

STAT 3743: Probability and Statistics

G. Jay Kerns, Youngstown State University

Fall 2010

Types of Data

datum: any piece of information

data set: collection of data related to each other somehow

- Categories of Data
 - **quantitative:** associated with a measurement of some quantity on an observational unit,
 - **qualitative:** associated with some quality or property of the observational unit,
 - **logical:** represents true/false, important later
 - **missing:** should be there but aren't
 - **other types:** everything else

Quantitative Data

Quantitative data: any that measure the quantity of something

- invariably assume numerical values
- can be further subdivided:
 - *Discrete data* take values in a finite or countably infinite set of numbers
 - *Continuous data* take values in an interval of numbers.
AKA scale, interval, measurement
- distinction between discrete and continuous data not always clear-cut

Example

Annual Precipitation in US Cities. (precip) avg amount rainfall (in.) for 70 cities in US and Puerto Rico.

```
> str(precip)
```

```
Named num [1:70] 67 54.7 7 48.5 14 17.2 20.7 13 43.4 4  
- attr(*, "names")= chr [1:70] "Mobile" "Juneau" "Phoe
```

```
> precip[1:4]
```

Mobile	Juneau	Phoenix
67.0	54.7	7.0
Little Rock		
48.5		

quantitative, continuous

Example

Lengths of Major North American Rivers. (rivers)

lengths (mi) of rivers in North America. See ?rivers.

```
> str(rivers)
```

```
num [1:141] 735 320 325 392 524 ...
```

```
> rivers[1:4]
```

```
[1] 735 320 325 392
```

Example

Yearly Numbers of Important Discoveries.

(discoveries) numbers of “great” inventions/discoveries in each year from 1860 to 1959 (from 1975 World Almanac)

```
> str(discoveries)
```

```
Time-Series [1:100] from 1860 to 1959: 5 3 0 2 0 3 2 3
```

```
> discoveries[1:4]
```

```
[1] 5 3 0 2
```

Displaying Quantitative Data

- Strip charts (or Dot plots):
 - for either discrete or continuous data
 - usually best when data not too large.
- the `stripchart` function
 - three methods:
 - `overplot` - only distinct values
 - `jitter` - add noise in *y* direction
 - `stack` - repeats on top of one another

Displaying Quantitative Data

- Strip charts (or Dot plots):
 - for either discrete or continuous data
 - usually best when data not too large.
- the `stripchart` function
 - three methods:
 - overplot - only distinct values
 - jitter - add noise in y direction
 - stack - repeats on top of one another


```
> stripchart(precip, xlab = "rainfall")  
> stripchart(rivers, method = "jitter",  
+   xlab = "length")  
> stripchart(discoveries, method = "stack",  
+   xlab = "number")
```

