Statistics Review I: Visualizing and Describing Data

Paul Vos

vosp@mail.ecu.edu

February 9, 2004

1. Preliminaries

1.1. Basic Terms

– **Individuals:** (cases, subjects, units) People, animals, or any object being studied.

– **Variable:** Some characteristic of the individuals (that varies among individuals)

– **Data:** The values of one or more variables for each of a group of individuals.

1.2. Types of Data

- **Examples**

  Consider a collection of n people: obtain data by "observing" **BP**, **marital status**, **sex**, **age**, **height**, **no. days absent**, **opinion**

- Classifying Data/Variables
  - Numerical (or, Quantitative)
    - discrete (counts)
    - continuous
  - Categorical (or, Qualitative)
    - nominal
    - ordinal

- **NOTE:** Numerical data can always be categorized; eg, income or age on consumer questionnaires.
- **CAUTION:** Considering data without regard to the variables being measured – both how they are defined and how they are measured – can lead to incorrect conclusions. Especially variables such as reading ability, psychological traits, and economic indicators

◦ **Example**: (Moore & McCabe, 2000) In 1989, 5426 drivers aged 65+ were involved in fatal accidents, while 2900 drivers aged 16–17 were involved in fatal accidents. Therefore, older drivers have more fatal accidents. *Discuss.*

fatality rate is more relevant

65+: 26 deaths per 100,000; 16–17: 70 deaths per 100,000

- **Example**: Data for States

State	Region	Pop. (1,000)	SAT Verbal	SAT Math	Percent Taking	Percent No HS	T.Pay ($1,000)
AL	ESC	4,273	565	558	8	33.1	31.3
AK	PAC	607	521	513	47	13.4	49.6
AZ	MTN	4,428	525	521	28	21.3	32.5
AR	WSC	2,510	566	550	6	33.7	29.3
CA	PAC	31,878	495	511	45	23.8	43.1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

  - **State** list of individuals; not a variable
  - **Region** categorical
  - **Population**,. . .,**Teacher's Pay** numerical

- **Example**: Survey Results

Person	Age	Gender	Vote	Attitude
1	20	0	0	2
2	27	0	0	1
3	19	1	1	1
4	38	1	0	3
5	38	1	1	3
⋮	⋮	⋮	⋮	⋮

  - **Person** label
  - **Age** numerical
  - **Gender** (0=F, 1=M) categorical
  - **Vote** (0=Dem, 1=Rep, 2=Other) categorical
  - **Attitude** (1=oppose, 2=neutral, 3=favor) categorical

1.3. Overview/Preview

- One Variable
  - Visualizing data
    - Numeric
    - Categorical
  - Summarizing data
    - Numeric (categorical)
  - Modeling data
    - Numeric – Normal
    - (Categorical – Binomial)

2. Visualize 1 Variable

2.1. Visualizing Numeric Data

- Stem plot
- Frequency and relative frequency tables
- Histogram (3 kinds)
  - Frequency - height gives count
  - Rel. Freq - height gives proportion
  - Density - *area* gives proportion
- Two things to look for:
  - Overall shape: skewed, symmetric, irregular
  - Departures from overall shape – outliers

- **Example**: Cavendish's (1798) density-of-earth:

5.50	5.61	4.88	5.07	5.26	5.55	5.36	5.29	5.58	5.65	5.57
5.53	5.62	5.29	5.44	5.34	5.79	5.10	5.27	5.39	5.42	5.47
5.63	5.34	5.46	5.30	5.75	5.68	5.85				

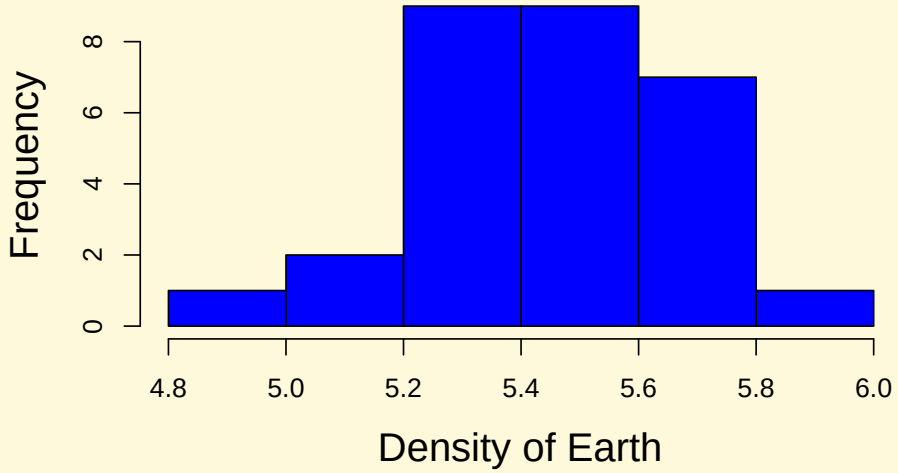
  – Stem (-and-leaf) plot

48	8
49	
50	7
51	0
52	6799
53	04469
54	2467
55	03578
56	12358
57	59
58	5

- **Example**: Cavendish (cont)
  - Frequency and relative frequency tables

Cell	Bndry	Freq.	Rel. Freq.	
4.8	5.0	1	.034	$= 1/29$
5.0	5.2	2	.069	$= 2/29$
5.2	5.4	9	.310	
5.4	5.6	9	.310	
5.6	5.8	7	.241	
5.8	6.0	1	.034	
		29	.998	

– Histogram



- **Example**: Head circumference at birth (cm) for male humans

33.1 34.6 34.2 36.1 34.2 35.6 34.5 35.8 34.5
34.2 34.3 35.2 36.0 34.2 34.7 34.6 34.3 33.7
33.4 34.9 33.8 33.6 35.2 34.6 33.7 34.8 33.9
34.7 35.1 34.2 36.5 34.1 34.0 35.1 35.3

Stem plot

33	1467789
34	012222233556667789
35	1122368
36	015

Split-Stem plot

33	14
33	67789
34	012222233
34	556667789
35	11223
35	68
36	01
36	5

- **Example**: Cardiac output

2.60	5.16	6.18	3.22	4.99	3.62	3.31	4.11
5.24	4.27	3.42	4.70	5.42	5.36	2.63	3.70
5.39	5.44	3.86	6.68	5.35	3.26	4.06	2.64
5.40	5.93	5.90	4.11	4.44			

  - Rounded Data:

2.6	5.2	6.2	3.2	5.0	3.6	3.3	4.1	5.2	4.3
3.4	4.7	5.4	5.4	2.6	3.7	5.4	5.4	3.9	6.7
5.4	3.3	4.1	2.6	5.4	5.9	5.9	4.1	4.4	

  - Stem plot of rounded data:

2	666
3	2334679
4	111347
5	02244444499
6	27

- **Example**: Honolulu Heart Study (Systolic BP)
  (Kuzma and Bohnenblust, 2001, pp 25-27)

  – Back-to-back stem plot:

Nonsmokers		Smokers
8642	9	8
888640	10	2248
888884422	11	2244666668
888888884422220	12	02266
844444422000	13	0468
644200	14	000026
6444422	15	00
22	16	2
20	17	68
	18	
	19	0
	20	8

- Histograms of six data sets (Figure 1)

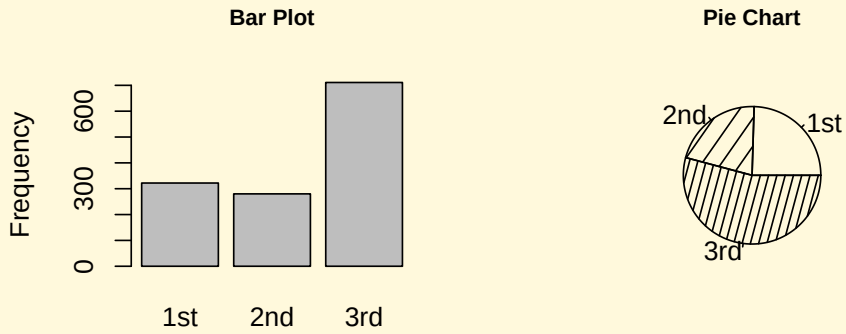
2.2. Visualizing Categorical Data

- Frequency and Relative Frequency tables
- Pie chart
- Bar plot

- **Example**: Titanic Data (Hinde, P., 1998. *Encyclopedia Titanica*, OzDASL)
  - Frequency and Relative Frequency table

Class	Frequency	Rel. Frequency	
1st	322	.245	$= 322/1313$
2nd	280	.213	$= 280/1313$
3rd	711	.542	
	1313	1.000	

- **Example**: Titanic (cont)



  - *Height* of bar gives Frequency (could give Rel. Freq.)
  - *Area* of slice gives Relative Frequency

- **Example**: Titanic (cont)
  - Tables for Variable **Class*Survival**

Class*Survival	Frequency	Rel. Frequency
1st&Died	129	.098
1st&Alive	193	.147
2nd&Died	161	.123
2nd&Alive	119	.091
3rd&Died	573	.436
3rd&Alive	138	.105
	1313	1.000

  - NOTE: There is a better way to compare **Class** and **Survival**

- **Example**: Titanic (cont)
  - 4 graphs to visualize this table
    - **(a) and previous graph**
      Height of stacked bars in (a) same as before.
    - **(a) and (b)**
      (b) shows Rel. Freq., (a) shows Freq.
    - **(b) and (c)**
      Stacked bars put along side.
    - **(c) and (d)**
      Re-group the bars.
  - Bar graphs for this table are in Figure 2.

3. Summarize 1 Variable

3.1. Measuring Center

– Notation:

data	$9, \quad 4, \quad 5$
generic data	$x_1, \quad x_2, \quad x_3$

In general, $x_1, x_2, \ldots, x_n$

– **Mean** $\bar{x}$ (average)

$\frac{9+4+5}{3} = 6$ or $\frac{x_1+x_2+\cdots+x_n}{n}$ or $\frac{\sum x_i}{n}$.

Notation: $\bar{x} = \frac{\sum x_i}{n} = \frac{1}{n}\sum x_i$.

– **Median** ('middle of ordered values')

n odd: $9, 4, 5 \longrightarrow 4, 5, 9 \longrightarrow$ median $= 5$

n even: $9, 4, 5, 9 \longrightarrow 4, 5, 9, 9 \longrightarrow$ median $= \frac{5+9}{2} = 7$

- (Mode) (most frequent observation(s))
  - **Not** a measure of center.
  - $9, 4, 5, 9$ has mode 9.
- **Mean** vs. **Median**
  - If the data are roughly symmetric and there are no outliers, mean and median are roughly the same. Mean is usually used.
  - For skewed data, median is often used.
  - Median is resistant to outliers; Mean is not.

    $9, 4, 5$ has mean 6
    $99, 4, 5$ has mean 36
    median is 5 in either case.

- **Examples**
  - Cavendish data:
    $\bar{x} = \frac{4.88+5.07+\cdots+5.85}{29} = 5.45$
    median is 5.46 (5.46 is 15th observation since $n = 29$,
    modes are at 5.34 and 5.29 (each occurs twice).
  - Passenger Class (Titanic data):
    mean and median? Meaningless.
    Mode is "3rd class" (**Not** 711 or .542).

3.2. Measures of Spread (Variability)

– **Variance** s^2 ("average" squared deviation)

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
9	$(9-6) = 3$	9
4	$(4-6) = -2$	4
5	$(5-6) = -1$	1
18	0	14

* mean $\bar{x} = \frac{18}{3} = 6$; variance $= \frac{14}{2}$.
* In general, variance $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$.
* Units: square of units of x

– **Standard Deviation** s (square root of variance)

* $s = \sqrt{s^2}$; $s = \sqrt{7} = 2.65$
* Units: same as x

- **Range**

  Maximum − Minimum

- **IQR** (Inter Quartile Range)

  * Q_1 is 1st Quartile (25th percentile)
  * Q_3 is 3rd Quartile (75th percentile)
  * IQR= $Q_3 - Q_1$
  * Units same as x

- Interpretation

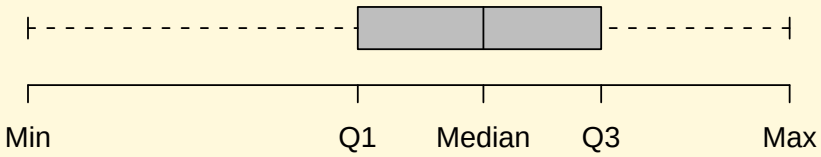
  * **IQR**: Length of the interval needed to contain the middle 50% of the data; resistant to outliers.
  * **standard deviation**: Difficult to interpret for non-Normal data; sensitive to outliers.

