Lecture Notes

Visualizing, Exploring and Summarizing Data

Roy Welsch

Spring 2017

SUTD-MIT

Visualization

Visualization is the conversion of data into a visual or tabular format so that the characteristics of the data and the relationships among data items or attributes can be analyzed or reported.

- Visualization of data is one of the most powerful and appealing techniques for data exploration.
  - Humans have a well developed ability to analyze large amounts of information that is presented visually
  - Can detect general patterns and trends
  - Can detect outliers and unusual patterns

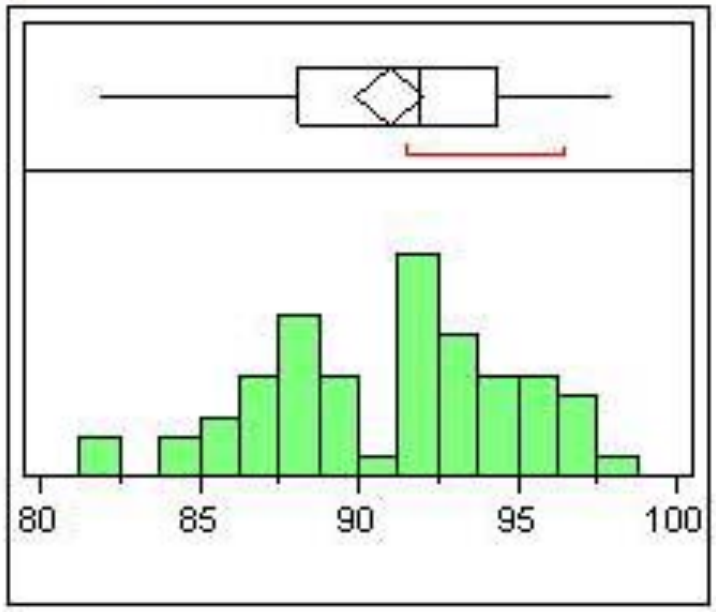
Graphical Excellence

"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

Distributions

Grades

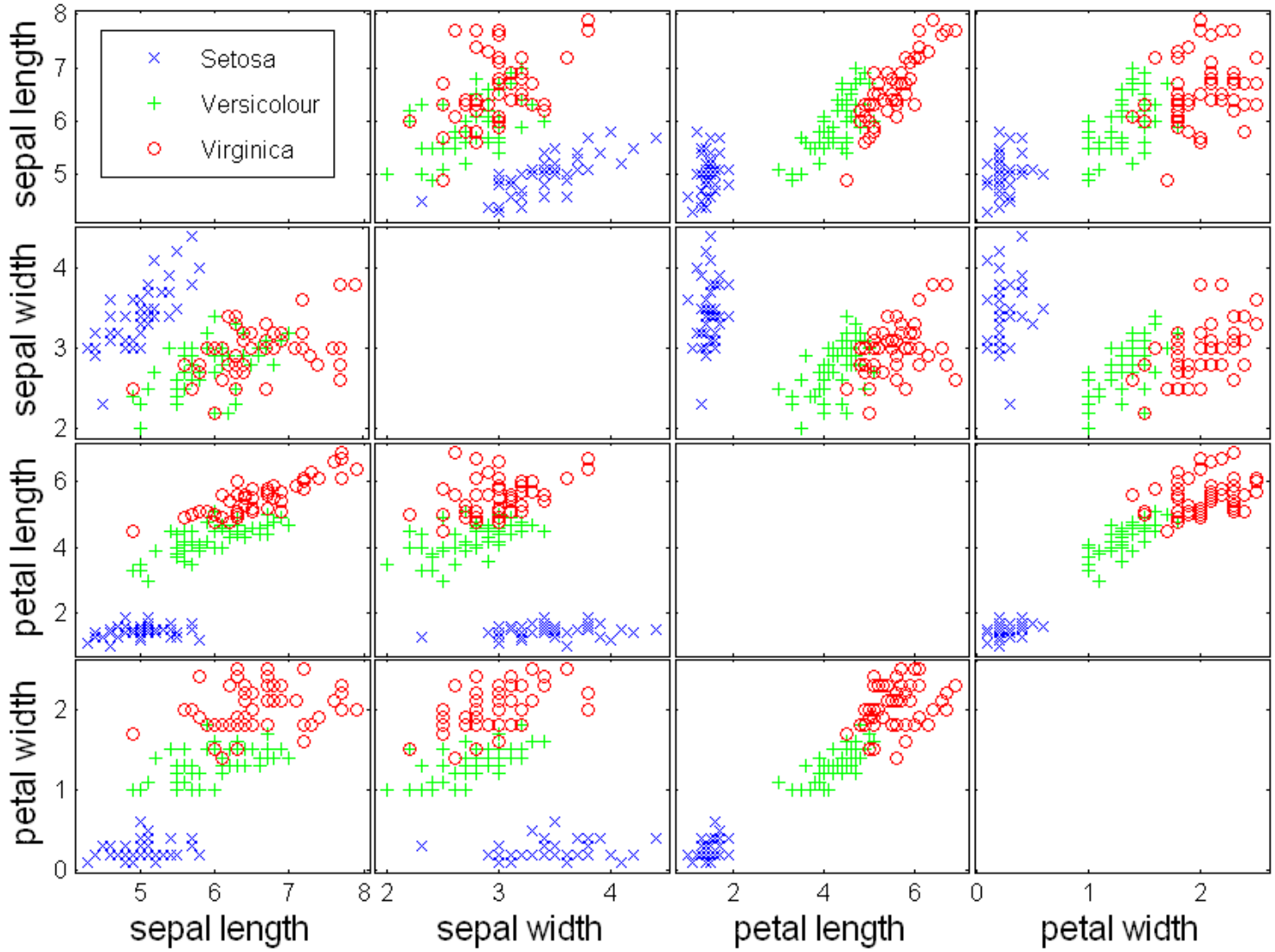


Stem and Leaf

Stem	Leaf	Count
98	0	1
97	1	1
96	345	3
95	33599	5
94	47999	5
93	116	3
92	00122448999	11
91	3569	4
90	9	1
89	029	3
88	15666789	8
87	24588	5
86	179	3
85	48	2
84	3	1
83	9	1
82	0	1
81	8	1

81|8 represents 81.8

Scatter Plot Array of Iris Attributes



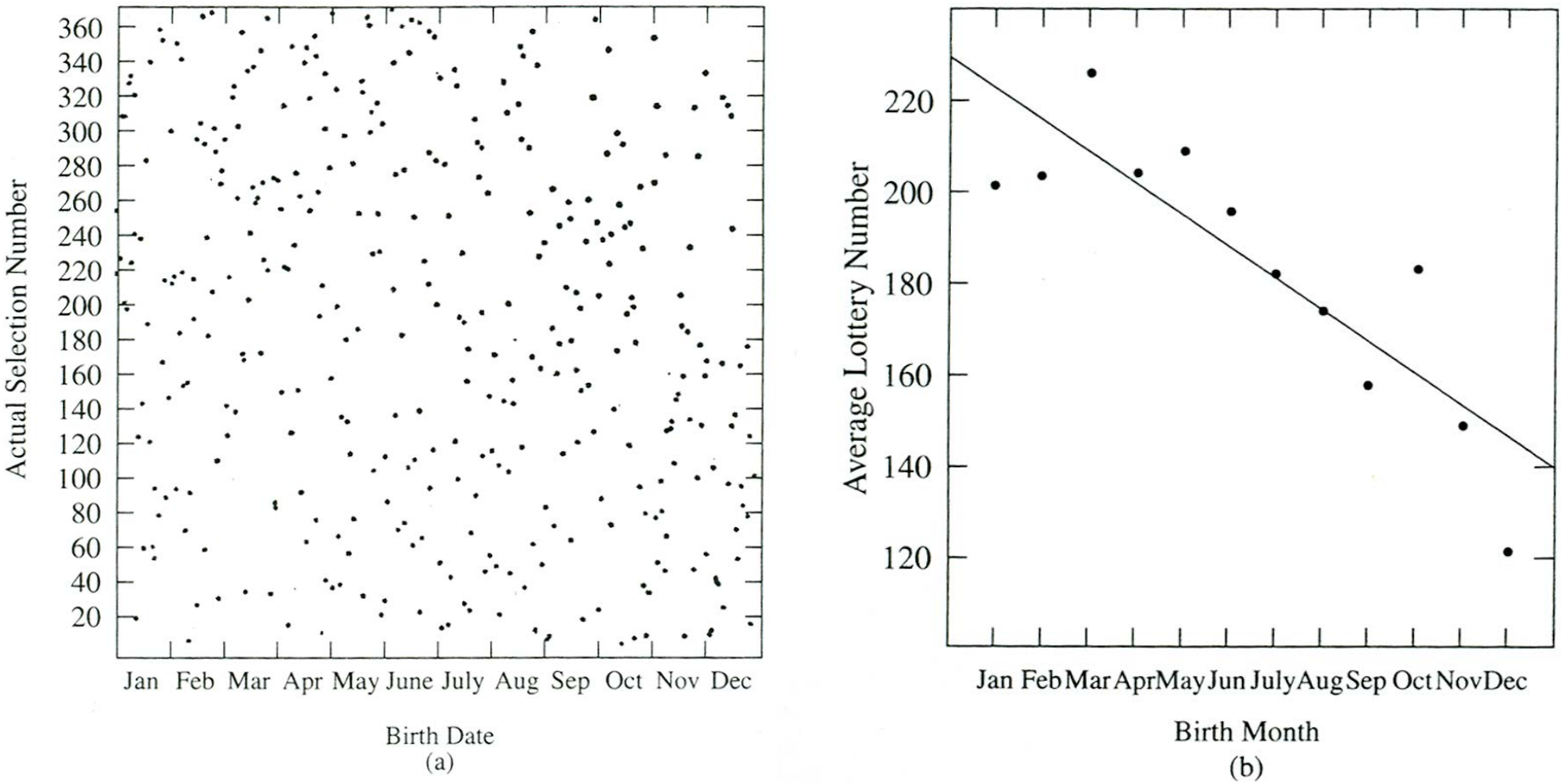


Figure 4.17 (a) Scatter Plot of Actual Selection Number vs. Birth Date; (b) Scatter Plot of Average Lottery Number vs. Birth Month (*Source*: S. E. Fienberg (1971), "Randomization and social affairs: The 1970 draft lottery," *Science*, **171**, pp. 255–261)

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.

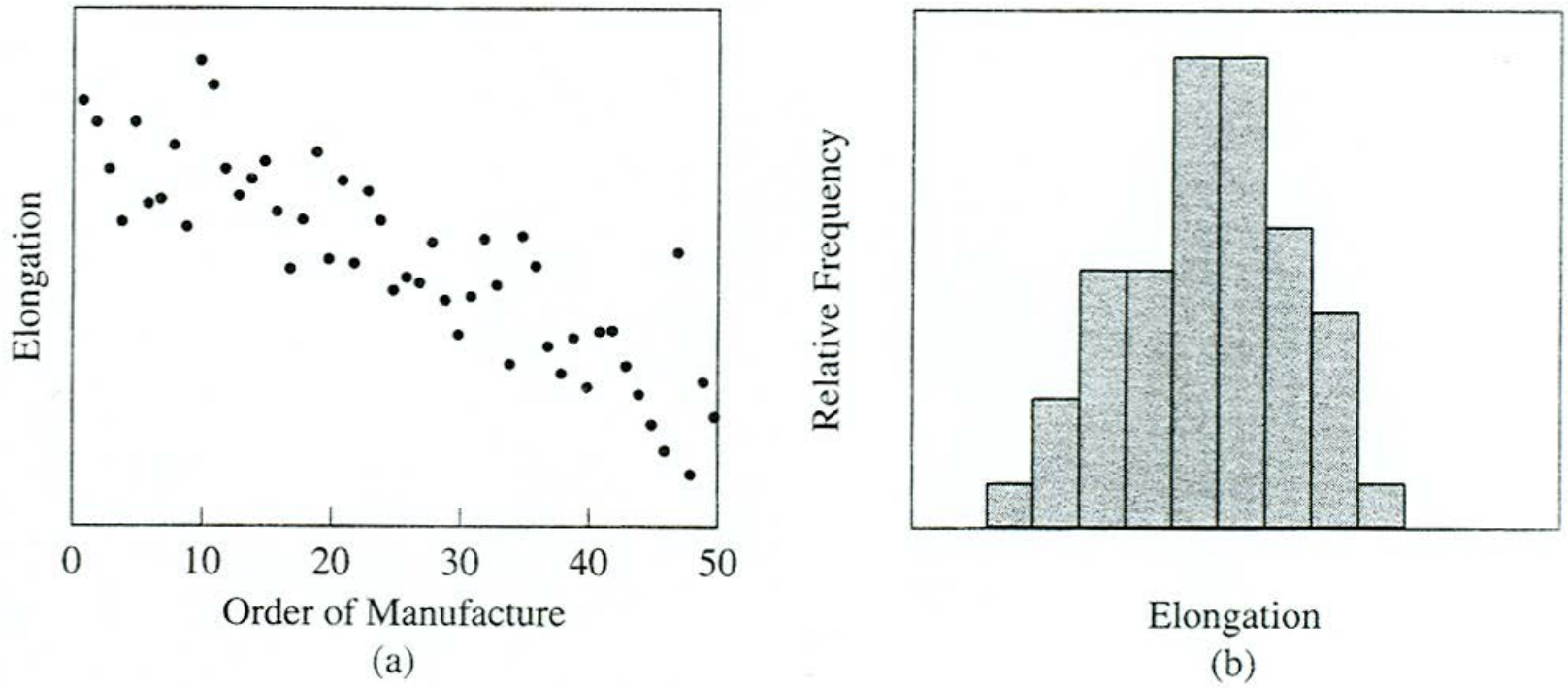


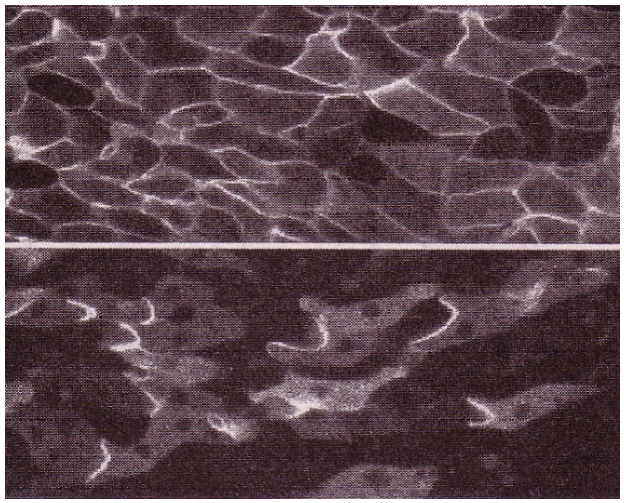
Figure (a) Run Chart and (b) Histogram of Time-Series Data

Always Plot Your Data Appropriately - Try Several Ways!

Other Visualization Techniques

- Star Plots
  - Axes (= # of dimensions) radiate from a central point
  - The line connecting the values of an object in each dimension is a polygon
- Chernoff Faces
  - Approach created by Herman Chernoff
  - This approach associates each attribute with a characteristic of a face
  - The values of each attribute determine the appearance of the corresponding facial characteristic
  - Each object becomes a separate face
  - Relies on human's ability to distinguish faces

Imaging Tissues

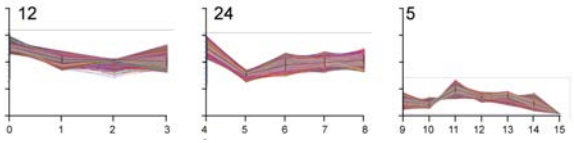


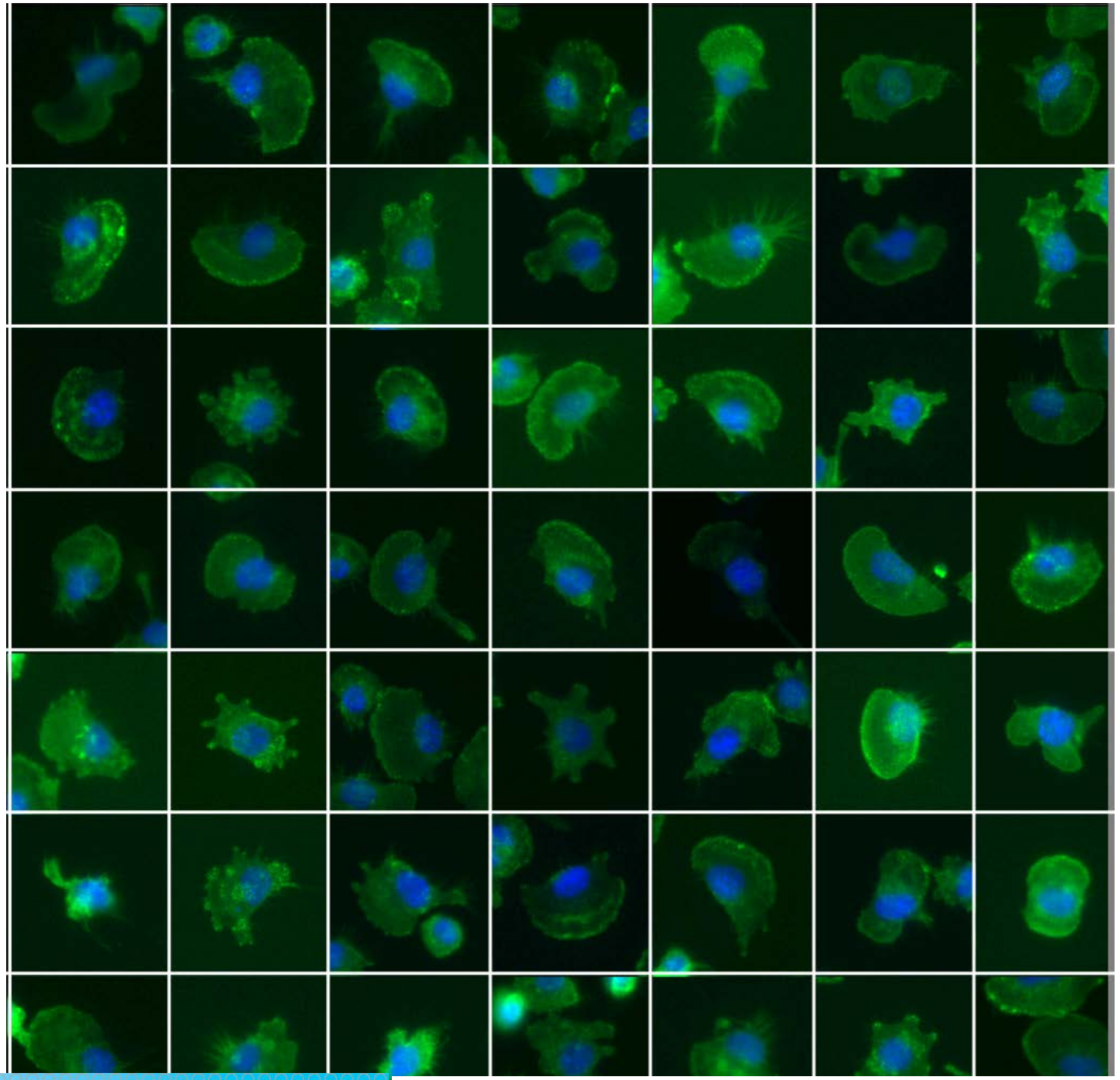
Features of almost any cellular components can now be imaged by automated light microscopy using selected fluorescence probes and filters.

It is difficult to follow a single cell over time. What we can have is a snapshot of thousands of cells at various stages in their life cycle and therefore the distribution of each feature is available.

If we induce a change in cell physiology, we would expect to see a change in the distribution of one or more of these features.

0-12-24-5





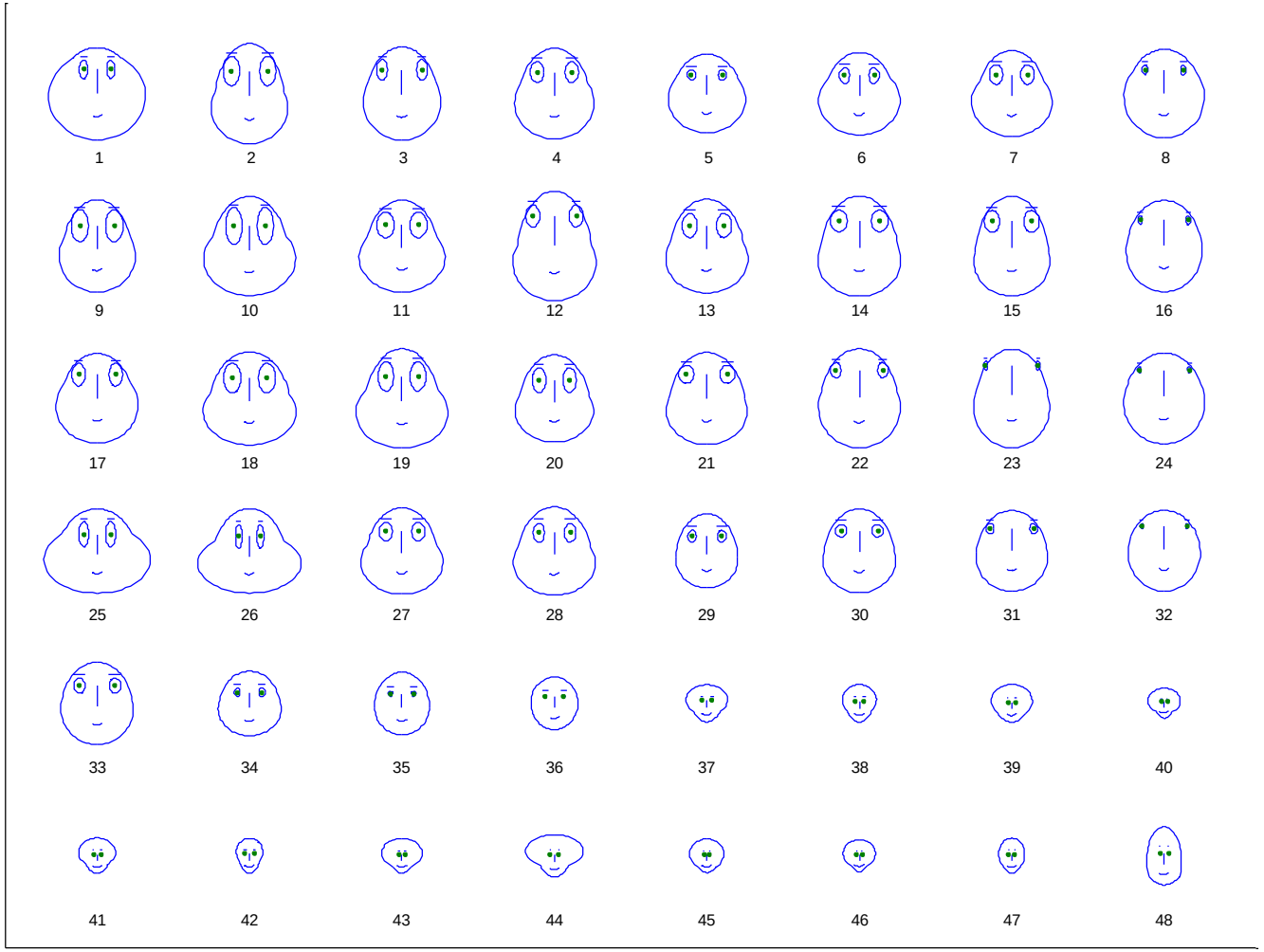
Actin = green
Nucleus = blue

Lots of data

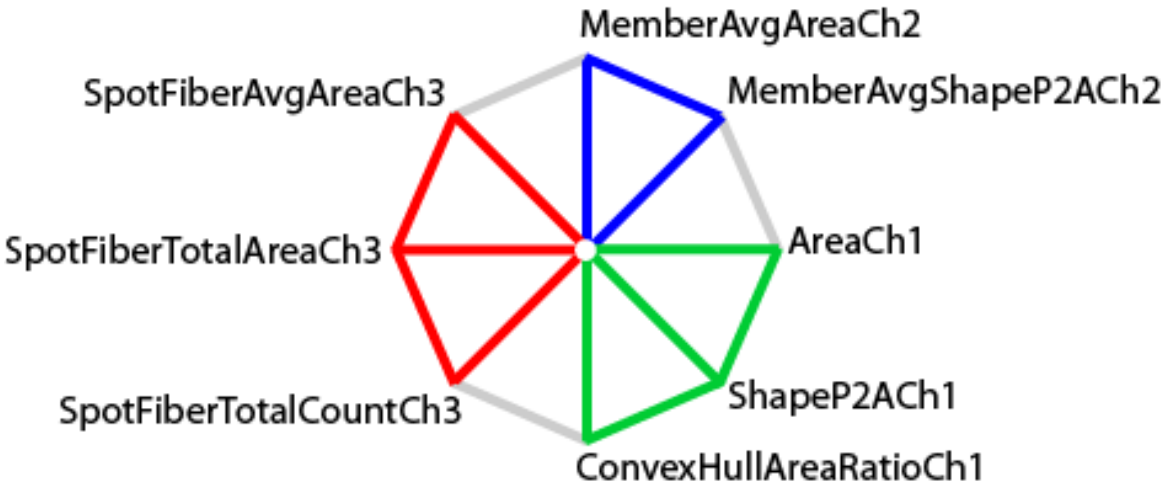
- 64 features per cell
- 48 dosage levels
- 1000s of cells per dosage level
- 4 time points

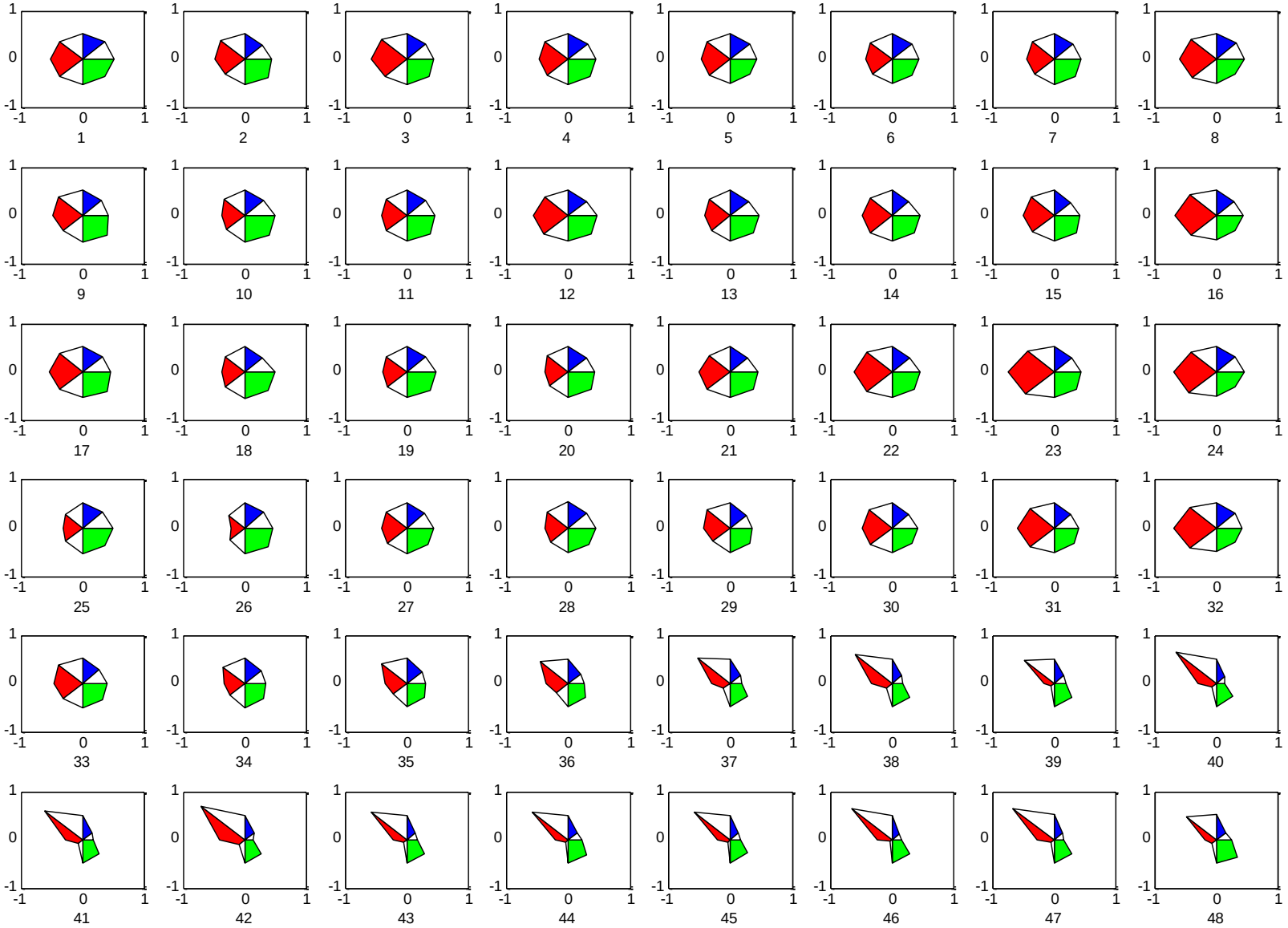
Face Legend

Column	Facial Feature
AreaCh1	Size of face
ShapeP2ACh1	Forehead/jaw relative arc length
ConvexHullAreaRatioCh1	Shape of forehead
MemberAvgAreaCh2	Shape of jaw
MemberAvgShapeP2ACh2	Width between eyes
SpotFiberCountCh3	Vertical position of eyes
SpotFiberTotalAreaCh3	Height of eyes
SpotFiberAvgAreaCh3	Width of eyes

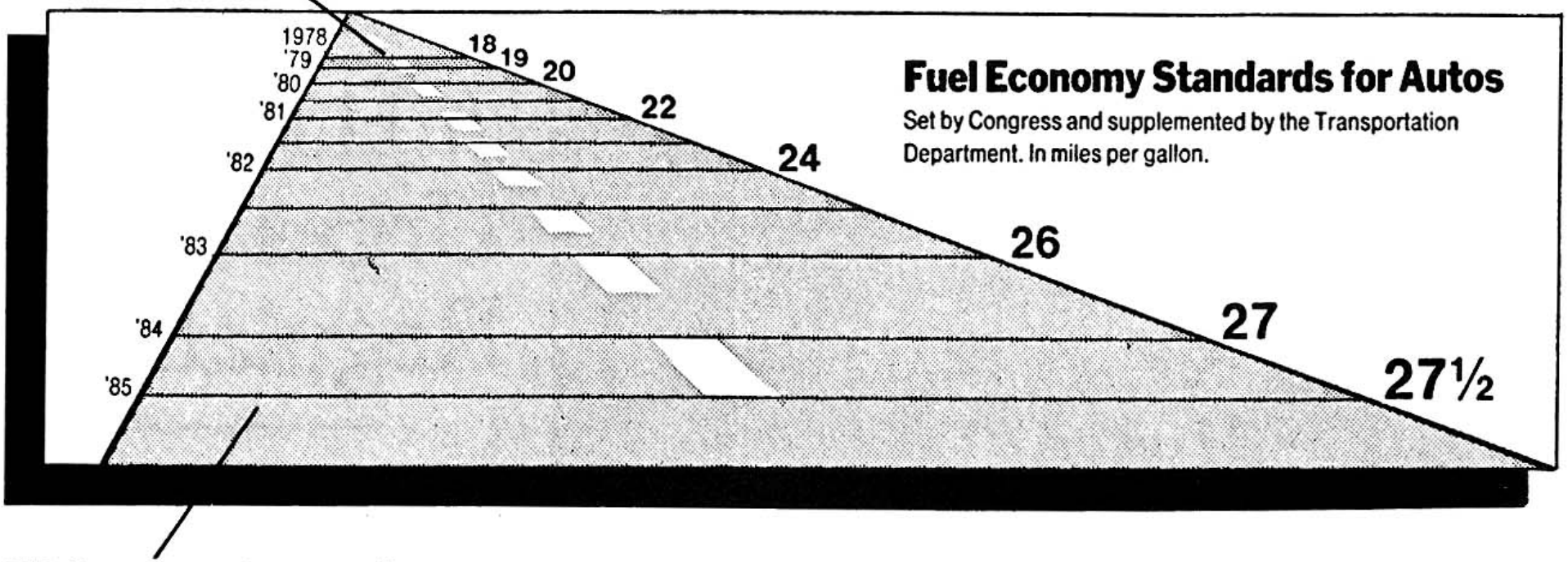

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48

Starplot Legend




1 2 3 4 5 6 7 8
9 10 11 12 13 14 15 16
17 18 19 20 21 22 23 24
25 26 27 28 29 30 31 32
33 34 35 36 37 38 39 40
41 42 43 44 45 46 47 48

This line, representing 18 miles per gallon in 1978, is 0.6 inches long.

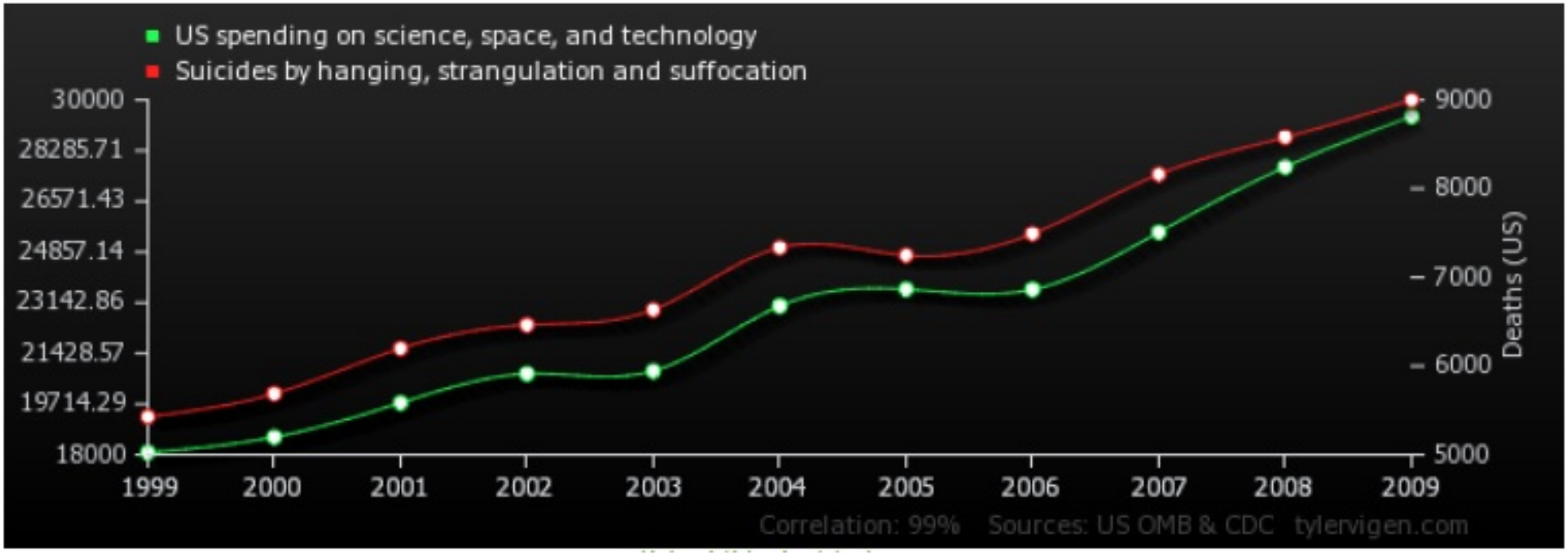


This line, representing 27.5 miles per gallon in 1985, is 5.3 inches long.