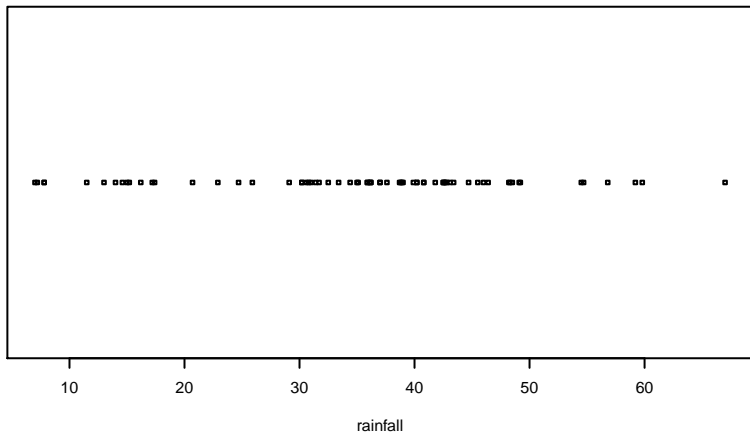


Figure: Stripchart of precip

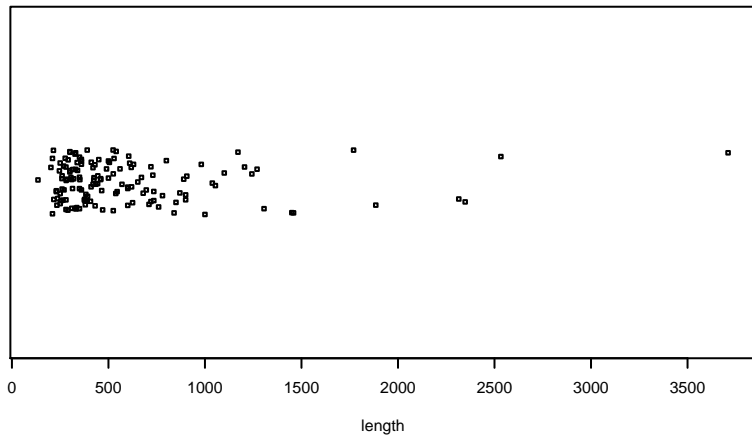


Figure: Stripchart of rivers

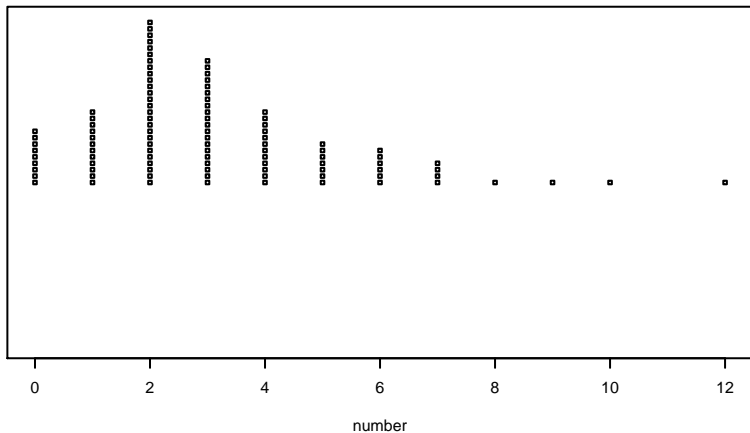


Figure: Stripchart of discoveries

Histograms

- Histograms
 - typically for continuous data
 - decide on bins/classes, make bars proportional to membership
 - often misidentified (bar graphs)

```
> hist(precip, main = "")
```

```
> hist(precip, freq = FALSE, main = "")
```

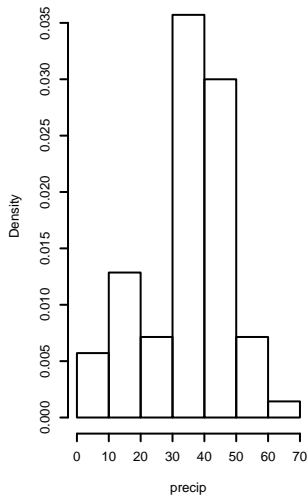
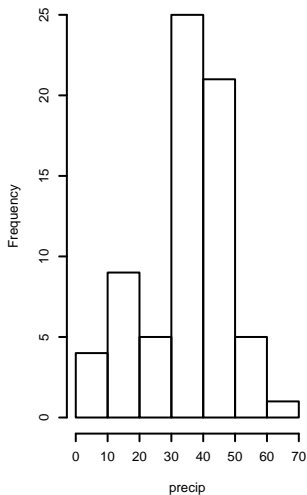


Figure: Histograms of precip

Remarks about histograms

- choose different bins, get a different histogram
- many algorithms for choosing bins automatically
- should investigate several bin choices
 - look for stability
 - try to capture underlying story of data

Stemplots

- Stemplots have two basic parts: *stems* and *leaves*
 - initial digit(s) taken for stem
 - trailing digits stand for leaves
 - leaves accumulate to the right

Example

Road Casualties in Great Britain 1969-84. A time series of total car drivers killed or seriously injured in Great Britain monthly from Jan 1969 to Dec 1984.

Stemplot of UK Driver Deaths

```
> library(aplpack)
> stem.leaf(UKDriverDeaths, depth = FALSE)
```

```
1 | 2: represents 120
leaf unit: 10
      n: 192
10 | 57
11 | 136678
12 | 123889
13 | 0255666888899
14 | 00001222344444555556667788889
15 | 000011111222222344445555566677779
16 | 012223334444445555555678888889
17 | 11233344566667799
18 | 00011235568
19 | 01234455667799
20 | 0000113557788899
21 | 145599
22 | 013467
23 | 9
24 | 7
HI: 2654
```

Code for stemplots

```
> UKDriverDeaths[1:4]
[1] 1687 1508 1507 1385
> stem.leaf(UKDriverDeaths, depth = FALSE)
1 | 2: represents 120
leaf unit: 10
      n: 192
10 | 57
11 | 136678
12 | 123889
13 | 0255666888899
14 | 00001222344444555556667788889
15 | 000011111222222344445555566677779
```

Index Plots

Good for plotting data ordered in time

- a 2-D plot, with index (observation number) on x -axis, value on y -axis
- two methods
 - spikes: draws vertical line up to value (type = "h")
 - points: simple dot at the observed height (type = "p")

Example

Level of Lake Huron 1875-1972. annual measurements of the level (in feet) of Lake Huron from 1875-1972.

Index Plots

Good for plotting data ordered in time

- a 2-D plot, with index (observation number) on x -axis, value on y -axis
- two methods
 - spikes: draws vertical line up to value (`type = "h"`)
 - points: simple dot at the observed height (`type = "p"`)

Example

Level of Lake Huron 1875-1972. annual measurements of the level (in feet) of Lake Huron from 1875–1972.

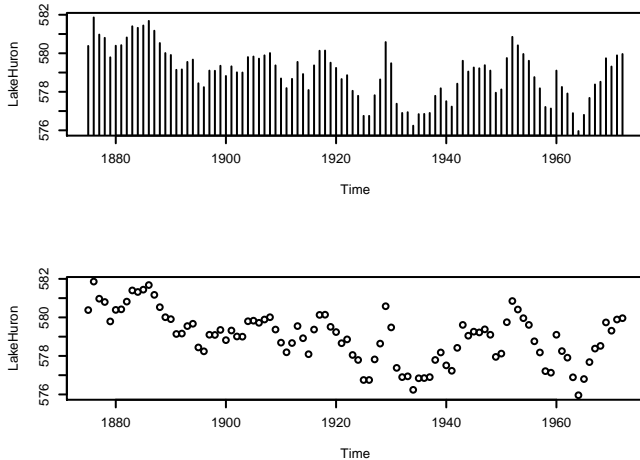


Figure: Index plots of LakeHuron

Qualitative Data, Categorical Data, Factors

- **Qualitative data:** any data that are not numerical, or do not represent numerical quantities
 - some data *look* qualitative. Example: shoe size
 - some data *identify* the observation, not of much interest
- **Factors** subdivide data into categories
 - possible values of a factor: *levels*
 - factors may be *nominal* or *ordinal*
 - **nominal:** levels are names, only (gender, political party, ethnicity)
 - **ordinal:** levels are ordered (SES, class rank, shoe size)