- **Examples**
  - Cavendish data
    - $s^2 = .0488, s = \sqrt{.0488} = .22$
    - $\text{IQR} = Q_3 - Q_1 = 5.61 - 5.30 = .31$
  - Systolic BP for Non-smokers
    - $s^2 = 344.0, s = \sqrt{334.0} = 18.55$ mmHg
    - $\text{IQR} = Q_3 - Q_1 = 140 - 118 = 22$ mmHg
  - Systolic BP for Smokers
    - $s^2 = 639.1, s = \sqrt{639.1} = 25.28$ mmHg
    - $\text{IQR} = Q_3 - Q_1 = 140 - 116 = 24$ mmHg
  - Notice spread is greater for smokers, especially when measured by s (see slide 1.).

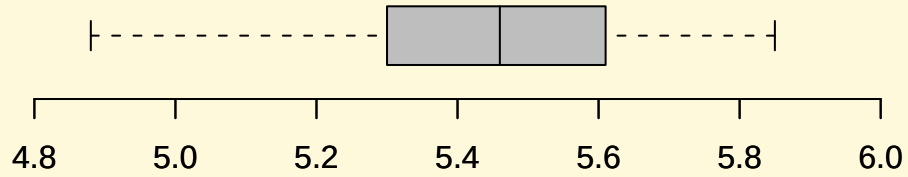
- Visualizing/Summarizing Data – Box Plot
  - **Five Number Summary**

Generic		Cavendish	
Median		5.46	
Q_1	Q_3	5.30	5.61
Min	Max	4.88	5.85

  - **Box-and-Whiskers Plot** (unmodified)



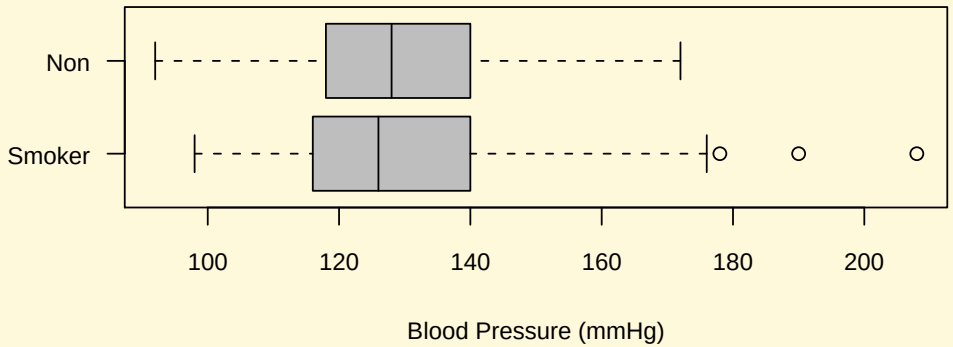
- **Example**: Cavendish Data



- Modified **Box Plot** (default)
  - Whisker at most $1.5 \times$ IQR
  - One definition of Outlier:
    Points more than $1.5 \times$ IQR below Q_1 or above Q_3
  - Outliers drawn outside whiskers

- **Example**: Systolic BPs

Non Smokers		Smokers	
128.0		126.0	
118.0	140.0	116.0	140.0
92.0	172.0	98.0	208.0



- What to look for in a Box plot:
  - Center (median)
  - Spread (IQR)
  - Outliers
  - Shape (see Figure 3)

- Sample Percentiles
  - Generalization of Quartiles (Q_1 is 25th percentile).
  - 90th percentile is value such that 90% of the data are smaller.
  - Knowing all percentiles $\longleftrightarrow$ data

◦ **Example**: Systolic BPs

- 90th Percentile for Non Smokers

  $x_{.90} = 154$ `mmHG` (63 nonsmokers; $.90 \times 63 = 57$; 57th ordered observation from stemplot on slide 1.)
- 90th Percentile for Smokers

  $x_{.90} = 162$ `mmHG` (37 smokers; $.90 \times 37 = 33$; 33rd ordered observation from stemplot on slide 1.)

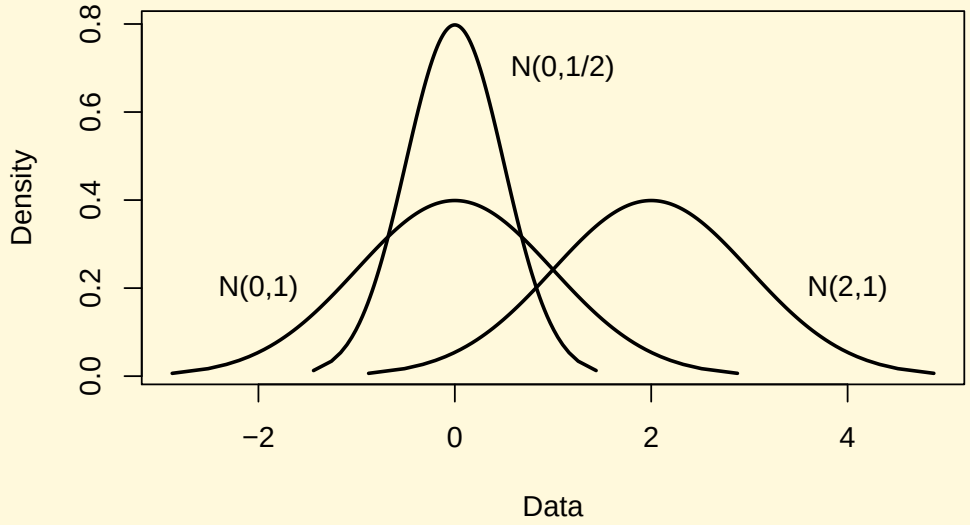
Review/Preview

- Data Classification
- Visualizing Data
- Numerical Summaries of Data
- **Modeling Data**
  - **Numeric Data**
    - **Normal Model**
    - Other Models
  - Categorical Data
    - Binomial Model
    - Other Models

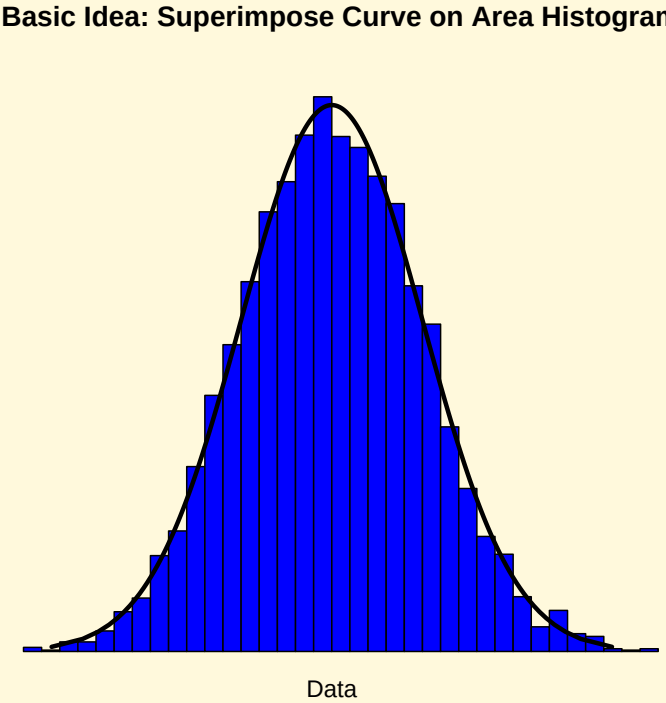
4. Model 1 Variable

- Normal Model (Distn)
  - Area (Density) Histograms
    - Histograms where Area of bar = Rel. Freq.
    - Total Area of all bars = 1.
  - Family of Bell-shaped (Normal) Curves
    - Area under each curve = 1.
    - Each curve is symmetric
    - center denoted μ (called *mean*)
    - spread denoted σ (called *standard deviation*)
    - Each curve completely specified by μ and σ
    - Notation: $N(\mu, \sigma)$

- 3 Curves from the Normal Family



- Basic Idea – Superimpose Normal Curve on Area Histogram



- 68–95–99.7 Rule

  If the "normal model holds":

  – 68% of the data fall within 1 SD of $\bar{x}$.

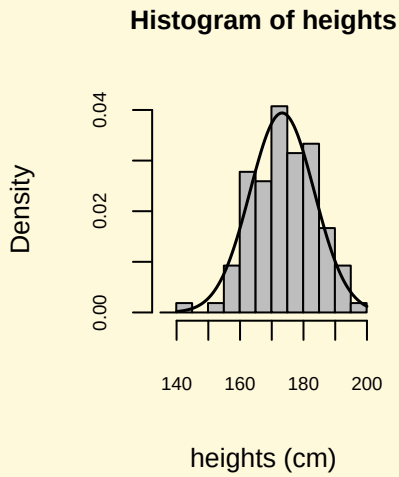
  – 95% of the data fall within 2 SD of $\bar{x}$.

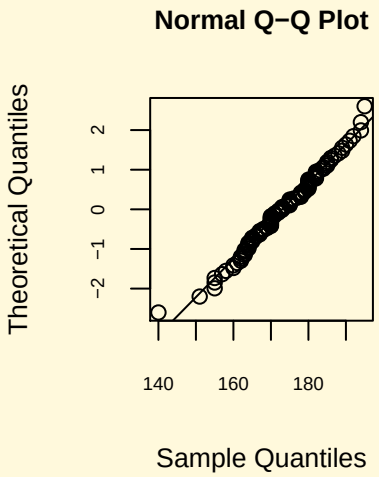
  – 99.7% of the data fall within 3 SD of $\bar{x}$.

◦ **Example**: Heights

  – Heights from Study on Pulse Rates and Exercise (Dr. Richard J. Wilson, Department of Mathematics, University of Queensland, OzDASL)

- **Example**: Heights (cont)





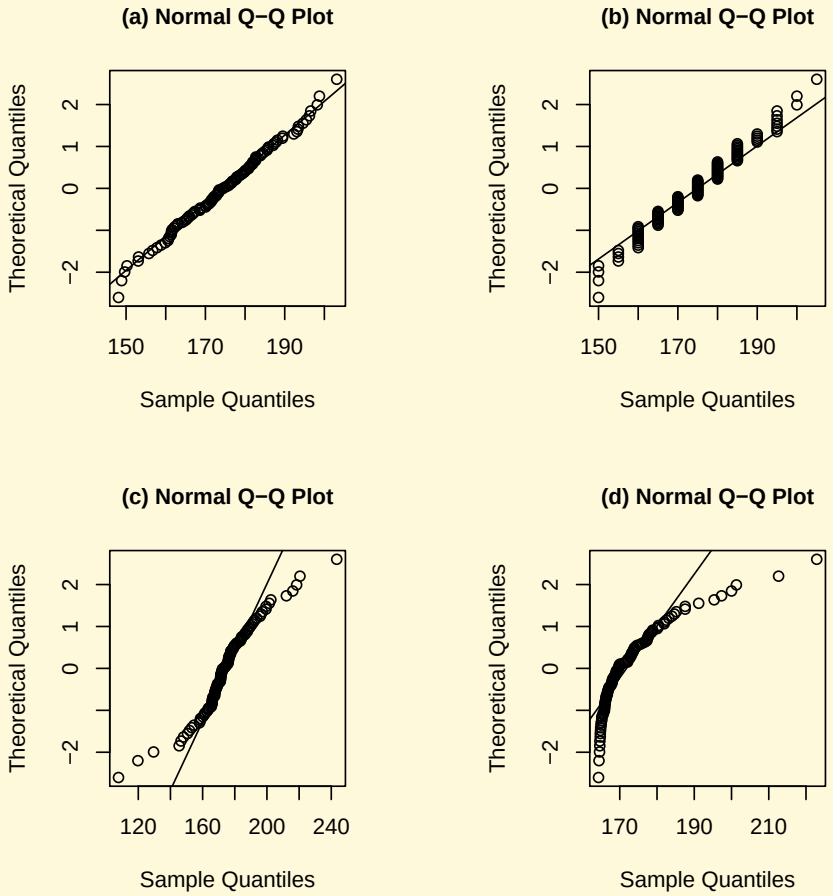
- Checking the 68–95–99.7 Rule for these data
  - Mean of Heights = 173.3; SD of Heights = 10.1; n=108
  - Mean $\pm$ SD = $173.3 \pm 10.1 = (163.2, 183.4)$
    - Data: 74 fall in this range (see next slide)
    - Rule: 68% of 108 is 73.44
  - Mean $\pm$ 2 SD = $173.3 \pm 20.2 = (153.1, 193.5)$
    - Data: 103 fall in this range
    - Rule: 95% of 108 is 102.60
  - Mean $\pm$ 3 SD = $173.3 \pm 30.3 = (143.0, 203.6)$
    - Data: 107 fall in this range
    - Rule: 99.7% of 108 is 107.68

