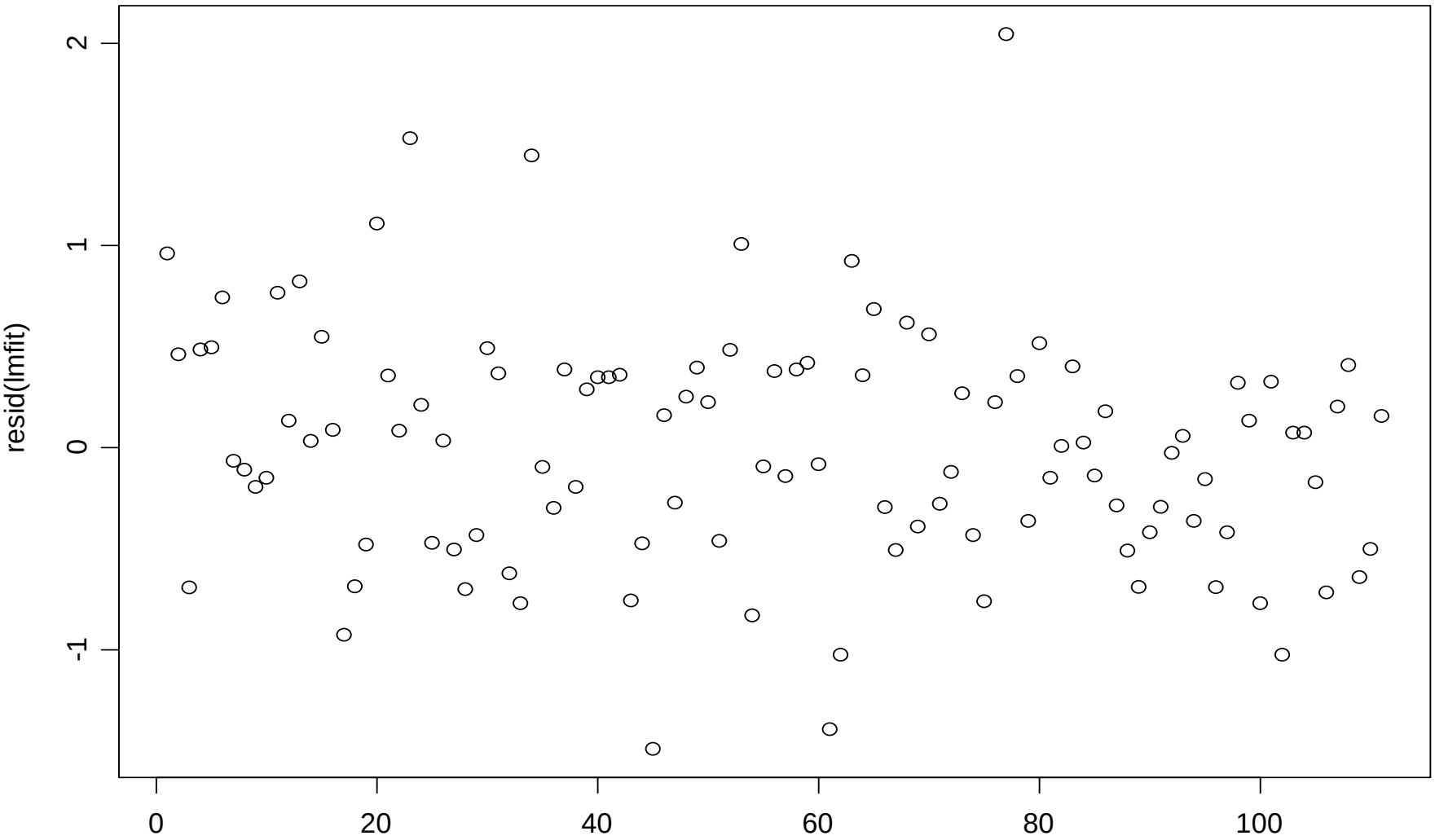
We want these pictures to look random — no pattern.