Example

U.S. State Facts and Features. postal abbreviations

```
> str(state.abb)
```

```
chr [1:50] "AL" "AK" "AZ" "AR" ...
```

Example

U.S. State Facts and Features. The region in which a state resides

```
> state.region[1:4]
```

```
[1] South West West South
```

```
4 Levels: Northeast South ... West
```

Qualitative Data

- Factors have special status in R
 - represented internally by numbers, but not always printed that way
 - constructed with `factor` command
- Displaying Qualitative Data
 - first try: make a (contingency) table with `table` function
 - `prop.table` makes a relative frequency table

Example

U.S. State Facts and Features. State division

Displaying Qualitative Data

```
> Tbl <- table(state.division)
> Tbl                                # frequencies
state.division
      New England      Middle Atlantic
              6              3
South Atlantic East South Central
              8              4
West South Central East North Central
              4              5
West North Central      Mountain
              7              8
      Pacific
```

Displaying Qualitative Data

```
> Tbl/sum(Tbl)      # relative frequencies
```

```
state.division
```

New England	Middle Atlantic
-------------	-----------------

0.12	0.06
------	------

South Atlantic	East South Central
----------------	--------------------

0.16	0.08
------	------

West South Central	East North Central
--------------------	--------------------

0.08	0.10
------	------

West North Central	Mountain
--------------------	----------

0.14	0.16
------	------

Pacific

0.10

Displaying Qualitative Data

```
> prop.table(Tbl)    # same thing
```

```
state.division
```

New England	Middle Atlantic
0.12	0.06
South Atlantic	East South Central
0.16	0.08
West South Central	East North Central
0.08	0.10
West North Central	Mountain
0.14	0.16
Pacific	
0.10	

Bar Graphs

- discrete analogue of the histogram
- make bar for each level of a factor
- may show frequencies or relative frequencies
- impression given depends on order of bars (default: alphabetical)

Example

U.S. State Facts and Features. State region

```
> barplot(table(state.region))  
> barplot(prop.table(table(state.region)))
```

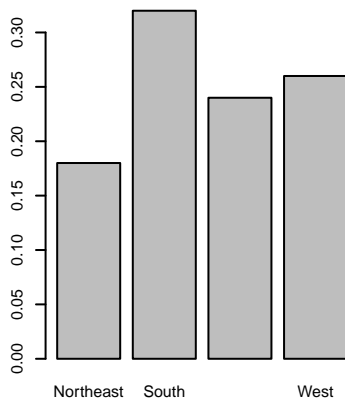
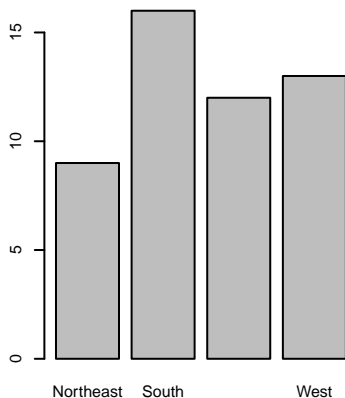


Figure: (Relative) frequency bar graphs of `state.region`

Pareto Diagram

- a bar graph with ordered bars
- bar with highest (relative) frequency goes on left
- bars drop from left to right
- can sometimes help discern hidden structure

Example

U.S. State Facts and Features. State division

```
> library(qcc)
> pareto.chart(table(state.division),
+             ylab = "Frequency")
```

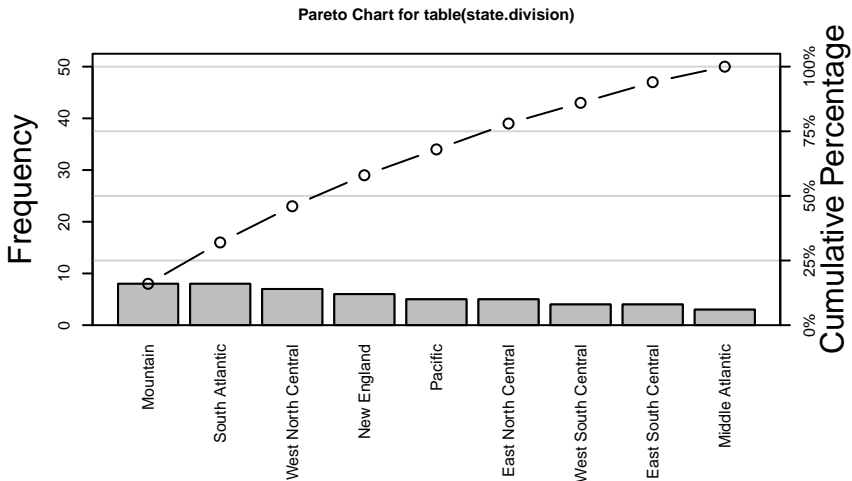


Figure: Pareto diagram of state.division

Dot Charts

- a bar graph on its side
- has dots instead of bars
- can show complicated multivariate relationships