- Height Data(in cm; sorted, two extreme outliers removed)

```
[  1] 140 151 155 155 155 157 158 160 160 161
[ 11] 162 162 162 163 163 163 163 164 164 164
[ 21] 164 164 165 165 165 165 166 166 167 167
[ 31] 167 167 168 168 169 169 170 170 170 170
[ 41] 170 170 170 170 170 170 170 171 171 171
[ 51] 172 172 172 173 173 173 173 174 175 175
[ 61] 175 175 175 175 175 176 177 178 178 178
[ 71] 178 178 179 179 179 180 180 180 180 180
[ 81] 180 180 180 180 182 182 182 182 182 182
[ 91] 183 184 184 185 185 185 186 186 187 188

[101] 189 189 190 191 192 194 194 195
```

- Normal Quantile-Quantile Plots



- Discussion of Normal Q-Q Plots on slide 1.

  x-axis Ordered values as predicted by the Normal model

  y-axis Ordered values actually observed

  (a) 108 observations generated from the Normal model

  (b) Same data in (a) except rounded to nearest 5cm

  (c) Not normal; heavy tails (t-distn w/ 2df)

  (d) Not normal; skewed to the right (χ^2 w/ 2df)

- Refining the 68–95–99.7 Rule
  - Notation:
    - $N(\mu, \sigma)$ is the curve from the **Normal** family that has **center** μ and **spread** σ.
    - Uppercase letters (eg, X, Y) will stand for collections of values.
    - X is the sample; ie, $X = x_1, x_2, \ldots, x_n$.
    - $X \sim N(110, 20)$ means
      AREA histogram for the data X follows a bell-shaped curve with center (mean) at 110 and spread (standard deviation) 20
    - $\Pr(X \leq 110)$ is shorthand for
      Proportion of the sample X that is less than 110

- Refining the 68–95–99.7 Rule (cont)
  - The area under the curve $N(\mu, \sigma)$ between two points a and b (or, the proportion of the data X between a and b) depends only on how many standard deviations a and b are from μ.
  - z is the number of standard deviations x is from its mean.

Let $X \sim N(110, 20)$

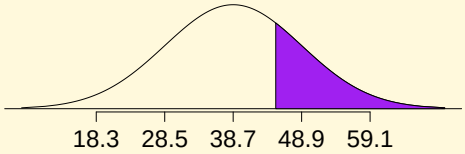
x	z
150	2
130	1
110	0
90	-1

  - In general, $z = \frac{x-\mu}{\sigma}$
  - This area can be obtained from a Table for the Standard Normal distribution.

- Normal Model Calculations:

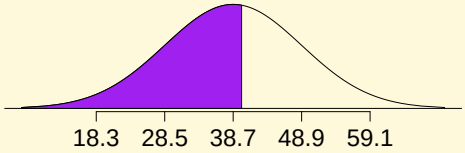
  The amount of time necessary for people to take a certain test has a normal distribution with mean 38.7 minutes and standard deviation 10.2 minutes.

  – What proportion of people need more than 45 minutes to finish this test?



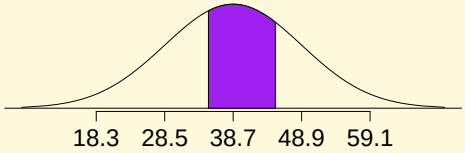
$z = \frac{45-38.7}{10.2} = 0.62;$
Answer = Area = .2676.

  – What proportion of people take less than 40 minutes to finish this test?



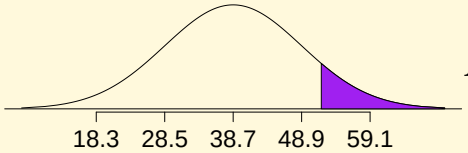
$z = \frac{40-38.7}{10.2} = 0.13;$
Answer = Area = .5517.

- Normal Model Calculations (cont)
  - What proportion of people take between 35 and 45 minutes to finish this test?



$$z = \frac{45-38.7}{10.2} = 0.62;$$
$$z = \frac{35-38.7}{10.2} = -0.36;$$
$$\text{Answer} = \text{Area} = .3730.$$

  - The slowest 10% take at least how long to finish the exam?



$$\text{Area} = .10; z_{.90} = 1.28$$

$$\text{Answer} = x_{.90} = 38.7 + 1.28 \times 10.2 = 51.8 \text{ min.}$$

Review/Preview

One Variable	Two Variables
Data Classification	
• Visualizing Data	• Visualizing Relationships
• Numerical Summaries of Data	• Numerical Summaries of Relationships
• Modeling Data	• Modeling Relationships
Model Checking	

5. Visualize 2+ Variables

- 'Tools' to Visualize Relationships
  - (Two-Way Frequency Tables)
  - Segmented Barplots; Side-by-Side Barplots
  - Scatter Plots
  - Side-by-Side Boxplots

5.1. Visualizing Categorical–Categorical Relationships

- Two-Way Frequency Tables
  - Frequency
  - Relative Frequency
  - Relative Frequency within Group

- **Example**: Titanic Data

(a) Frequency Table

	1st	2nd	3rd	Total
Died	129	161	573	863
Alive	193	119	138	450
Total	322	280	711	1313

(b) Rel. Freq. w/in Row

	1st	2nd	3rd	Total
Died	$\frac{129}{863}$	$\frac{161}{863}$	$\frac{573}{863}$	$\frac{863}{863}$
Alive	$\frac{193}{450}$	$\frac{119}{450}$	$\frac{138}{450}$	$\frac{450}{450}$

(c) Rel. Freq. w/in Column

	1st	2nd	3rd
Died	$\frac{129}{322}$	$\frac{161}{280}$	$\frac{573}{711}$
Alive	$\frac{193}{322}$	$\frac{119}{280}$	$\frac{138}{711}$
Total	$\frac{322}{322}$	$\frac{280}{280}$	$\frac{711}{711}$

(d) Rel. Freq. Table

	1st	2nd	3rd	Total
Died	$\frac{129}{1313}$	$\frac{161}{1313}$	$\frac{573}{1313}$	$\frac{863}{1313}$
Alive	$\frac{193}{1313}$	$\frac{119}{1313}$	$\frac{138}{1313}$	$\frac{450}{1313}$
Total	$\frac{322}{1313}$	$\frac{280}{1313}$	$\frac{711}{1313}$	$\frac{1313}{1313}$

- **Example**: Titanic Data (cont1)

(a) Frequency Table

	1st	2nd	3rd	Total
Died	129	161	573	863
Alive	193	119	138	450
Total	322	280	711	1313

(b) Rel. Freq. w/in Row

	1st	2nd	3rd	Total
Died	.149	.187	.664	1.000
Alive	.429	.264	.307	1.000

(c) Rel. Freq. w/in Column

	1st	2nd	3rd
Died	.401	.575	.806
Alive	.599	.425	.194
Total	1.000	1.000	1.000

(d) Rel. Freq. Table

	1st	2nd	3rd	Total
Died	.098	.123	.436	.657
Alive	.147	.091	.105	.343
Total	.245	.213	.542	1.000

- Statistical Jeopardy
  - Table (b) Answer: .264

    What proportion of survivors came from 2nd class?
  - Table (b) Answer: .307

    What proportion of survivors came from 3rd class?
  - Table (c) Answer: .425

    What proportion of 2nd class passengers survived?
  - Table (c) Answer: .194

    What proportion of 3rd class passengers survived?
  - Table (d) Answer: .091

    What proportion of **Titanic passengers** were in 2nd class and survived?
  - Table (d) Answer: .105

    What proportion of **Titanic passengers** were in 3rd class and survived?

- Visualization of Table (c) (Figure 4)
  - **(c-1)** Segmented Bar Graph of Survived within Class

    Heights of segments give proportion Dead and Alive within each class.
  - **(c-2)** Side-by-Side Bar Graph of Survived within Class

    Bar segments from (c-1) moved to baseline to make height easier to read.
  - **(c-3)** Side-by-Side Bar Graph of Survived within Class

    Same bars as in (c-2) but grouped by the variable survived.
  - **(c-4)** Bar Graph of Proportion Alive

    Three bars from right side of (c-3); since survived has only two categories, proportion Dead is redundant information.

5.2. Visualizing Numeric–Numeric Relationships (Scatter Plot)

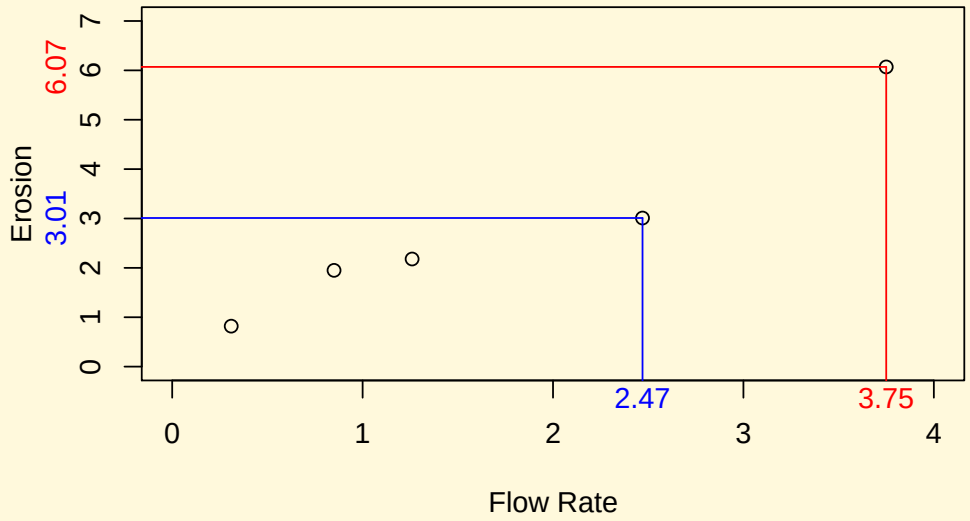
– Response and Explanatory Variables

* **Response variable** is the variable that *responds* to changes in another variable, called the **explanatory variable**.
* Response variable also called dependent variable; explanatory variable also called independent variable.
* **NOTE:** Nomenclature is to distinguish how the variables are used in visualizing (plotting) and modeling the relation. There need be no cause-and-effect relationship.
* In some cases, either variable can be treated as the explanatory variable.