Outliers and Influence

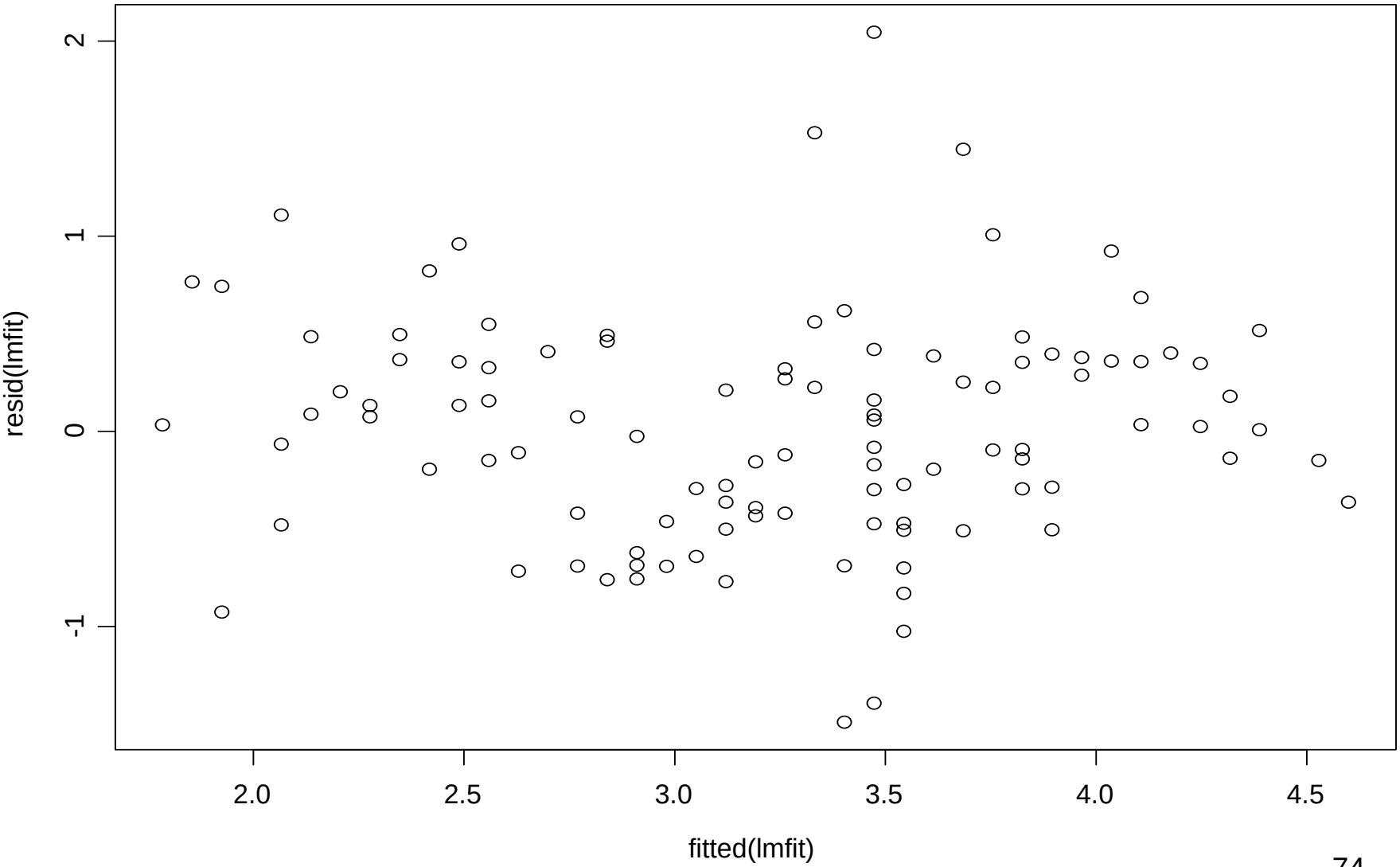
Values of *x* far away from the line have a lot of leverage on the line. Values of *y* with large residuals at high leverage points will usually be quite influential on the fitted line.

We can check by setting influential points aside and comparing fits and residuals.