US Spending on science, space, & technology

correlates with

Suicides by hanging, strangulation, & suffocation

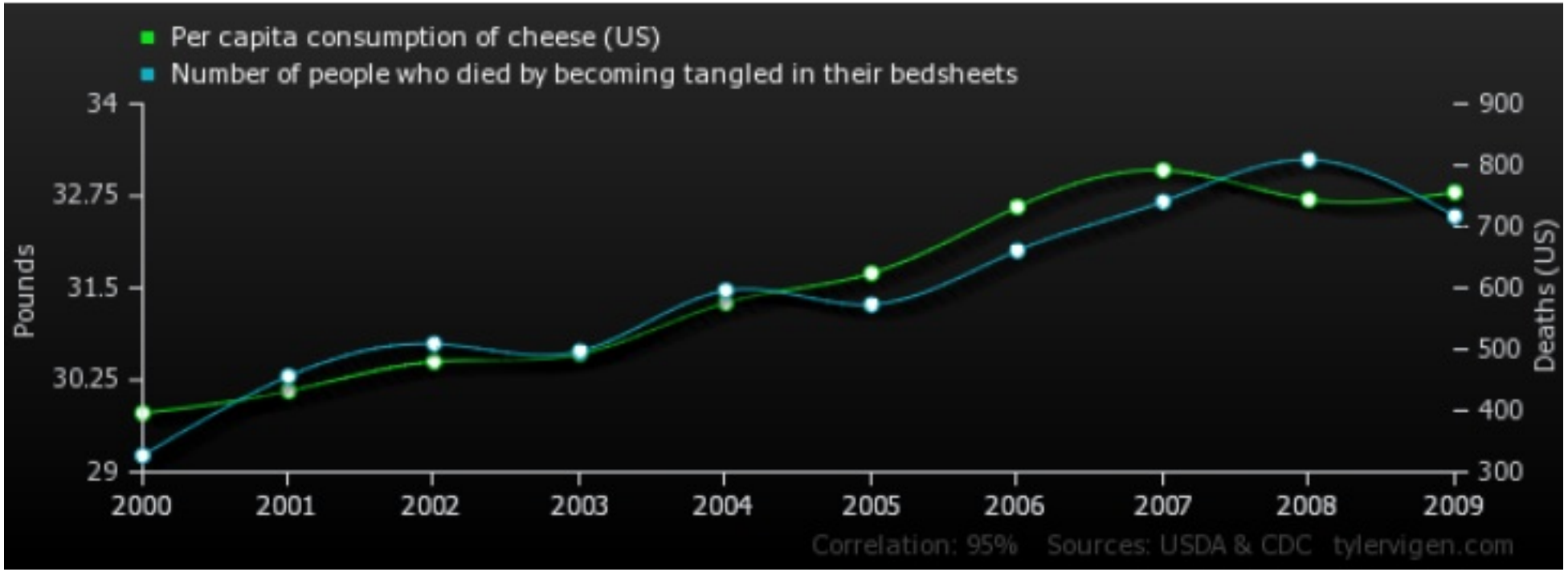


	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
US spending on science, space, & technology Millions of todays dollars (US OMB)	18,079	18,594	19,753	20,734	20,831	23,029	23,597	23,584	25,525	27,731	29,449
Suicides by hanging, strangulation & suffocation Deaths (US) (CDC)	5,427	5,688	6,198	6,462	6,635	7,336	7,248	7,491	8,161	8,578	9,000
Correlation: 0.992082											

Per capita consumption of cheese (US)

correlates with

Number of people who died by becoming tangled in their bed sheets



	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Per capita consumption of cheese (US) Pounds (USDA)	29.8	30.1	30.5	30.6	31.3	31.7	32.6	33.1	32.7	32.8
Number of people who died by becoming tangled in bedsheets Deaths (US) (CDC)	327	456	509	497	596	573	661	741	809	717
Correlation: 0.947091										

Lecture Notes

Statistical Reasoning and One Sample Analyses

Roy Welsch

Spring 2017

SUTD-MIT

Statistical View of the World

- Data are imperfect
  - We do the best we can -- Statistics helps!
- Events are random
  - Can't be right 100% of the time
- Use statistical methods
  - Along with **common sense** and **good judgment**
- Be skeptical!
  - Statistics can be used to support contradictory conclusions
  - Look at **who funded the study**?

Is It Real?

Engineering and management require:

1. data collection
2. data analysis

Problem:

Data always exhibits *variability*. Variability obscures our ability to make decisions.

Example:

Thicknesses of a particular type of silicon wafer
Target value: 244 μm
Consider a sample of five wafers from one batch.

245 250 250 247 248

All are larger than the target.
Is there real evidence of a problem?

Conventional Statistics

Assumptions of “conventional” statistics:

- Data randomly sampled from population described by a data generating process
- The process modeled by a hypothetical probability distribution (function) with a few population parameters such as the population mean and variance

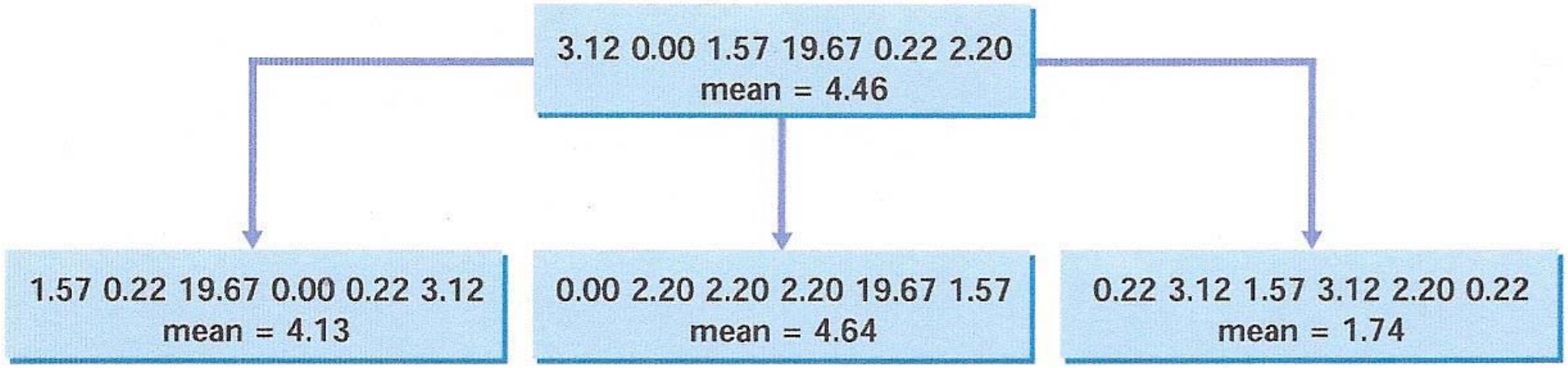
Thus, the basis of “conventional” inference is that samples are drawn at random from a larger population and the observations in the sample are then presumed to reflect the population (e.g., mean and variance).

Bootstrap Resampling Statistics

In resampling statistics, statistical estimates are formed by taking random samples directly from the data at hand.

In other words, you randomly sample your random sample!

Bootstrap Re-sampling with Replacement



The resampling idea. The top box is a sample of size $n = 6$ from the Verizon data. The three lower boxes are three resamples from this original sample. Some values from the original sample occur more than once in the resamples because each resample is formed by sampling with replacement. We calculate the statistic of interest—the sample mean in this example—for the original sample and each resample.

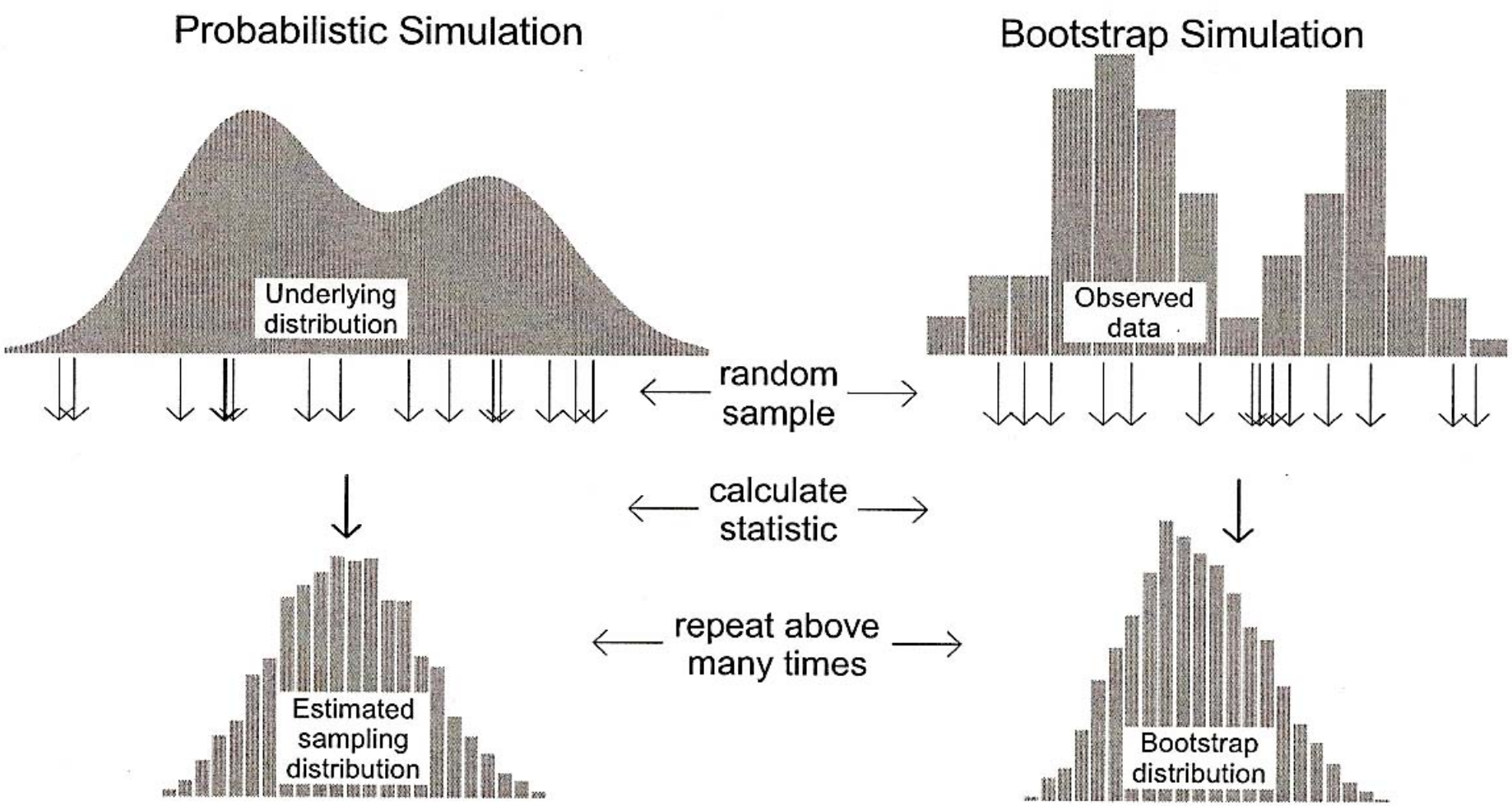


Figure 3: Diagram of probabilistic simulation and bootstrap sampling estimates of sampling distributions.

Example

TABLE 18.1 Selling prices (in $1000) for an SRS of 50 Seattle real estate sales in 2002

142	232	132.5	200	362	244.95	335	324.5	222	225
175	50	215	260	307	210.95	1370	215.5	179.8	217
197.5	146.5	116.7	449.9	266	265	256	684.5	257	570
149.4	155	244.9	66.407	166	296	148.5	270	252.95	507
705	1850	290	164.95	375	335	987.5	330	149.95	190

SRS = Simple Random Sample

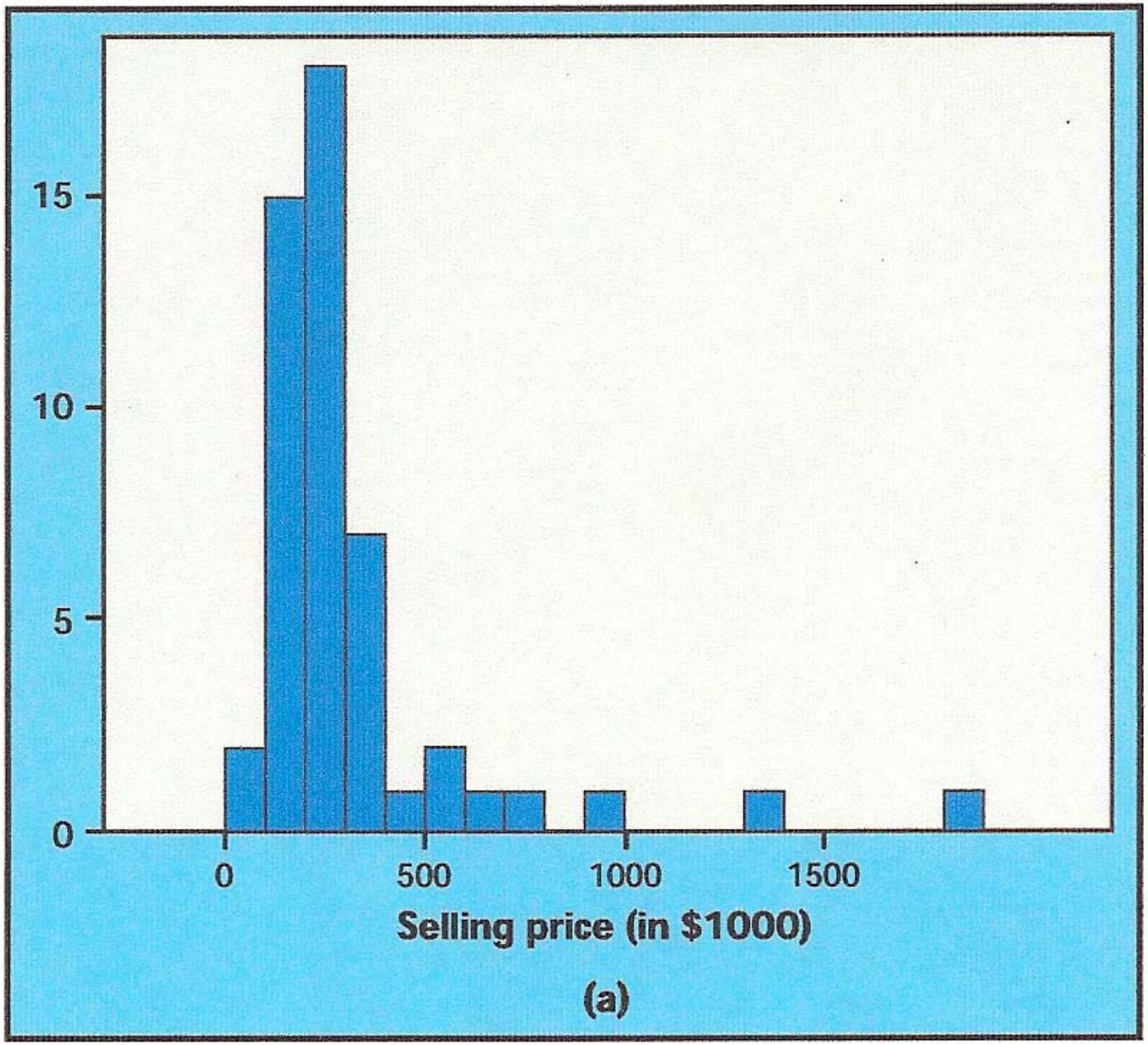
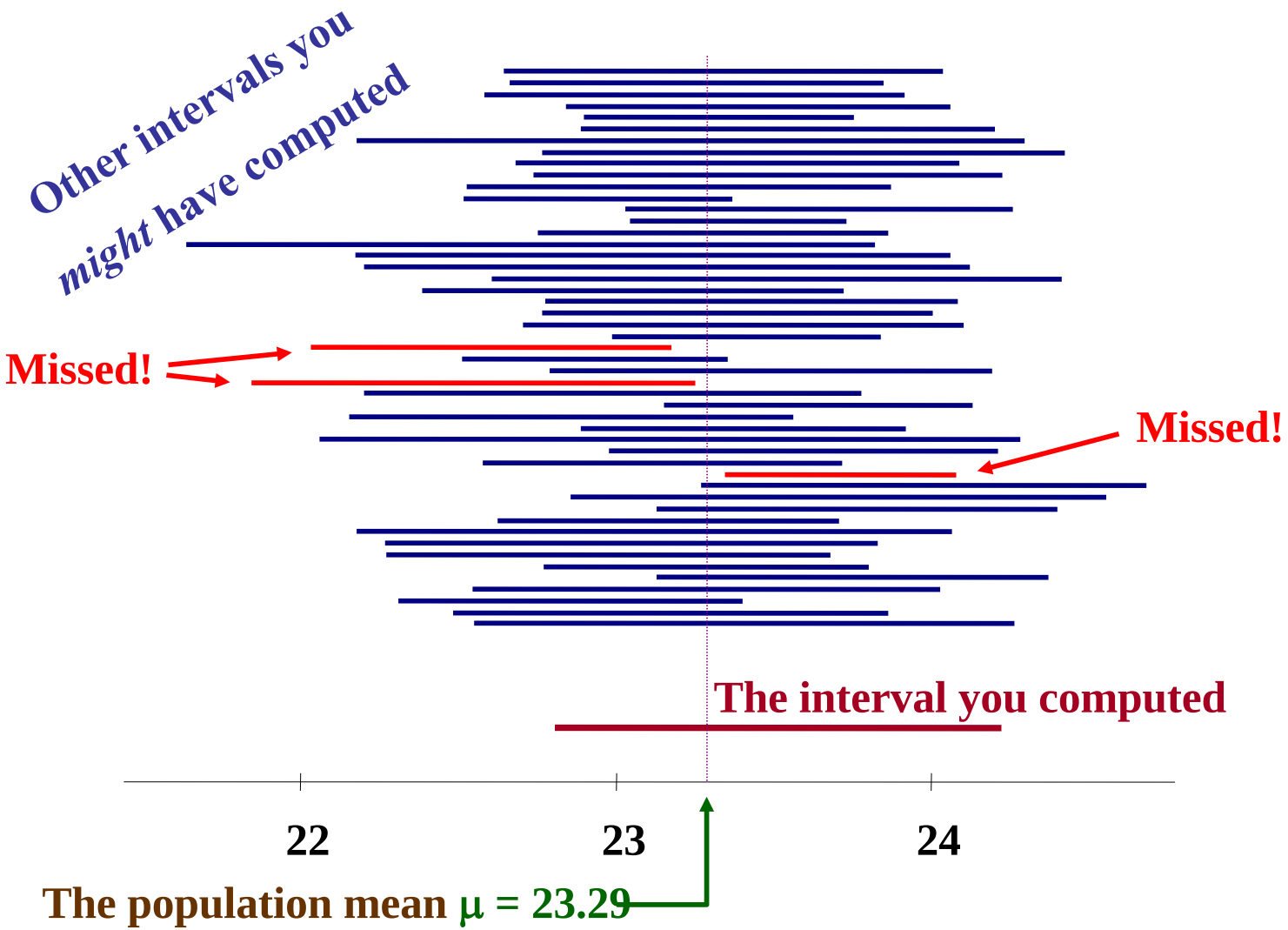


FIGURE 18.6 Graphical displays of the 50 selling prices in Table 18.1. The distribution is strongly skewed, with high outliers.

Imagine Many Samples



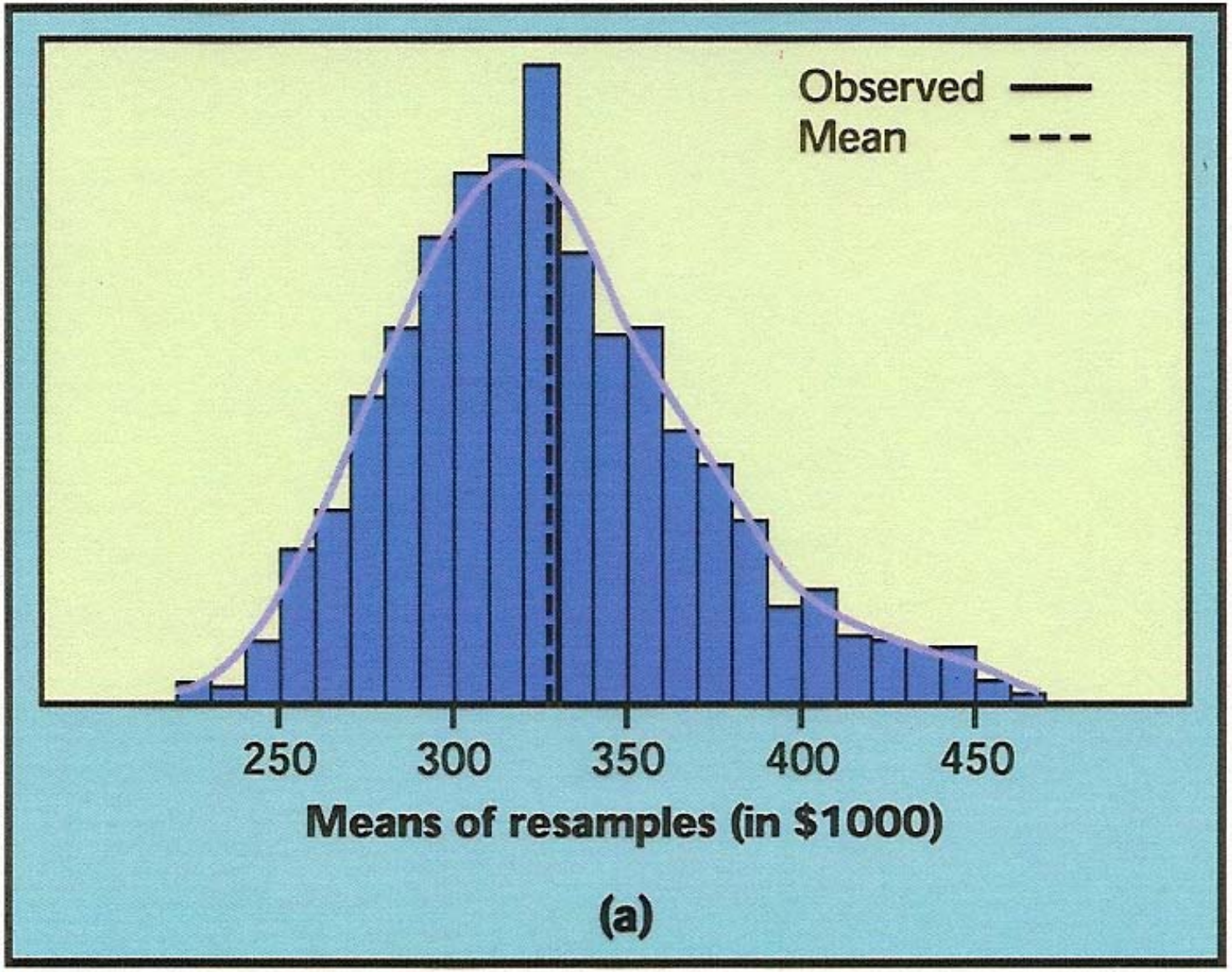


FIGURE 18.7 The bootstrap distribution of the sample means of 1000 resamples from the data in Table 18.1. The bootstrap distribution is right-skewed, so we conclude that the sampling distribution of $\overline{x}$ is right-skewed as well.

25% Trimmed Mean

Order the sample data from smallest to largest and remove the lowest 25% of the data and the highest 25% of the data. Then take the mean of the rest (middle 50%).

This is known as a robust estimate of the sample mean with a 25% breakdown point.

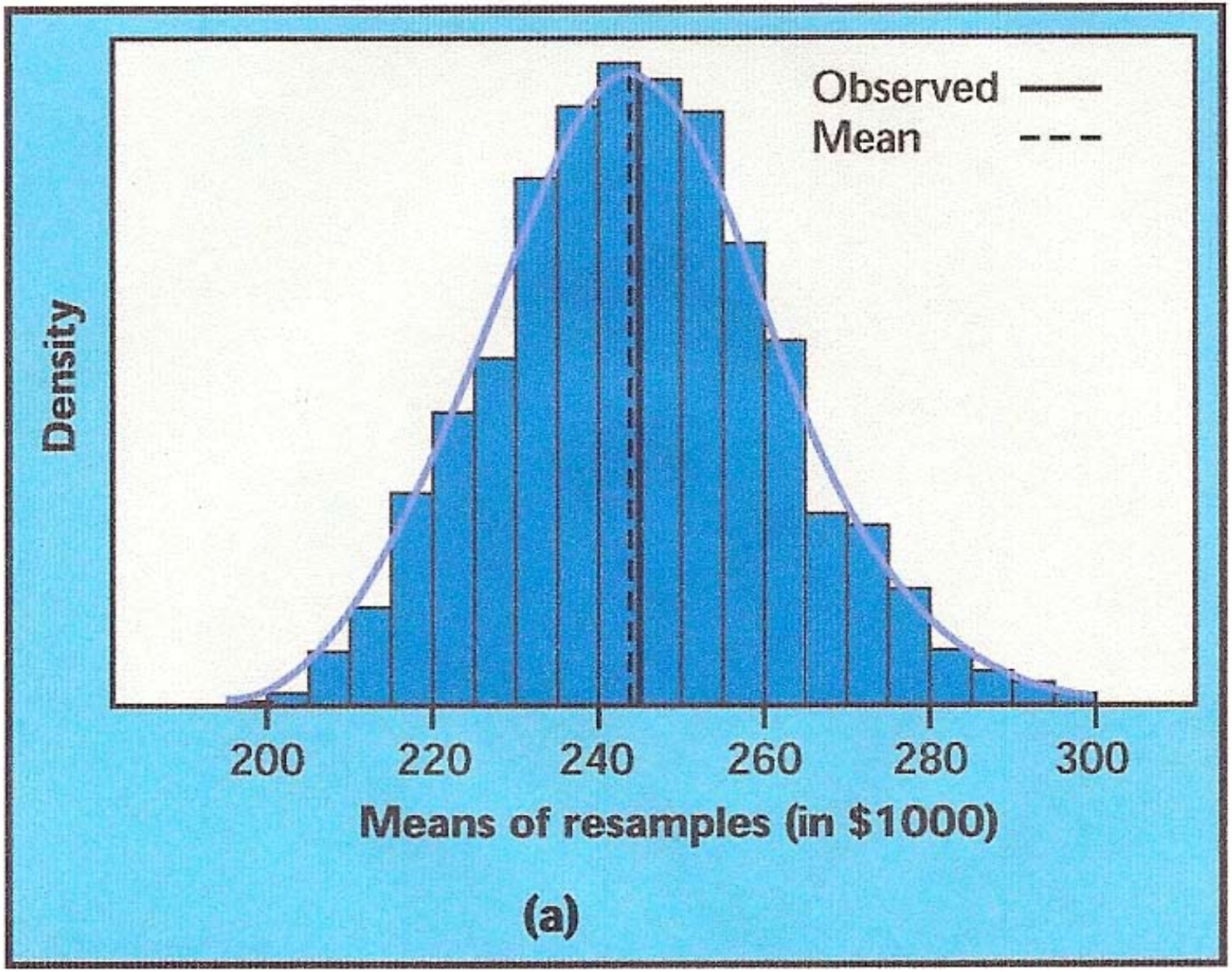


FIGURE 18.8 The bootstrap distribution of the 25% trimmed means of 1000 resamples from the data in Table 18.1. The bootstrap distribution is roughly Normal.

Lecture Notes

Bootstrap and Permutation Confidence Intervals and Tests For Two Samples

Roy Welsch

Spring 2017

SUTD-MIT

Comparing Two Treatments

1. Independent samples — divide subjects randomly into two groups. One group gets treatment 1, the other treatment 2.

2. Matched pair design — subjects chosen in pairs so that members in each pair are alike or even the same person (self-pairing), but different pairs may be substantially different. Treatments are assigned to pair members by coin toss.

It is often useful to create a control group where no treatment is given (placebo). This eliminates the placebo effect where if you know you have been given a new drug, you just think you feel better.

Double-blind — person giving treatments does not know which one you get and you don't either.

Independent Samples

- Sampling brake linings:
  - Two types of brake linings, A and B
  - Two car types, Lexus and Ford
  - Two drivers, Joe and Ed

- Randomly assign brake linings A and B among the four possible driver-car combinations for *n* trials.

- Via randomization (say 50-50 chance) we arrive at n_1 observations for population A and n_2 for population B.

Independent Random Sampling

- For independent sampling as described, the brake lining A sample contains *a mixture of observations with both Joe and Ed as drivers*. So does the brake lining B sample.

- We now decide to compare the sample mean of brake lining A wear with the sample mean of brake lining B wear.

- If Joe and Ed drive differently and even drive differently in different cars, each car-driver combination may induce a different rate of wear on brake linings.

- Then observations will include variations due to driver effects, car effects and driver-car interaction effects as well as variation between type A and B brake lining!

- We wish to “filter out” the driver-car effects.

Matched Pairs

Suppose that we do our study as follows:

- Put both type A and B on front wheels, but randomly assign them to either the left or right front wheel (stratify so 50% each).
- Don't let Joe or Ed know which wheel has which brake type.
- Don't let the mechanic who installs them tell *you* which side is which [Double Blind].

- We have *blocked* our sampling procedure on car-driver pairs by assigning both A and B brake linings to each driver-car pair.

- This will filter out the three effects we identified at first
  - Car effect
  - Driver effect
  - Interaction between car and driver effect

Pros and Cons of Each Design

- In the independent sample design the two groups may not be quite equal on some attribute, especially if the sample sizes are small, because randomization assures equality only on average. Does one group have a higher pretest score than the other? Difference in the groups could be the result of this difference and not the treatment factor.
- The matched pair design enables more precise comparisons to be made between treatment groups because of smaller experimental error.
- However, it can be difficult to form matched pairs, especially when dealing with human subjects
- The two types of designs can be used in observational studies though the two groups are not formed by randomization.

BOOTSTRAP FOR COMPARING TWO POPULATIONS

Given independent SRSs of sizes n and m from two populations:

1. Draw a resample of size n with replacement from the first sample and a separate resample of size m from the second sample. Compute a statistic that compares the two groups, such as the difference between the two sample means.
2. Repeat this resampling process hundreds of times.
3. Construct the bootstrap distribution of the statistic. Inspect its shape, bias, and bootstrap standard error in the usual way.

SRS = Simple Random Sample

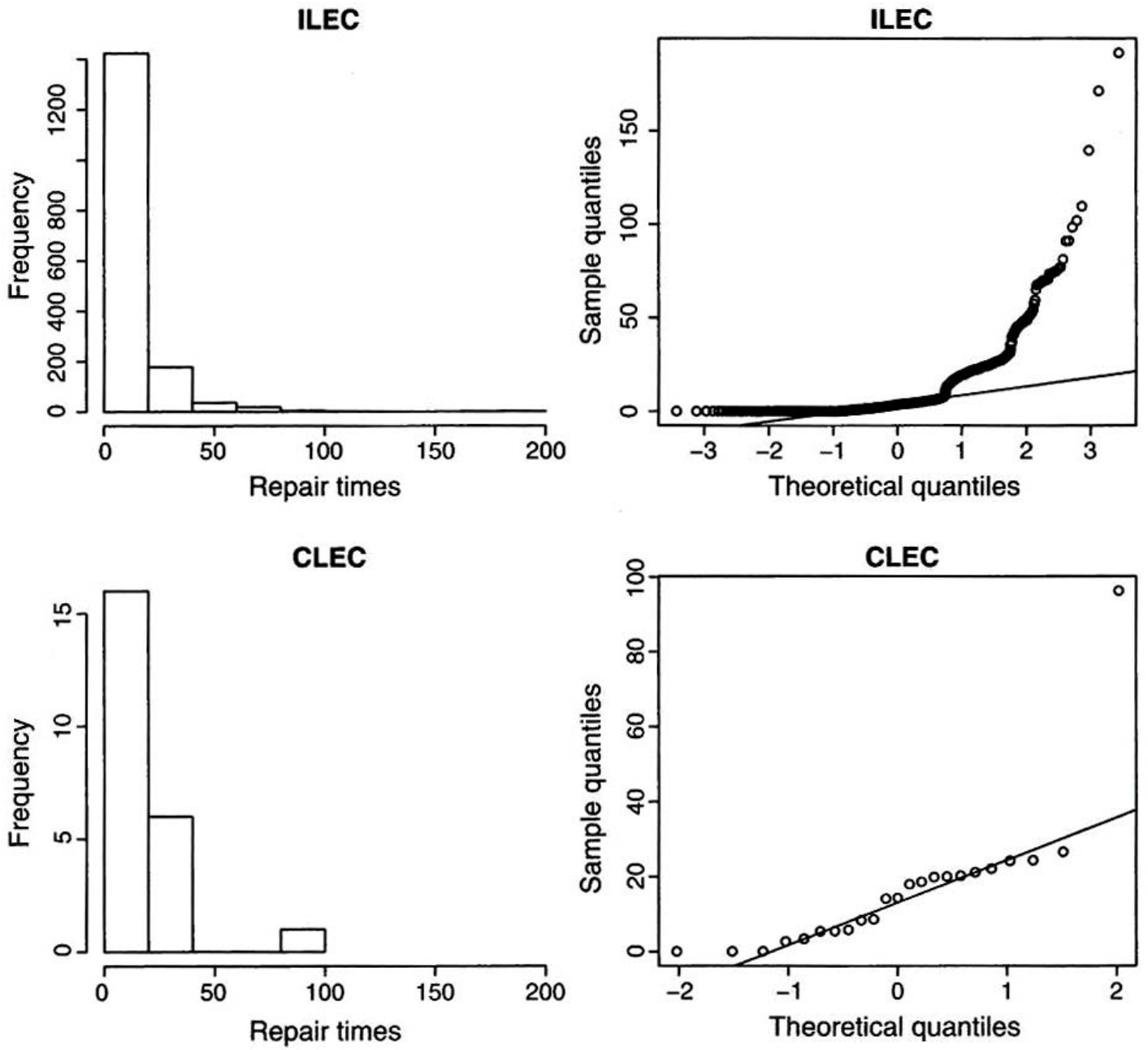


FIGURE 3.4 Distribution of repair times for Verizon (ILEC) and competitor (CLEC) customers. Note that the Y-axis scales are different.