- **Example**: Erosion and Flow Rate

Flow (liter/sec)	3.75	2.47	1.26	0.85	0.31
Erosion (kg)	6.07	3.01	2.18	1.95	0.82

  - **Pairs** of observations
  - Explanatory variable along **x**-axis



- What to look for in a scatter plot
  - Association
    - positive (Figure 5)
      Large x associated with Large y
      (Small x associated with Small y)
    - negative (Figure 6)
      Large x associated with Small y
      Small x associated with Large y
    - no association (Figure 7)
  - Shape
    - Linear - line can be place in *center* of the point cloud
    - Non-linear
  - Departures from Overall Shape
    - Outliers
    - Heteroscedasticity

- Discussion of Scatter Plots in Figures 5, 6, and 7.
  - Linear Model (Regression) can be used for data in top six graphs
  - Requirement is Linear shape with no departures from shape.
  - Pos Assoc: lower plots show nonlinearity and heteroscedasticity.
  - Neg Assoc: lower plots show outliers (one has high leverage).
  - No Assoc: lower plots show no assoc $\neq$ no relationship.

- **Example**: Fisher's Iris Data
  - Reference: Fisher, R. A. (1936). The Use of Multiple Measurements in Axonomic Problems. *Annals of Eugenics* 7, 179-188.
  - Description: Data set containing five variables with 150 observations (50 from each of three species of iris).
  - Variable Names:

    Species: *Iris setosa, I. versicolor, and I. virginica*

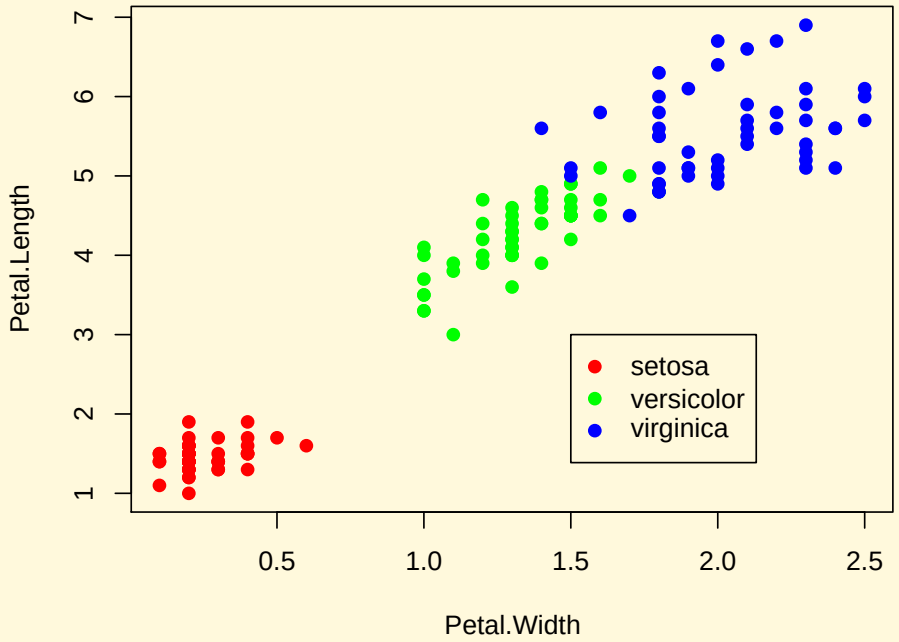
    Petal.Width: Petal Width (cm)

    Petal.Length: Petal Length (cm)

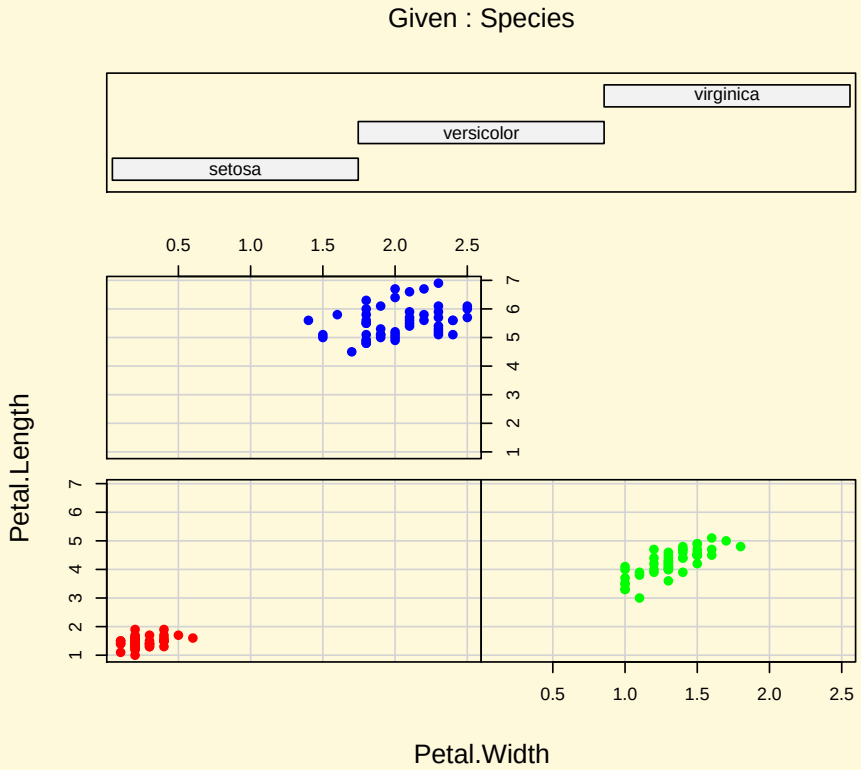
    Sepal.Width: Sepal Width (cm)

    Sepal.Length: Sepal Length (cm)

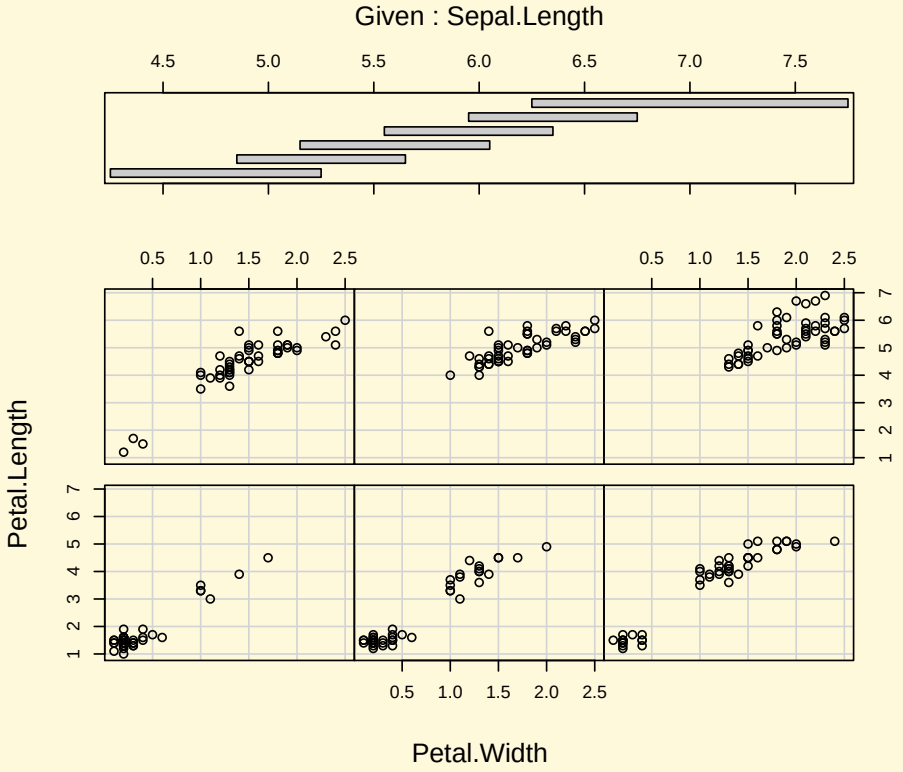
- **Example**: Iris Data – Scatter Plot w/ Color



- **Example**: Iris Data – Conditioning Plot (Trellis Plot)



◦ **Example**: Iris Data – Conditioning Plot (cont)



- Advantages of Conditioning Plot
  - Separates groups of overlaid points.
  - Works better than colors when there are many categories.
  - Can be used to condition on continuous data.

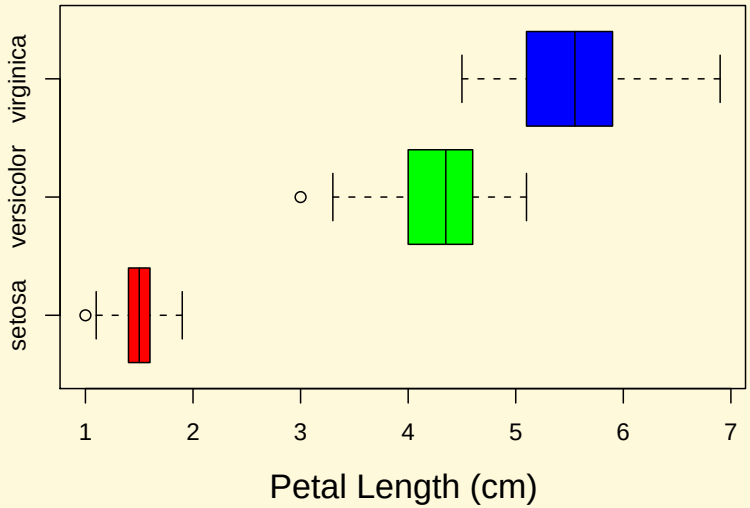
5.3. Visualizing Numeric–Categorical Relationships (Box Plot)

– Categorical variable partitions data

– One boxplot for each subgroup

– All boxplots share common reference line

◦ **Example**: Iris Data (cont)



6. Lurking Variables

- **Example**: Salary and Experience at University A
  - Salary: Faculty Salaries
  - Experience: Years in profession
  - Association: expect positive
  - (fictitious data)
  - Salary versus Experience (Figure 8)
  - Salary vs Experience given Dept (Figures 9, 10)
  - Salary vs Dept and Experience vs Dept (Figure 13)

- Discussion
  - This is an example of **Simpson's Paradox**.
  - The relationship between two variables can *reverse* when a third variable is taken into account.
  - In this example, we measured the variable (Dept).
  - But what about **Lurking variables** – variables that affect the relationship but are not measured.
  - The problem of lurking variables is important even if they do not reverse the relationship (see next example).
- **Example**: Salary and Experience at University B
  - Figures 11 and 12
  - Figure 14

7. Summarize 2+ Variables

- Informal definition

  Measure of the strength of the **linear** relationship between two variables.

- Formal definition

$$r = r_{xy} = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

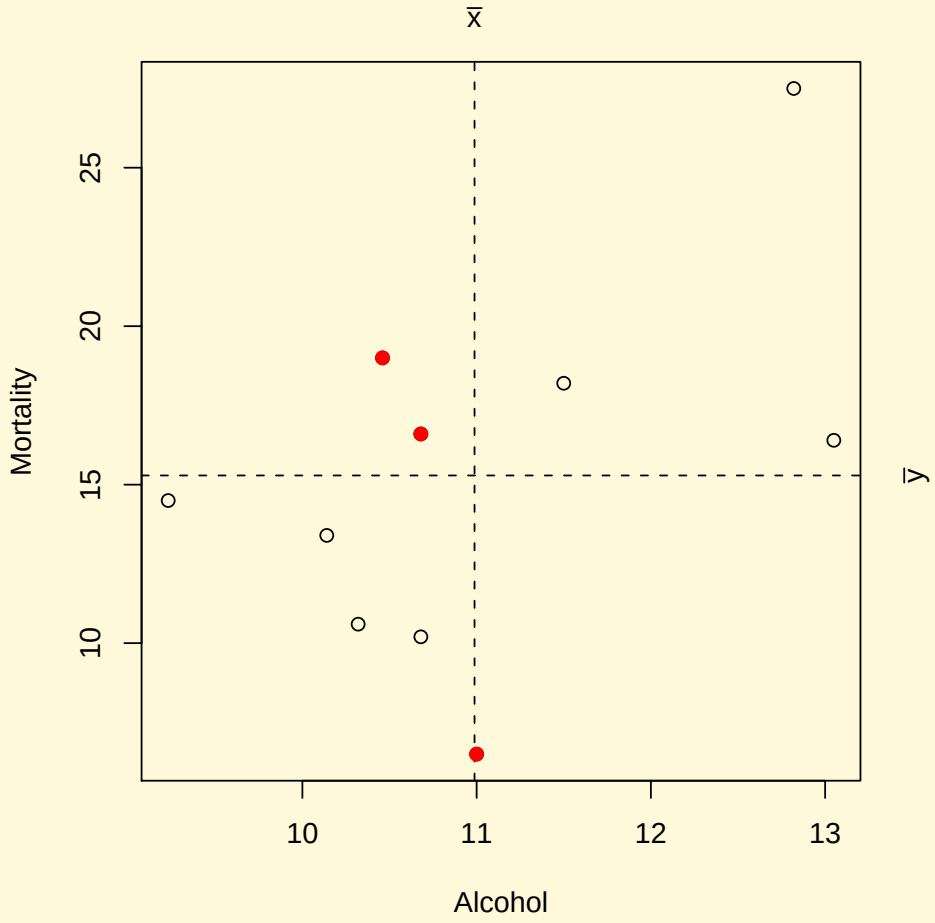
where s_x is the standard deviation of x and s_y is the standard deviation of y.