Plot of residuals vs. observation number for ozone data



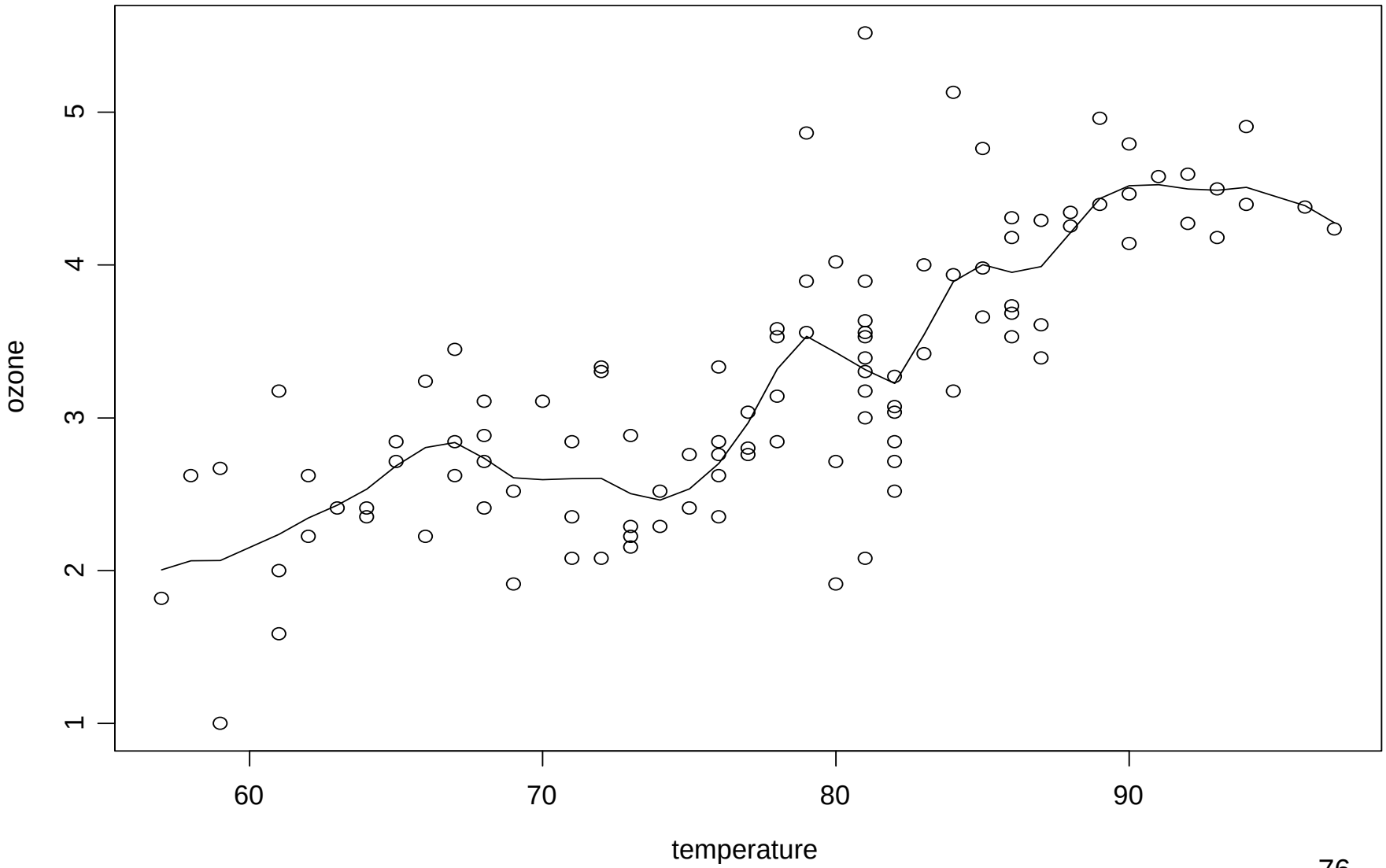
Residuals vs. Fitted Values for ozone data