Example

U.S. State Facts and Features. State region

```
> x <- table(state.region)
> dotchart(as.vector(x), labels = names(x))
```

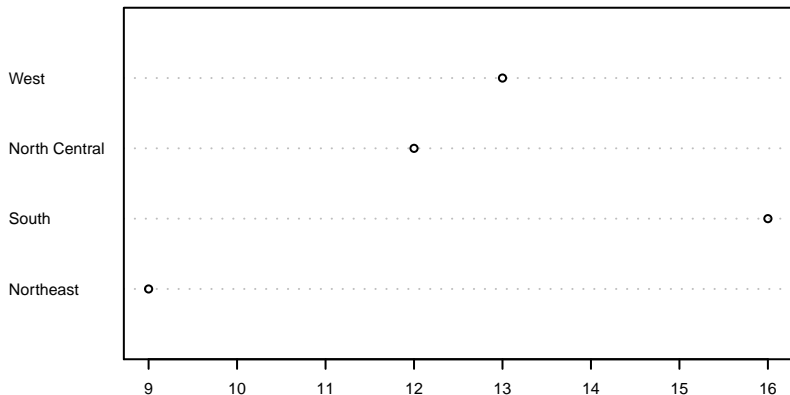



Figure: Dot chart of state.region

Other Data Types

- Logical

```
> x <- 5:9
```

```
> y <- (x < 7.3)
```

```
> y
```

```
[1] TRUE TRUE TRUE FALSE FALSE
```

```
> !y
```

```
[1] FALSE FALSE FALSE TRUE TRUE
```

- Missing

Other Data Types

- Missing: represented by NA

```
> x <- c(3, 7, NA, 4, 7)
```

```
> y <- c(5, NA, 1, 2, 2)
```

```
> x + y
```

```
[1] 8 NA NA 6 9
```

- Some functions have `na.rm` argument

```
> is.na(x)
```

```
[1] FALSE FALSE TRUE FALSE FALSE
```

```
> z <- x[!is.na(x)]
```

```
> sum(z)
```

```
[1] 21
```

Features of Data Distributions

Four Basic Features

- ① **Center:** middle or general tendency
- ② **Spread:** small means tightly clustered, large means highly variable
- ③ **Shape:** symmetry versus skewness, kurtosis
- ④ **Unusual Features:** anything else that pops out at you about the data

More about shape

Symmetry versus Skewness

- symmetric
- right (positive) and left (negative) skewness

Kurtosis

- leptokurtic - steep peak, heavy tails
- platykurtic - flatter, thin tails
- mesokurtic - right in the middle

Unusual features: clusters or gaps

```
> stem.leaf(faithful$eruptions)
```

```

1 | 2: represents 1.2
leaf unit: 0.1
      n: 272
12   s | 667777777777
51   1. | 8888888888888888888888888888889999999999
71   2* | 00000000000011111111
87   t | 2222222222333333
92   f | 44444
94   s | 66
97   2. | 889
98   3* | 0
102  t | 3333
108  f | 445555
118  s | 6666677777
(16) 3. | 8888888889999999
138  4* | 00000000000000001111111111111111
107  t | 222222222222333333333333333333
78   f | 4444444444444445555555555555555555555555
43   s | 6666666666667777777777777777
21   4. | 888888888888899999
4     5* | 0001

```

Unusual features: extreme observations

- Extreme observation: falls far from the rest of the data
 - possible sources
 - could be typo
 - could be in wrong study
 - could be indicative of something deeper
- Quantitatively measure features: **Descriptive Statistics**
 - qualitative data: frequencies or relative frequencies
 - quantitative data: measures of CUSS

Measures of center: sample mean \bar{x} (read "x-bar"):

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

- Good: natural, easy to compute, nice properties
- Bad: sensitive to extreme values

How to do it with R

```
> stack.loss      # built-in data  
[1] 42 37 37 28 18 18 19 20 15 14 14 13 11  
[14] 12  8  7  8  8  9 15 15  
  
> mean(stack.loss)  
[1] 17.52381
```


Measures of center: sample median \tilde{x}

How to find it

- 1 sort the data into an increasing sequence of n numbers
- 2 \tilde{x} lies in position $(n + 1)/2$
 - Good: resistant to extreme values, easy to describe
 - Bad: not as mathematically tractable, need to sort the data to calculate

How to do it with R

```
> median(stack.loss)
[1] 15
```

Measures of center: trimmed mean \bar{X}_t

How to find it

- 1 “trim” a proportion of data from both ends of the ordered list
 - 2 find the sample mean of what's left
- Good: also resistant to extreme values, has good properties, too
 - Bad: still need to sort data to get rid of outliers

How to do it with R

```
> mean(stack.loss, trim = 0.05)  
[1] 16.78947
```

Order statistics

Given data x_1, x_2, \dots, x_n , sort in an increasing sequence

$$x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \dots \leq x_{(n)} \quad (2)$$

- $x_{(k)}$ is the k^{th} order statistic
- approx $100(k/n)\%$ of the observations fall below $x_{(k)}$