- **Example**: Alcohol and Cirrhosis Mortality (Anderson & Finn)

Observation	Alcohol x_i	Mortality y_i	$\frac{x_i-\bar{x}}{s_x}$	$\frac{y_i-\bar{y}}{s_y}$	$\frac{x_i-\bar{x}}{s_x}\frac{y_i-\bar{y}}{s_y}$
1 Pr. Edward Is.	11.00	6.5	0.010	-1.513	-0.015
2 Newfoundland	10.68	10.2	-0.260	-0.876	0.228
3 Nova Scotia	10.32	10.6	-0.565	-0.807	0.456
4 Saskatchewan	10.14	13.4	-0.717	-0.325	0.233
5 New Brunswick	9.23	14.5	-1.486	-0.136	0.202
6 Alberta	13.05	16.4	1.743	0.191	0.333
7 Manitoba	10.68	16.6	-0.260	0.226	-0.059
8 Ontario	11.50	18.2	0.433	0.501	0.217
9 Quebec	10.46	19.0	-0.446	0.639	-0.285
10 Brit. Columbia	12.82	27.5	1.549	2.102	3.255
Total:	109.88	152.9	0	0	4.566
Mean*:	10.988	15.29	0	0	.507

Correlation $r_{xy} = .507$.

- Scatter Plot for Correlation



- Properties of r_{xy}
  - r is symmetric: $r_{xy} = r_{yx} = r$
  - Range of r:$-1 \leq r \leq 1$
  - Meaning of $r = -1$ or $r = 1$:

    Points fall *exactly* on a line with positive slope ($r = 1$) or with negative slope ($r = -1$).
  - What $r = 0$ does NOT mean:

    It does **NOT** mean there is no relationship; rather, there is no linear relationship.
  - r has no units.
  - r is not resistant to outliers.

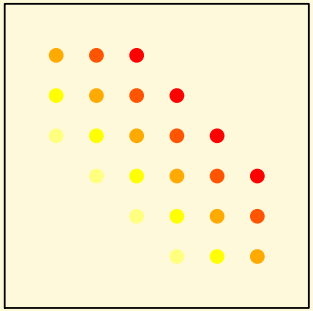
- Ecologic Correlation

  When correlations are calculated on grouped data, the correlation is *often stronger* than when calculated on individuals

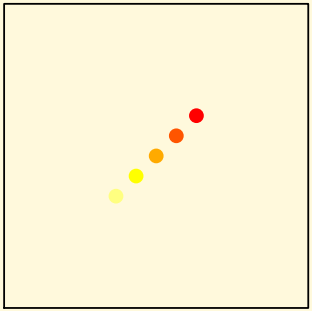
- Ecologic Fallacy

  It is possible that the correlation obtained from grouped data has opposite sign of the ungrouped data.

Plot of Ungrouped Data
Color Shows Group



Plot of Group Means



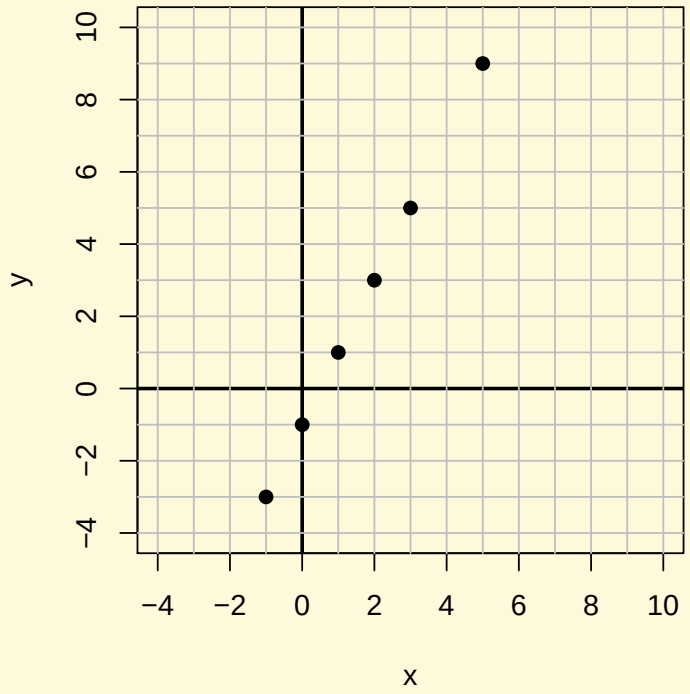
8. Model 2+ Variables

- Preview
  - Review of linear equations
  - Least Squares Line
  - Interpretation
  - Model Assumptions/Model Checking

8.1. Review of linear equations

Example: $y = 2x - 1$

x	-1	0	1	2	3	5
y	-3	-1	1	3	5	9



- General Equation: $y = bx + a$
- b is slope; $\dfrac{\text{Rise}}{\text{Run}}$
- a is y-intercept
- Every (non-vertical) line described by particular values for b and a.
- Usually more interested in b than in a.
- b shows how *changes* in x are related to *changes* in y.
- Units of b: $\dfrac{\text{units of } y}{\text{units of } x}$

8.2. Least Squares Regression Line

– slope $\hat{b} = r\frac{S_y}{S_x}$

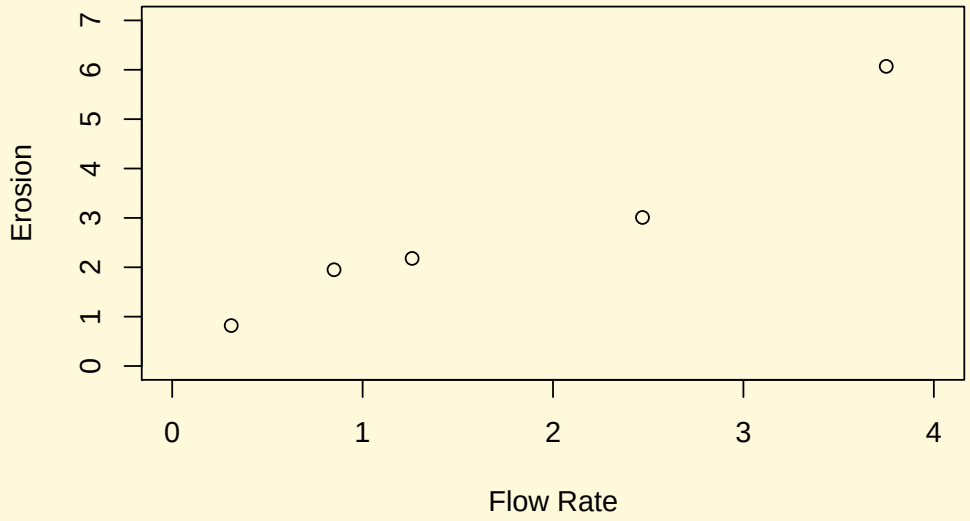
– y-intercept $\hat{a} = \bar{y} - \hat{b}\bar{x}$

– equation of Least Squares Line: $y = \hat{b}x + \hat{a}$

– **fitted** (model) **values** $\hat{y}_i = \hat{b}x_i + \hat{a}$ where x_i is the x-value of the ith pair (x_i, y_i).

– **residual** $r_i = y_i - \hat{y}_i$.

- **Example**: Erosion (cont from slide 1.)

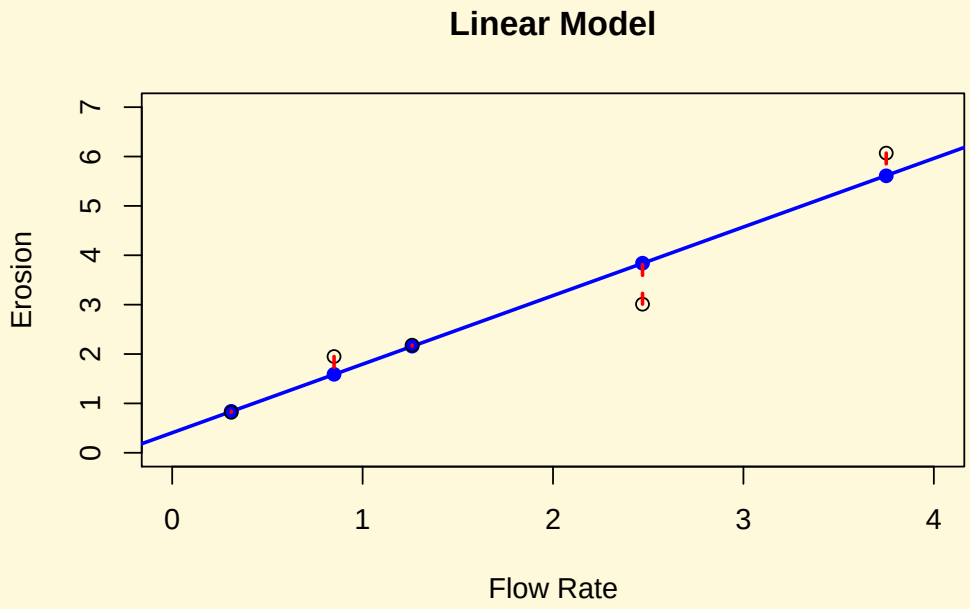


x_i = Flow (liter/sec)	3.75	2.47	1.26	0.85	0.31
y_i = Erosion (kg)	6.07	3.01	2.18	1.95	0.82

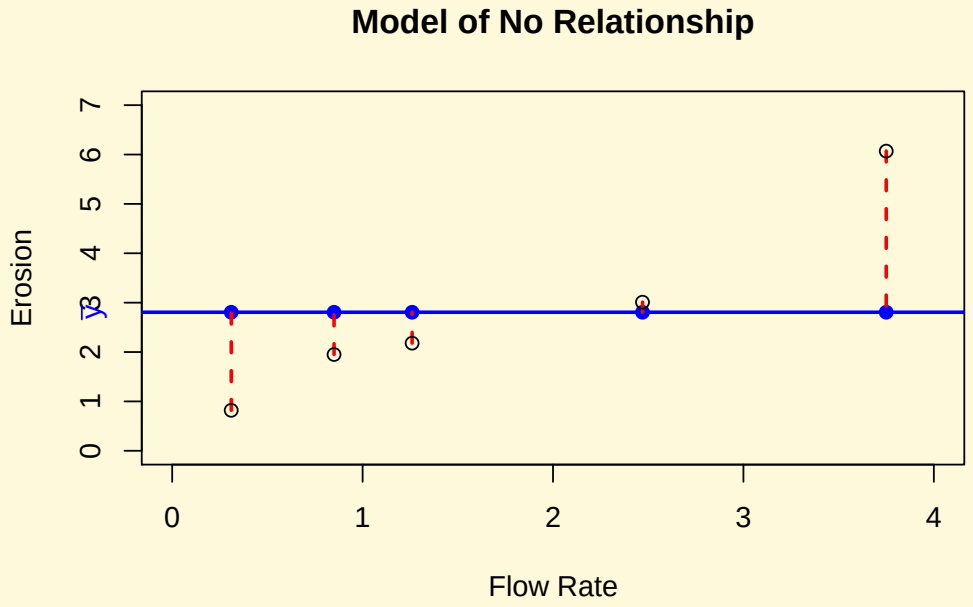
- **Example**: Erosion (cont)
  - parameters of l.s. line: $\hat{b} = 1.389, \ \hat{a} = .4057$
  - Equation of l.s. line: $y = 1.389x + .4057$
    (Or, Erosion $= 1.389 \times$ Flow $+ .4057$)
  - First fitted value $\hat{y}_1 = 1.389 \times 3.75 + .4057 = 5.61$
  - First Residual $r_1 = y_1 - \hat{y}_1 = 6.07 - 5.61 = .46$

i	1	2	3	4	5
$x_i =$ Flow (liter/sec)	3.75	2.47	1.26	0.85	0.31
$y_i =$ Erosion (kg)	6.07	3.01	2.18	1.95	0.82
$\hat{y}_i$	5.61	3.84	2.16	1.59	0.84
r_i	0.46	-0.83	0.02	0.36	-0.02

- **Example**: Erosion (cont)



- **Example**: Erosion (cont)



- Properties of Least Squares Line
  - Line must contain $(\bar{x}, \bar{y})$
  - Line minimizes the sum of *squared* residuals
  - Line is unique
  - Sensitive to outliers, high leverage points.
  - Relationship to r^2: Proportion of variability *explained* by x.