Service provider	n	$\bar{x}$	s
Verizon	1664	8.4	14.7
CLEC	23	16.5	19.5
Difference		−8.1	

```
Number of Replications: 1000

Summary Statistics:
          Observed   Mean     Bias     SE
meanDiff    -8.098 -8.251 -0.1534 4.052
```

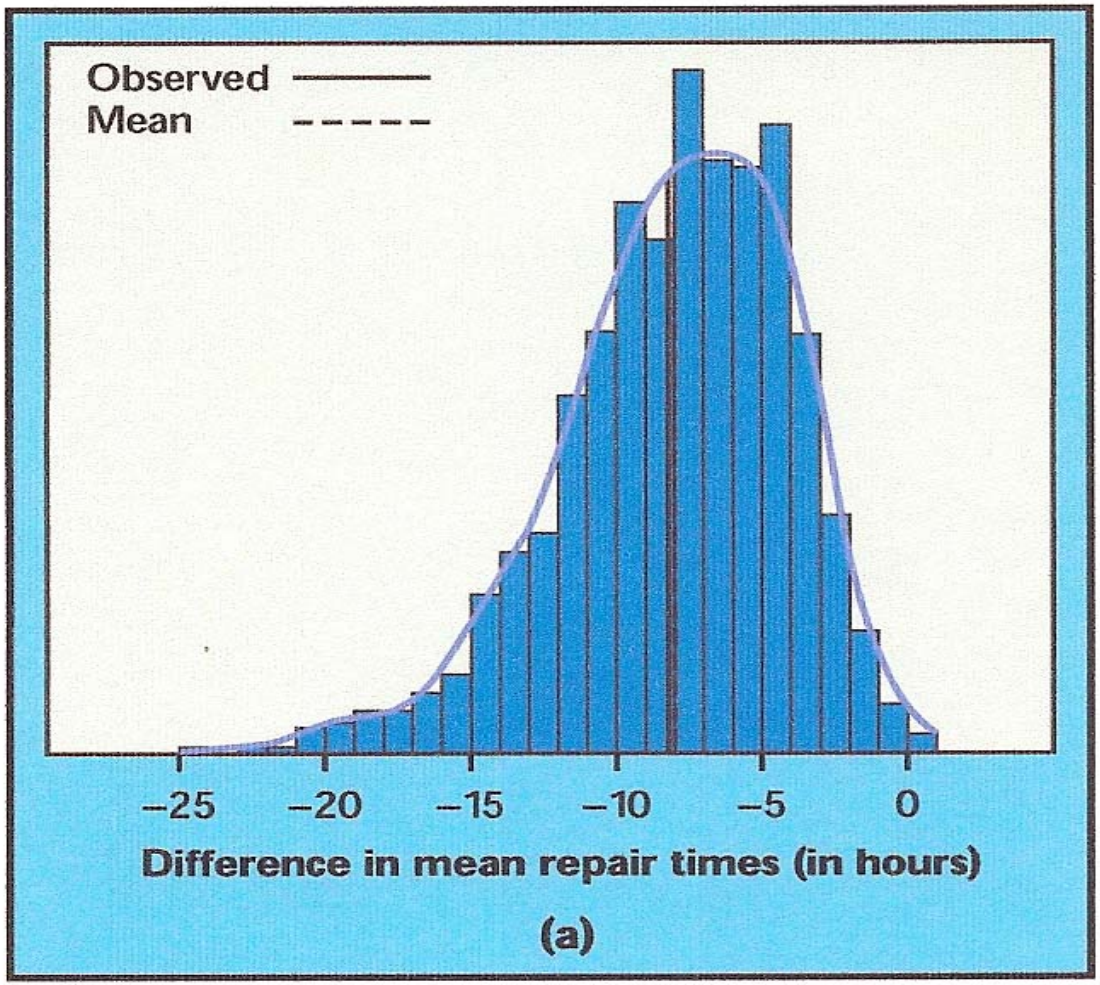


FIGURE 18.10 The bootstrap distribution of the difference in means for the Verizon and CLEC repair time data.

Note where the null hypothesis of zero difference sits on this plot.

Another Idea: The Permutation Bootstrap or Resampling Without Replacement

Degree of Reading Power Scores

TABLE 18.4 DRP scores for third-graders

Treatment group				Control group			
24	61	59	46	42	33	46	37
43	44	52	43	43	41	10	42
58	67	62	57	55	19	17	55
71	49	54		26	54	60	28
43	53	57		62	20	53	48
49	56	33		37	85	42	

Permutation Resampling for Two Independent Groups

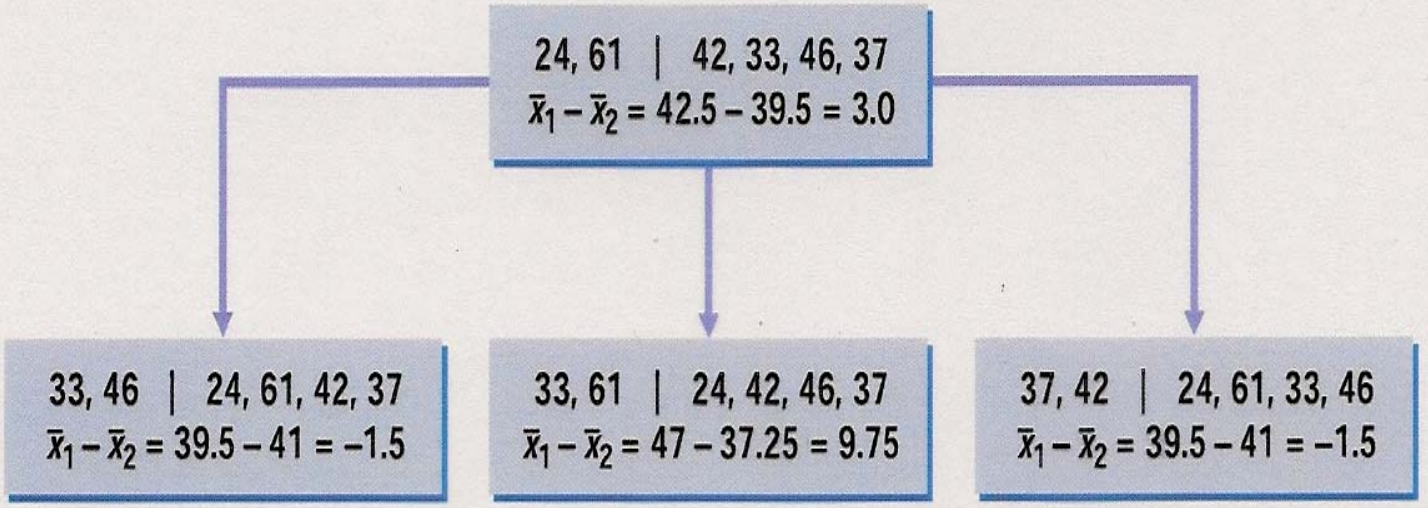


FIGURE 18.20 The idea of permutation resampling. The top box shows the outcomes of a study with four subjects in one group and two in the other. The boxes below show three permutation resamples. The values of the statistic for many such resamples form the permutation distribution.

Could also be viewed as sampling without replacement.

GENERAL PROCEDURE FOR PERMUTATION TESTS

To carry out a permutation test based on a statistic that measures the size of an effect of interest:

1. Compute the statistic for the original data.
2. Choose permutation resamples from the data without replacement in a way that is consistent with the null hypothesis of the test and with the study design. Construct the permutation distribution of the statistic from its values in a large number of resamples.
3. Find the *P*-value by locating the original statistic on the permutation distribution.

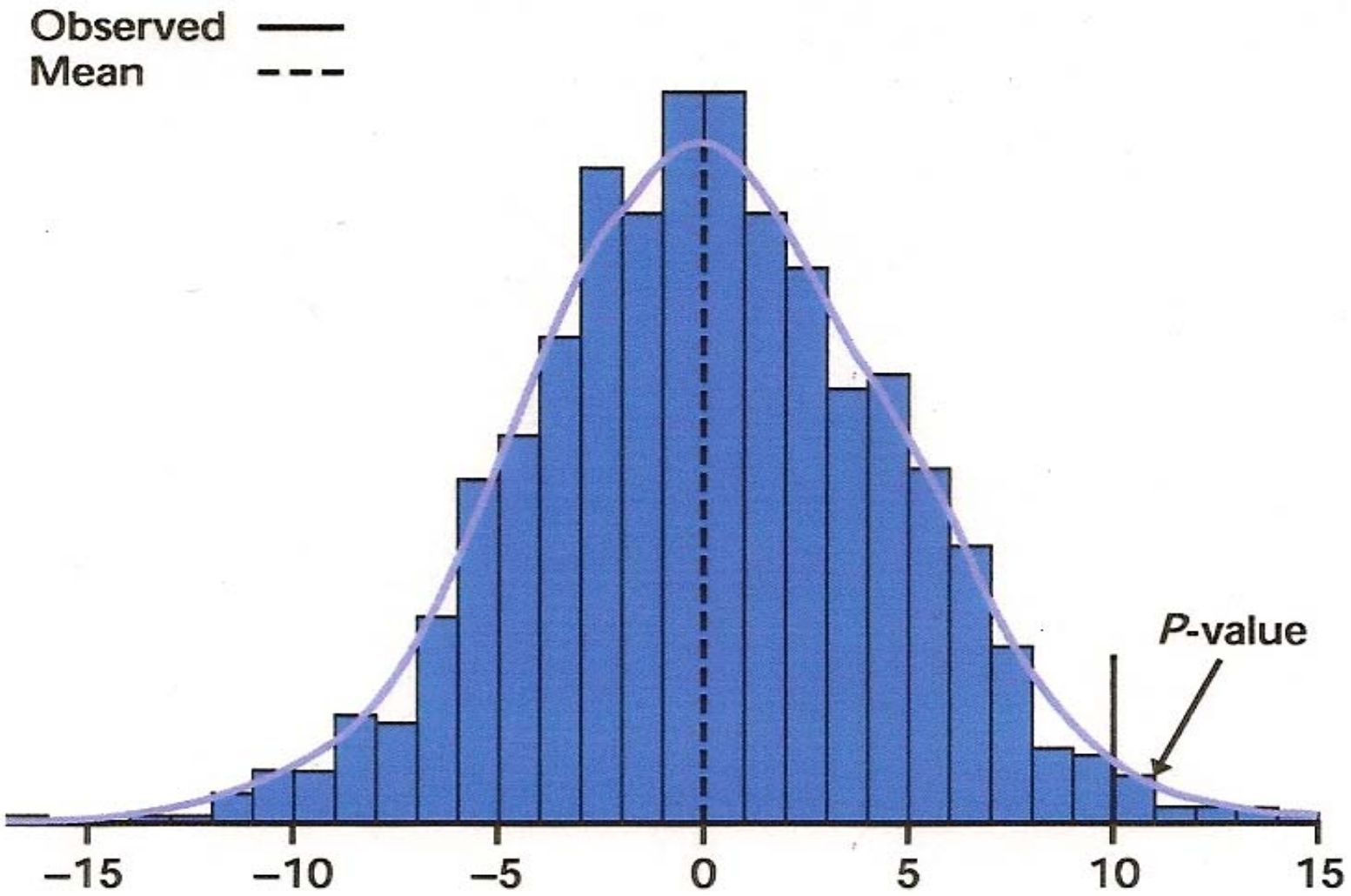


FIGURE 18.21 The permutation distribution of the statistic $\overline{x}_{\text{treatment}} - \overline{x}_{\text{control}}$ based on the DRP scores of 44 students. The observed difference in means, 9.954, is in the right tail. p = 0.014

Note where zero is in this plot. That is the null hypothesis of no difference.

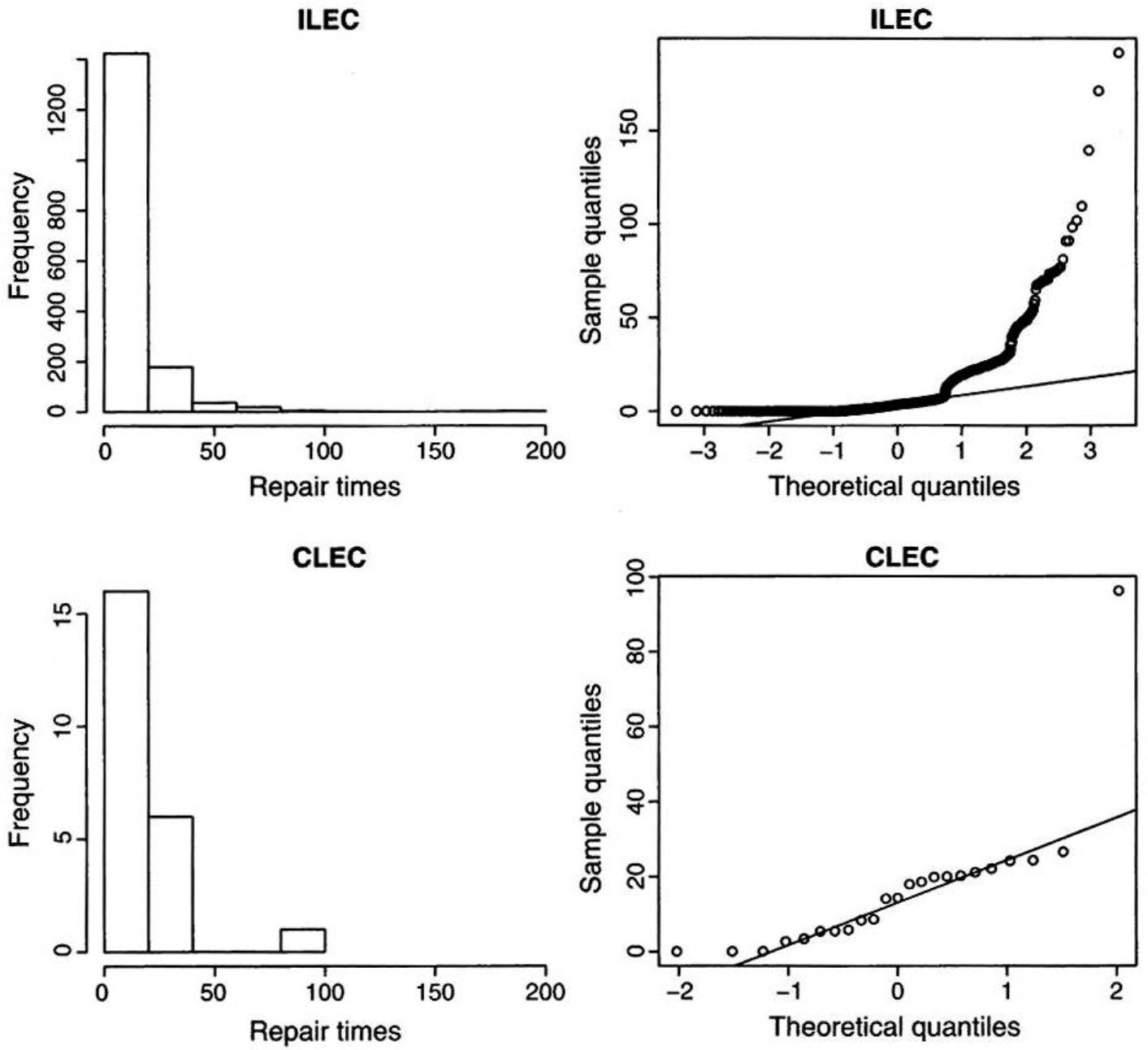


FIGURE 3.4 Distribution of repair times for Verizon (ILEC) and competitor (CLEC) customers. Note that the Y-axis scales are different.

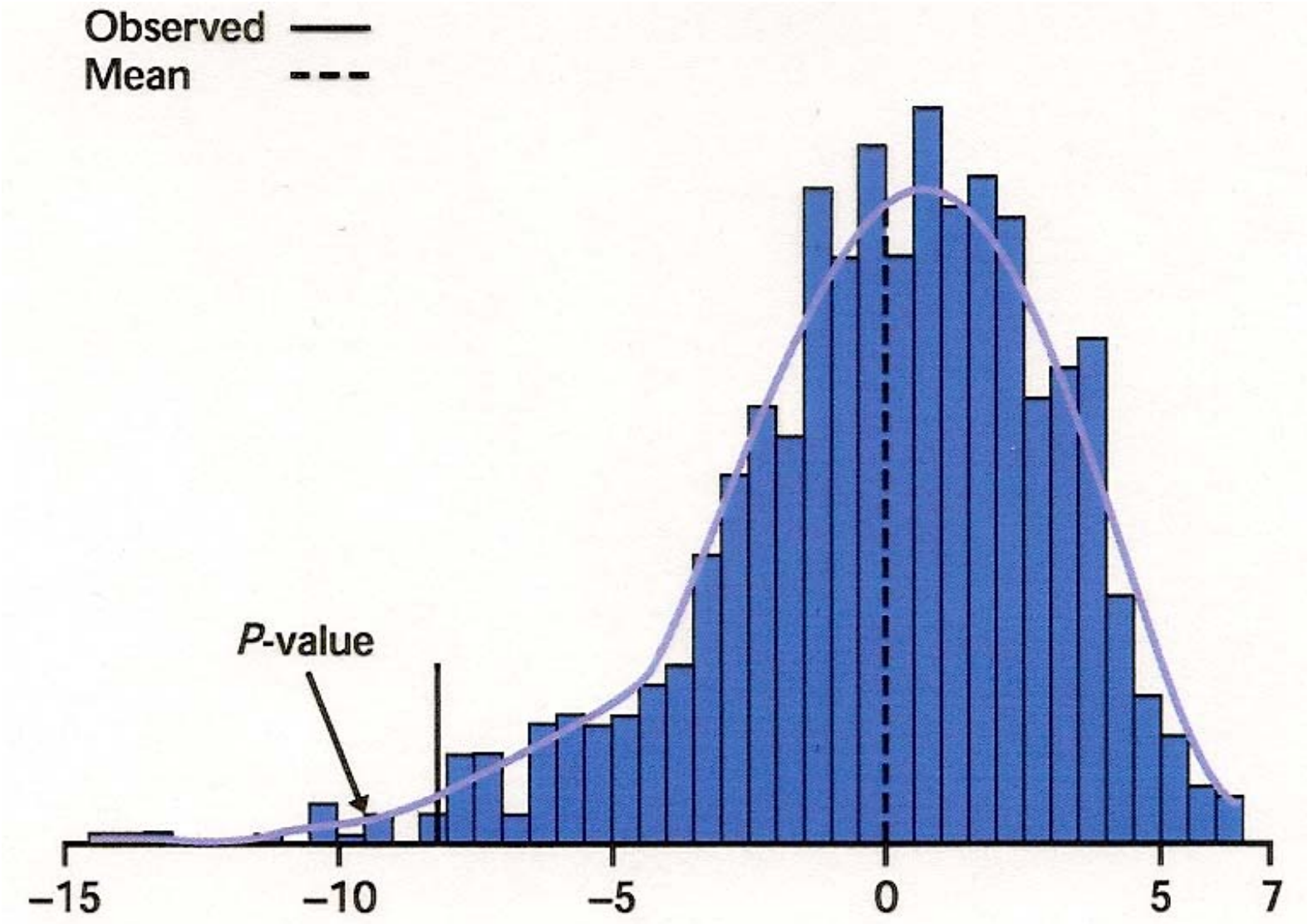


FIGURE 18.22 The permutation distribution of the difference of means $\overline{x}_1 - \overline{x}_2$ for the Verizon repair time data. The distribution is skewed left. The observed difference in means, −8.098, is in the left tail. p = 0.0183

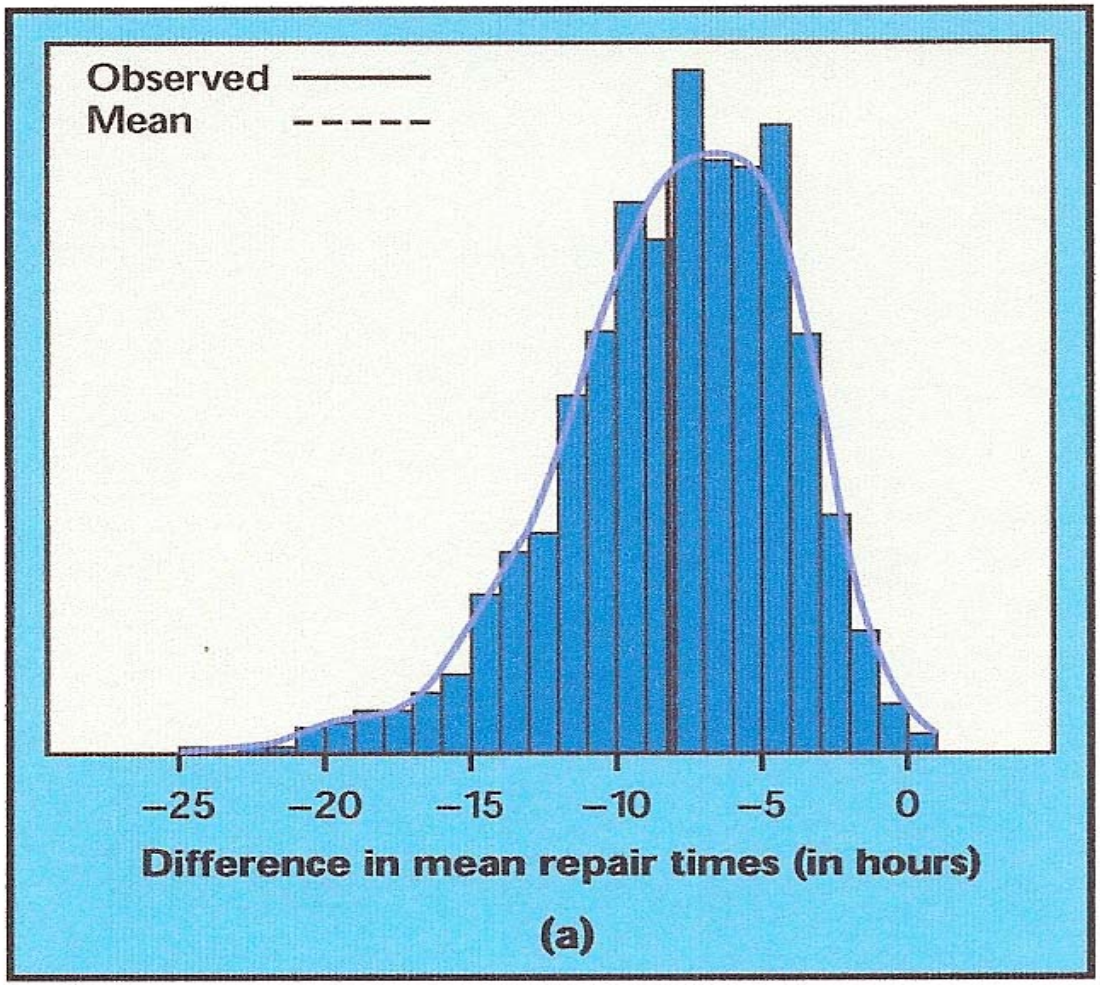


FIGURE 18.10 The bootstrap distribution of the difference in means for the Verizon and CLEC repair time data.

Note where the null hypothesis of zero difference sits on this plot.

Bootstrapping a Paired Sample

1. Compute each paired difference.

2. Bootstrap the mean of these differences.

3. Find confidence interval for the mean differences from the bootstrap sampling distribution.

4. Why not just do an independent sample (not paired) analysis?

Paired Sample Permutation Tests

- Null hypothesis says training has no effect on scores
- “Before” and “after” for each pair have no meaning since no effect
- Resampling randomly assigns one of each executive’s two scores to “before” and “after”
- Do not mix scores from different people since that would not be consistent with paired study design
- Compute mean differences and repeat many times

Paired French Listening Scores

TABLE 7.2 French listening scores for executives

Executive	Pretest	Posttest	Gain	Executive	Pretest	Posttest	Gain
1	32	34	2	11	30	36	6
2	31	31	0	12	20	26	6
3	29	35	6	13	24	27	3
4	10	16	6	14	24	24	0
5	30	33	3	15	31	32	1
6	33	36	3	16	30	31	1
7	22	24	2	17	15	15	0
8	25	28	3	18	32	34	2
9	32	26	−6	19	23	26	3
10	20	26	6	20	23	26	3

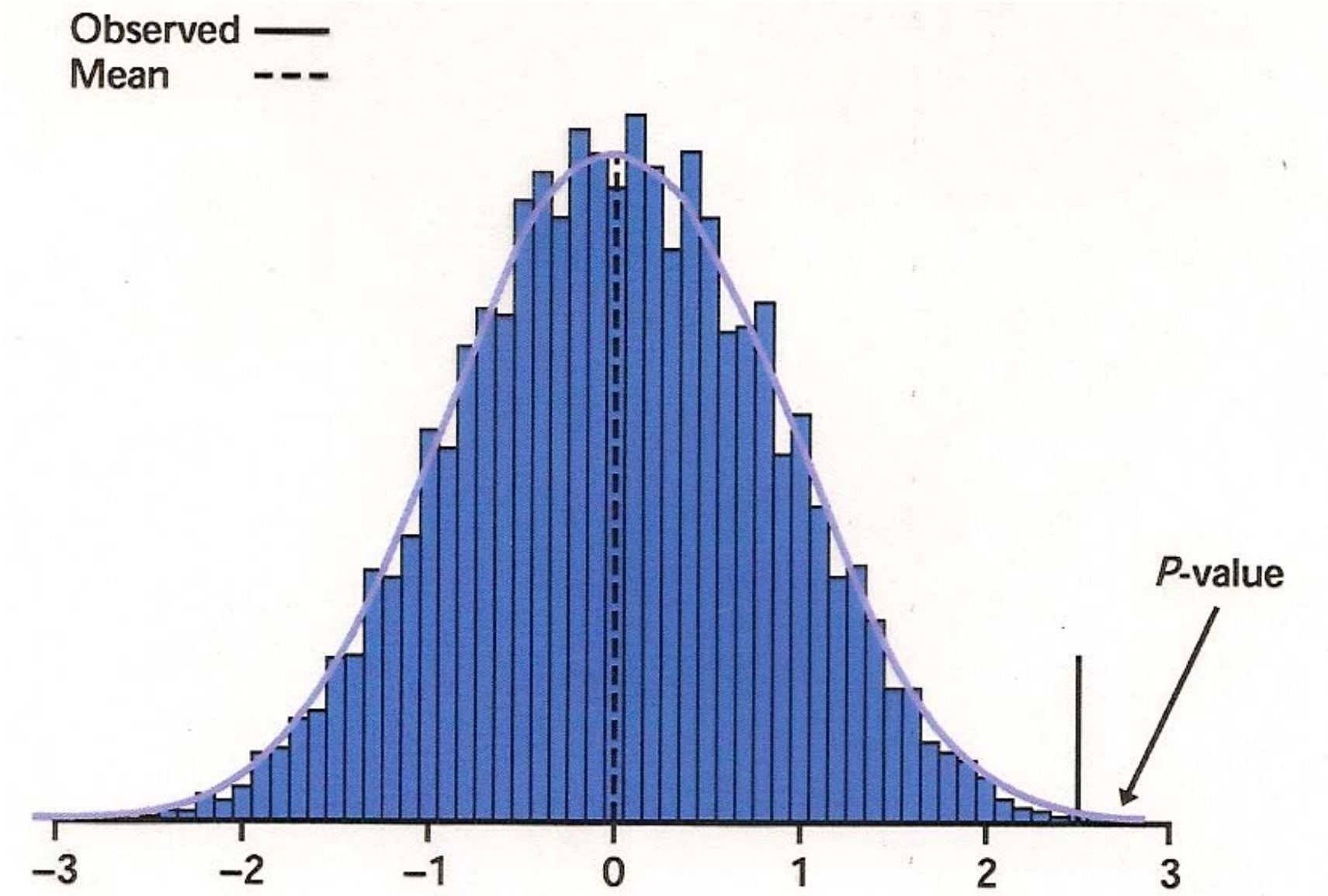


FIGURE 18.23 The permutation distribution for the mean difference (score after instruction minus score before instruction) from 9999 paired resamples from the data in Table 7.2. The observed difference in means, 2.5, is in the right tail. p = 0.0015

Lecture Notes

Categorical Data Analysis

Roy Welsch

Spring 2017

SUTD-MIT

Chi-Squared Tests

- Hypothesis Tests for Qualitative Data
  - Categories instead of numbers
  - Based on counts
    - the number of sampled items falling into each category
  - Chi-squared statistic
    - Measures the difference between ***Actual*** counts and ***Expected*** counts (as expected under the null hypothesis H_0)

$$\text{Chi-squared statistic} = \text{Sum of } \frac{(\text{Observed Count} - \text{Expected Count})^2}{\text{Expected Count}} = \sum \frac{(O_i - E_i)^2}{E_i}$$

where the sum extends over all categories or combinations of categories

  - ***Significant*** if the chi-squared statistic is large enough

Independence (No Association)

- Two Qualitative Variables are ***Independent*** if:
  - knowledge about the value (i.e., category) of one variable does not help you predict the other variable
- For Example: Background Sales Information
  - The two variables are
    - where the customer lives, and
    - the customer's favorite product
  - Independence would say that
    - customers tend to have the same pattern (distribution) of favorite products, regardless of where they live, and
    - where customers live is not associated with which product is their favorite

Testing for Association

- The chi-squared test for independence
  - ***The data***: A table indicating the counts for each combination of categories for two qualitative variables
  - ***The hypotheses***:

    H_0: The two variables are ***independent*** of one another

    H_a: The two variables are ***associated***; they are ***not independent***
  - ***The expected table***:

$$\text{Expected count} = \frac{n\left(\begin{array}{c}\text{Count for category}\\ \text{for one variable}\end{array}\right)\left(\begin{array}{c}\text{Count for category}\\ \text{for other variable}\end{array}\right)}{n^2}$$

    - Tells you what the counts *would have been*, on average, if the variables were independent and there were no randomness
    - Expected count is n times the product of the estimated probabilities. Note product rule for independence.

Testing Association (continued)

– ***The assumptions***:

1. Data set is a random sample from the population of interest
2. At least 5 counts are expected in each combination of categories

– ***The chi-squared statistic***:

$$\text{Sum of } \frac{(\text{Observed Count} - \text{Expected Count})^2}{\text{Expected Count}} = \sum \frac{(O_i - E_i)^2}{E_i}$$

- where the sum extends over all combinations of categories

– ***The degrees of freedom***:

$$\left(\begin{array}{c}\text{Number of categories} \\ \text{for first variable}\end{array} - 1\right)\left(\begin{array}{c}\text{Number of categories} \\ \text{for second variable}\end{array} - 1\right)$$

– ***The test result***: ***Significant*** if the chi-squared statistic is larger than the critical value from the table

What can we learn from simulation?

- At this point let’s assume we do not know anything about the distribution of the chi-squared statistic
- Therefore, no way to see if we should reject the null hypothesis of independence
- We can try permuting some of the data in a way that assumes independence and see if the chi-squared statistic for the unpermuted data is far away from the chi-squared statistics for many permutations assuming independence

Permutation Test for Independence of Two Variables

Store the data in a table with one row per observation and one column per variable. Calculate a test statistic for the original data. Normally large values of the test statistic suggest dependence.
repeat
    Randomly permute the rows in one of the columns.
    Calculate the test statistic for the permuted data.
until we have enough samples
Calculate the *P*-value as the fraction of times the random statistics exceed the original statistic.
Optionally, plot a histogram of the resampled statistic values.

Chihara & Hesterberg, 2011

Table 3.4 Counts of Death Penalty Opinions Grouped by Education

	Death penalty for murder (opinion)?			
Education	Favor	Oppose	Row Sum	% Favor
Bachelors	135	71	206	65.6%
Graduate	64	50	114	56.1%
HS	511	200	711	71.9%
JrColl	71	16	87	81.6%
Left HS	117	72	189	61.9%
Column sum	898	409	n = 1307	68.7%

Only the opinion variable is permuted for 1307 people. The education variable is held fixed. Think about an urn with 898 favor balls and 409 opposed balls and sample from it without replacement. For Bachelors you would draw out 206 balls, for Graduate, 114 balls, etc.

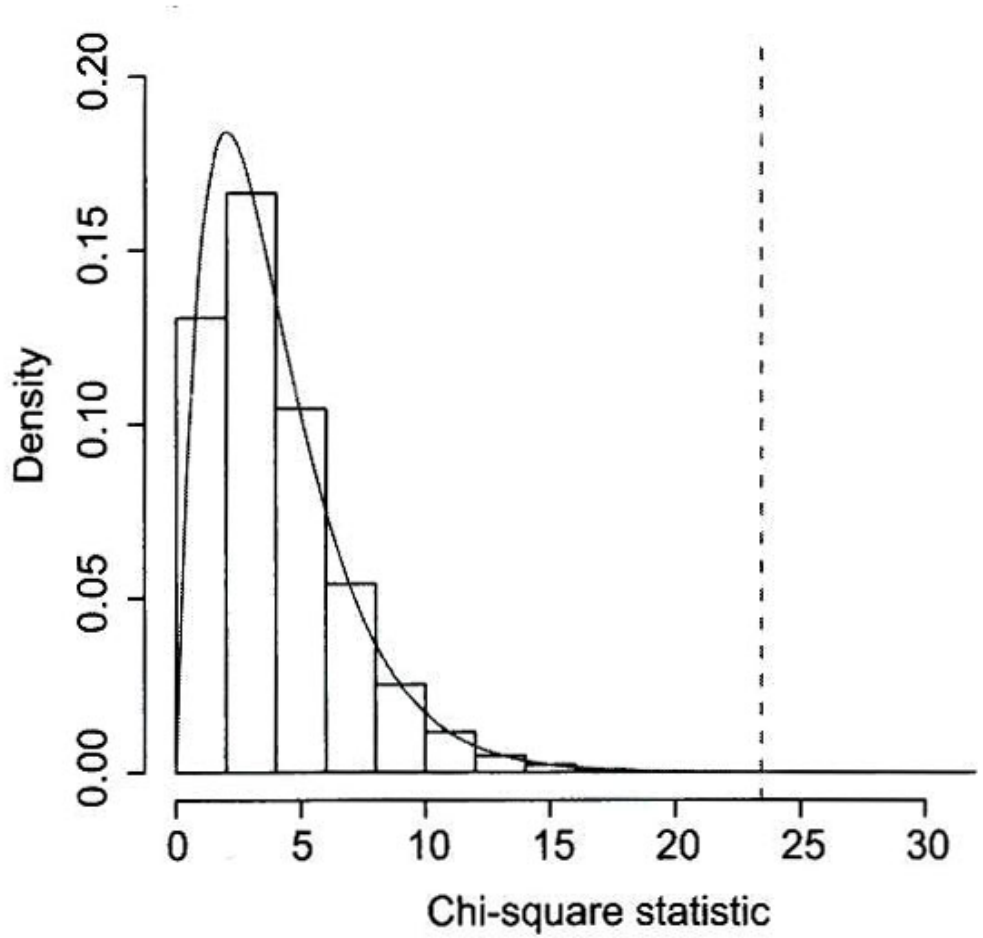


FIGURE 3.8 Null distribution for chi-square statistic for death penalty opinions; the overlaid density is a chi-square distribution with 4 degrees of freedom.

The dotted line is the computed chi-square statistic from the original data. Clearly the null hypothesis of independence would have been rejected.

Summary

- Tabulate the data in a contingency table.
- Sum the number of individuals in each row and each column and figure the percentage of all individuals who fall in each row and column, independent of the column or row in which they fall.
- Use these percentages to compute the number of people that would be expected in each cell of the table if the treatment had no effect or there were no association.
- Summarize the differences between these expected frequencies and the observed frequencies by computing χ^2.
- Use a permutation statistic (an exact test) to find the proportion of tables with the given row and column totals that have χ^2 greater than or equal to the observed χ^2 or compute the number of degrees of freedom and use tables of the χ^2 distribution (an approximate test). Both agree for large samples.
- Always look at which cells are causing large χ^2.

Rejection Does Not Mean Dependence

If we reject χ^2 test for independence (no association), then does not mean dependent (associated).

We need consistency, responsiveness, and mechanism. Coffee drinking related to heart attacks in 2×2 table. However, what if most coffee drinkers are also smokers? Hard to argue coffee drinking causes heart attacks (spurious dependence or lurking variable).