How to do it with R

```
> sort(stack.loss)
```

```
[1] 7 8 8 8 9 11 12 13 14 14 15 15 15  
[14] 18 18 19 20 28 37 37 42
```

Sample quantile, order p ($0 \leq p \leq 1$), denoted \tilde{q}_p

We describe the default (type = 7)

- 1 get the order statistics $x_{(1)}, x_{(2)}, \dots, x_{(n)}$.
- 2 calculate $(n - 1)p + 1$, write in form $k.d$, with k an integer and d a decimal

3

$$\tilde{q}_p = x_{(k)} + d(x_{(k+1)} - x_{(k)}). \quad (3)$$

- approximately $100p\%$ of the data fall below the value \tilde{q}_p .

How to do it with R

```
> quantile(stack.loss, probs = c(0, 0.25, 0.37))  
0%  25%  37%  
7.0 11.0 13.4
```

Measures of spread: sample variance, std. deviation

The *sample variance* s^2

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (4)$$

The *sample standard deviation* is $s = \sqrt{s^2}$.

- Good: tractable, nice mathematical/statistical properties
- Bad: sensitive to extreme values

How to do it with R

```
> var(stack.loss); sd(stack.loss)
```

```
[1] 103.4619
```

```
[1] 10.17162
```

Interpretation of s

Chebychev's Rule:

The proportion of observations within k standard deviations of the mean is at least $1 - 1/k^2$, *i.e.*, at least 75%, 89%, and 94% of the data are within 2, 3, and 4 standard deviations of the mean, respectively.

Empirical Rule:

If data follow a bell-shaped curve, then approximately 68%, 95%, and 99.7% of the data are within 1, 2, and 3 standard deviations of the mean, respectively.

Measures of spread: interquartile range

The *Interquartile range IQR*

$$IQR = \tilde{q}_{0.75} - \tilde{q}_{0.25} \quad (5)$$

- Good: resistant to outliers
- Bad: only considers middle 50% of the data

How to do it with R

```
> IQR(stack.loss)
```

```
[1] 8
```

Measures of spread: median absolute deviation

The *median absolute deviation* *MAD*:

- 1 get the order statistics, find the median \tilde{x} .
- 2 calculate the *absolute deviations*:

$$|x_1 - \tilde{x}|, |x_2 - \tilde{x}|, \dots, |x_n - \tilde{x}|$$

- 3 the *MAD* \propto median $\{|x_1 - \tilde{x}|, |x_2 - \tilde{x}|, \dots, |x_n - \tilde{x}|\}$
 - Good: excellently robust
 - Bad: not as popular, not as intuitive

How to do it with R

```
> mad(stack.loss)
```

```
[1] 5.9304
```


Measures of spread: the range

The *range* R :

$$R = x_{(n)} - x_{(1)} \quad (6)$$

- Good (not so much): easy to describe and calculate
- Bad: ignores everything but the most extreme observations

How to do it with R

```
> range(stack.loss)
[1] 7 42
> diff(range(stack.loss))
[1] 35
```

Measures of shape: sample skewness

The *sample skewness* g_1 :

$$g_1 = \frac{1}{n} \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{s^3}. \quad (7)$$

Things to notice:

- invariant w.r.t. location and scale
- $-\infty < g_1 < \infty$
- sign of g_1 indicates direction of skewness (\pm)

How to do it with R

```
> library(e1071)
> skewness(stack.loss)
[1] 1.156401
```

Measures of shape: sample skewness

How big is BIG?

4.34 versus 0.434?? (8)

Rule of thumb:

If $|g_1| > 2\sqrt{6/n}$, then the data distribution is substantially skewed (in the direction of the sign of g_1).

```
> skewness(discoveries)
```

```
[1] 1.207600
```

```
> 2 * sqrt(6/length(discoveries))
```

```
[1] 0.4898979
```

Measures of shape: sample excess kurtosis

The *sample excess kurtosis* g_2 :

$$g_2 = \frac{1}{n} \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{s^4} - 3. \quad (9)$$

Things to note:

- invariant w.r.t. location and scale
- $-2 \leq g_2 < \infty$
- $g_2 > 0$ indicates leptokurtosis, $g_2 < 0$ indicates platykurtosis

How to do it with R

```
> library(e1071)
> kurtosis(stack.loss)
```

```
[1] 0.1343524
```

Measures of shape: sample excess kurtosis

Again, how big is BIG?

Rule of thumb:

If $|g_2| > 4\sqrt{6/n}$, then the data distribution is substantially kurtic.

```
> kurtosis(UKDriverDeaths)
```

```
[1] 0.07133848
```

```
> 4 * sqrt(6/length(UKDriverDeaths))
```

```
[1] 0.7071068
```

Exploratory data analysis: more on stemplots

- Trim Outliers: observations that fall far from the bulk of the other data often obscure structure to the data and are best left out. Use the `trim.outliers` argument to `stem.leaf`.
- Split Stems: we sometimes fix “skyscraper” stemplots by increasing the number of lines available for a given stem. The end result is a more spread out stemplot which often looks better. Use the `m` argument to `stem.leaf`
- Depths: give insight into balance of the data around the median. Frequencies are accumulated from the outside inward, including outliers. Use `depths = TRUE`.