Total Variability	Sum of squared lengths in 1.	15.76
Variability Unexplained by L. S. line	Sum of squared lengths in 1.	1.03
Variability Explained by L. S. line	Difference	14.73
Prop. Explained	Difference/Total	.935

  - Two regression lines: $Y \mid x$ and $X \mid y$.

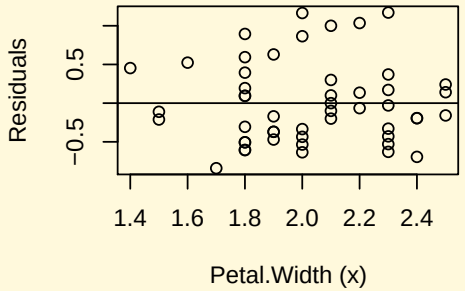
8.3. Interpretation of the Regression Model

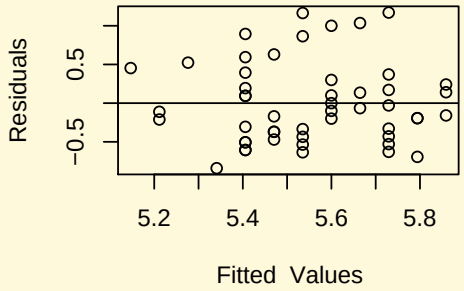
- **Example**: SAT Verbal and Math Scores (made-up data) See Figure 15.
  - X = Verbal SAT; Y = Math SAT
  - $\bar{X} = 503.0$, $\bar{Y} = 499.3$, $s_X = 93.34$, $s_Y = 86.85$, $r = .6926$
  - equation of Least Squares line: $y = .64x + 177.4$.
  - Regression Model: Mean (Math SAT) $= .64 \times$ (Verbal SAT) $+ 177.4$
  - At $x = 450$ (i.e., 450 on verbal), model predicts Mean Math SAT $= .64 \times (350) + 177.4 = 465$
  - At $x = 600$ (i.e., 600 on verbal), model predicts Mean Math SAT $= .64 \times (600) + 177.4 = 561$

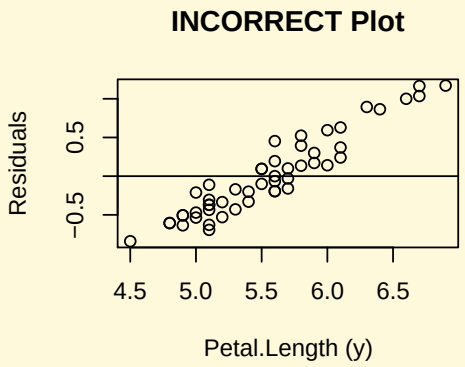
8.4. Model Checking

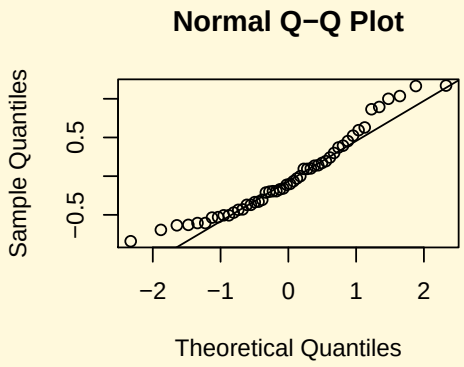
- Regression Model Assumptions.
  - Assumptions about $Y \mid x$ for different values of x
  - Center: $\text{Mean}(Y \mid x)$ is linear in x; $\text{Mean}(Y \mid x) = bx + a$ for some values a and b.
  - Spread: $\text{SD}(Y \mid x)$ is constant
  - No outliers
  - (Shape: Normal; $Y \mid x \sim N(bx + a, \sigma)$)
- Residual Plots
  - Residual = Observed value − Model value
  - $r_i = y_i - \hat{y}_i$
  - Plot (x_i, r_i) or $(\hat{y}_i, r_i)$; **NOT** (y_i, r_i)

- **Example**: Iris Virginica Data (slide 1.)
  Regress Petal.Length on Petal.Width





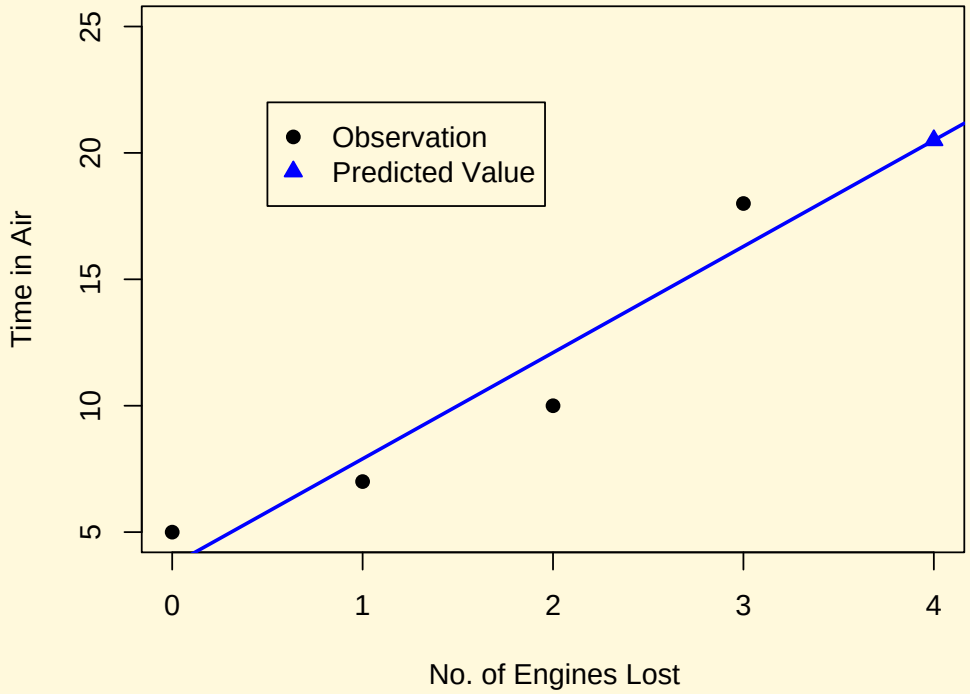




- Indications of Problems:
  - Nonlinear shapes
  - Outliers
  - Changing Spread
- Danger of Extrapolation

Two statisticians were traveling in an airplane from LA to New York. About an hour into the flight, the pilot announced that they had lost an engine, but don't worry, there are three left. However, instead of 5 hours it would take 7 hours to get to New York. A little later, he announced that a second engine failed, and they still had two left, but it would take 10 hours to get to New York. Somewhat later, the pilot again came on the intercom and announced that a third engine had died. Never fear, he announced, because the plane could fly on a single engine. However, it would now take 18 hours to get to new York. At this point, one statistician turned to the other and said,"Gee, I hope we don't lose that last engine, or we'll be up here forever!"

- Danger of Extrapolation (cont)



9. Review/Omissions

One Variable	Two or More Variables
• Data Classification – (n) numeric data – (c) categorical data	
• Visualizing Data – freq tables (2) – (n) stemplot, histogram, boxplot – (c) pie, bar plots	• Visualizing Relationships – freq tables (4) – (n-n) scatter plot – (n-c) side-by-side boxplots – (c-c) segmented barplots

Descriptive Statistics Review (cont)

One Variable	Two or More Variables
• Summarizing Data – center: mean, median – spread: st. dev., IQR – 5 Number Summary – (c) mode	• Summarizing Relationships – (n-n) correlation – (n-c) – (c-c) odds ratio relative risk

Descriptive Statistics Review (cont)

One Variable	Two or More Variables
• Modeling Data – (n) Normal Distn – (n) Exponential Distn – (c) Binomial Distn – (c) Poisson Distn	• Modeling Relationships – (n∼nc) linear regression – (n∼nc) nonlinear regression – (n∼c) ANOVA – (n∼cn) ANCOVA – (c∼nc) logistic regression – (c∼c) contingency tables

• Model Checking
– Q-Q plots
– Residual plots

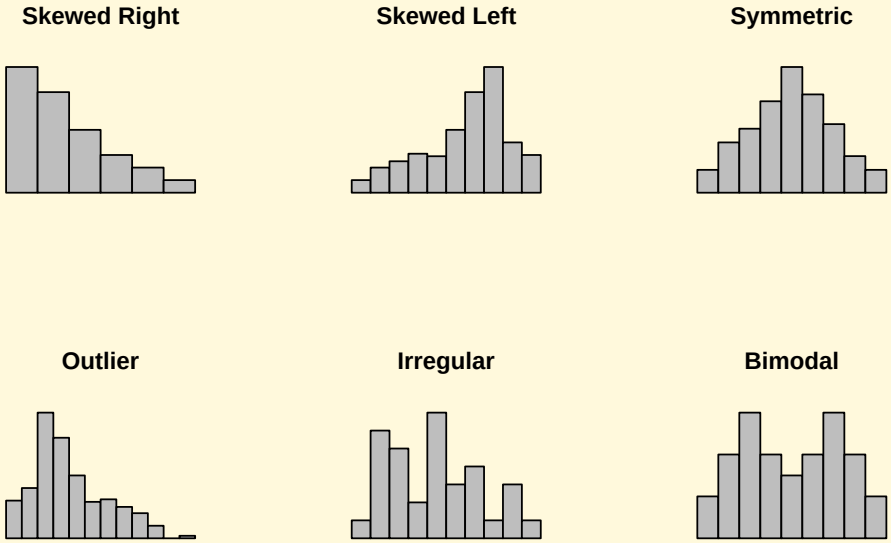


Figure 1: Histograms for six data sets.

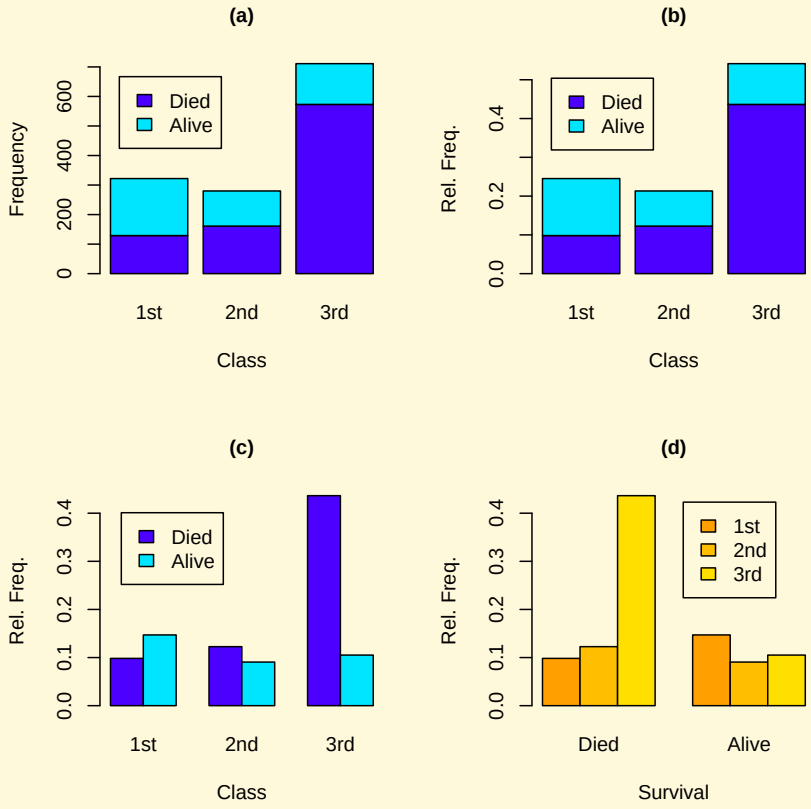


Figure 2: Bar graphs for the variable `Class`*`Survived` (Titanic Data).