Lecture Notes

Regression and the Bootstrap

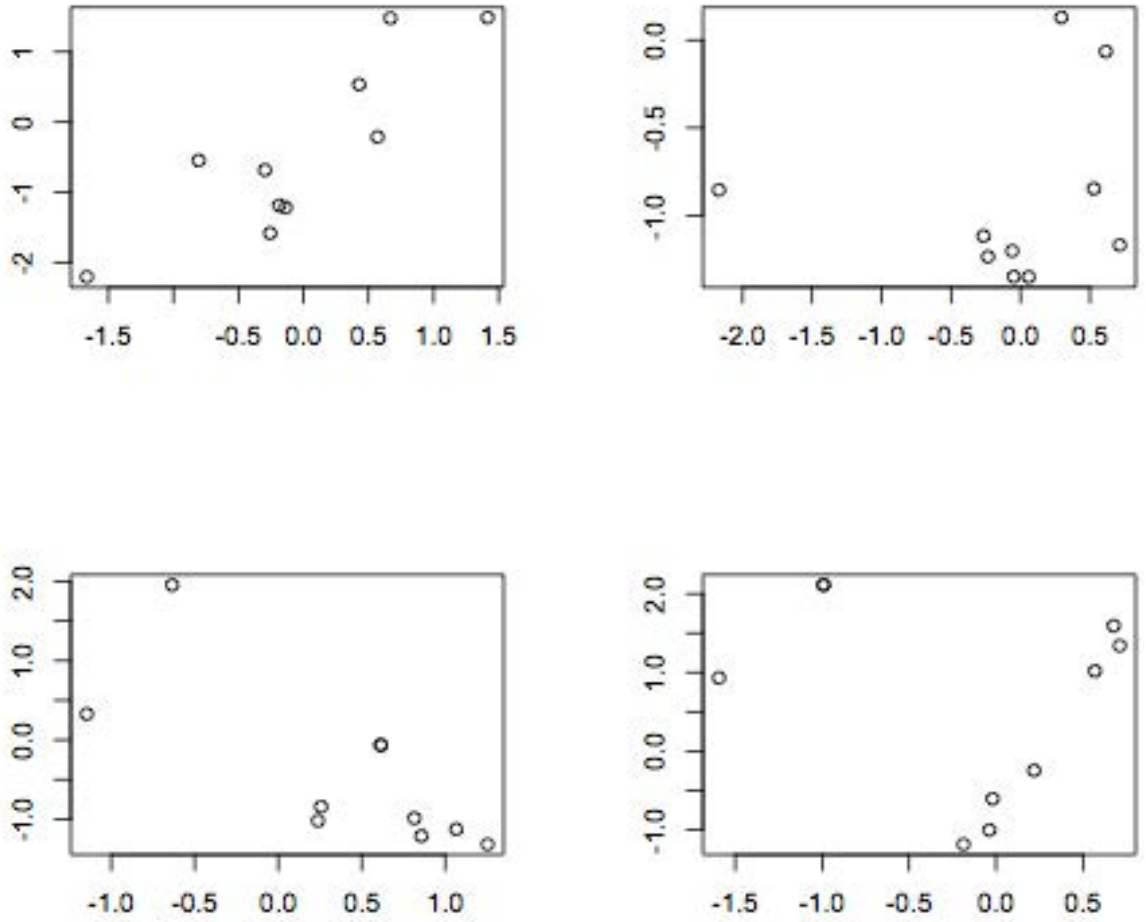
Roy Welsch

Spring 2017

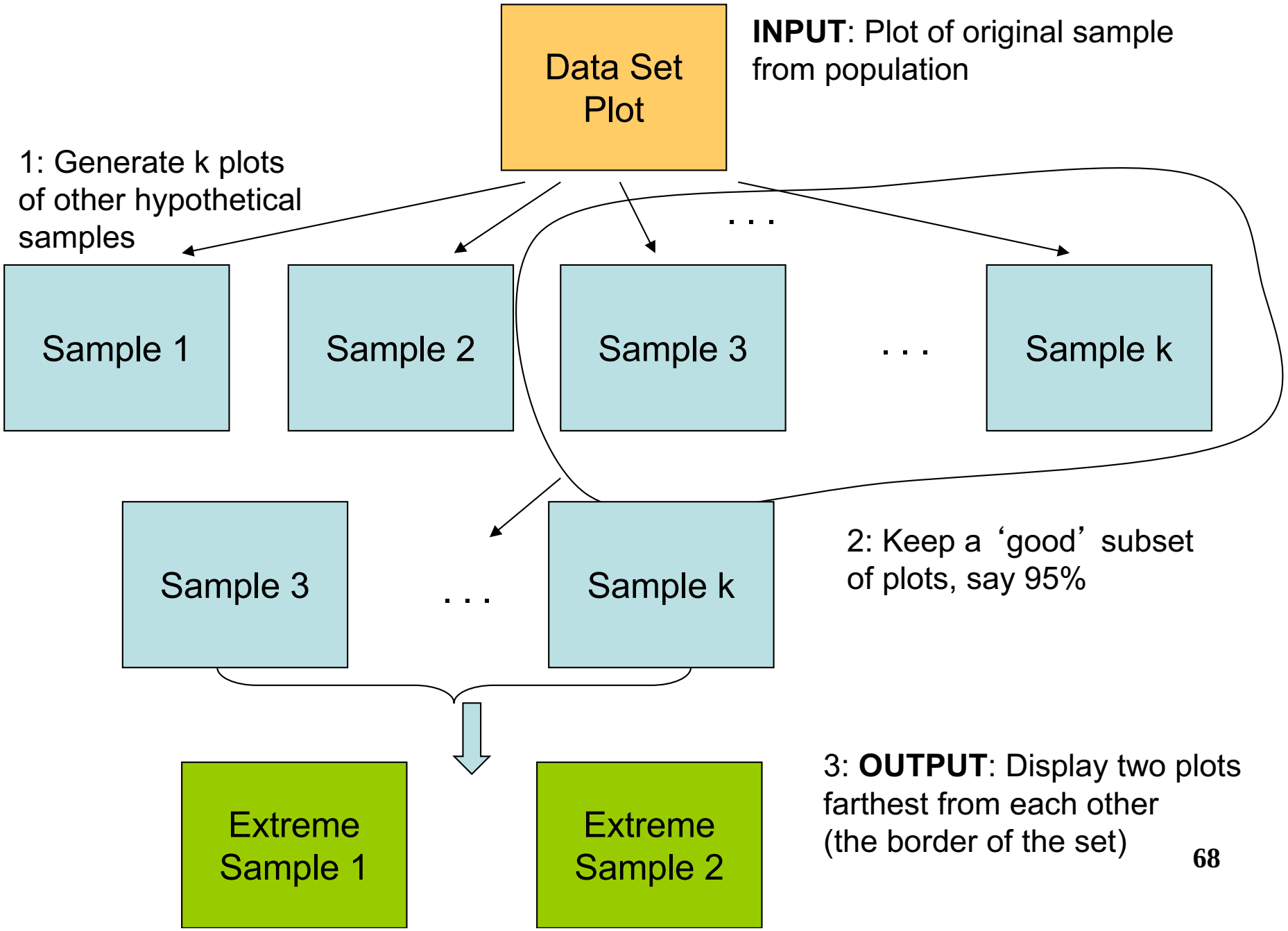
SUTD-MIT

Plots of a data set can look different than the plot of the population they are from!

A simulated illustration: Four data sets, sampled from the same population.

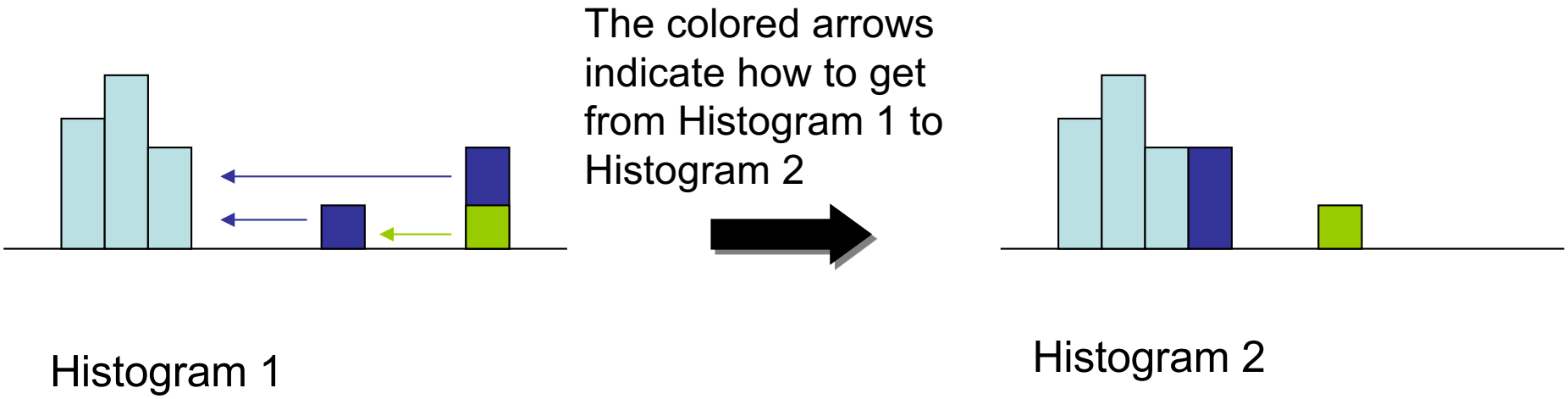


A Picture of the Methodology



68

The Earth Mover's Distance

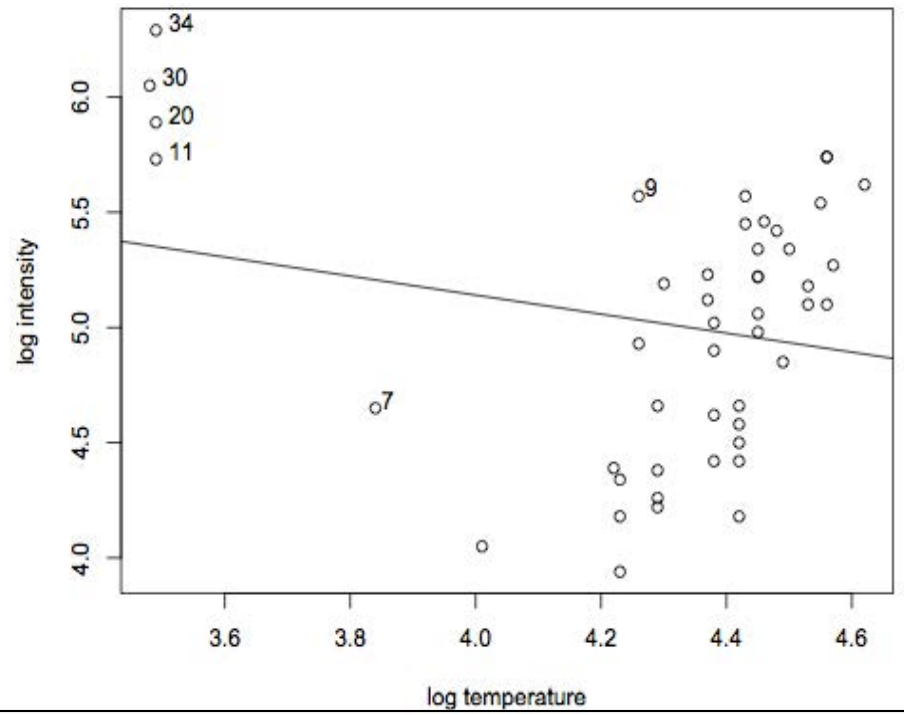


- Histograms are viewed as piles of 'Earth' or dirt
- Earth Mover's distance equals the amount of work ((amount moved) * (distance moved)) required to turn one pile into another pile
- Computing the Earth Mover's distance requires solving an assignment problem (a network flow problem)
- Earth Mover's distance generalizes to several types of plots:
  - Scatter plots, parallel coordinate plots, biplots, … etc.
- Ordering plots is related to the traveling salesman problem (Touring a set of cities with smallest total distance.)

Ex. 1: Our method depicts variability of relationships in data

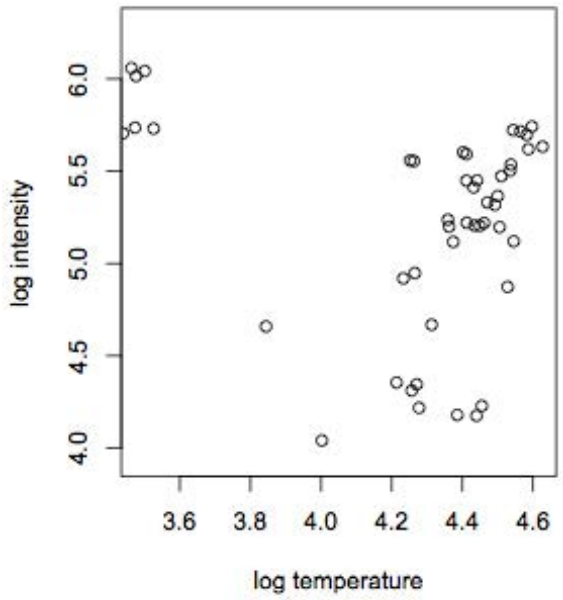
INPUT

Plot of the original data **(Hertzsprung Russell Star Data)**

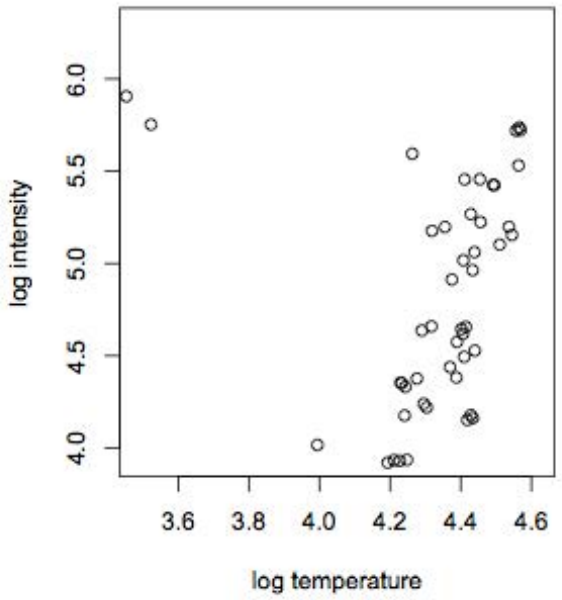


OUTPUT

Plot of one extreme

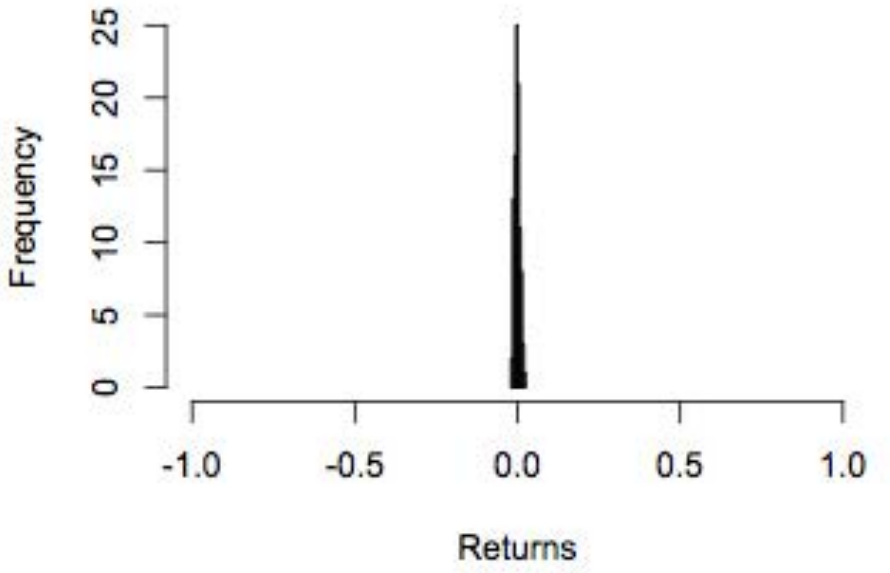


Plot of the other extreme

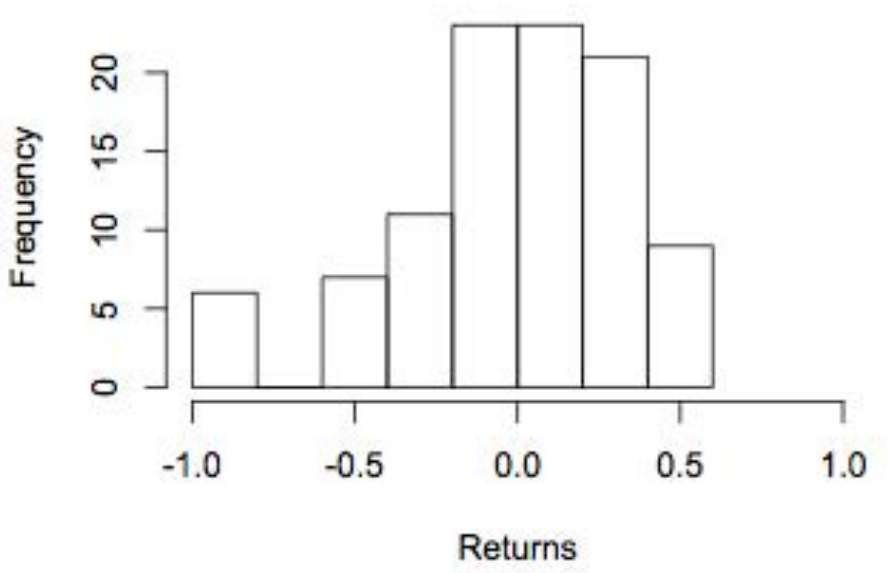


Ex. 2: Our method demonstrates the highly variable results of portfolio optimization!

Plot of one extreme



Plot of the other extreme



- Data are daily returns on 50 industries among the MSCI US Equity indices 01/03/1995 - 02/07/2005
- Portfolio weights trained on first 100 days, in order to maximize Sharpe ratio
- Object of interest is the histogram of portfolio returns for the next 100 days

Advantages of the approach

- Generalizes to several types of plots
- Only two plots are necessary to convey the message
- Can report the most interesting plots in a data set while remaining statistically sound
- Improves validity of visualization in statistics

Sample Correlation Coefficient

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

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

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.

TABLE 18.2 Major League Baseball salaries and batting averages

Name	Salary	Average	Name	Salary	Average
Matt Williams	$9,500,000	.269	Greg Colbrunn	$1,800,000	.307
Jim Thome	8,000,000	.282	Dave Martinez	1,500,000	.276
Jim Edmonds	7,333,333	.327	Einar Diaz	1,087,500	.216
Fred McGriff	7,250,000	.259	Brian L. Hunter	1,000,000	.289
Jermaine Dye	7,166,667	.240	David Ortiz	950,000	.237
Edgar Martinez	7,086,668	.270	Luis Alicea	800,000	.202
Jeff Cirillo	6,375,000	.253	Ron Coomer	750,000	.344
Rey Ordonez	6,250,000	.238	Enrique Wilson	720,000	.185
Edgardo Alfonzo	6,200,000	.300	Dave Hansen	675,000	.234
Moises Alou	6,000,000	.247	Alfonso Soriano	630,000	.324
Travis Fryman	5,825,000	.213	Keith Lockhart	600,000	.200
Kevin Young	5,625,000	.238	Mike Mordecai	500,000	.214
M. Grudzielanek	5,000,000	.245	Julio Lugo	325,000	.262
Tony Batista	4,900,000	.276	Mark L. Johnson	320,000	.207
Fernando Tatis	4,500,000	.268	Jason LaRue	305,000	.233
Doug Glanville	4,000,000	.221	Doug Mientkiewicz	285,000	.259
Miguel Tejada	3,625,000	.301	Jay Gibbons	232,500	.250
Bill Mueller	3,450,000	.242	Corey Patterson	227,500	.278
Mark McLemore	3,150,000	.273	Felipe Lopez	221,000	.237
Vinny Castilla	3,000,000	.250	Nick Johnson	220,650	.235
Brook Fordyce	2,500,000	.208	Thomas Wilson	220,000	.243
Torii Hunter	2,400,000	.306	Dave Roberts	217,500	.297
Michael Tucker	2,250,000	.235	Pablo Ozuna	202,000	.333
Eric Chavez	2,125,000	.277	Alexis Sanchez	202,000	.301
Aaron Boone	2,100,000	.227	Abraham Nunez	200,000	.224

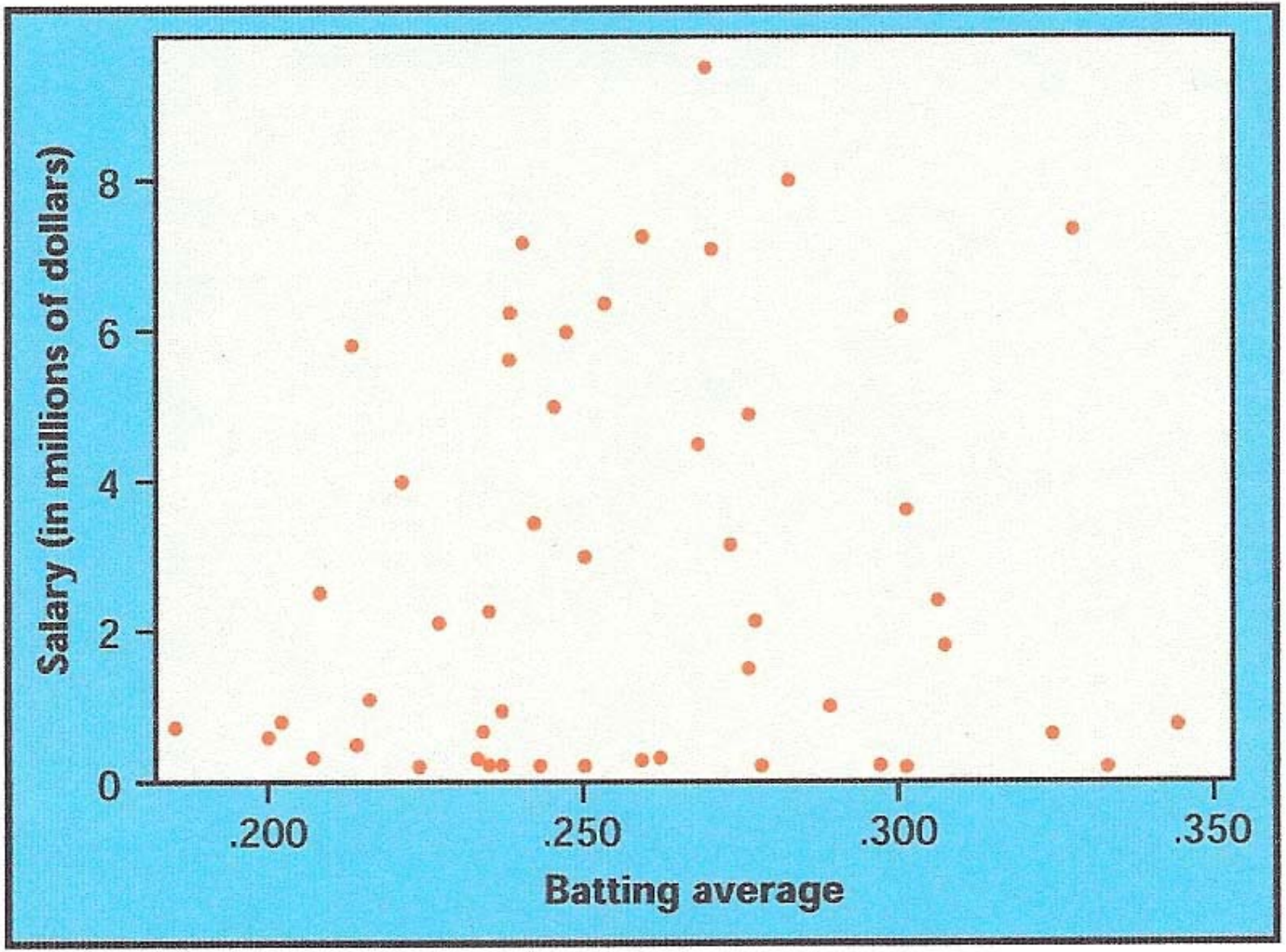


FIGURE 18.16 Batting average and salary for a random sample of 50 Major League Baseball players.

Resample Bootstrap for Correlation

- Have *n* sets of observations (x_i , y_i) , i = 1 to n.

- Simply bootstrap these *n* sets (with replacement) many times, compute correlation coefficient (or other measure of association) and create bootstrap histograms (sampling distributions) of the correlation coefficients

- From the bootstrap distributions find confidence intervals and *p*-values

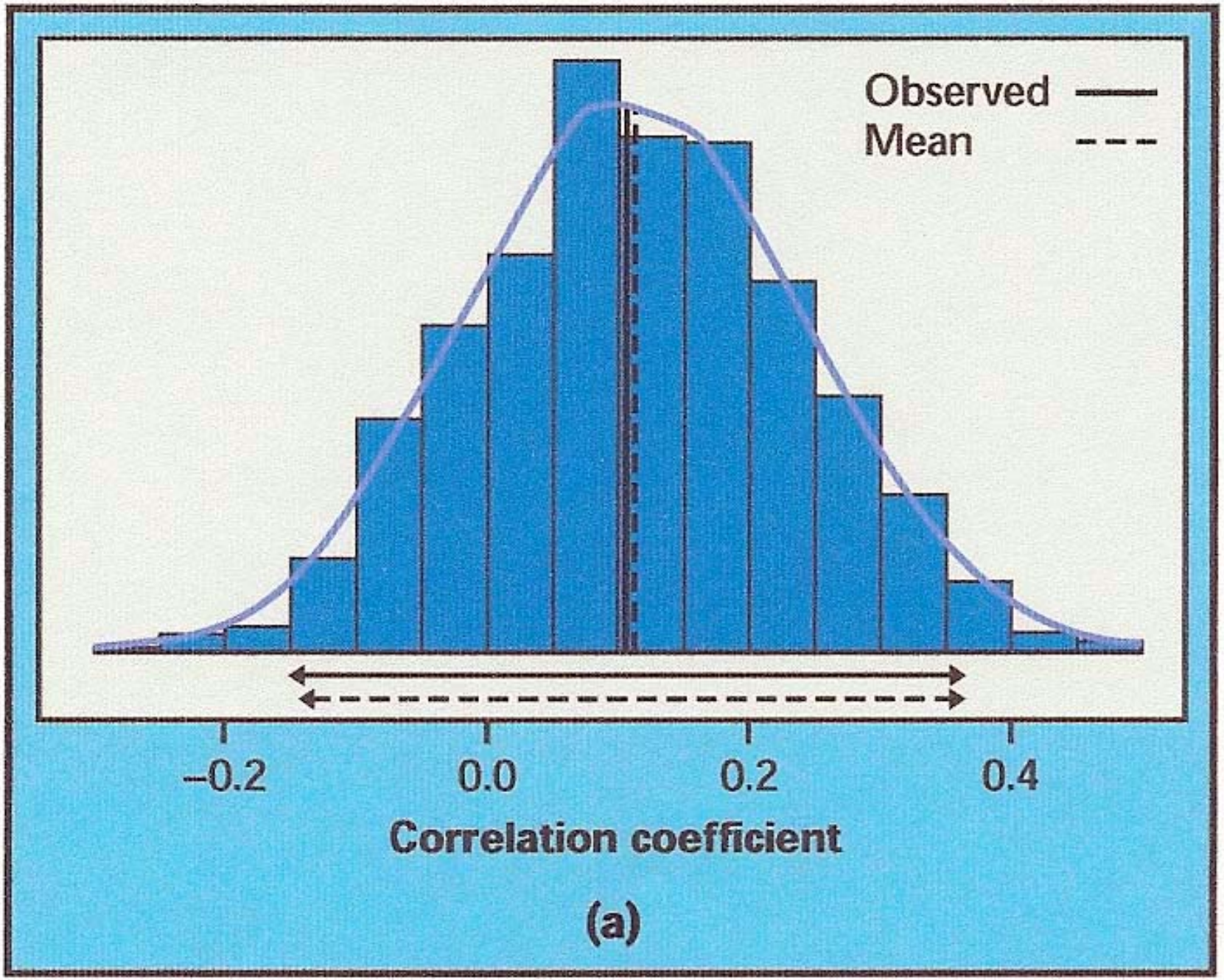


FIGURE 18.17

Baseball Correlation Results

The sample correlation is 0.107.

The bootstrap standard error is $SE_{boot,\, r} = 0.125$.

The bootstrap percentile interval is

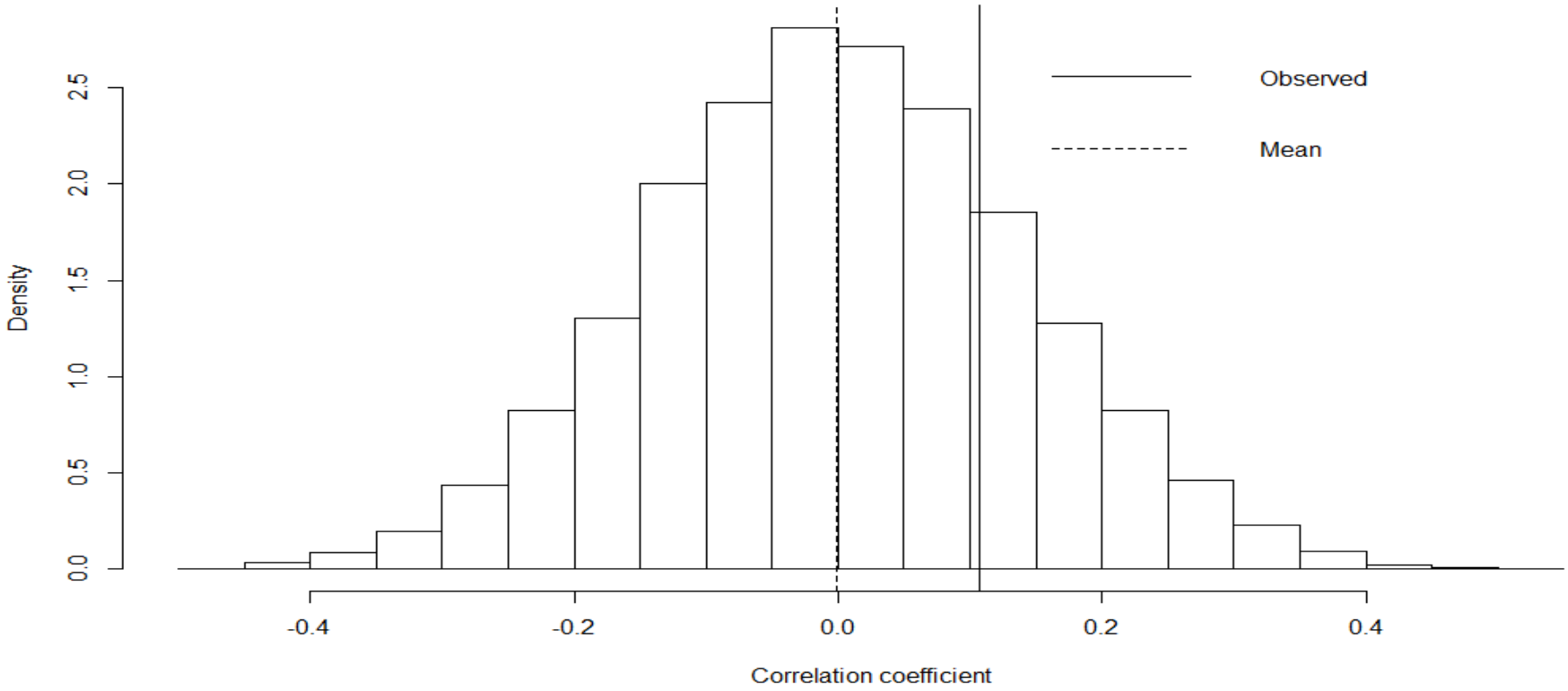
(2.5th percentile, 97.5th percentile) = (–0.128, 0.356).

Note that the interval contains zero.

Permutation Test for Significance of a Relationship

- Null hypothesis is no relationship between two variables (salaries and batting averages)
- Resample in a way that is consistent with the null hypothesis by permuting the observed salaries among the players at random
- If correlation is the test statistic, calculate the correlation between the batting averages (in their original order) and salaries (in the reshuffled order)
- The *p*-value is the proportion of resamples with correlation larger than the original correlation

Baseball Permutation Test Results



Actual sample correlation: 0.107
P-value: 0.2243
2.5% quantile: -0.275
Note where zero is in this plot.

Permutation test mean: 0.0
Permutation test standard dev.: 0.1416
97.5% quantile: 0.280
That represents the null hypothesis.