More about stemplots

```
> stem.leaf(faithful$eruptions)
```

1 | 2: represents 1.2

```
leaf unit: 0.1
```

n: 272

12 s | 667777777777

51 1. | 88888888888888888888888888889999999999

```
71      2* | 000000000000011111111
```

87 t | 2222222222333333

92 f | 44444

94 s | 66

97 2. | 889

98 3* | 0

102 t | 3333

108 f | 445555

118 s | 6666677777

(16) 3. | 888888888899999999

```
138      4* | 0000000000000000001111111111111111
```

```
107      t | 222222222223333333333333333333
```

78 f | 44444444444444455555555555555555

43 s | 6666666666667777777777

```
21      4. | 88888888888899999
```

4 5* | 0001

Hinges and the 5NS

- Find the order statistics $x_{(1)}, x_{(2)}, \dots, x_{(n)}$.
- The *lower hinge* h_L is in position $L = \lfloor (n+3)/2 \rfloor / 2$
- The *upper hinge* h_U is in position $n+1-L$.

Given the hinges, the *five number summary* (5NS) is

$$5NS = (x_{(1)}, h_L, \tilde{x}, h_U, x_{(n)}). \quad (10)$$

How to do it with R

```
> fivenum(stack.loss)
```

```
[1] 7 11 15 19 42
```


Boxplots

Boxplot: a visual display of the 5NS. Can visually assess multiple features of the data set:

- *Center:* estimated by the sample median, \tilde{x}
- *Spread:* judged by the width of the box, $h_U - h_L$
- *Shape:* indicated by the relative lengths of the whiskers, position of the median inside box.
- *Extreme observations:* identified by open circles

How to do it with R

```
> boxplot(rivers, horizontal = TRUE)
```

Outliers

- *potential*: falls beyond 1.5 times the width of the box

less than $h_L - 1.5(h_U - h_L)$ or greater than $h_U + 1.5(h_U - h_L)$

- *suspected*: falls beyond 3 times the width of the box

less than $h_L - 3(h_U - h_L)$ or greater than $h_U + 3(h_U - h_L)$

How to do it with R

```
> boxplot.stats(rivers)$out
```

```
[1] 1459 1450 1243 2348 3710 2315 2533 1306
```

```
[9] 1270 1885 1770
```

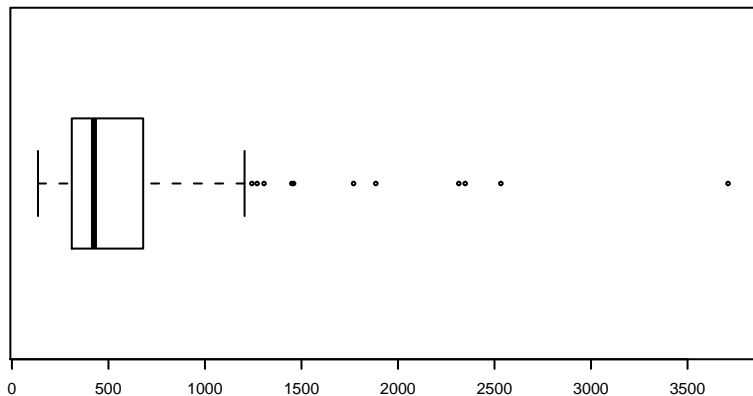


Figure: Boxplot of rivers

Standardizing variables

- useful to see how observation relates to other observations
- AKA measure of relative standing, z-score

$$z_i = \frac{x_i - \bar{x}}{s}, \quad i = 1, 2, \dots, n$$

- unitless
- positive (negative) z-score falls above (below) mean

How to do it with R

```
> scale(precip)[1:3]  
[1] 2.342971 1.445597 -2.034466
```

Multivariate data: data frames

- usually have two (or more) measurements associated with each subject
- display in rectangular array
 - each row corresponds to a subject
 - columns contain the measurements for each variable

How to do it with R

```
> x <- 5:6; y <- letters[3:4]; z <- c(0.1, 3.8)
> data.frame(v1 = x, v2 = y, v3 = z)
```

	v1	v2	v3
1	5	c	0.1
2	6	d	3.8

More on data frames

- must have same number of rows in each column
- all measurements in single column must be same type
- indexing is two-dimensional; the columns have names

How to do it with R

```
> A <- data.frame(v1 = x, v2 = y, v3 = z)
```

```
> A[2, 1]; A[1,]; A[, 3]
```

```
[1] 6
```

```
  v1 v2  v3
```

```
1  5  c 0.1
```

```
[1] 0.1 3.8
```

Bivariate data: qualitative versus qualitative

Two categorical variables

- usually make a two-way contingency table
- in the R Commander with Statistics ▸ Contingency Tables
▸ Two-way Tables

How to do it with R

```
> library(RcmdrPlugin.IPSUR)
> data(RcmdrTestDrive)
> xtabs(~ gender + smoking, data = RcmdrTestDrive)
```

```
      smoking
gender  Nonsmoker  Smoker
Female          61      9
Male           75     23
```

Bivariate data: more on tables

- Descriptive statistics: for now, marginal totals/percentages
- more to talk about later: odds ratio, relative risk