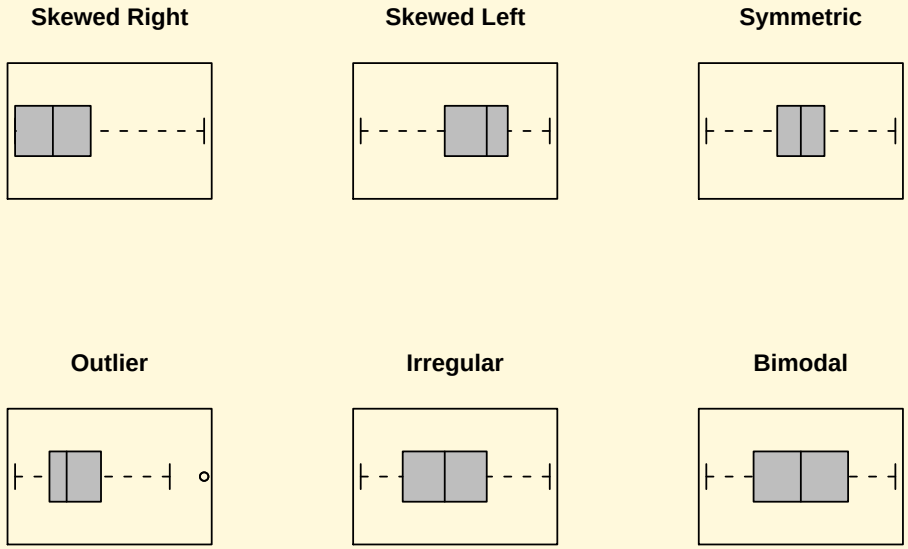


Figure 3: Shapes of six distributions (same data as histograms in Figure 1).

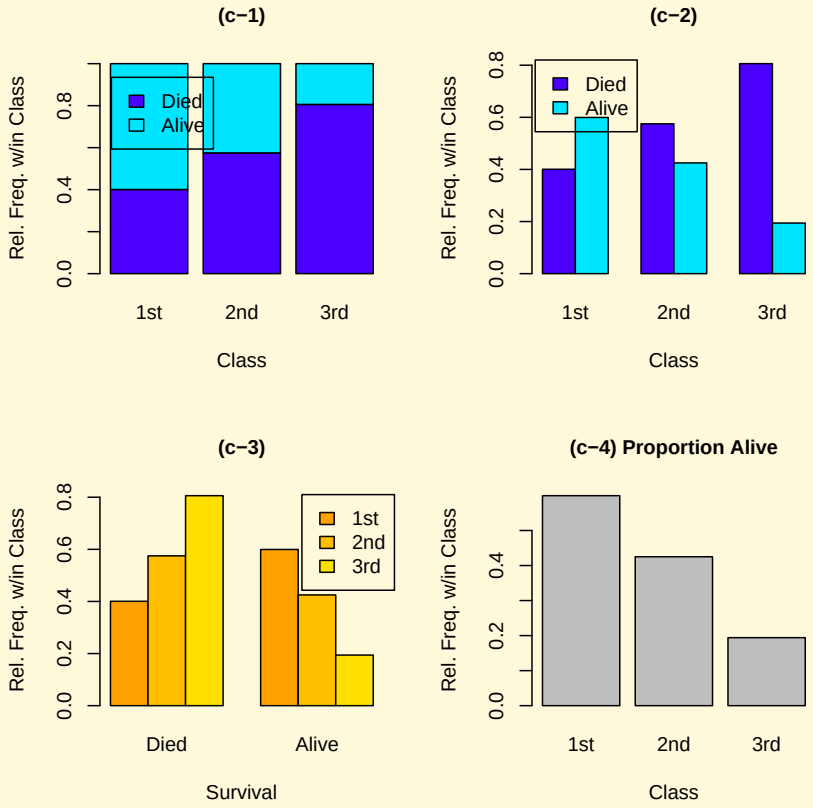


Figure 4: Bar graphs of the 2-way table of `Survived` within `Class`. Compare with Figure 2.

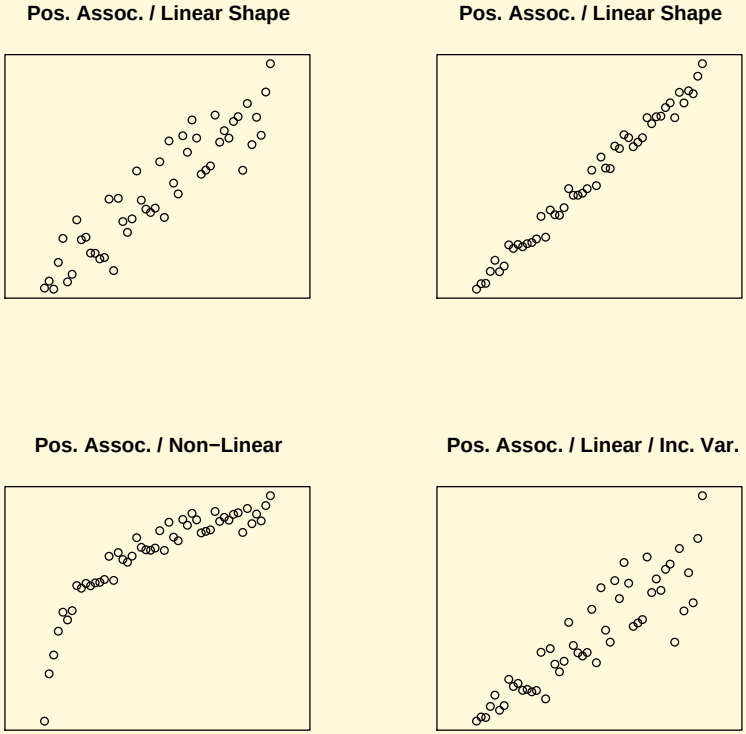


Figure 5: Scatter plots having positive association.

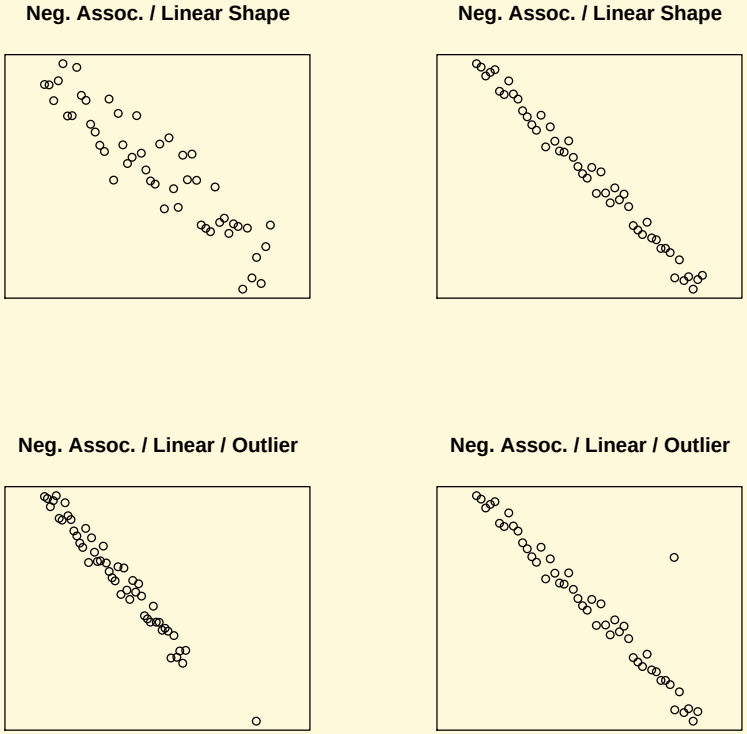


Figure 6: Scatter plots having negative association.

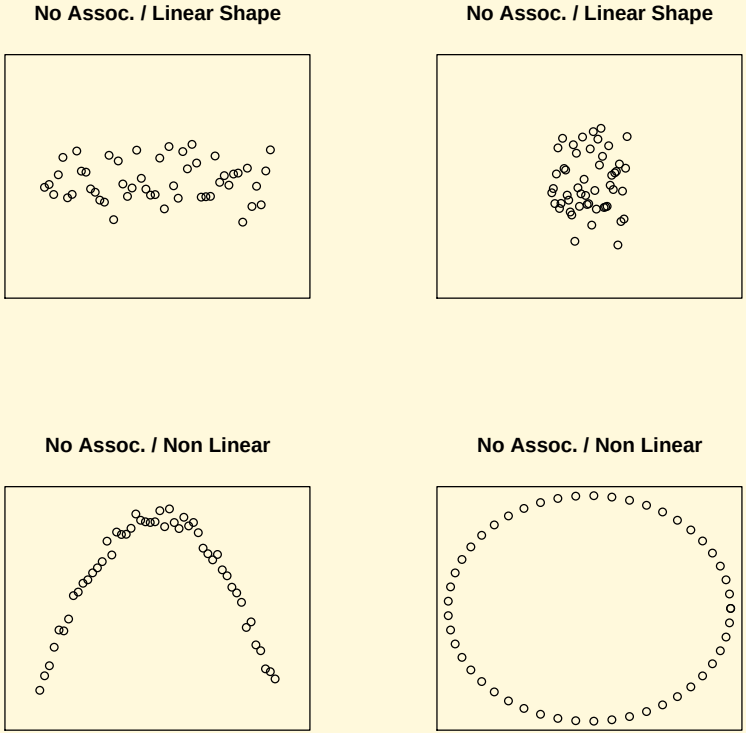


Figure 7: Scatter plots having no association.

Figure 8: Data from University A.

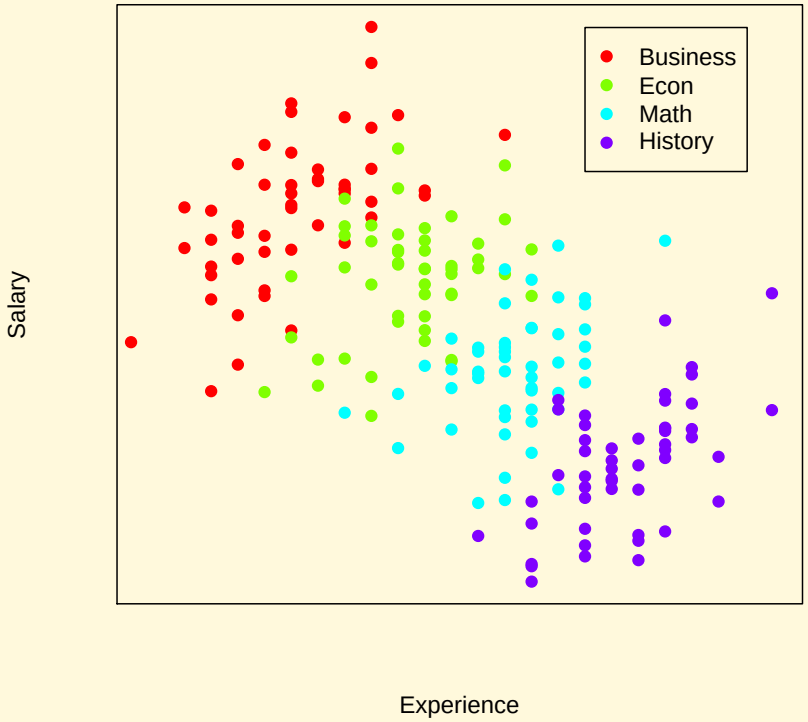


Figure 9: Data from University A. (Compare with University B in Figure 11.)