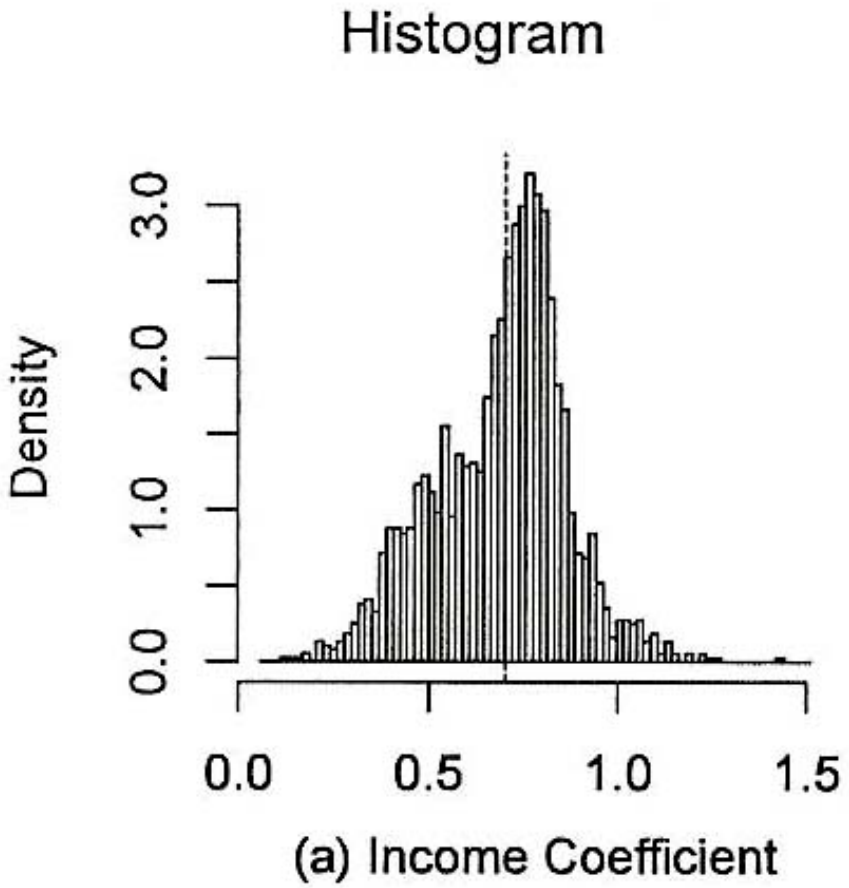
Full Resample Bootstrap for Regression

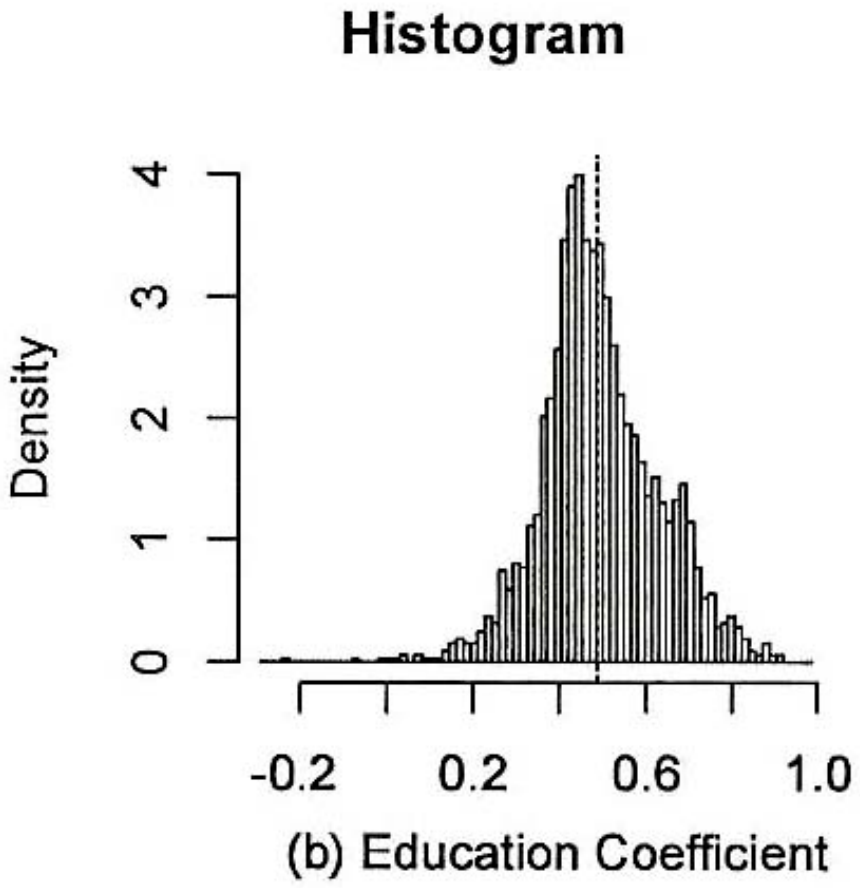
- Have *n* sets of observations on the response variable and the values of the explanatory variables that go with that value of the response variable

- Simply bootstrap these *n* sets (with replacement) many times, compute least-squares regression (or other type of regression) and create bootstrap histograms (sampling distributions) of the regression coefficients and related statistics of interest

- From the bootstrap distributions find confidence intervals and *p*-values

Example


Histogram
Density
0.0
1.0
2.0
3.0
0.0
0.5
1.0
1.5
(a) Income Coefficient


Histogram
Density
0
1
2
3
4
-0.2
0.2
0.6
1.0
(b) Education Coefficient

Lecture Notes

Regression Diagnostics

Roy Welsch

Spring 2017

SUTD-MIT

Figure
Regression Lines for Four Analysis

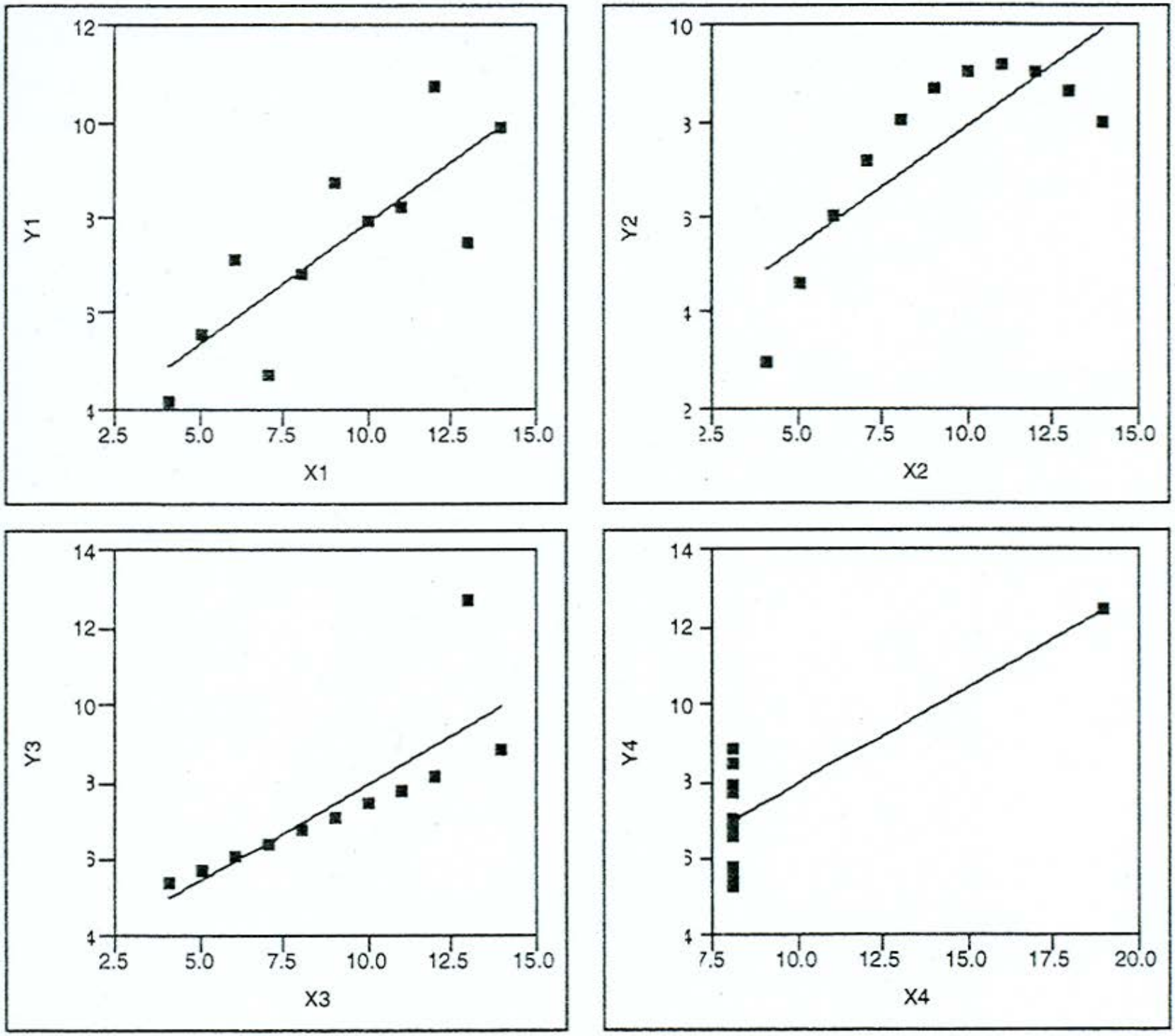
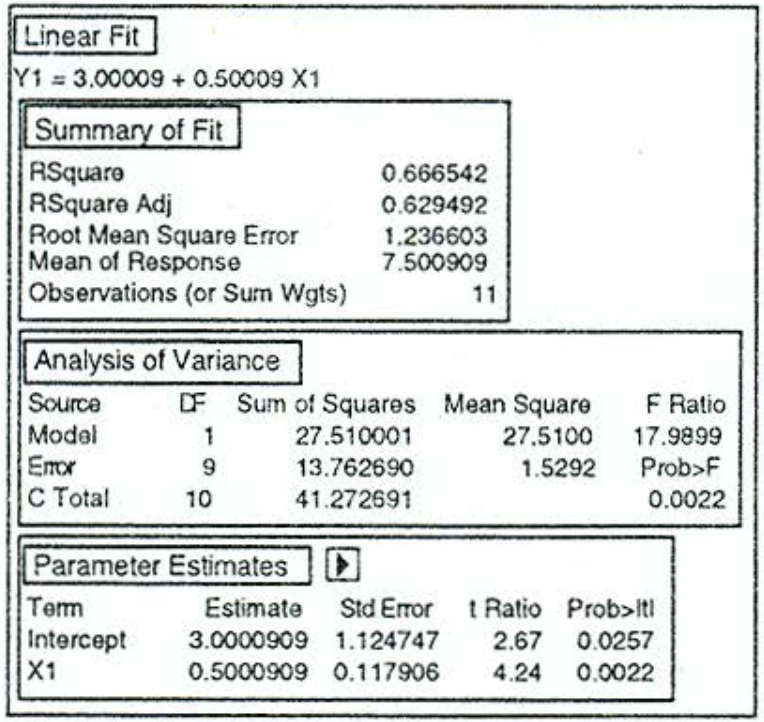


Figure Statistical Reports for Four Analyses



Linear Fit

Y1 = 3.00009 + 0.50009 X1

Summary of Fit

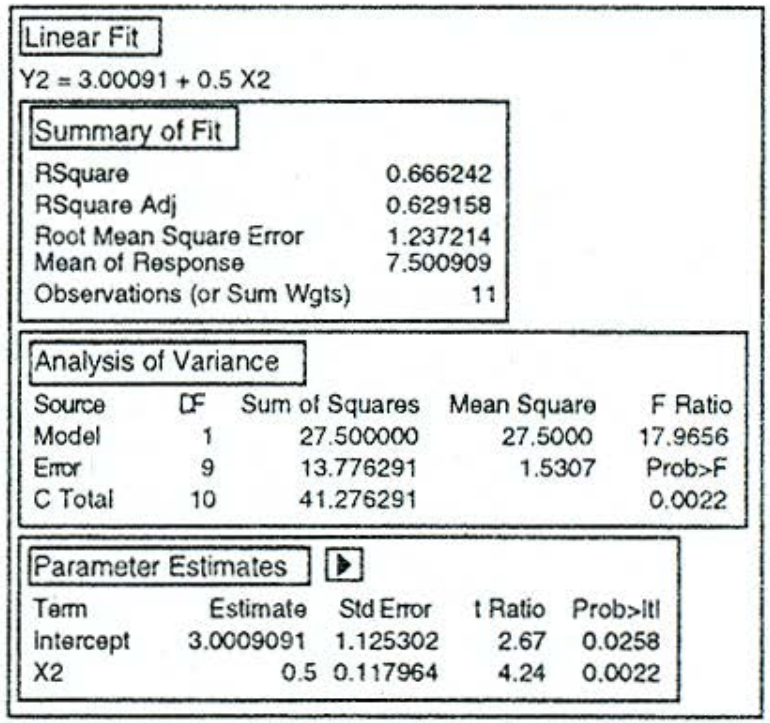
RSquare	0.666542
RSquare Adj	0.629492
Root Mean Square Error	1.236603
Mean of Response	7.500909
Observations (or Sum Wgts)	11

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	27.510001	27.5100	17.9899
Error	9	13.762690	1.5292	Prob>F
C Total	10	41.272691		0.0022

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3.0000909	1.124747	2.67	0.0257
X1	0.5000909	0.117906	4.24	0.0022



Linear Fit

Y2 = 3.00091 + 0.5 X2

Summary of Fit

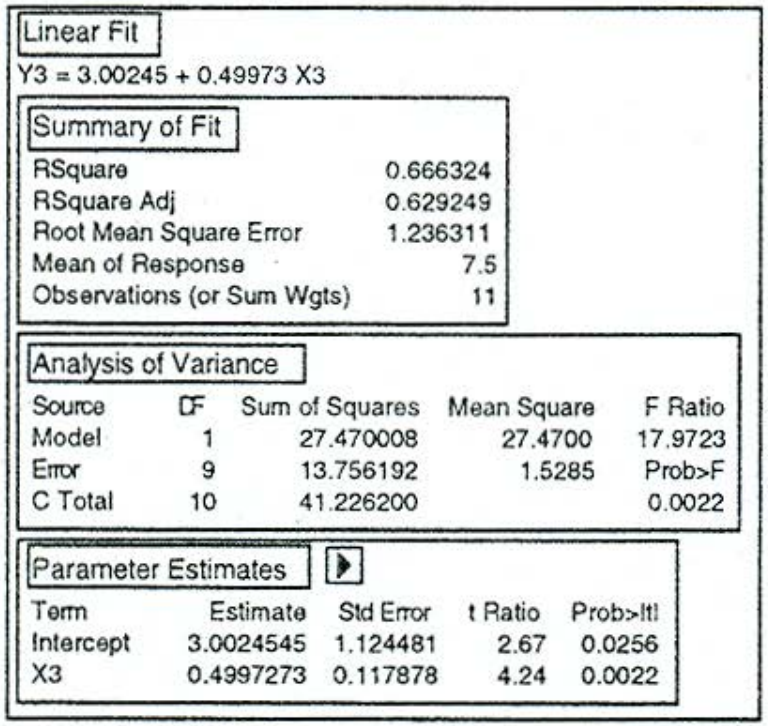
RSquare	0.666242
RSquare Adj	0.629158
Root Mean Square Error	1.237214
Mean of Response	7.500909
Observations (or Sum Wgts)	11

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	27.500000	27.5000	17.9656
Error	9	13.776291	1.5307	Prob>F
C Total	10	41.276291		0.0022

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3.0009091	1.125302	2.67	0.0258
X2	0.5	0.117964	4.24	0.0022



Linear Fit

Y3 = 3.00245 + 0.49973 X3

Summary of Fit

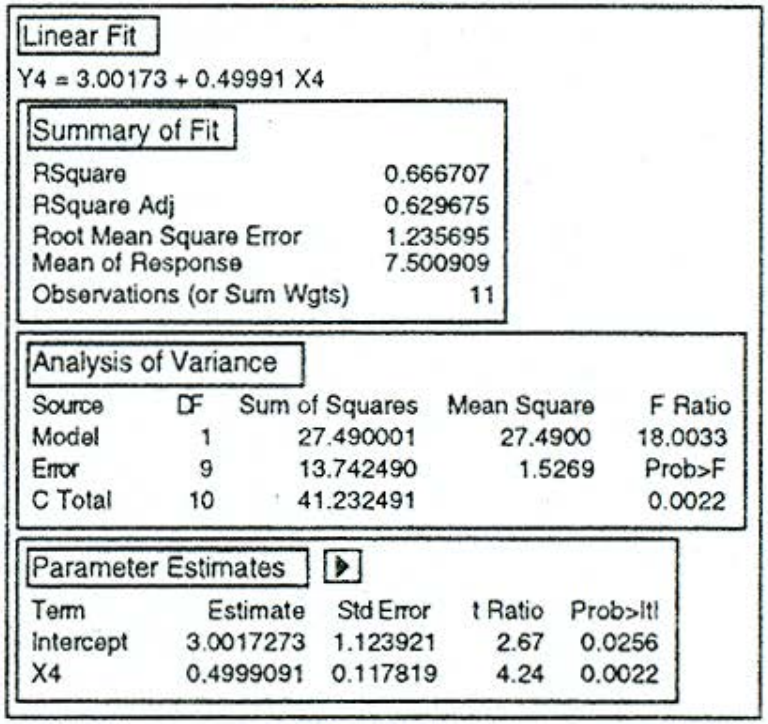
RSquare	0.666324
RSquare Adj	0.629249
Root Mean Square Error	1.236311
Mean of Response	7.5
Observations (or Sum Wgts)	11

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	27.470008	27.4700	17.9723
Error	9	13.756192	1.5285	Prob>F
C Total	10	41.226200		0.0022

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3.0024545	1.124481	2.67	0.0256
X3	0.4997273	0.117878	4.24	0.0022



Linear Fit

Y4 = 3.00173 + 0.49991 X4

Summary of Fit

RSquare	0.666707
RSquare Adj	0.629675
Root Mean Square Error	1.235695
Mean of Response	7.500909
Observations (or Sum Wgts)	11

Analysis of Variance

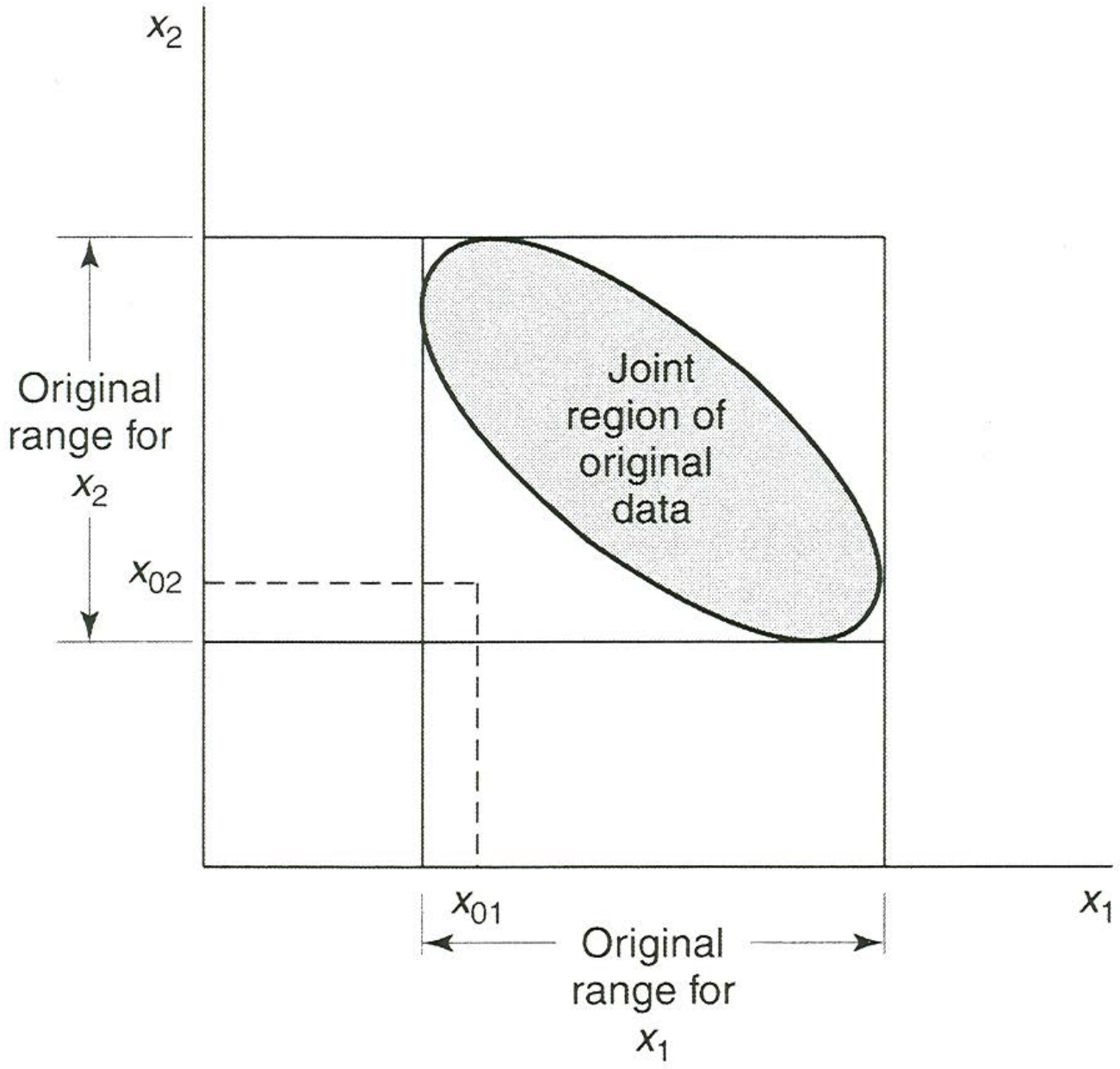
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	27.490001	27.4900	18.0033
Error	9	13.742490	1.5269	Prob>F
C Total	10	41.232491		0.0022

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3.0017273	1.123921	2.67	0.0256
X4	0.4999091	0.117819	4.24	0.0022

Extrapolation

We are concerned about extrapolation too far beyond the data we have on the explanatory variables. In the case of simple linear regression, we just look at how far away a point x_i is from $\bar{x}$. In higher dimensions, we need to be concerned about hidden extrapolation.



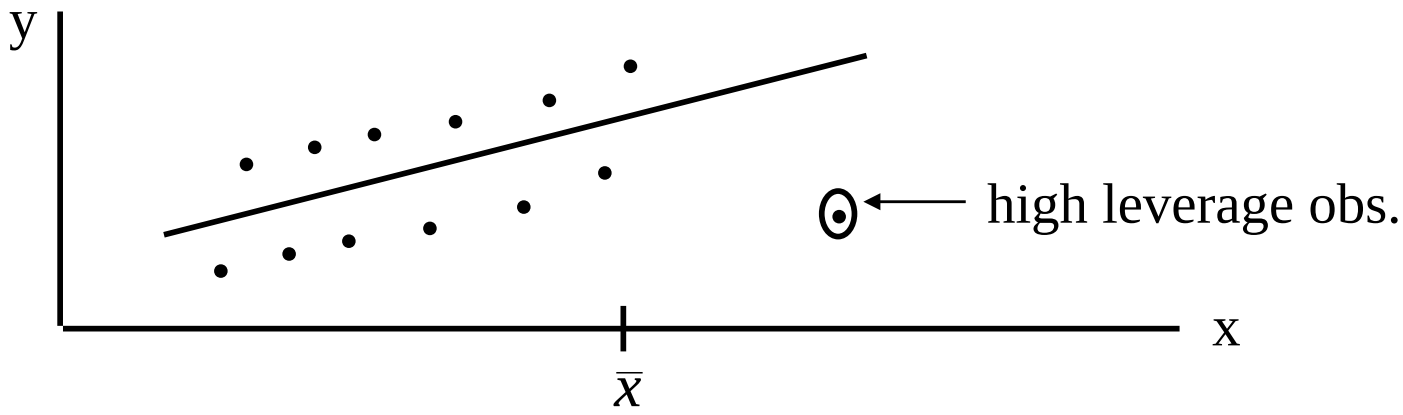
An example of extrapolation in multiple regression.

Leverage

In simple linear reg., define leverage as $h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_k (x_k - \bar{x})^2}$

Note that when $|x_i - \bar{x}|$ large we are likely to have high leverage.

[Turns out that $\frac{1}{n} \le h_{ii} \le 1$.]



Leverage tells us about outlying points in *x*. Pays to look at high leverage values and check for errors, etc.

In multiple regression:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \mathbf{b} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{H}\mathbf{y}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$ ("hat" matrix)

h_{ii} are the diagonal elements of H.

We can show that:

$$\frac{1}{n} \le h_{ii} \le 1,$$

$$\sum_{i=1}^{n} h_{ii} = k + 1 \quad (\#\text{ of coefficients})$$

Considered large when $h_{ii} > 2(k + 1)/n$ or $3(k + 1)/n$.
Best to print out or plot the h_{ii} and look for the larger values.

Two Measures of Influence of an Observation

- Influence on the Fitted Value of the dependent variable [DFFITS]
- Influence on the Regression Coefficients [DFBETAS]
- Only look at DFFITS now

DFFITS

- DFFITS is

$$\text{DFFITS}_i = \frac{\left(\hat{y}_i - \hat{y}_i(i)\right)}{s(i)\sqrt{h_{ii}}}$$

- "DF" indicates the difference in $\hat{y}_i$ with and without the *i*th observation.
- The denominator standardizes the difference:

$$\text{Var}(\hat{y}_i) = \sigma^2 h_{ii}$$

- DFFITS represents *the number of standard errors that the fitted value $\hat{y}_i$ changes if the ith observation is removed from the data set.*
- $\hat{y}_i(i) = \hat{\beta}_0(i) + \hat{\beta}_1(i)x_{i1} + \hat{\beta}_2(i)x_{i2} + \cdots + \hat{\beta}_k(i)x_{ik}$

How Large?

- First of all, *there are no natural critical values* called for when doing diagnostics
- We are NOT doing significance testing
- There are several schools of thought:
  - The choice should be a function of sample size
  - No need for an omnibus formula: clear understanding of the statistics allows interpretation based on experience

Cut-offs

These measures should be used informally since no theory exists without assumptions about the explanatory variables. Generally studentized residuals larger than 3 (in absolute value) should be noticed and h_{ii} (leverage) values greater than $3(k+1)/n$. DFFITS values greater than $3\sqrt{k+1}/\sqrt{n}$ in absolute value are of interest.

These values indicate possible errors, etc. You may want to refit without these data to see what happens. If you can find nothing wrong with these observations, then you should leave them in, but warn your client that they are possibly very influential.

NOTE: Cook's D = $\text{DFFITS}^2/(k+1)$

Some algebra shows that

$$\text{DFFITS}_i = \frac{e_i}{s(i)\sqrt{1-h_{ii}}}\left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2}$$

which is a product of the studentized residual and a measure of leverage.

Good diagnostics include looking at residuals and leverage separately, and studying influence which is a product of the two.

2003 MLB team salaries and wins

Obs	team	salary03	wins03	lnsal03
1	New York Yankees*	152.75	101	5.02880
2	New York Mets	117.18	66	4.76371
3	Atlanta Braves*	106.24	101	4.66570
4	Los Angeles Dodgers	105.87	85	4.66221
5	Texas Rangers	103.49	71	4.63947
6	Boston Red Sox*	99.95	95	4.60467
7	Seattle Mariners	86.96	93	4.46545
8	St. Louis Cardinals	83.49	85	4.42473
9	San Francisco Giants*	82.85	100	4.41703
10	Arizona Diamondbacks	80.64	84	4.38999
11	Chicago Cubs*	79.87	88	4.38040
12	Anaheim Angels	79.03	77	4.36983
13	Baltimore Orioles	73.88	71	4.30244
14	Houston Astros	71.04	87	4.26324
15	Philadelphia Phillies	70.78	86	4.25958
16	Colorado Rockies	67.18	74	4.20738
17	Cincinnati Reds	59.36	69	4.08362
18	Minnesota Twins*	55.51	90	4.01656
19	Pittsburgh Pirates	54.81	75	4.00387
20	Montreal Expos	51.95	83	3.95028
21	Toronto Blue Jays	51.27	86	3.93711
22	Chicago White Sox	51.01	86	3.93202
23	Oakland Athletics*	50.26	96	3.91721
24	Detroit Tigers	49.17	43	3.89528
25	Florida Marlins*	49.05	91	3.89284
26	Cleveland Indians	48.59	68	3.88342
27	San Diego Padres	47.93	64	3.86974
28	Milwaukee Brewers	40.63	68	3.70451
29	Kansas City Royals	40.52	83	3.70180
30	Tampa Bay Devil Rays	19.63	63	2.97706

Regression to Predict Wins based on Salary (03)

The REG Procedure
Model: MODEL1
Dependent Variable: wins03

Number of Observations Read	30
Number of Observations Used	30

Analysis of Variance

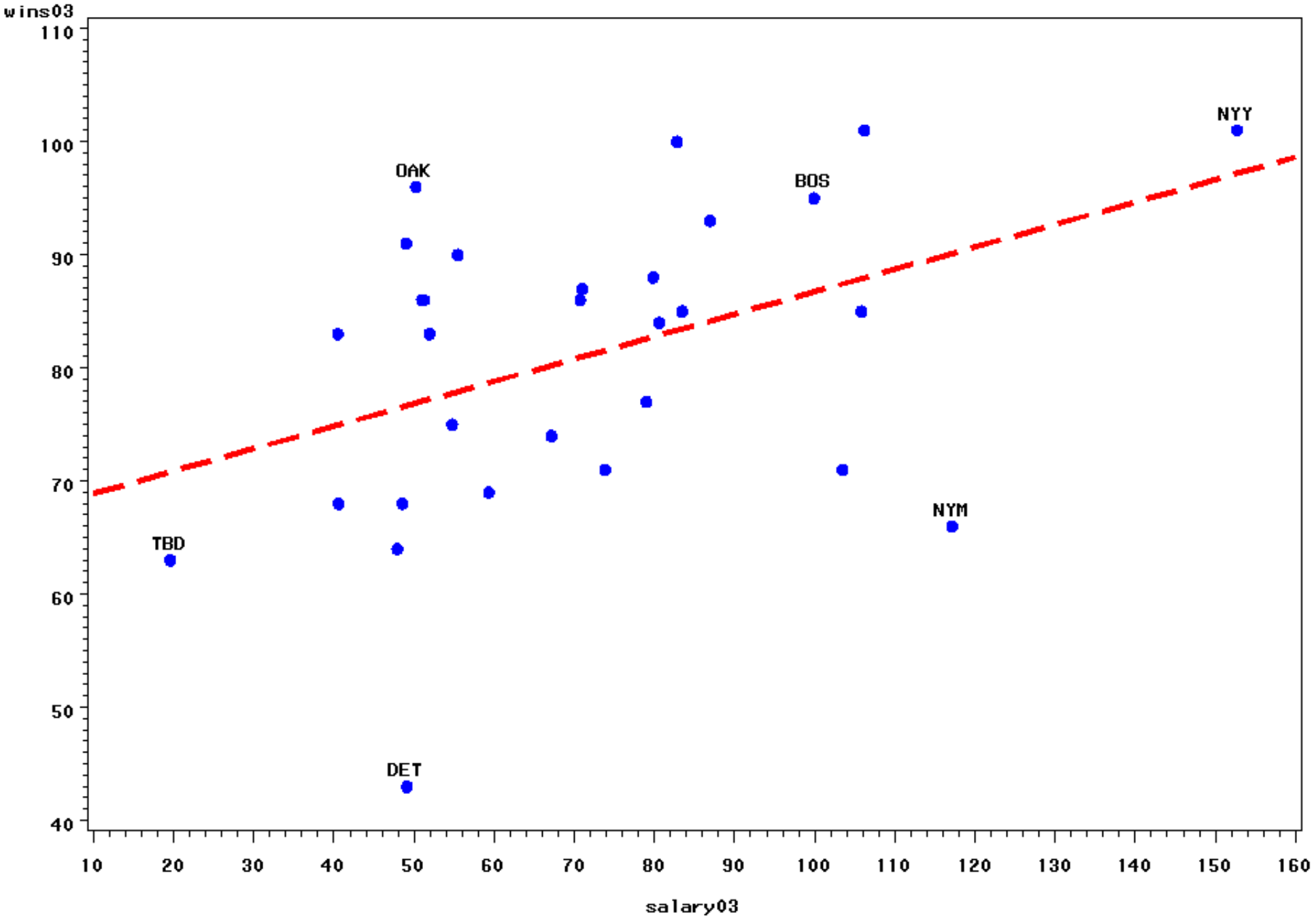
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	889.41297	889.41297	5.80	0.0228
Error	28	4291.55370	153.26977		
Corrected Total	29	5180.96667			

Root MSE	12.38022	R-Square	0.1717
Dependent Mean	80.96667	Adj R-Sq	0.1421
Coeff Var	15.29051		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	66.90062	6.26135	10.68	<.0001	0
salary03	1	0.19803	0.08221	2.41	0.0228	1.00000

Regression to Predict Wins based on Salary (03)



The REG Procedure
Model: MODEL1
Dependent Variable: wins03

Output Statistics

Obs	-2-1 0 1 2	Cook's D	RStudent	Hat Diag H	DFFITS
1	| |	0.035	0.3734	0.3278	0.2608
2	|**** |	0.317	-2.2267	0.1272	-0.8502
3	| |**	0.059	1.1092	0.0880	0.3446
4	| |	0.003	-0.2382	0.0869	-0.0734
5	|** |	0.083	-1.4043	0.0798	-0.4135
6	| |*	0.018	0.6892	0.0702	0.1894
7	| |*	0.013	0.7275	0.0445	0.1570
8	| |	0.000	0.1268	0.0402	0.0259
9	| |**	0.039	1.3991	0.0395	0.2837
10	| |	0.000	0.0914	0.0374	0.0180
11	| |	0.004	0.4284	0.0368	0.0837
12	| |	0.004	-0.4502	0.0362	-0.0872
13	* |	0.013	-0.8613	0.0337	-0.1608
14	| |	0.004	0.4887	0.0333	0.0908
15	| |	0.003	0.4113	0.0333	0.0764
16	* |	0.005	-0.5030	0.0340	-0.0944
17	* |	0.013	-0.7904	0.0393	-0.1599
18	| |**	0.023	1.0001	0.0440	0.2144
19	| |	0.001	-0.2238	0.0449	-0.0485
20	| |	0.006	0.4748	0.0494	0.1082
21	| |*	0.015	0.7355	0.0505	0.1697
22	| |*	0.015	0.7400	0.0510	0.1716
23	| |***	0.070	1.6355	0.0524	0.3844
24	|***** |	0.225	-3.2309	0.0544	-0.7750
25	| |**	0.041	1.2047	0.0546	0.2896
26	* |	0.015	-0.7019	0.0555	-0.1702
27	|** |	0.032	-1.0319	0.0569	-0.2534
28	* |	0.014	-0.5761	0.0741	-0.1630
29	| |*	0.018	0.6713	0.0744	0.1903
30	* |	0.041	-0.6756	0.1498	-0.2836

Regression to Predict Wins based on Salary (03)
No New York Yankees

The REG Procedure
Model: MODEL1
Dependent Variable: wins03

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	496.29214	496.29214	3.14	0.0878
Error	27	4269.50096	158.12967		
Corrected Total	28	4765.79310			

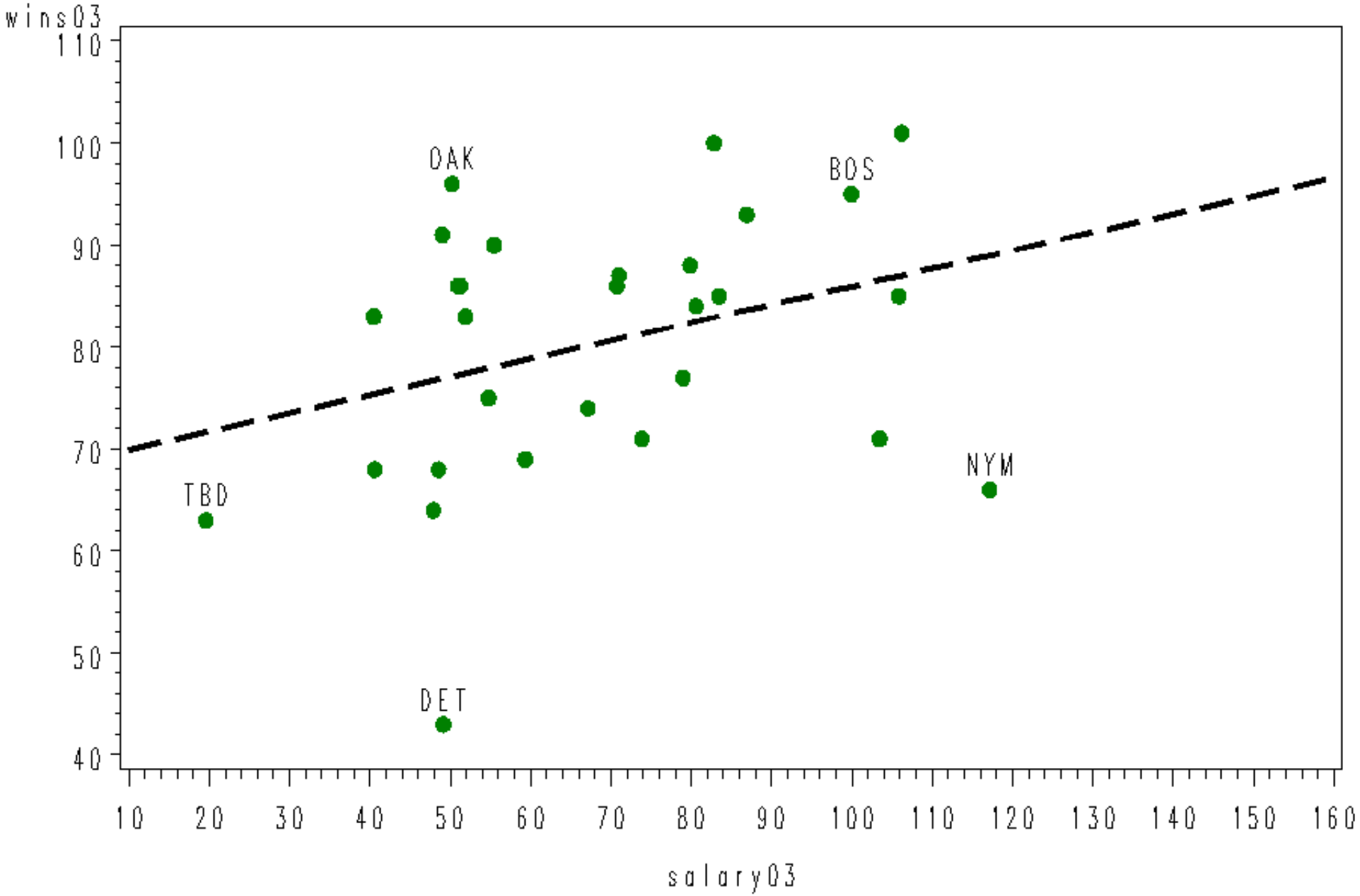
Root MSE	12.57496	R-Square	0.1041
Dependent Mean	80.27586	Adj R-Sq	0.0710
Coeff Var	15.66469		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	68.17561	7.21832	9.44	<.0001
salary03	1	0.17739	0.10013	1.77	0.0878

Regression to Predict Wins based on Salary (03)

No New York Yankees



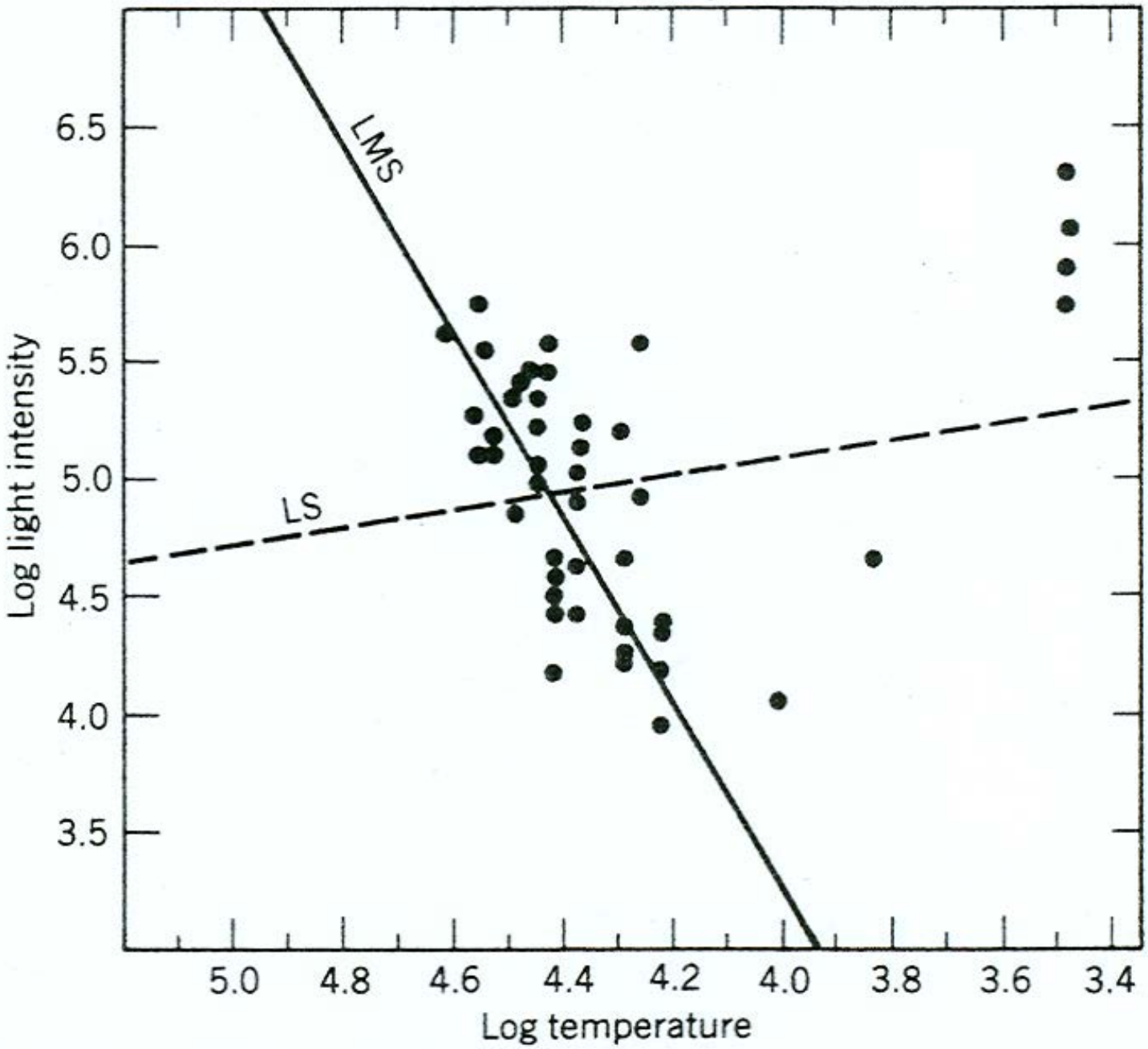


Figure 4. Hertzsprung–Russell diagram of the star cluster CYG OB1 with the LS (dashed line) and LMS fit (solid line).