How to do it with R

```
> A <- xtabs(Freq ~ Survived + Class, data = Titanic)
> addmargins(A)
```

	Class				
Survived	1st	2nd	3rd	Crew	Sum
No	122	167	528	673	1490
Yes	203	118	178	212	711
Sum	325	285	706	885	2201

Bivariate data: more on tables

```
> library(abind)
```

```
> colPercents(A)
```

Class

Survived	1st	2nd	3rd	Crew
No	37.5	58.6	74.8	76
Yes	62.5	41.4	25.2	24
Total	100.0	100.0	100.0	100
Count	325.0	285.0	706.0	885

```
> rowPercents(A)
```

Class

Survived	1st	2nd	3rd	Crew	Total	Count
No	8.2	11.2	35.4	45.2	100	1490
Yes	28.6	16.6	25.0	29.8	100	711

Plotting two categorical variables

- Stacked bar charts
- Side-by-side bar charts
- Spine plots

How to do it with R

```
> barplot(A, legend.text = TRUE)
> barplot(A, legend.text = TRUE, beside = TRUE)
> spineplot(A)
```

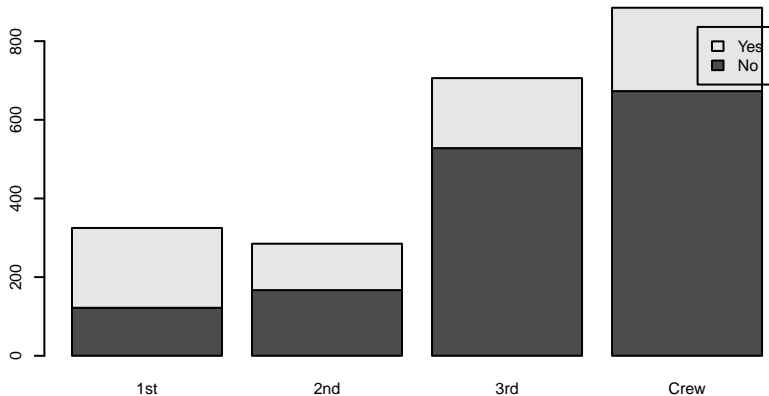


Figure: Stacked bar chart of Titanic data

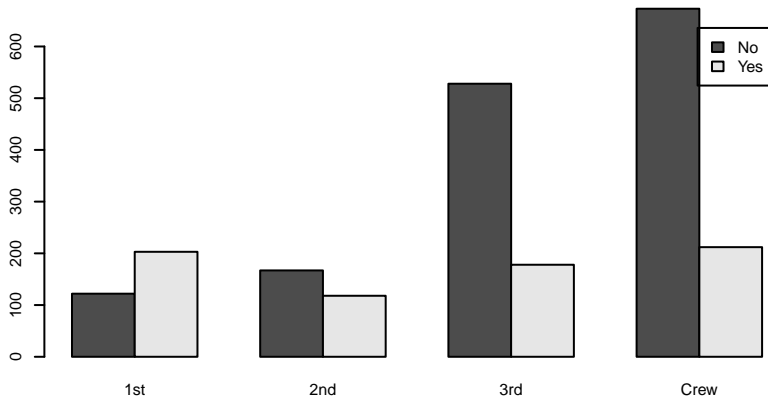


Figure: Side-by-side bar chart of Titanic data

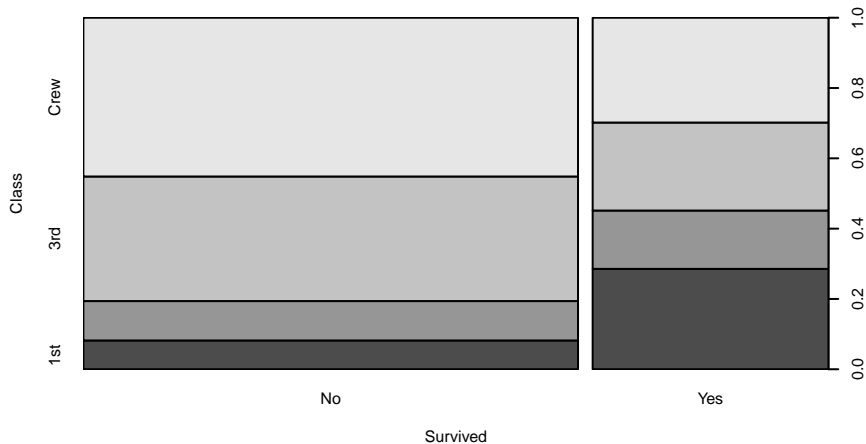


Figure: Spine plot of Titanic data

Bivariate data: quantitative versus quantitative

- Can do univariate graphs of both variables separately
- Make scatterplots for both variables simultaneously

How to do it with R

```
> plot(conc ~ rate, data = Puromycin)
> library(lattice)
> xyplot(conc ~ rate, data = Puromycin)
```