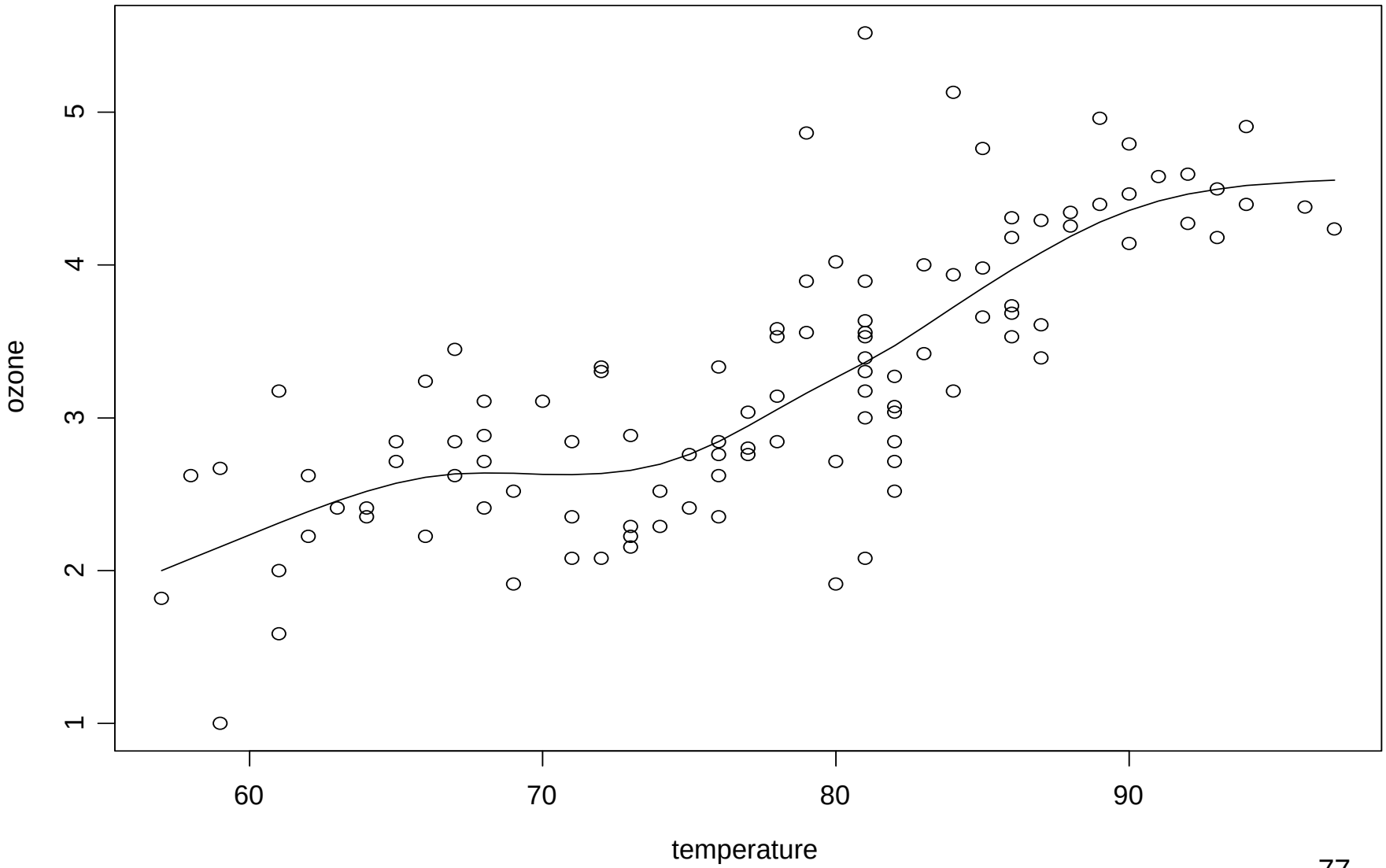
Smoothing

- **Fitting curves to data**
- **Separate Signal from noise**
- **Fitted values, $\hat{y}$, are a weighted average of the response y.**
- **Weights are a function of predictor x.**
- **Degrees of freedom indicate roughness**
- **Simple linear regression, df=2**