Lecture Notes

Data Mining

Roy Welsch

Spring 2017

SUTD-MIT

Data Mining

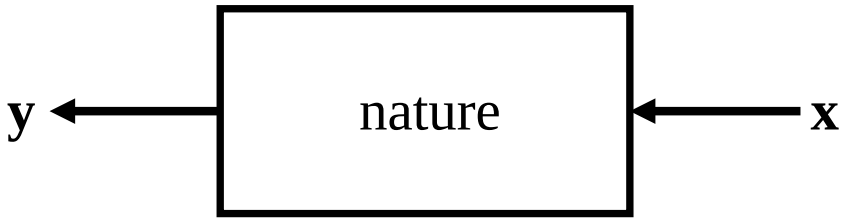
- Search for **patterns** in **large data sets**
  - Businesses data: marketing, finance, production ...
    - Collected for some purpose, often useful for others
    - From government or private companies
  - Makes use of
    - Statistics – all the basic activities, and
      - Prediction, classification, clustering
    - Computer science – efficient algorithms (instructions) for
      - Collecting, maintaining, organizing, analyzing data
    - Optimization – calculations to achieve a goal
      - Maximize or minimize (e.g. sales or costs)

Statistical Modeling

- Statistics starts with data.
- Data from a black box and subject to error.

x (vector of input or explanatory variable)

y (vector of response variables)



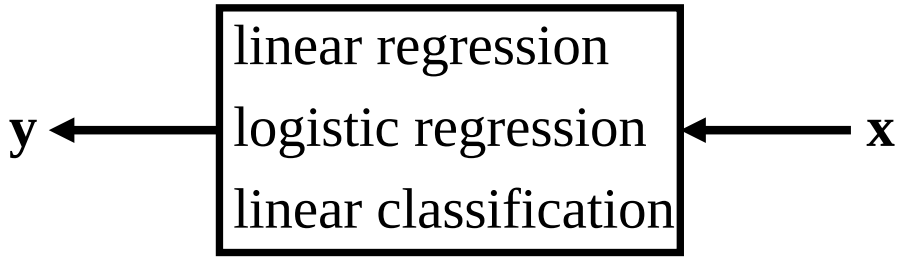
- Two goals for analyzing the data:
  - Prediction of response to future inputs.
  - Information about how nature associates inputs and response variables.
- Two approaches to achieving these goals:
  - The data modelers
  - The algorithmic modelers

The Data Modeling Culture

- Assumes a stochastic data model (data generating process) for the black box

  response = *f* (explanatory variables, random noise, parameters)

  Values of parameters are estimated from the data and the model used for information or prediction.

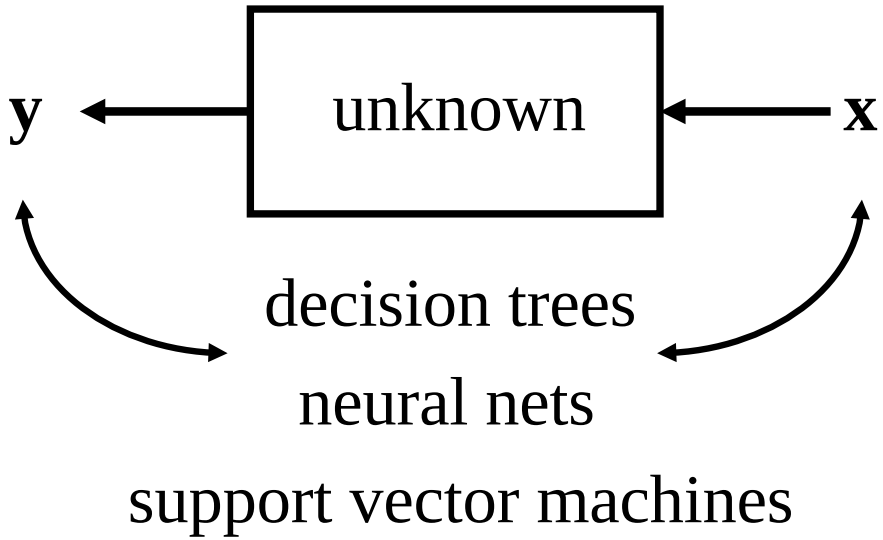


- Model validation: Yes-no using goodness-of-fit tests and residual examination.
- But, consider the bootstrap.

The Algorithmic Modeling Culture

- Inside of black box viewed as complex and unknown.

- Find a function $f(\mathbf{x})$ – an algorithm that operates on **x** to predict responses **y**.



- Model validation: measured by predictive accuracy (on holdout test sets).

Difficult Issues for Everyone

- The multiplicity of good models.
- The conflict between simplicity and accuracy.
- High dimensional data – curse or blessing?
- Data corruption (unintentional and intentional).
- Sensitivity to errors. Robustness.
- Specifying loss and risk functions.
- Communication.

Emerging Major Data Mining Applications

- Spam
- Bioinformatics/Genomics
- Medical History Data – Insurance Claims
- Personalization of services in e-commerce
- RFID Tags
- Security :
  - Container Shipments
  - Network Intrusion Detection
  - Stock trading fraud

108

Data Partitions

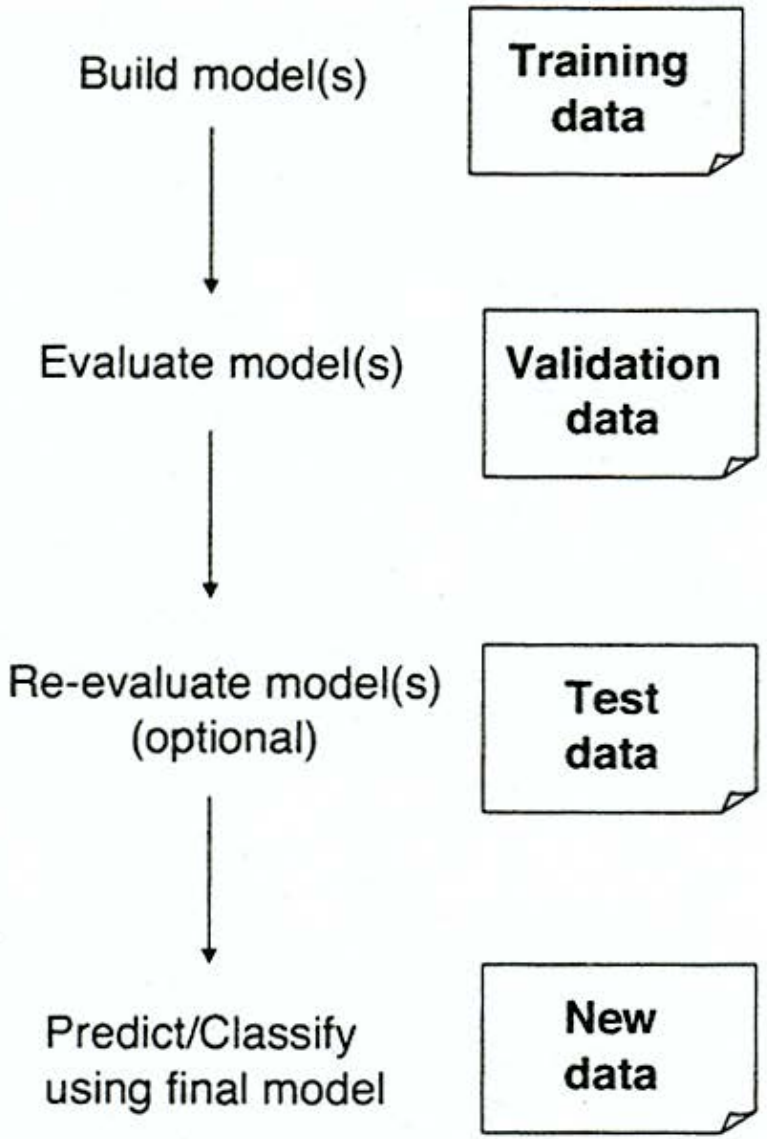
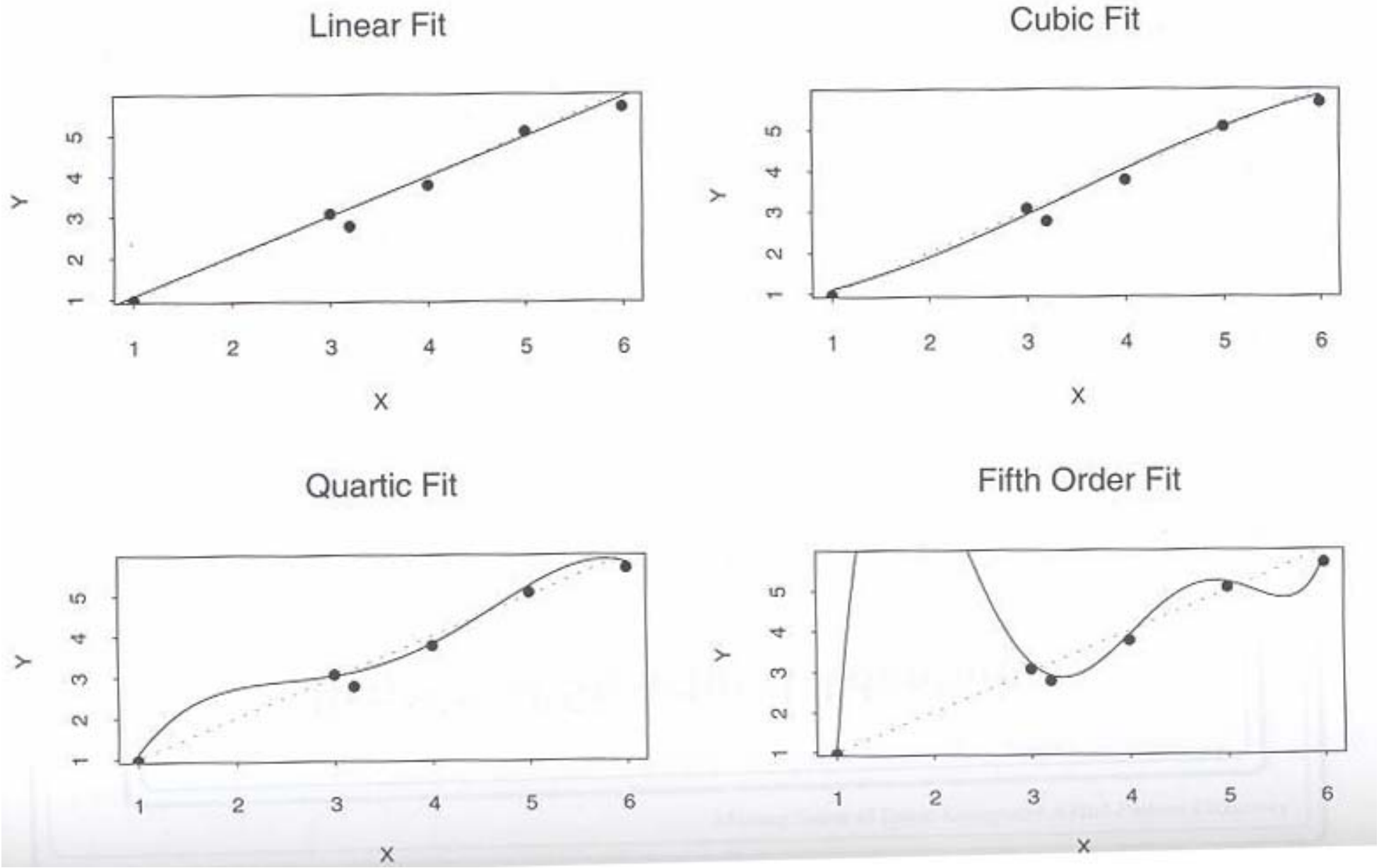
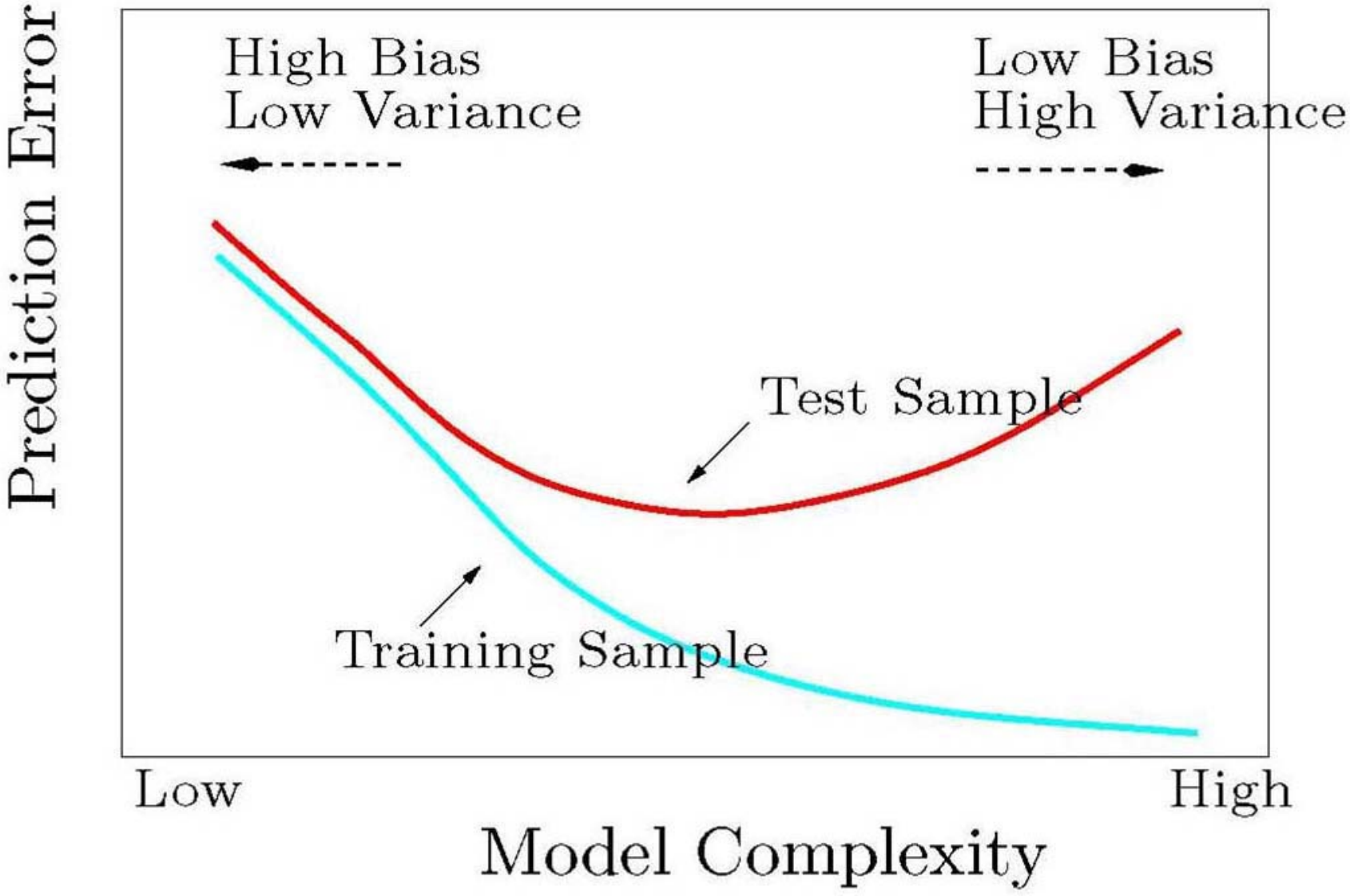


Figure : The Three Data Partitions and Their Role in the Data Mining Process

Regression Examples with Overfit





Test and training error as a function of model complexity.

Classification: Application

- Direct Marketing
  - Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
  - Approach:
    - Use the data for a similar product introduced before.
    - We know which customers decided to buy and which decided otherwise. This *{buy, don't buy}* decision forms the *class attribute*.
    - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
      - Type of business, where they stay, how much they earn, etc.
    - Use this information as input attributes to learn a classifier model.

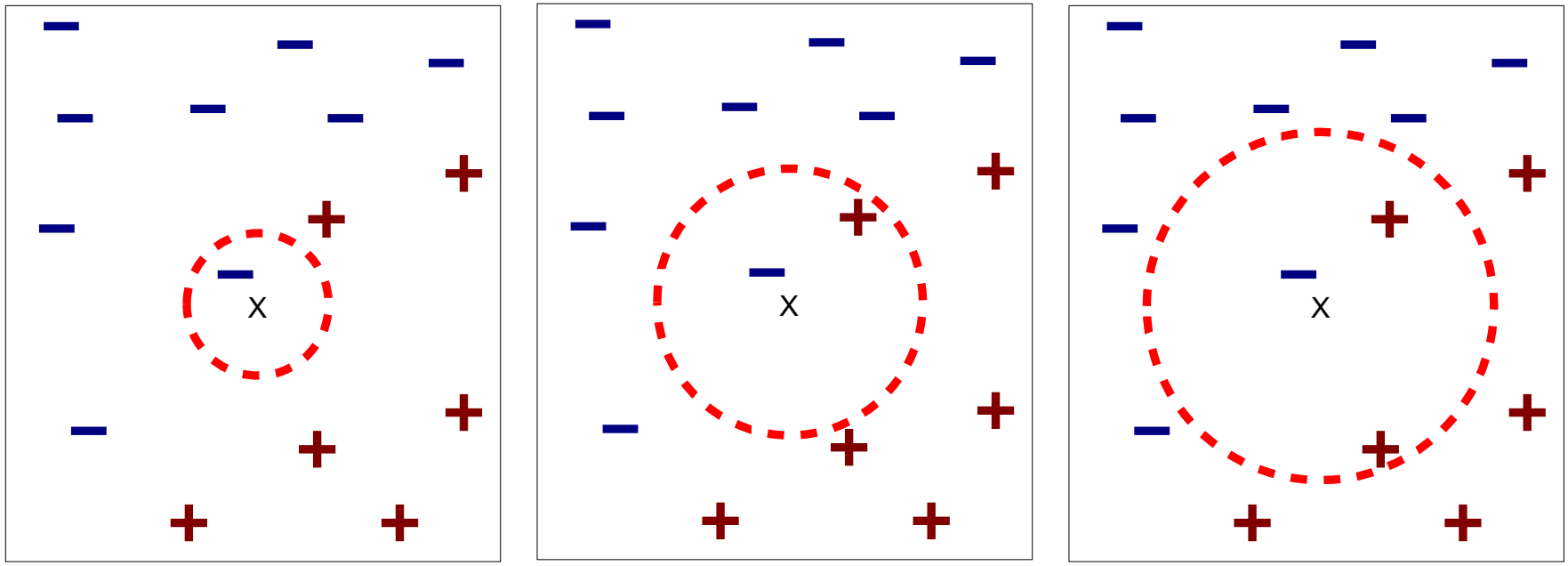
From [Berry & Linoff] Data Mining Techniques, 1997

112

Nearest Neighbor Methods

- No assumptions about $y = f(x_1, x_2, \ldots, x_p)$
- A non-parametric method
- Training data with y = class to which observation belongs.
- Given a new observation $(u_1, u_2, \ldots, u_p)$, look for "near by" values in training data to decide how to classify.
- Need a distance measure – Euclidean, or other.

Definition of Nearest Neighbor



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

K-nearest neighbors of a record x are data points that have the k smallest distance to x

Choosing *K*

- Divide data into training, validation, and test sets.
- Train with various values of *k*, compute error on validation data, choose best *k*.
- Still have test data to measure performance on fresh data not used in training or choosing best value for *k*.

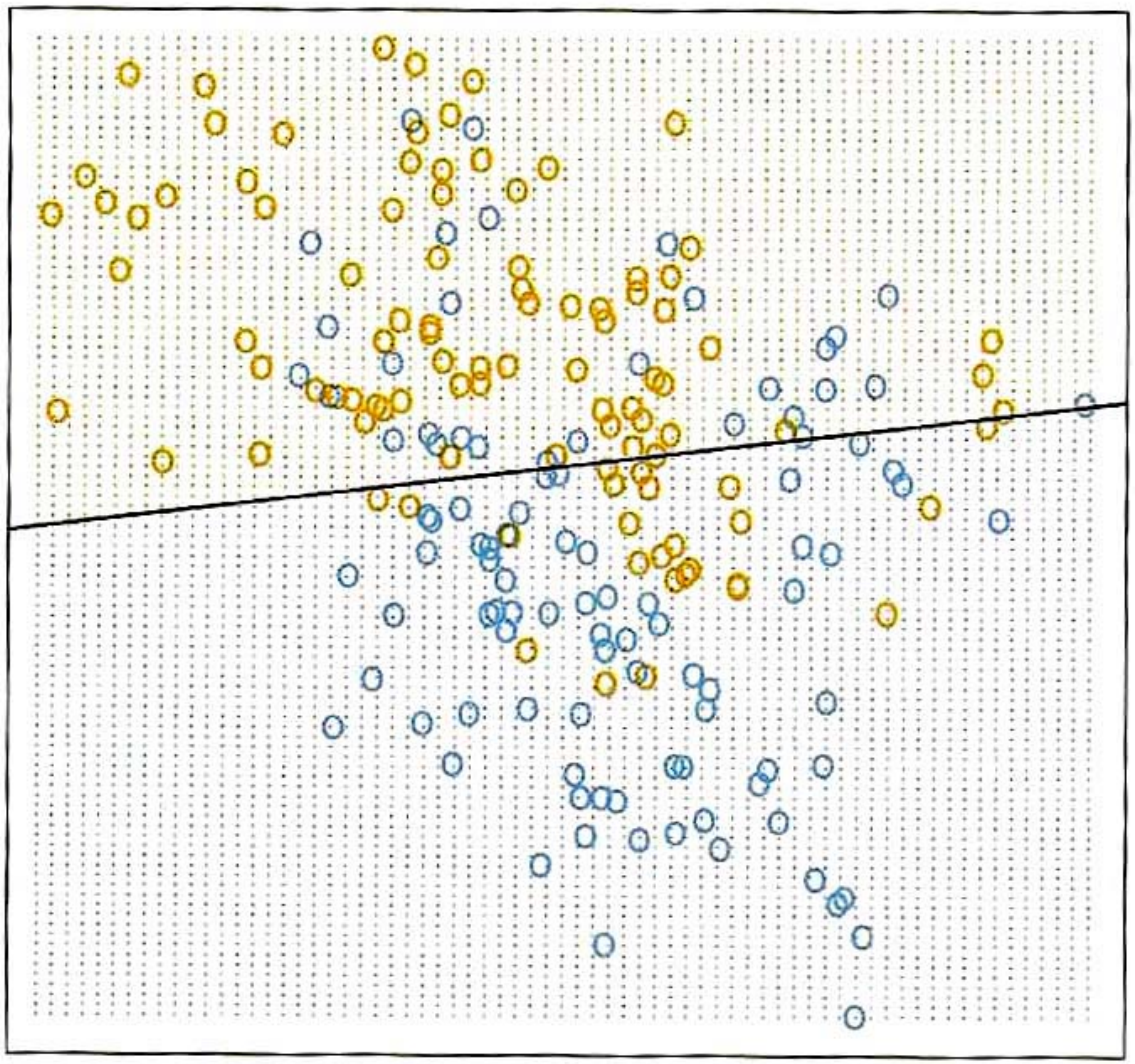


FIGURE A. *A classification example in two dimensions. The classes are coded as a binary variable (*BLUE $= 0$, ORANGE $= 1$*), and then fit by linear regression. The line is the decision boundary defined by* $x^T\hat{\beta} = 0.5$. *The orange shaded region denotes that part of input space classified as* ORANGE, *while the blue region is classified as* BLUE.

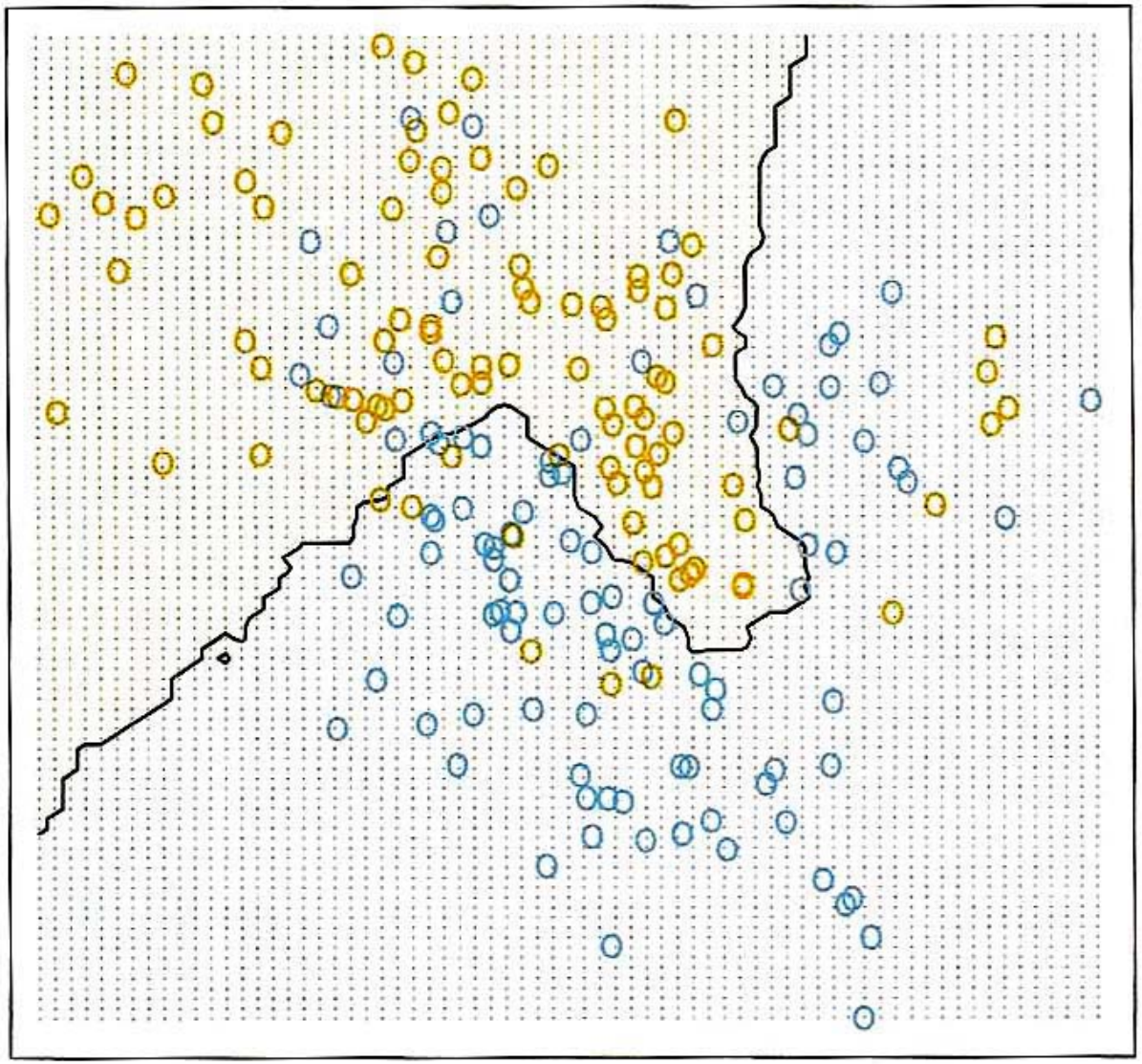


FIGURE. *The same classification example in two dimensions as in Figure* A. *The classes are coded as a binary variable* (BLUE $= 0$, ORANGE $= 1$) *and then fit by* 15*-nearest-neighbor averaging. The predicted class is hence chosen by majority vote amongst the* 15*-nearest neighbors.*