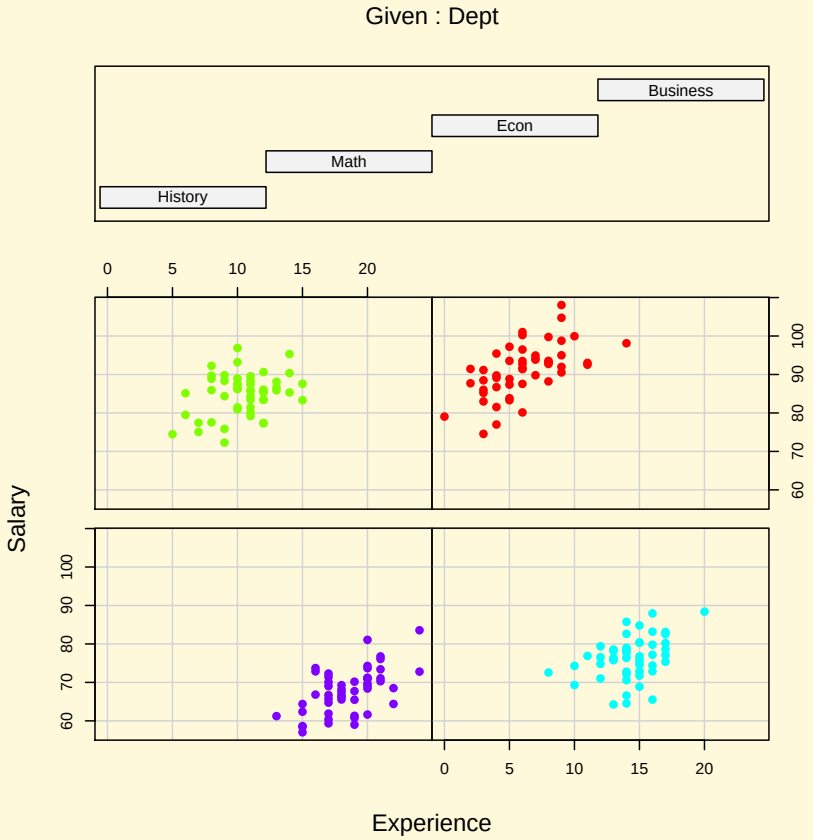


Figure 10: Data from University A. (Compare with University B in Figure 12.)

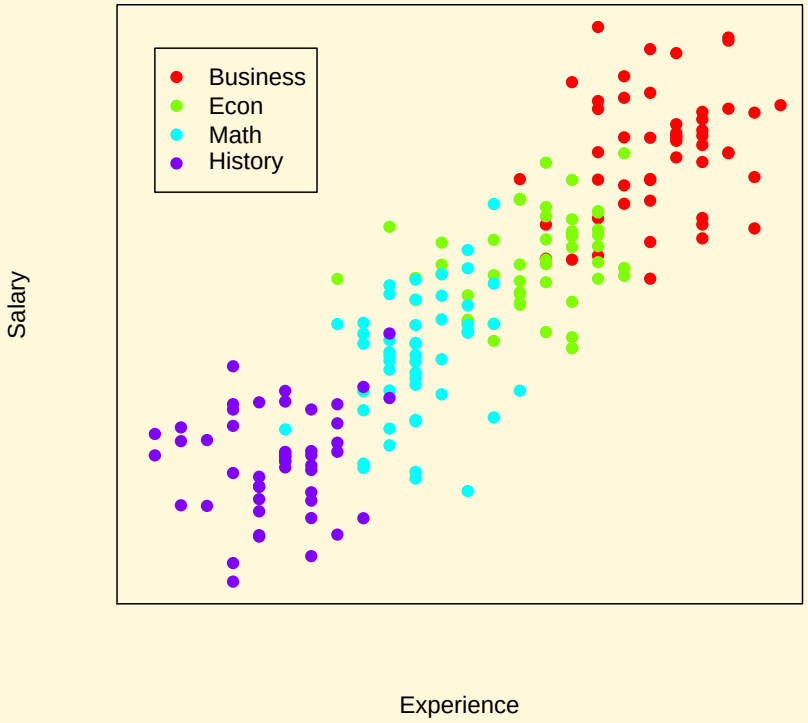


Figure 11: Data from University B. (Compare with University A in Figure 9.)

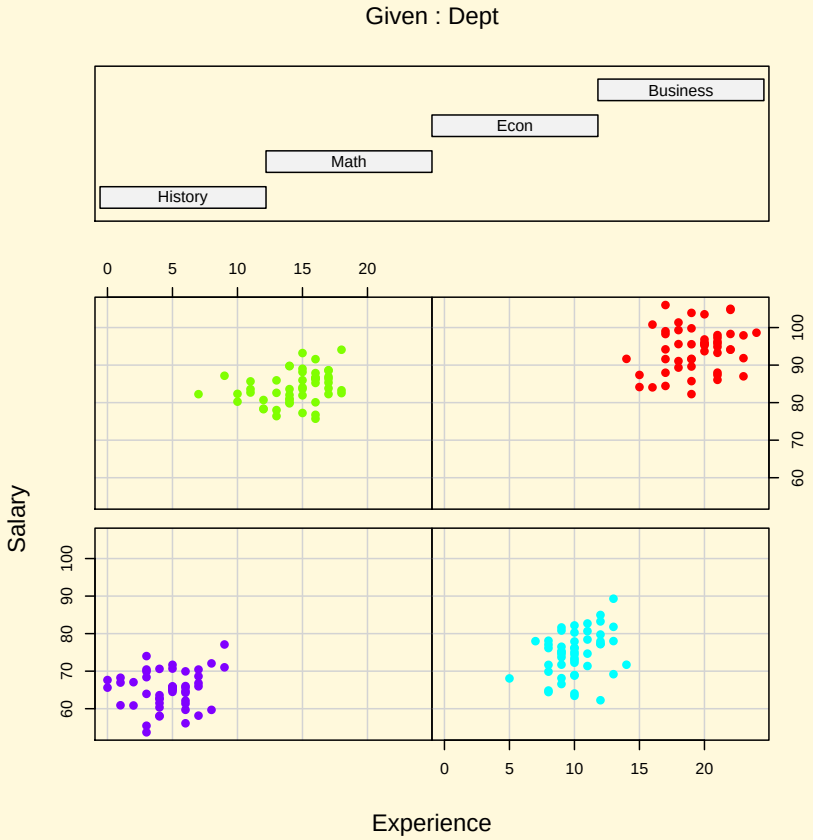


Figure 12: Data from University B.(Compare with University A in Figure 10.)

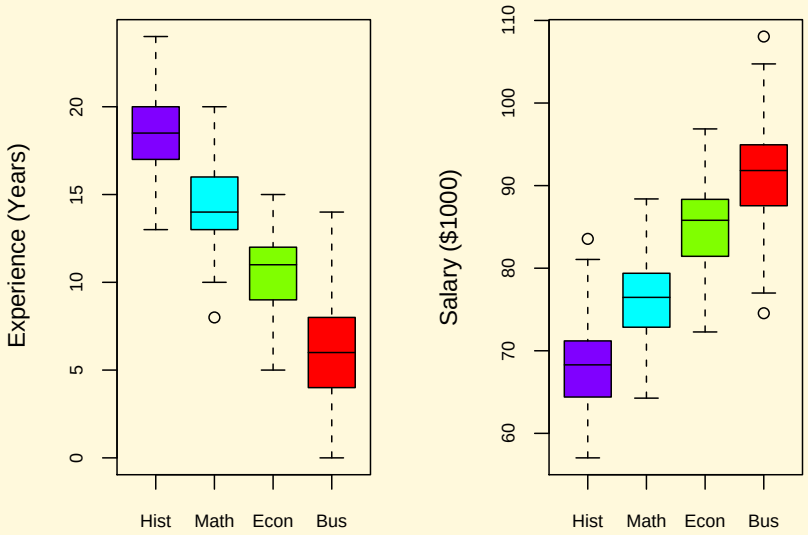


Figure 13: Data from University A. (Compare with University B in Figure 14.)

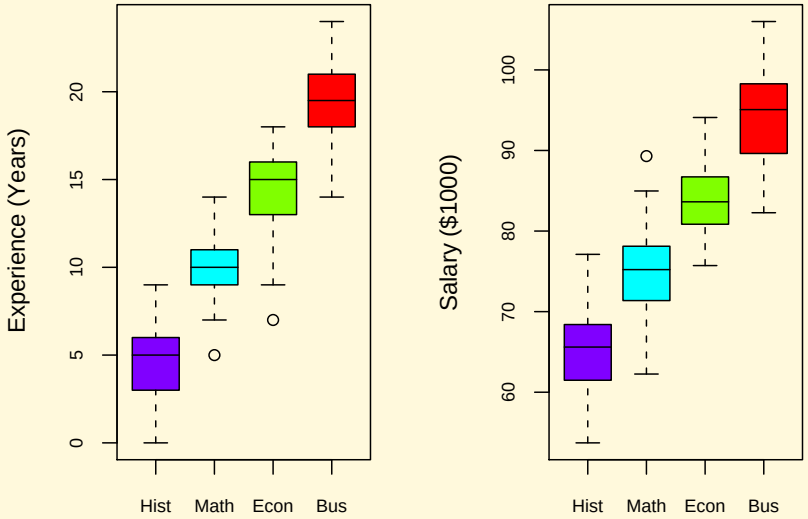


Figure 14: Data from University B. (Compare with University A in Figure 13.)

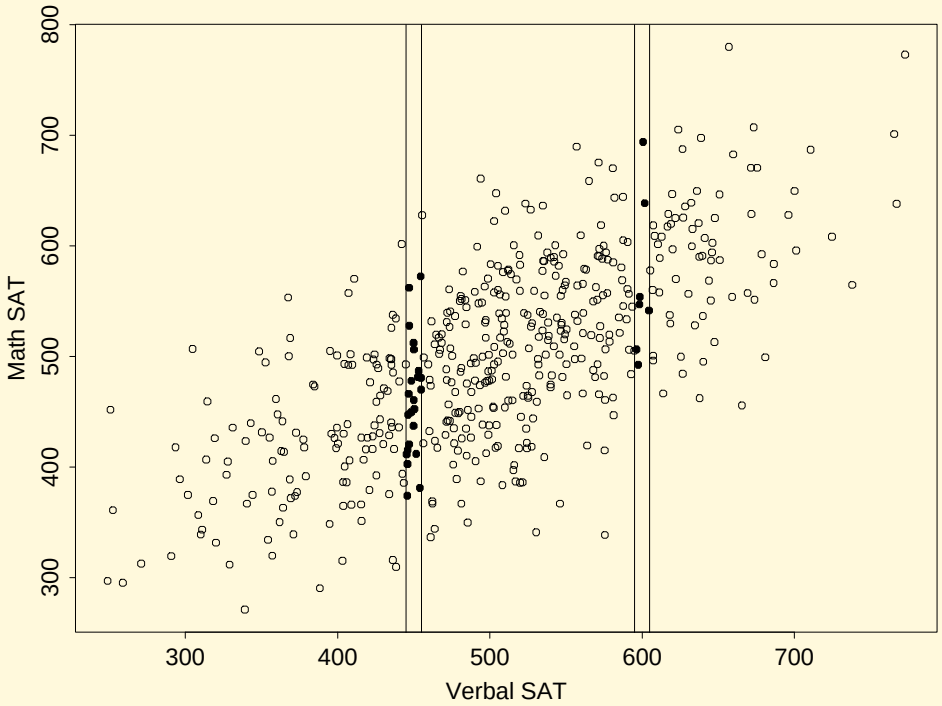


Figure 15: SAT scores.