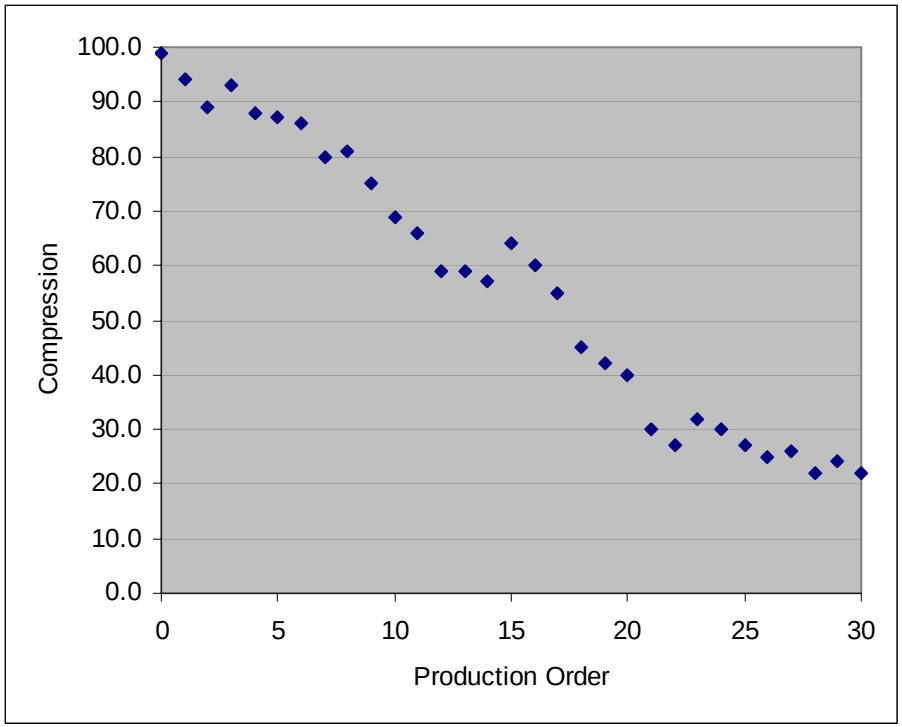
```
plot(temperature,ozone)
lines(smooth.spline(temperature,ozone,df=16.5))
```



plot(temperature,ozone)
lines(smooth.spline(temperature,ozone,df=6))



Time-Series/Runs Chart



Plot of Compression vs. Time (Order of Production)

This is example of a process not in "statistical control" as seen from the downward drift.

The usual statistics procedures (such as means, standard deviation, confidence interval, hypothesis testing) should NOT be applied until the process has been stabilized.

Time-Series Data

Data obtained at successive time points for the same sampling unit(s).

A time series typically consists of the following components.

1. Stable component
2. Trend component
3. Seasonal component
4. Random component
5. Cyclic (long term) component

Univariate time series $\{ x_t, t = 1, 2, \ldots, T \}$

Time-series plot: X_t vs. Time

Data Smoothing and Forecasting

Two types of averages for time-series data:

1. Moving averages
2. Exponentially weighted averages

These should be used only if mean is constant (process is in “statistical control” or is stationary) or mean varies slowly.

Regression techniques can be used to model trends.

More advanced methods are needed to model seasonality and dependence between successive observations (autocorrelation).

(Arithmetic) Moving Averages (MA)

The average of a set of *w* successive data values (called a window); the oldest data is successively dropped off.

$$MA_t = \frac{x_{t-w+1} + \ldots + x_t}{w} \quad \text{for} \quad \mathrm{t = w, w+1, \ldots, T}$$

The bigger the window (w), the more the smoothing.

MA forecast: $\hat{x}_t = MA_{t-1}$

Forecast error: $e_t = x_t - \hat{x}_t = x_t - MA_{t-1}, \quad \mathrm{t = 2, \cdots, T}$

Mean Absolute Percent Error: $\left(\frac{1}{T-1} \sum_{t-2}^{T} \left| \frac{e_t}{\mathbf{x}_t} \right| \right) \times 100\%$

(error in eqn 4.12 in textbook,
x not **y** in the denominator)

Exponentially Weighted Moving Averages

Uses all data, but the most recent data is weighted the heaviest.

$$EWMA_t = w\,x_t + (1-w)EWMA_{t-1}$$

where $0 < w < 1$ is the smoothing constant (usually 0.2 to 0.3).

EWMA forecast: $\hat{x}_t = EWMA_{t-1}$

Forecast error: $e_t = x_t - \hat{x}_t = x_t - EWMA_{t-1}$

Alternative formula: $EWMA_t = w\,e_t + EWMA_{t-1}$

Interpretation: If the forecast error is positive (forecast underestimated the actual value), the next period's forecast is adjusted upward by a fraction of the forecast error.

Autocorrelation Coefficient

For time-series data, observations separated by a specified time period (called a lag) are said to be lagged.