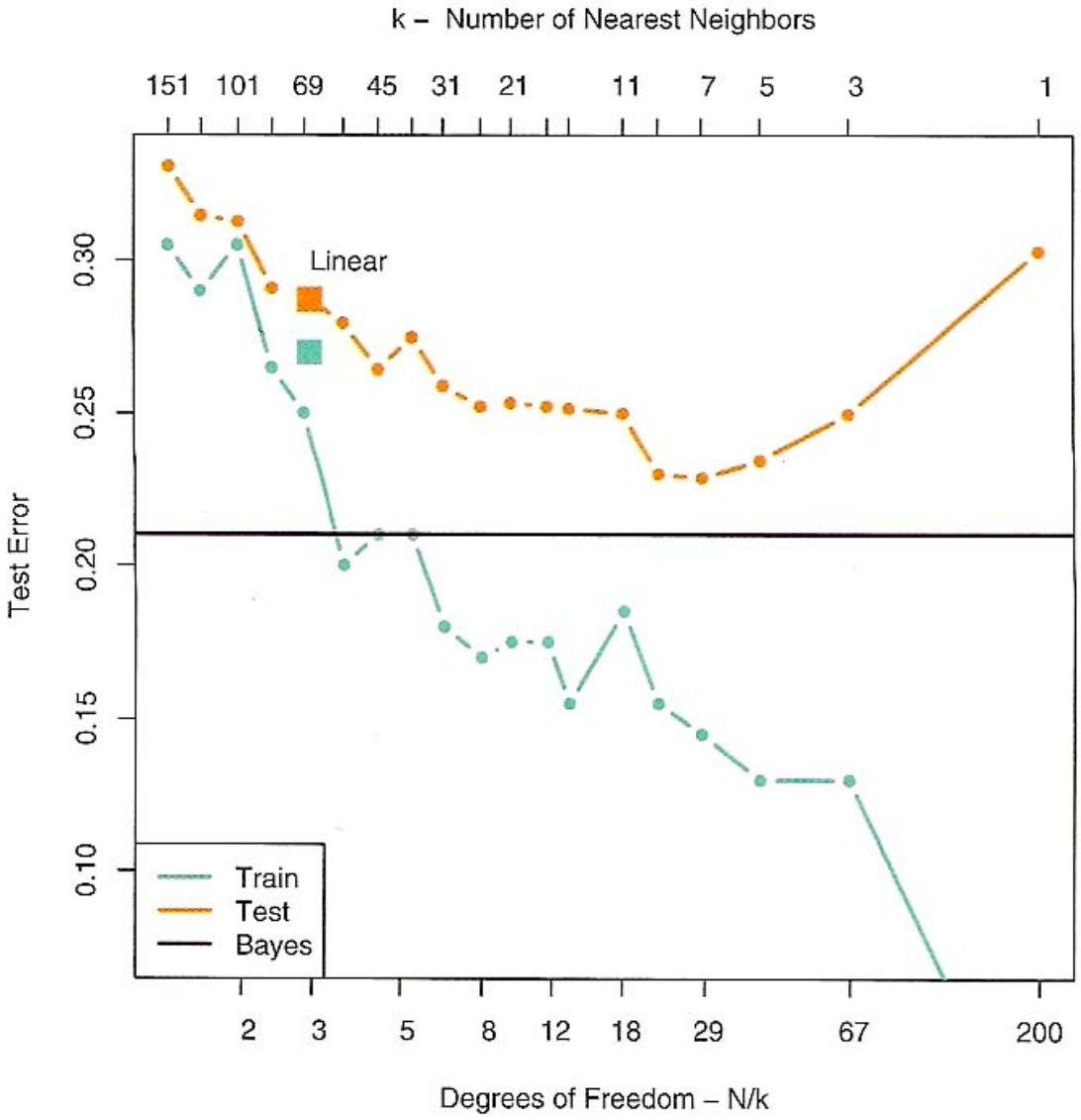
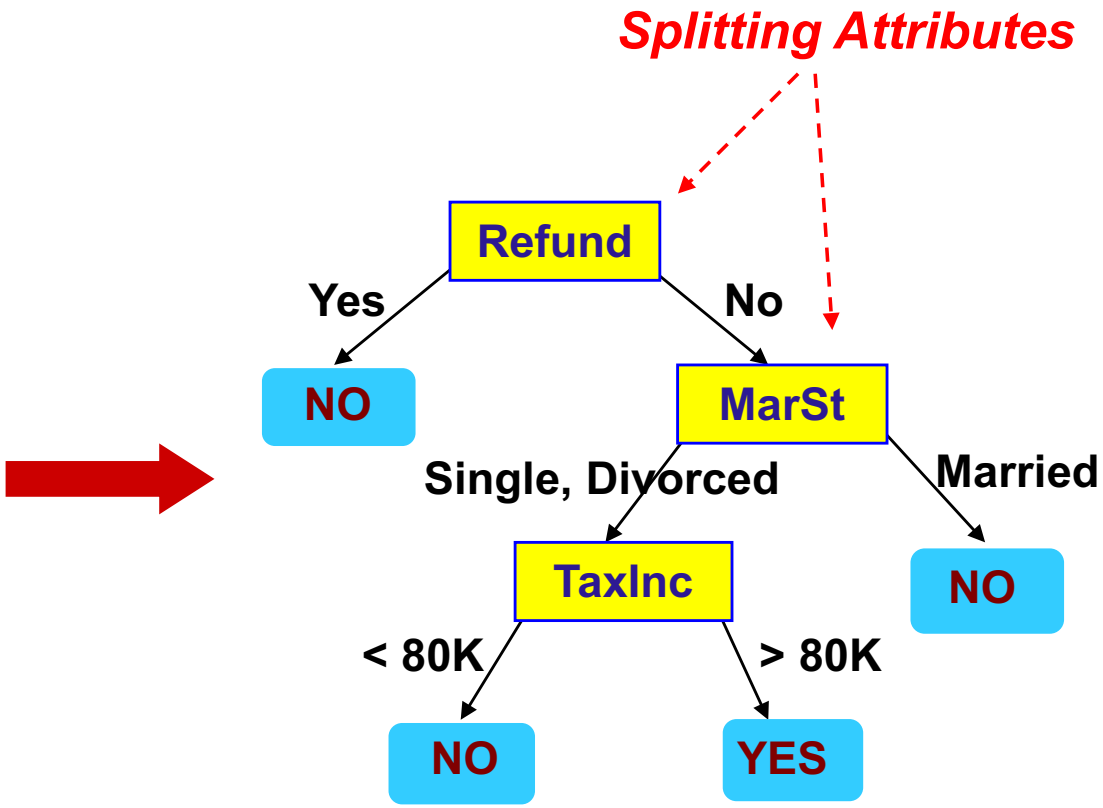


FIGURE 2.4. *Misclassification curves for the simulation example used in Figures 2.1, 2.2 and 2.3. A single training sample of size* 200 *was used, and a test sample of size* $10,000$. *The orange curves are test and the blue are training error for k-nearest-neighbor classification. The results for linear regression are the bigger orange and blue squares at three degrees of freedom. The purple line is the optimal Bayes error rate.*

Example of a Decision Tree

Tid	Refund	Marital Status	Taxable Income	Cheat
	categorical	categorical	continuous	class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Data

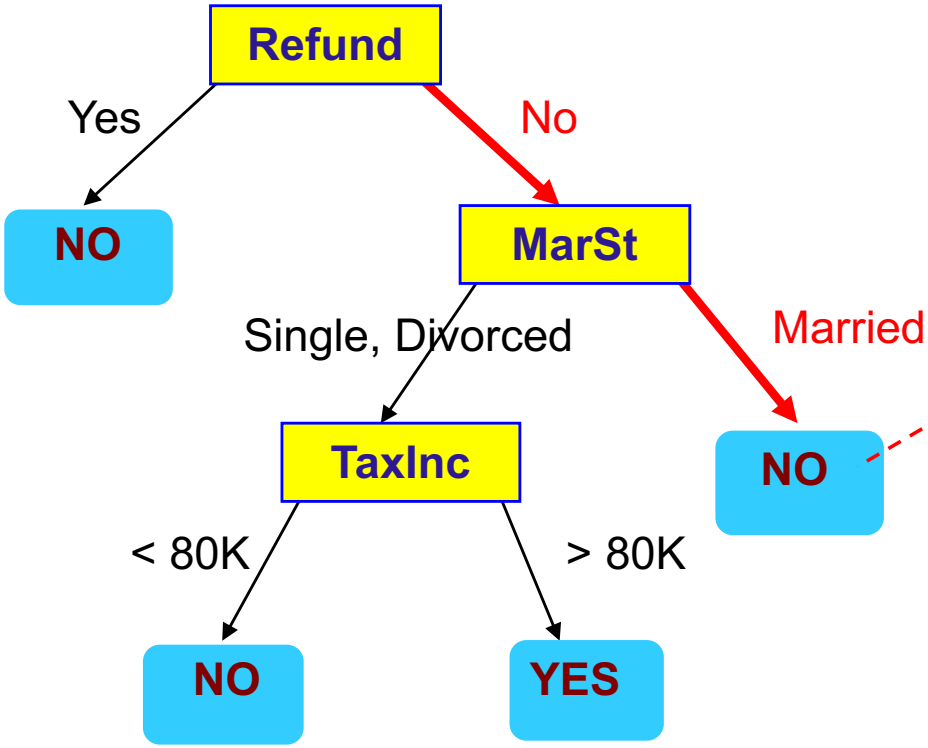


Model: Decision Tree

Apply Model to Test Data

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Assign Cheat to “No”

Recursive Partitioning

- Outcome variable *y*, predictor variables $x_1, \ldots, x_p$
- Outcome variable – categorical
- Recursive partitioning – divide *x-space* into (multi-dimensional) rectangles
  - First select, say x_i, and a value of x_i, say s_i
  - Split *x-space* into 2 parts: points with $x_i > s_i$ and $x_i \leq s_i$
  - Repeat, trying to divide space so that each rectangle is "pure" – has points belonging to one class

Partition 1

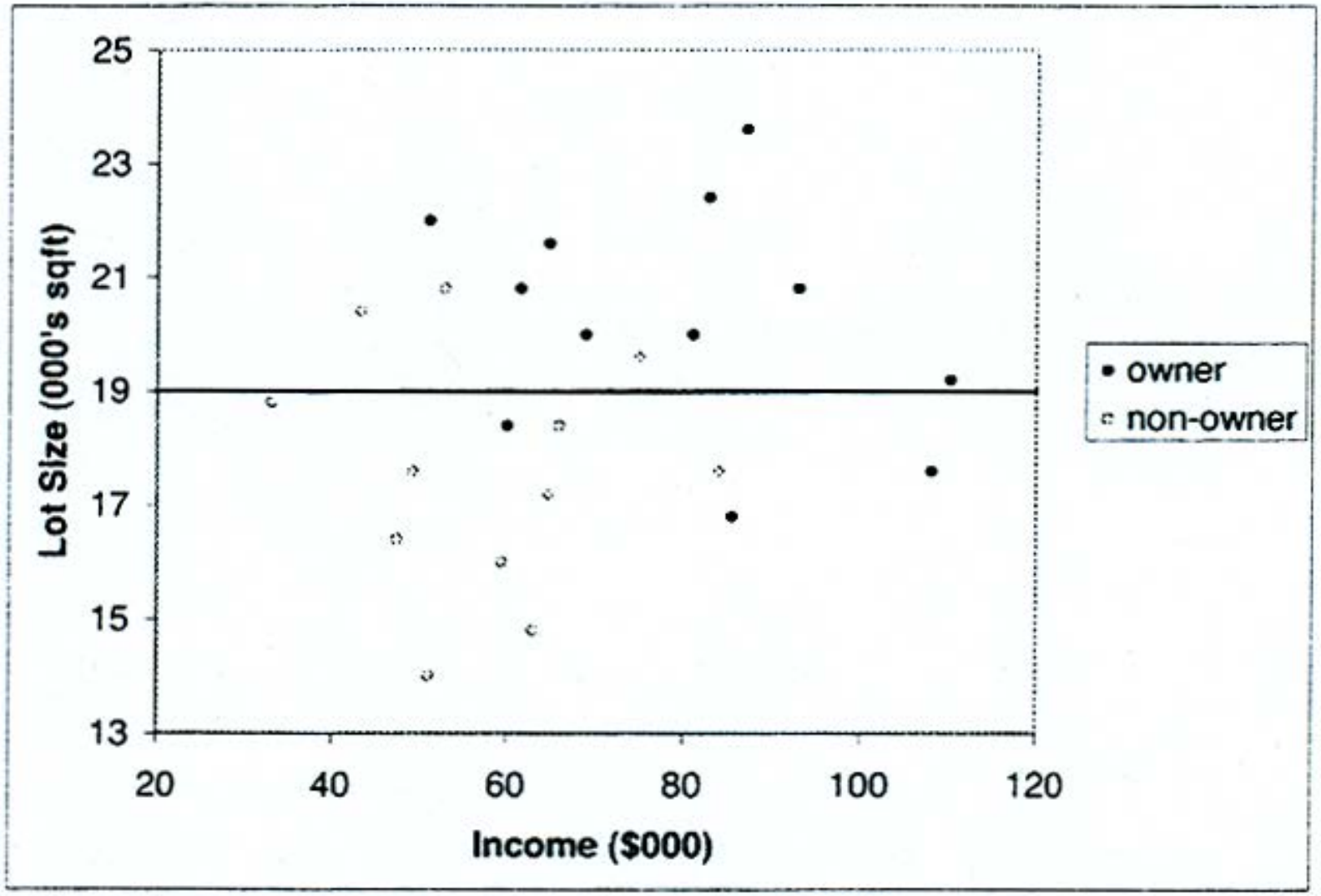


Figure B: Splitting the 24 Observations By Lot Size Value of 19

Partition 2

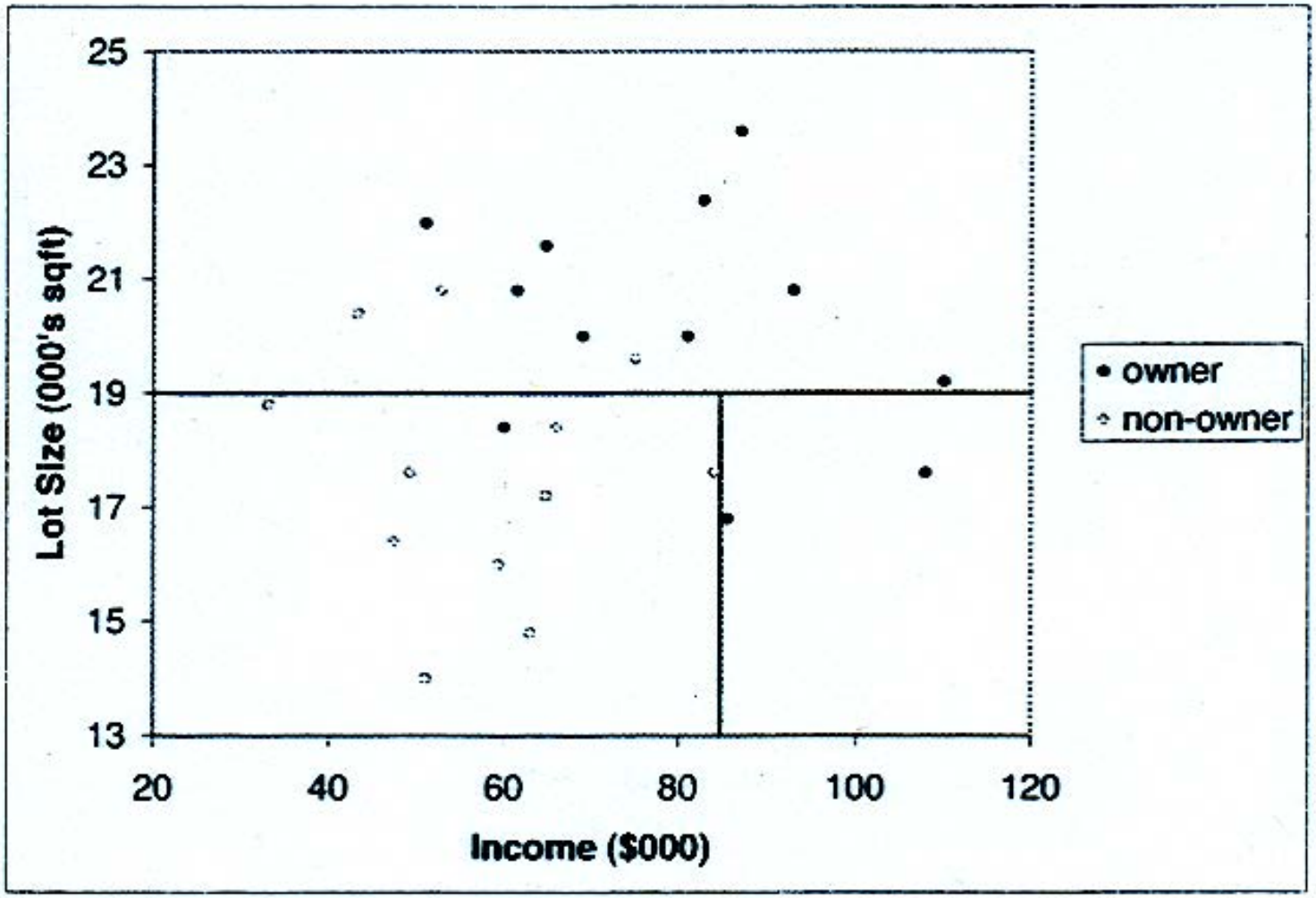


Figure D: Splitting the 24 Observations By Lot Size Value of 19K, and then Income Value of 84.75K

Final Partitioning

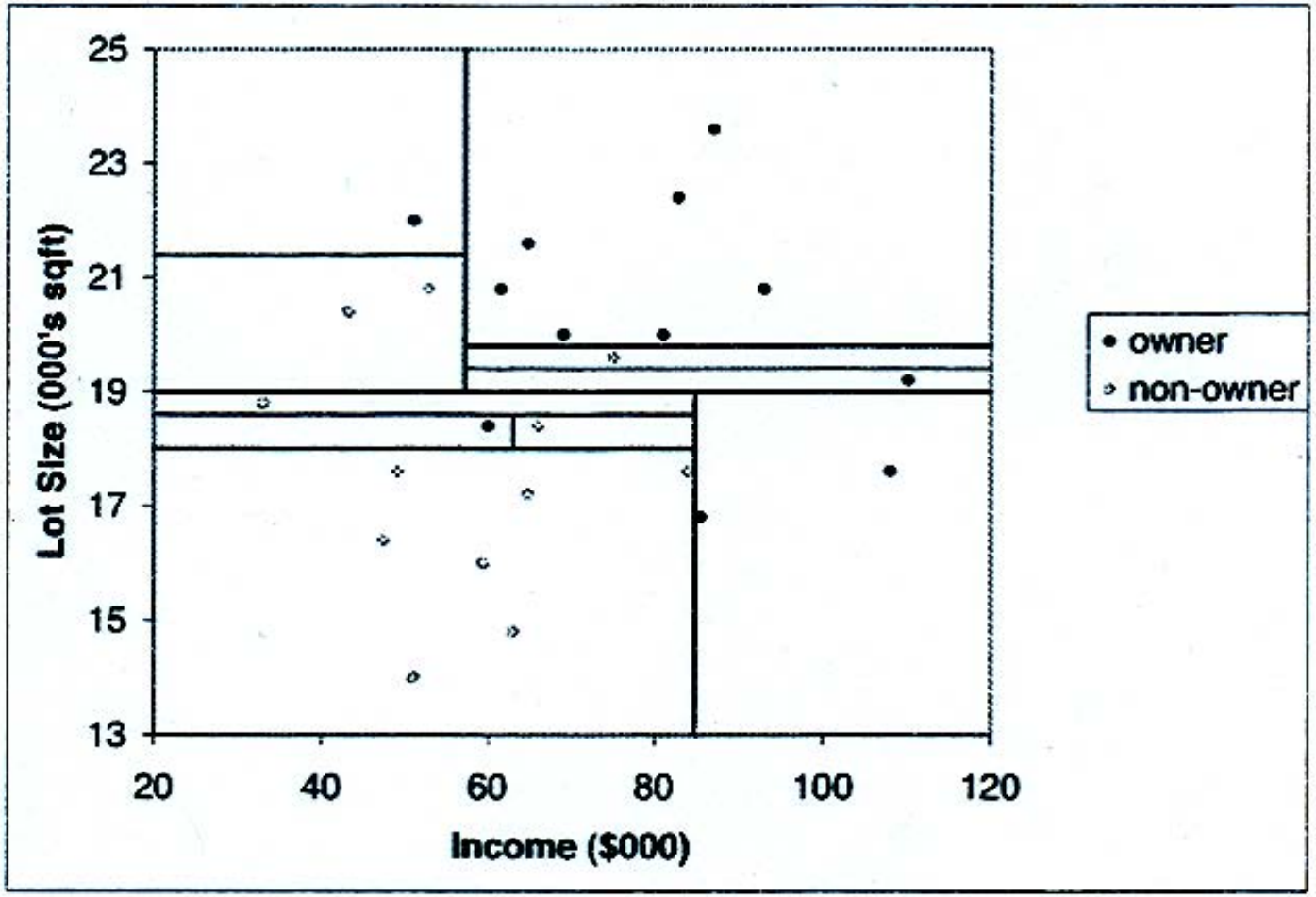


Figure F: Final Stage of Recursive Partitioning: Each Rectangle Consists of a Single Class (Owners or Non-Owners)

How to Address Overfitting…

- Post-pruning
  - Grow decision tree to its entirety
  - Trim the nodes of the decision tree in a bottom-up fashion
  - If generalization (validation set) error improves after trimming, replace sub-tree by a leaf node.
  - Class label of leaf node is determined from majority class of instances in the sub-tree

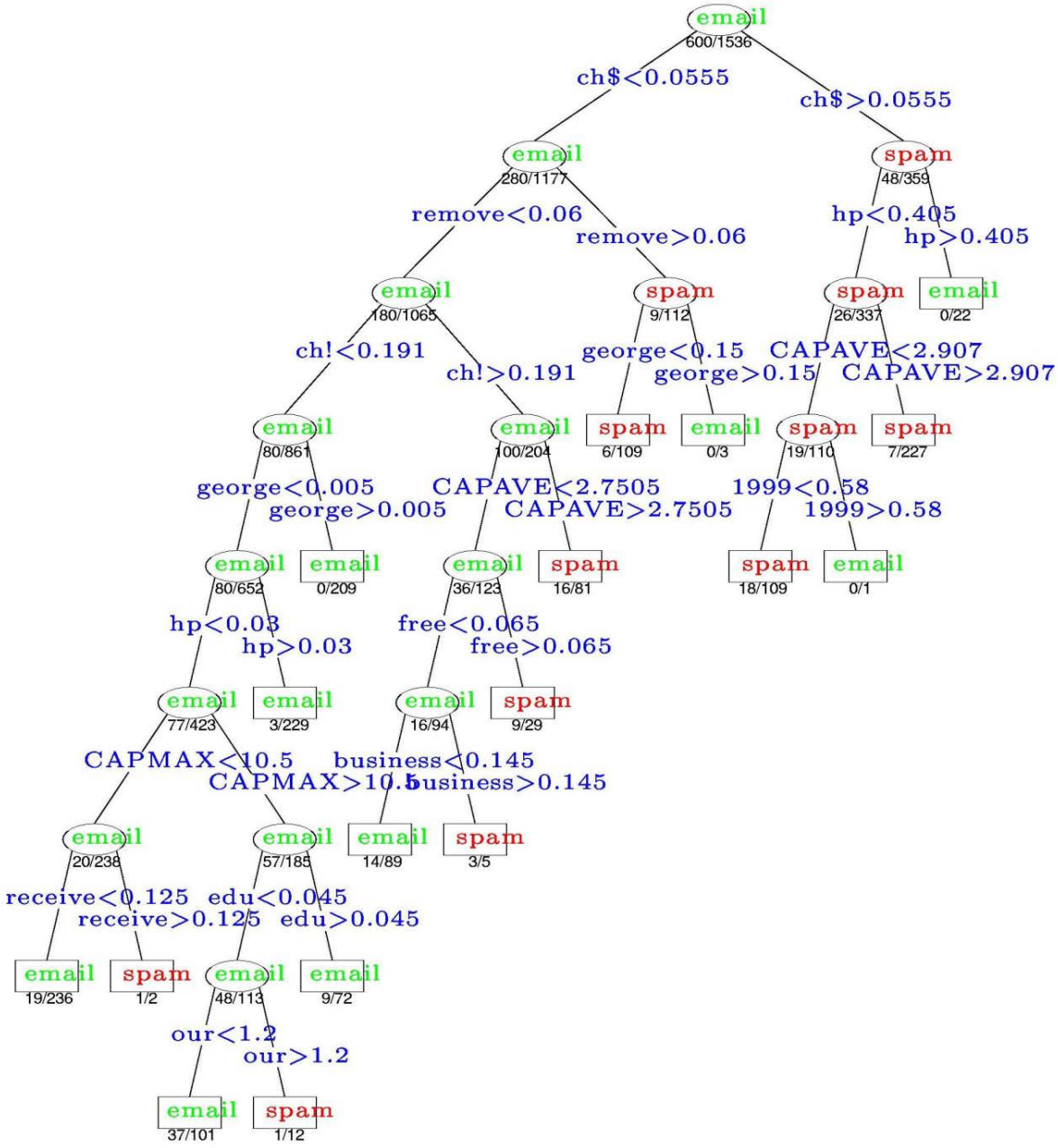
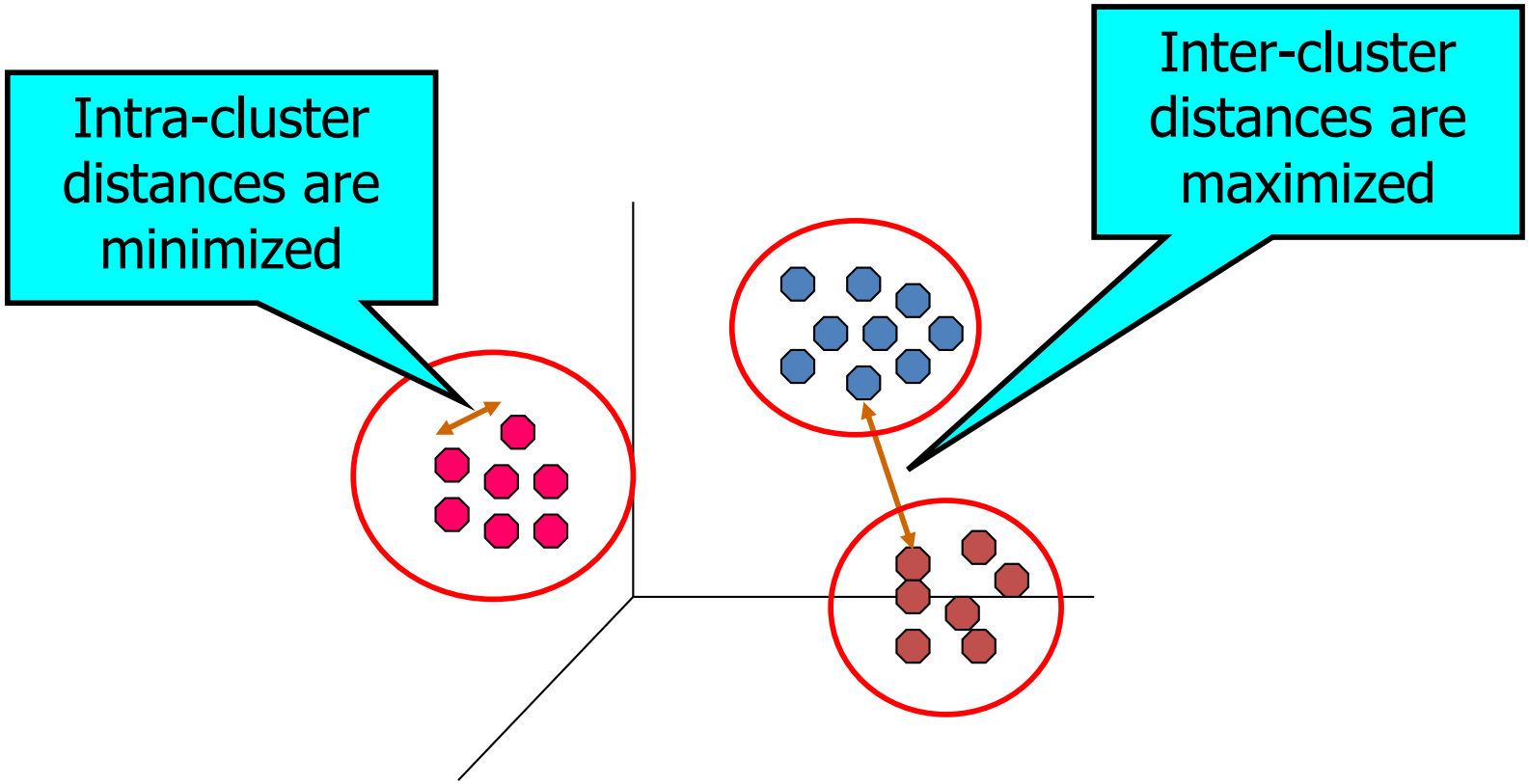


Figure B : *The pruned tree for the* spam *example. The split variables are shown in blue on the branches, and the classification is shown in every node. The numbers under the terminal nodes indicate misclassification rates on the test data.*

What is Cluster Analysis?

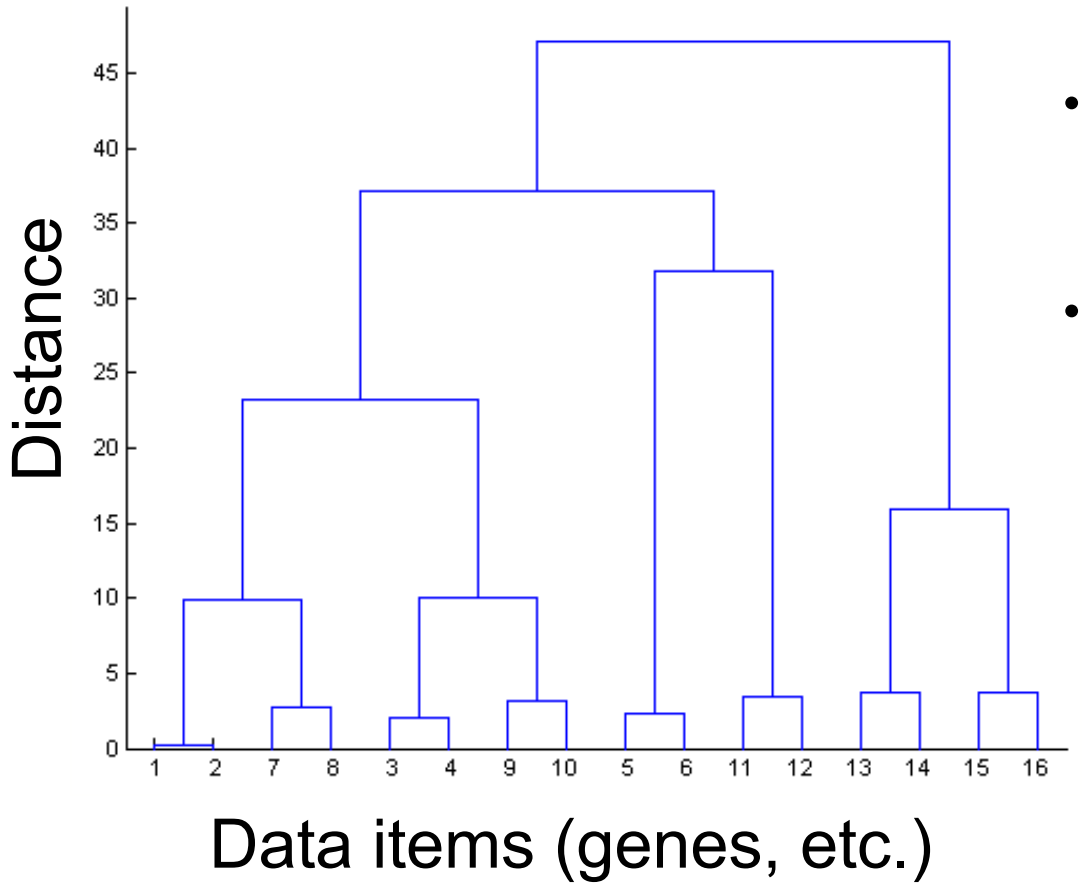
- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



Hierarchical Clustering

- Two main types of hierarchical clustering
  - Agglomerative:
    - Start with the points as individual clusters
    - At each step, merge the closest pair of clusters until only one cluster (or k clusters) left
  - Divisive:
    - Start with one, all-inclusive cluster
    - At each step, split a cluster until each cluster contains a point (or there are k clusters)
- Traditional hierarchical algorithms use a similarity or distance matrix
  - Merge or split one cluster at a time

Hierarchical Clustering



- This produces a binary tree or ***dendrogram***
- The final cluster is the root and each data item is a leaf
- The heights of the bars indicate how close the items are

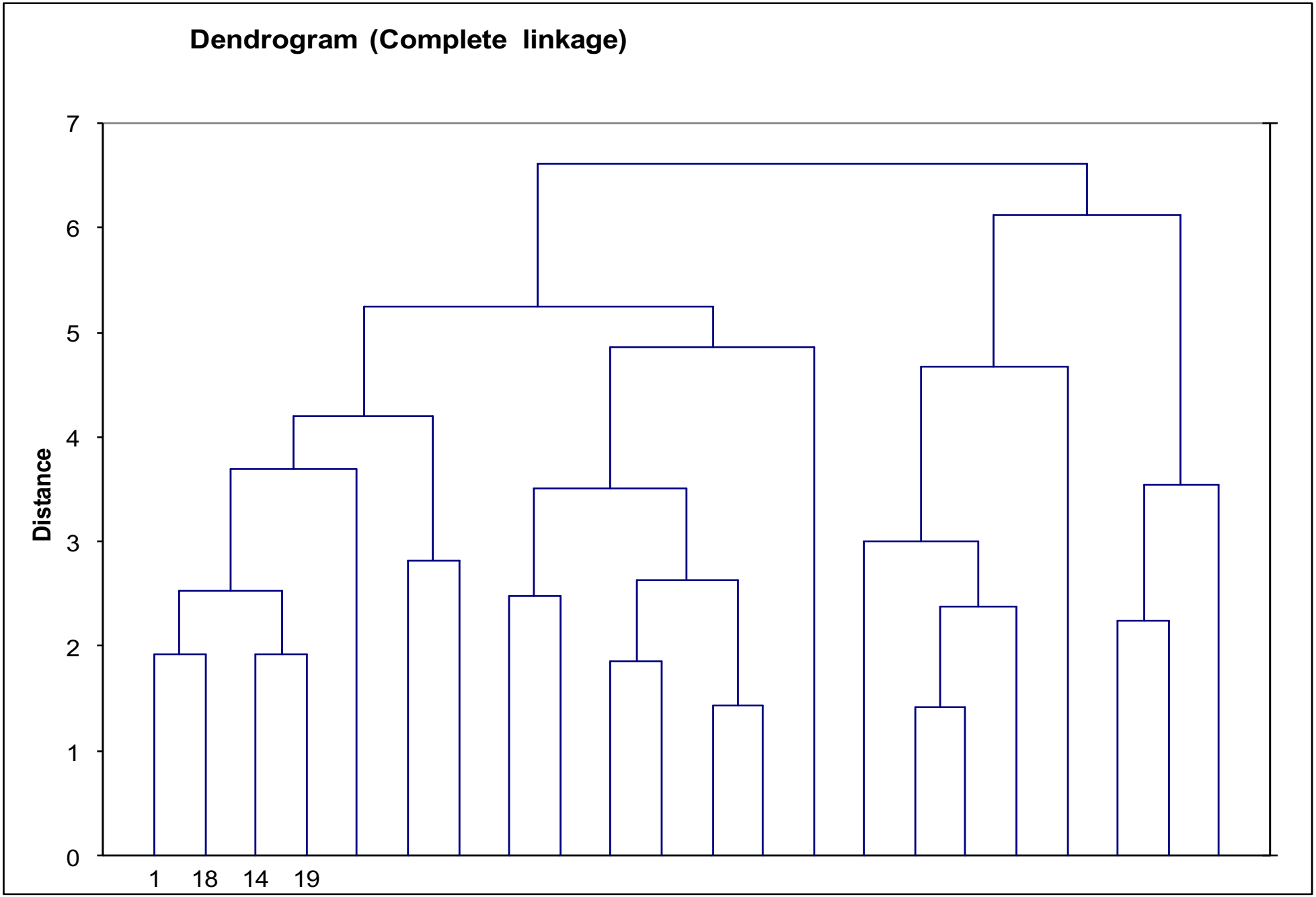


Figure C: Dendogram: Complete Linkage for All 22 Utilities, Using All 8 Measurements

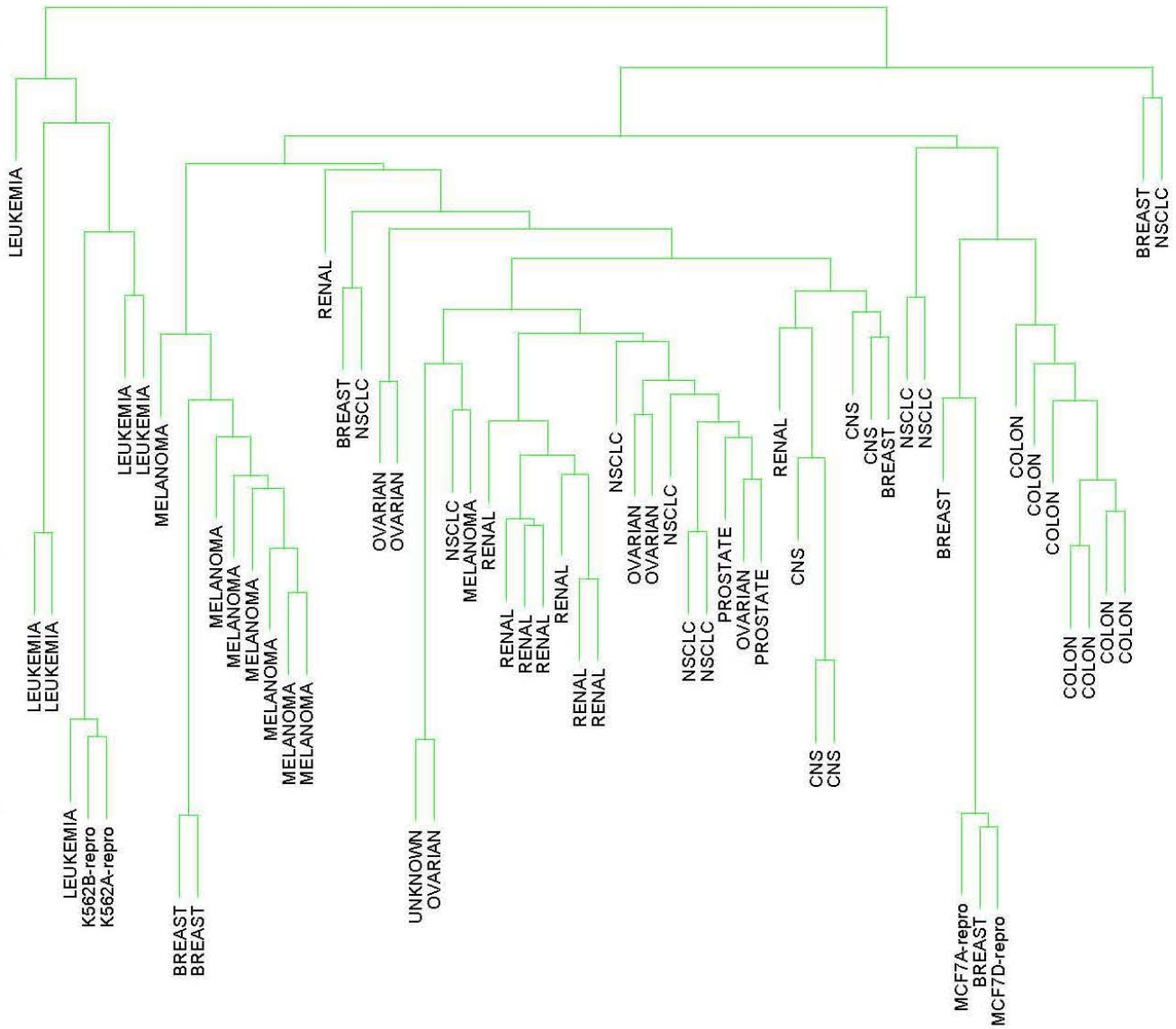
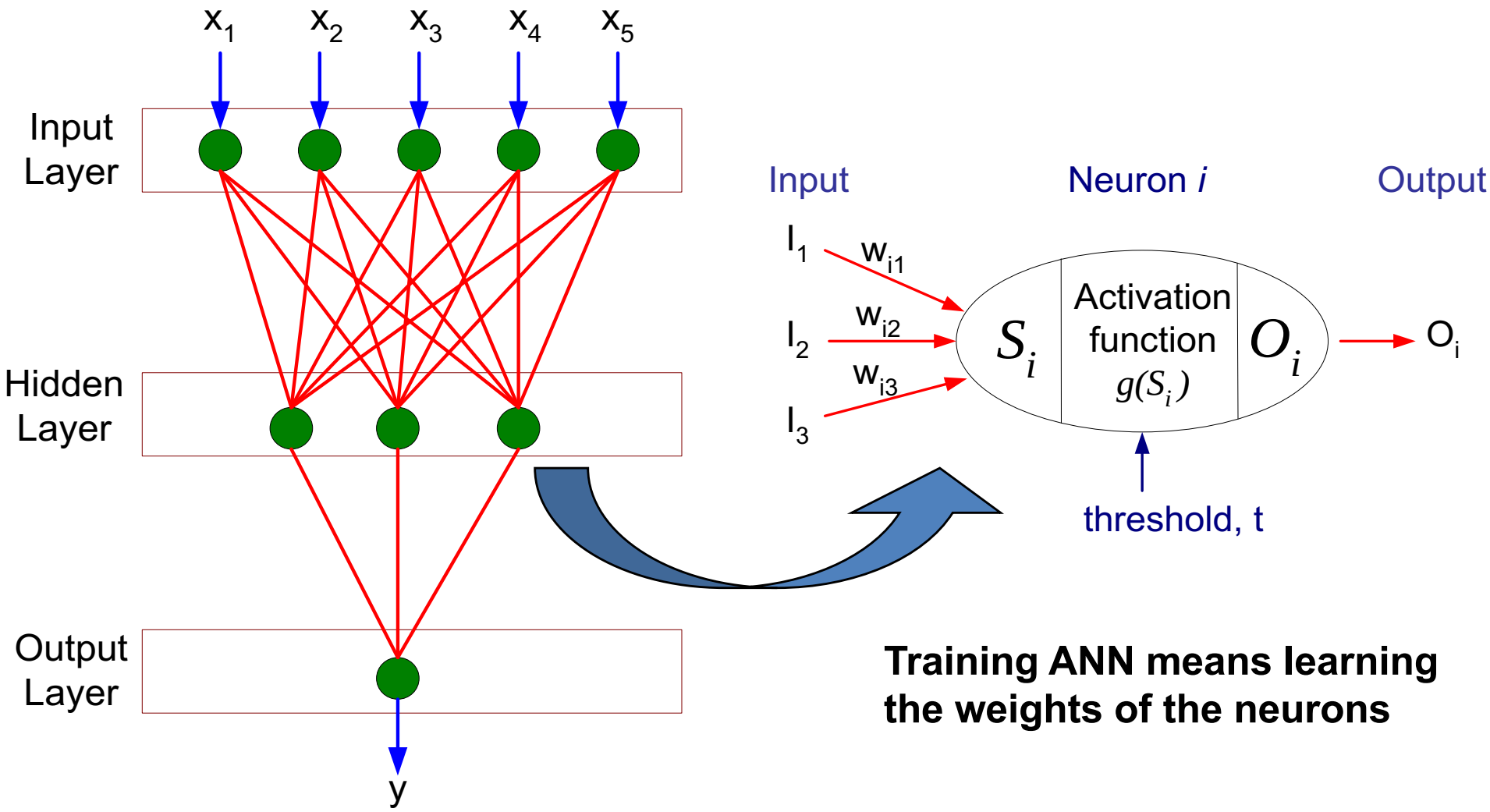


Figure D : *Dendrogram from agglomerative hierarchical clustering with average linkage to the human tumor microarray data.*

General Structure of a Neural Net



Training ANN means learning the weights of the neurons

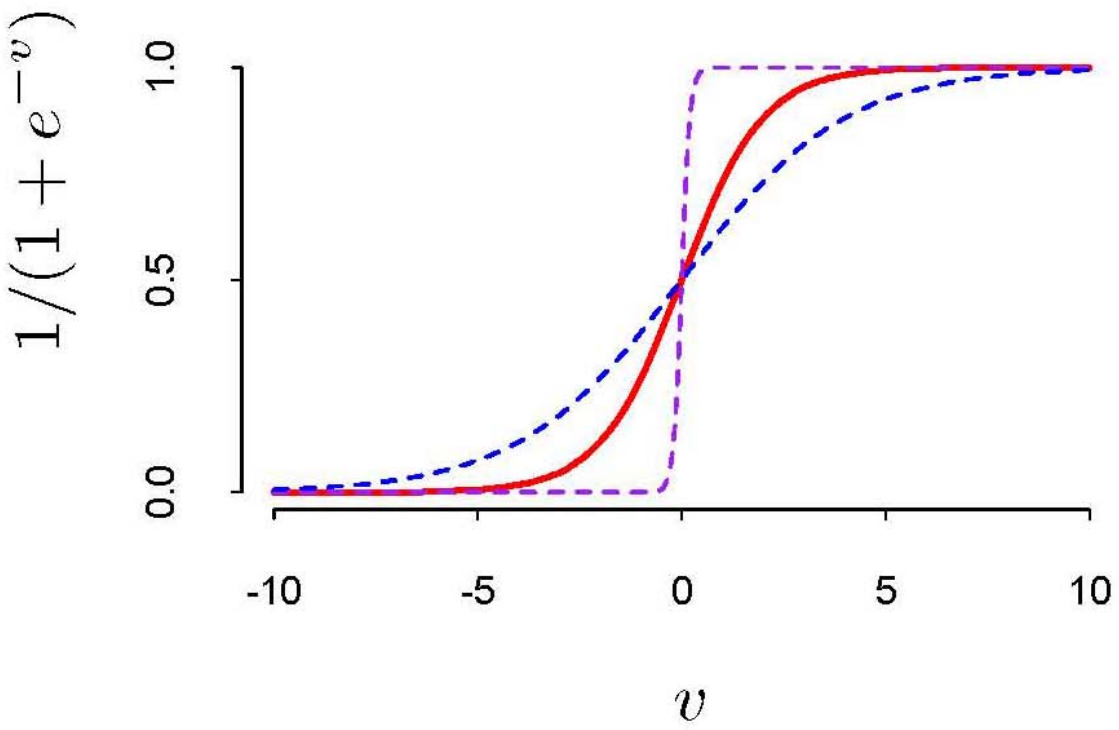
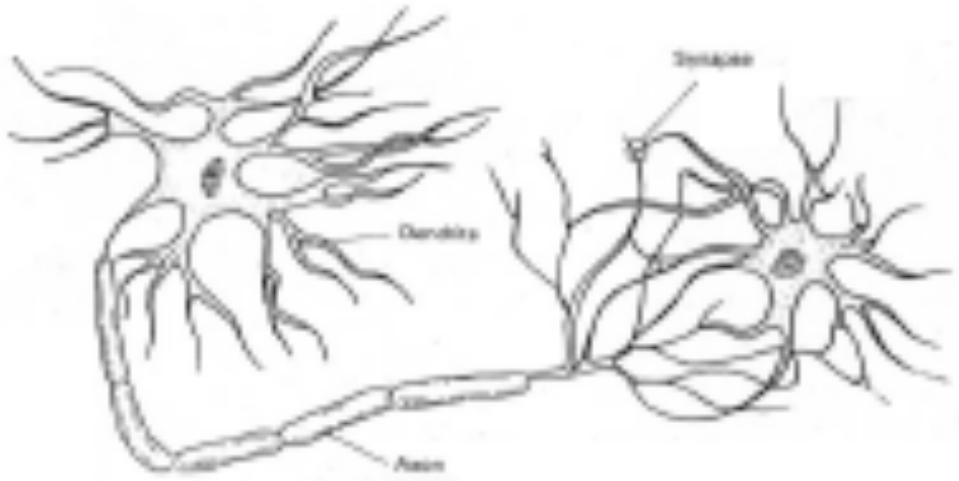
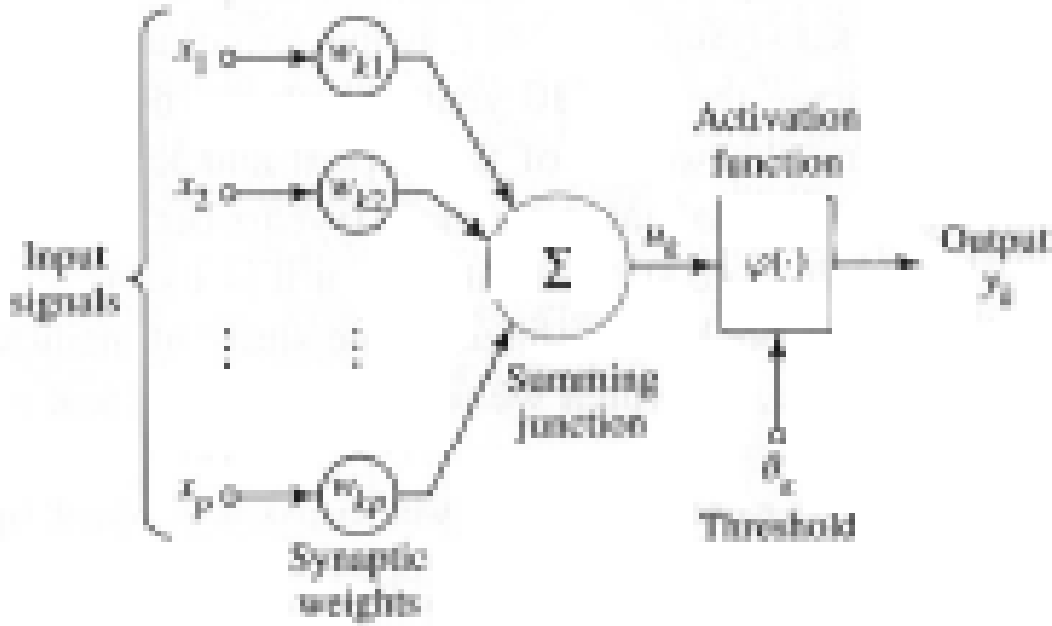


FIGURE 11.3. *Plot of the sigmoid function* $\sigma(v) = 1/(1+\exp(-v))$ *(red curve), commonly used in the hidden layer of a neural network. Included are* $\sigma(sv)$ *for* $s = \frac{1}{2}$ *(blue curve) and* $s = 10$ *(purple curve). The scale parameter* s *controls the activation rate, and we can see that large* s *amounts to a hard activation at* $v = 0$*. Note that* $\sigma(s(v - v_0))$ *shifts the activation threshold from* 0 *to* v_0*.*

Neurons


Synapses
Dendrites
Axon


Activation function
Input signals
Summing junction
Output
Threshold
Synaptic weights

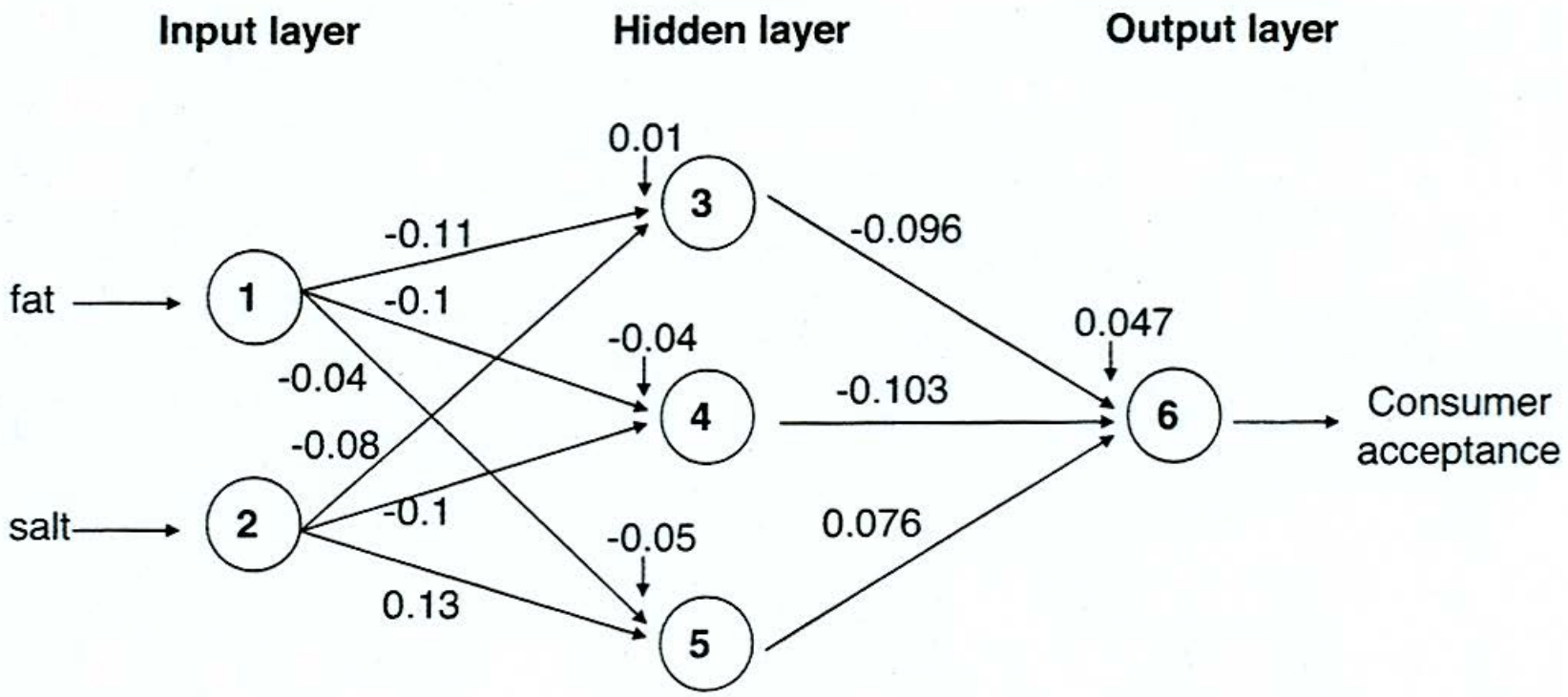
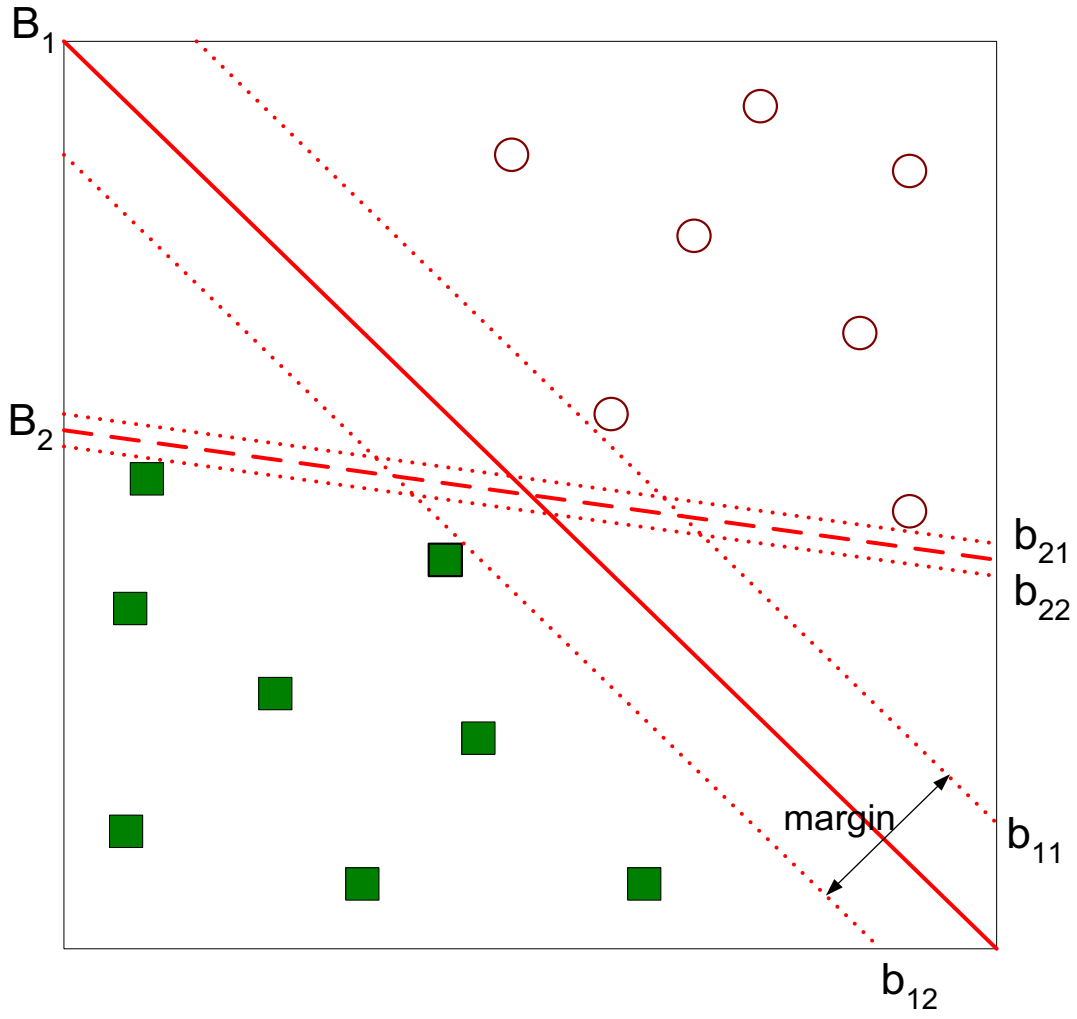


Figure 9.5: Diagram of a Neural Network for The Tiny Example with Weights from XLMiner Output (Figure 9.4)

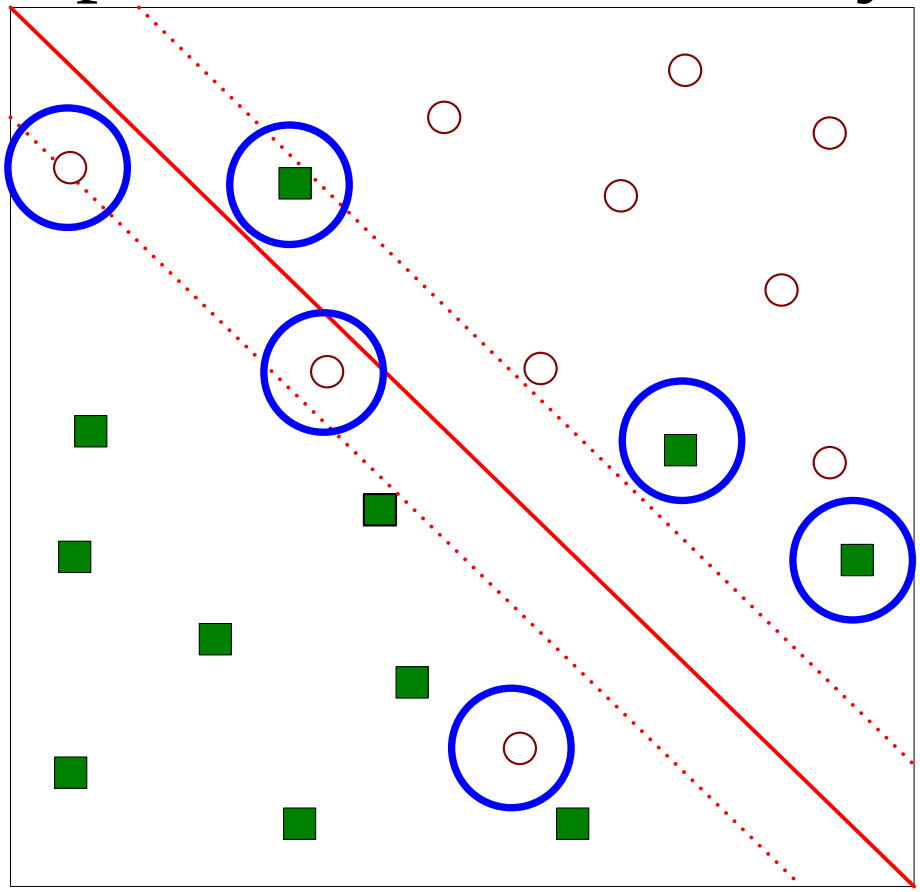
Support Vector Machines



- Find hyperplane maximizes the margin => B1 is better than B2

Support Vector Machines

- What if the problem is not linearly separable?



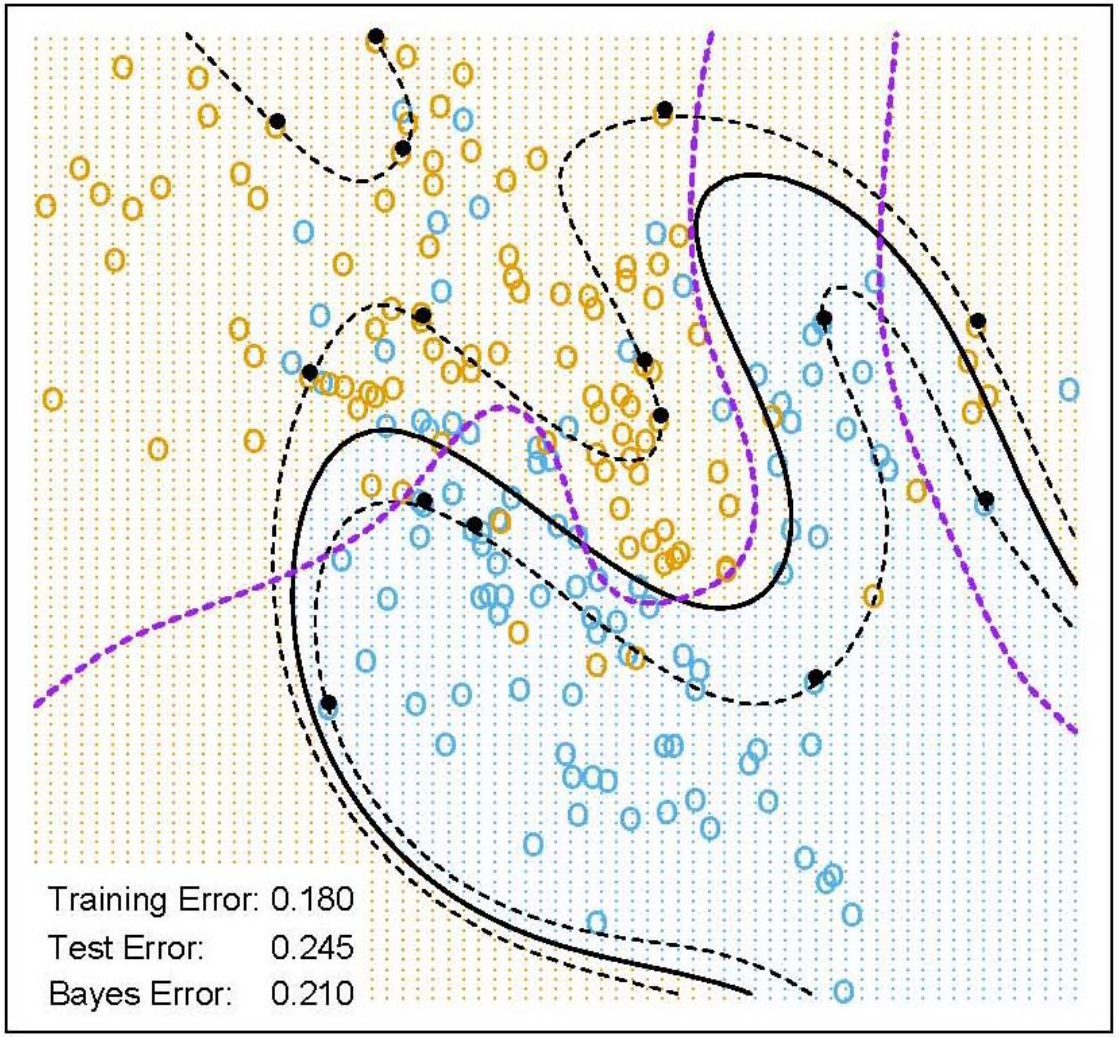


Figure 12.3. A nonlinear SVM for the mixture data. The plot uses a 4th degree polynomial kernel. C was tuned to approximately achieve the best test error performance, and C = 1 worked well. The broken purple curve in the background is the Bayes decision boundary.

Association Rule Mining (Filtering)

- Given a set of transactions, find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction. Netflix prize example.

Market-Basket transactions

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Example of Association Rules

{Diaper} $\rightarrow$ {Beer},
{Milk, Bread} $\rightarrow$ {Eggs,Coke},
{Beer, Bread} $\rightarrow$ {Milk},

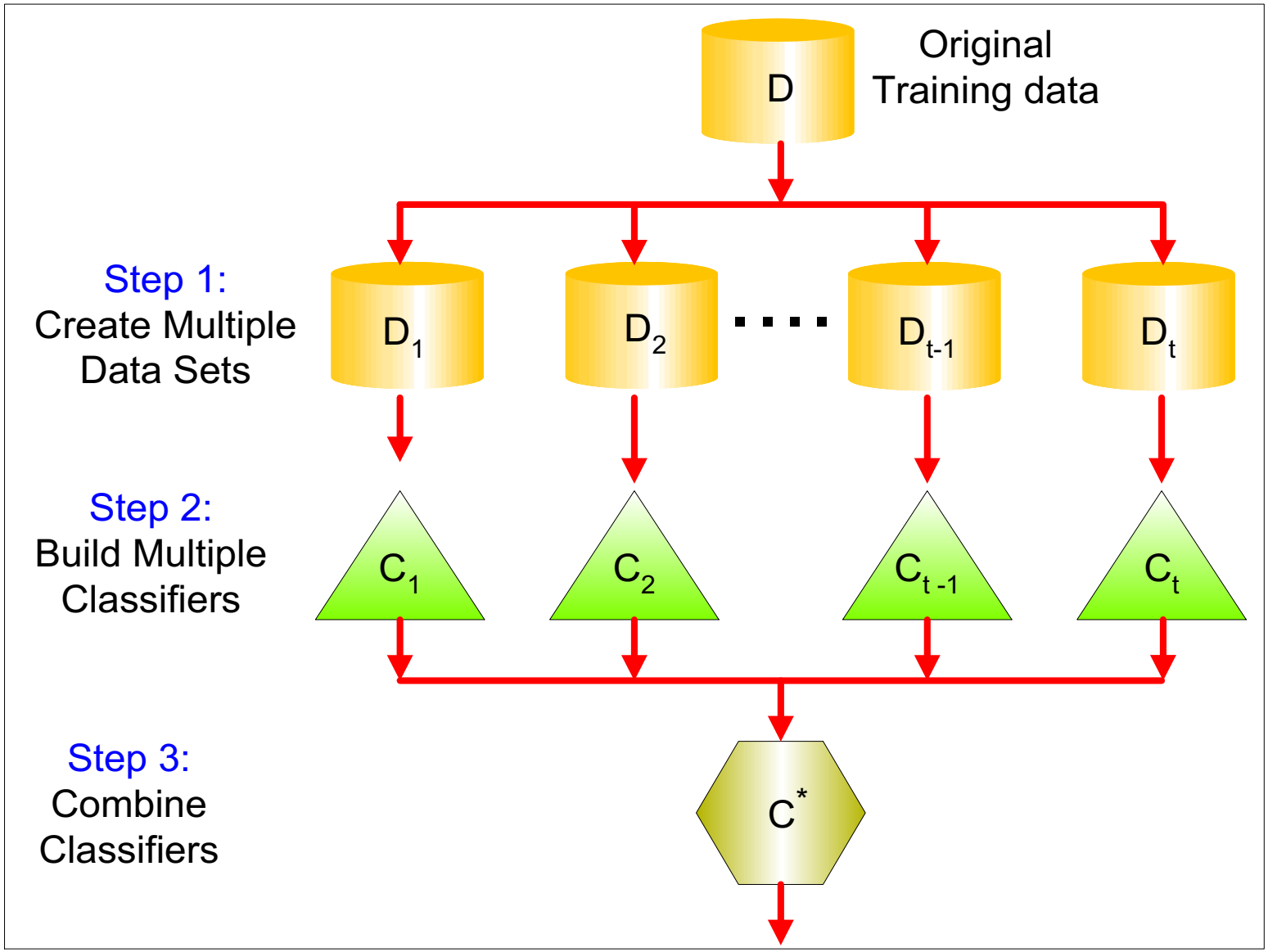
Implication means co-occurrence, not causality!

Ensemble Methods

- Construct a set of classifiers from the training data

- Predict class label of previously unseen records by aggregating predictions made by multiple classifiers

General Idea



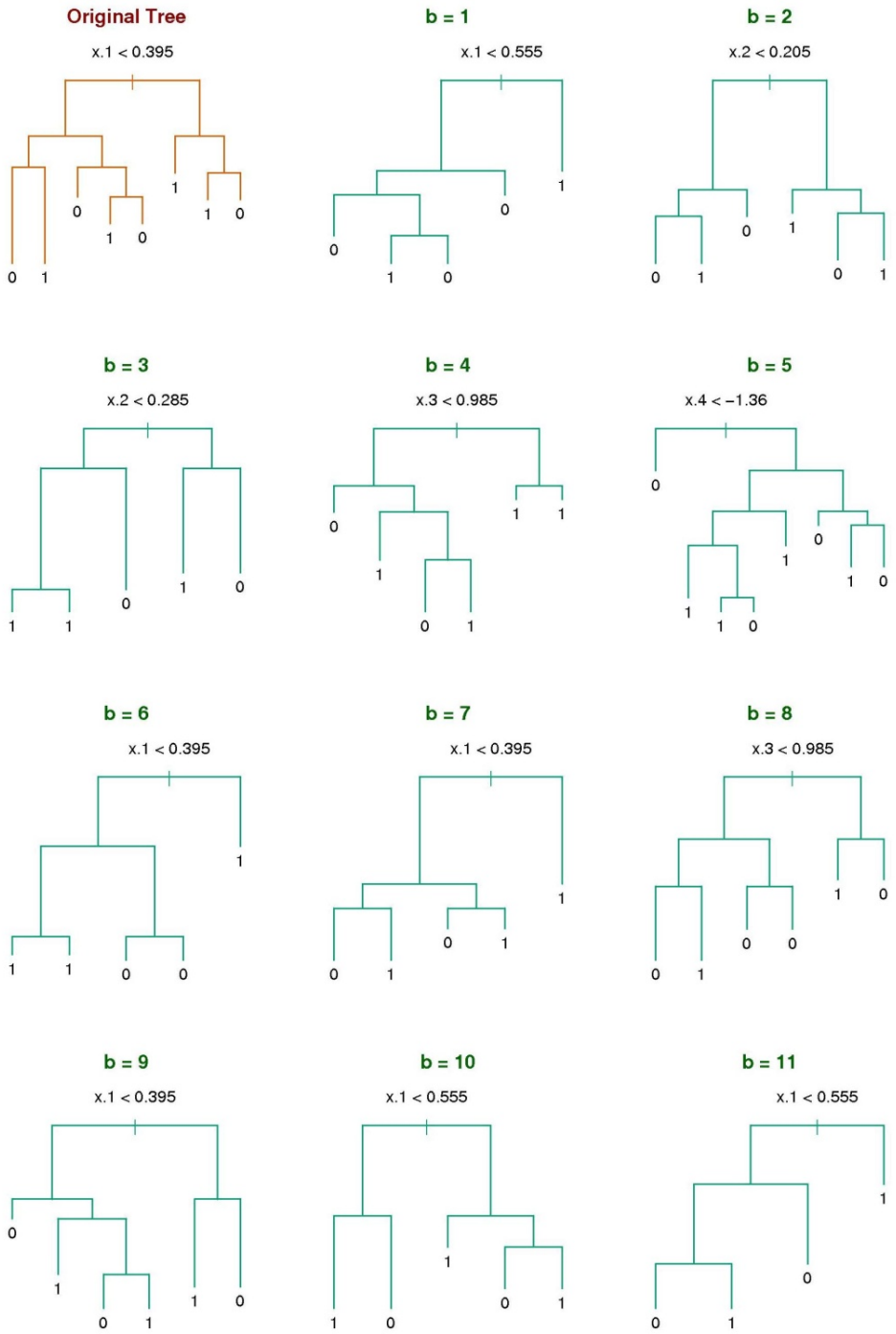


Figure 8.9: Bagging trees on simulated dataset. Top left panel shows original tree. Eleven trees grown on bootstrap samples are shown.

Why does it work?

- Suppose there are 25 base classifiers
  - Each classifier has error rate, $\varepsilon = 0.35$
  - Assume classifiers are independent
  - Probability that the ensemble classifier makes a wrong prediction:

$$\sum_{i=13}^{25} \binom{25}{i} \varepsilon^i (1-\varepsilon)^{25-i} = 0.06$$

Ridge Regression

Assume input data centered and scaled (standardized). A common form of shrinkage is

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

$$\text{or} \qquad = \arg\min_{\beta} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 \le C.$$

Clearly λ or C need to be chosen and that is often done by cross-validation.

The Lasso

The lasso estimate is defined by

$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

$$\text{subject to } \sum_{j=1}^{p} |\beta_j| \leq C.$$

Note that L_2 ridge penalty $\sum_1^p \beta_j^2$ is replaced by L_1 lasso penalty $\sum_1^p |\beta_j|$.

Generally forces small coefficients to be hard zeros. Fast algorithms available using Least Angle Regression (LARS).