First-order autocorrelation or the serial correlation coefficient between observations with lag = 1:

$$r_1 = \frac{\sum_{t=2}^{T}(x_{t-1} - \bar{x})(x_t - \bar{x})}{\sum_{t=1}^{T}(x_t - \bar{x})^2}$$

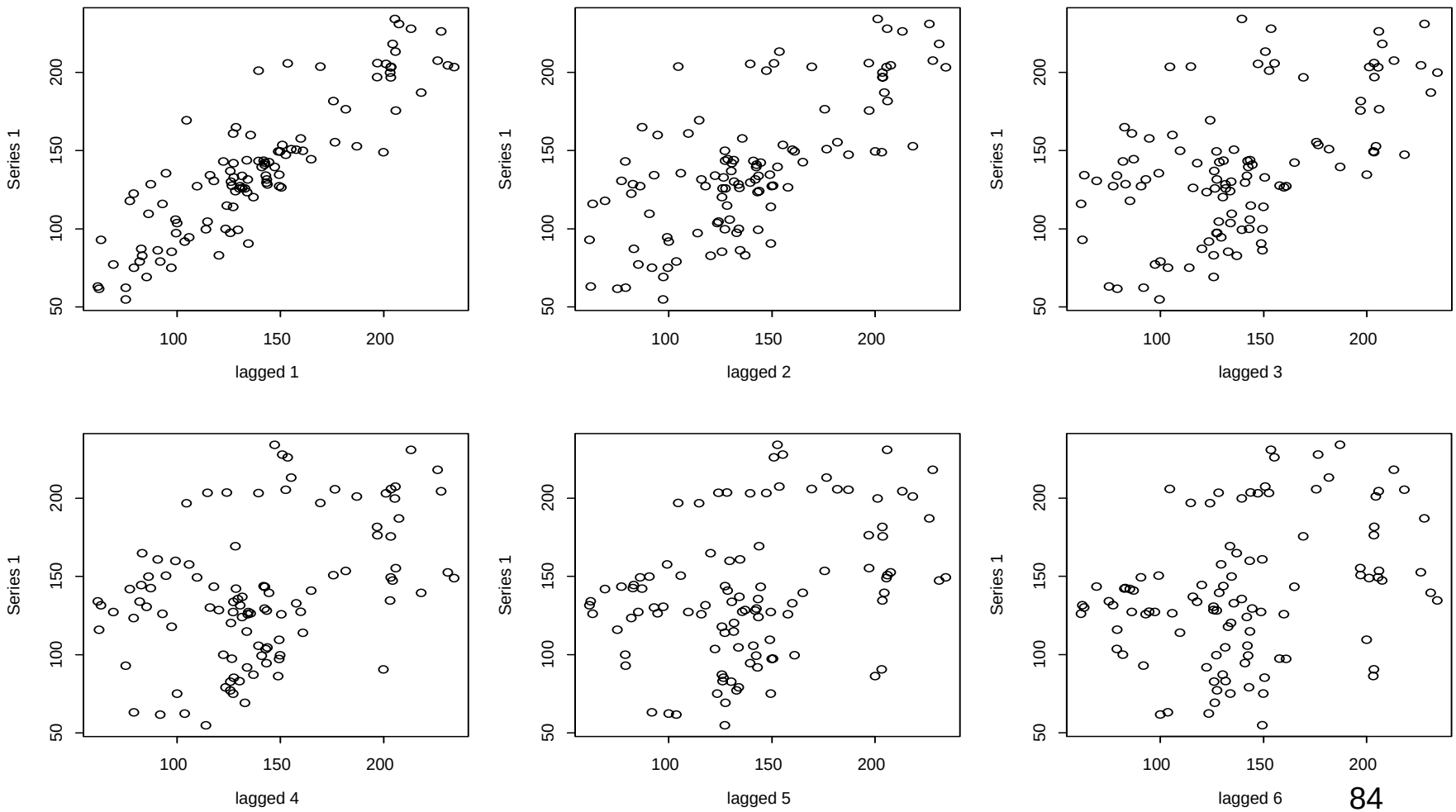
The k-th order autocorrelation coefficient:

$$r_k = \frac{\sum_{t=k+1}^{T}(x_{t-k} - \bar{x})(x_t - \bar{x})}{\sum_{t=1}^{T}(x_t - \bar{x})^2}$$

Lag Plots in S-Plus

lag.plot(x) or plot(x[1:(n-i)],x[(i+1):n])

Housing starts 1966:1974, lagged scatterplots



John W. Tukey (1915 - 2000)

Statistician at Princeton Univ. and Bell Labs

Co-developer of Fast Fourier Transform

Coined terms “bit” (binary digit) and “software”

“An approximate answer to the right problem is worth a great deal more than a precise answer to the wrong problem.”

Developed new graphical displays (stem-and-leaf and box plots) to examine the data, as a reaction to the “mathematization of statistics.”