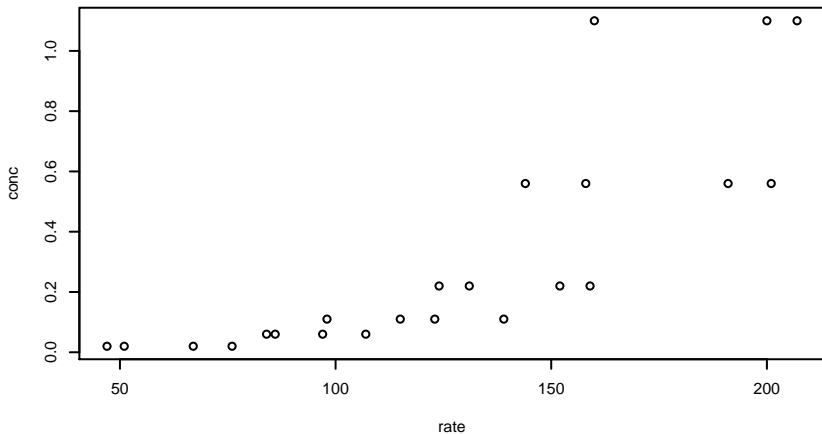


Figure: Scatterplot of Puromycin data

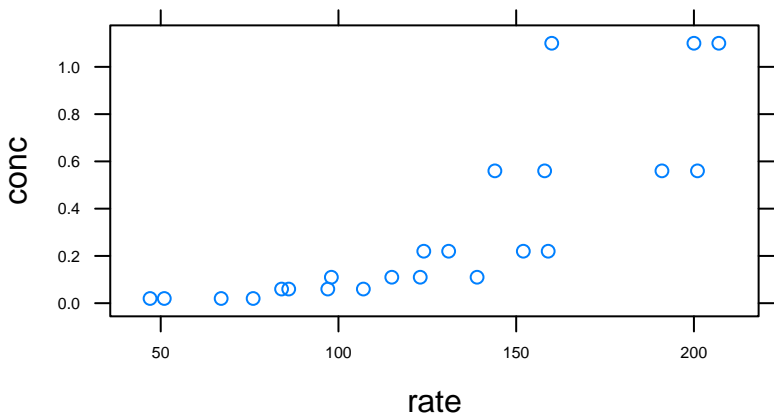


Figure: Scatterplot of Puromycin data

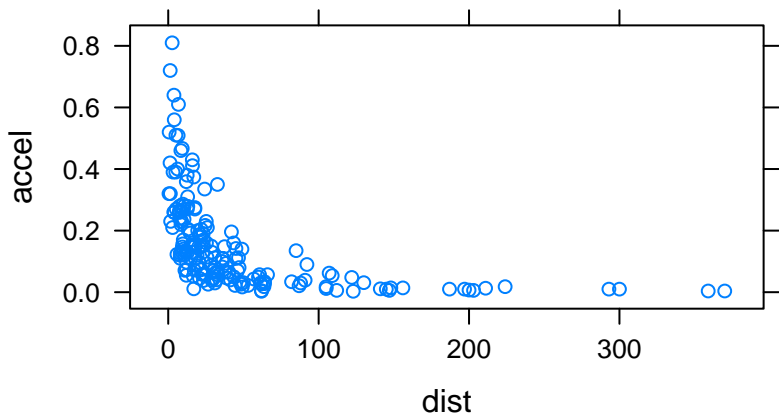


Figure: Scatterplot of attenu data

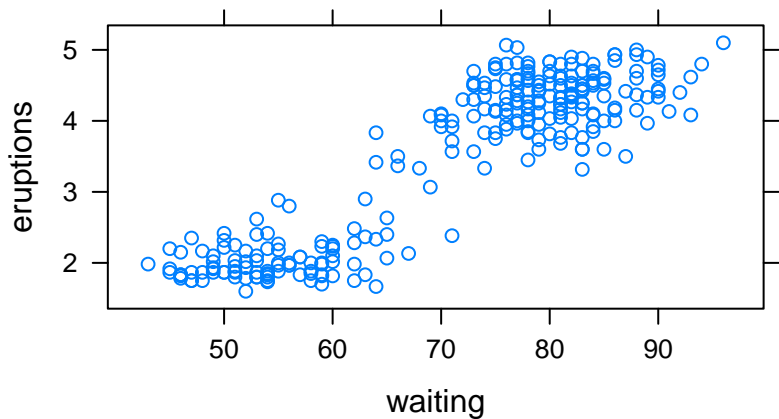


Figure: Scatterplot of faithful data

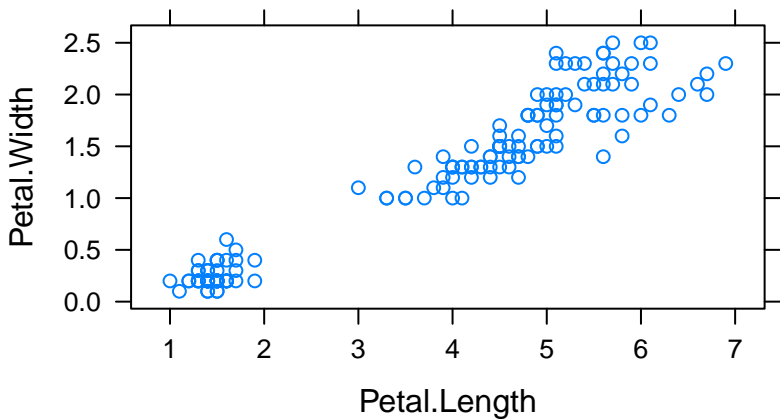


Figure: Scatterplot of iris data

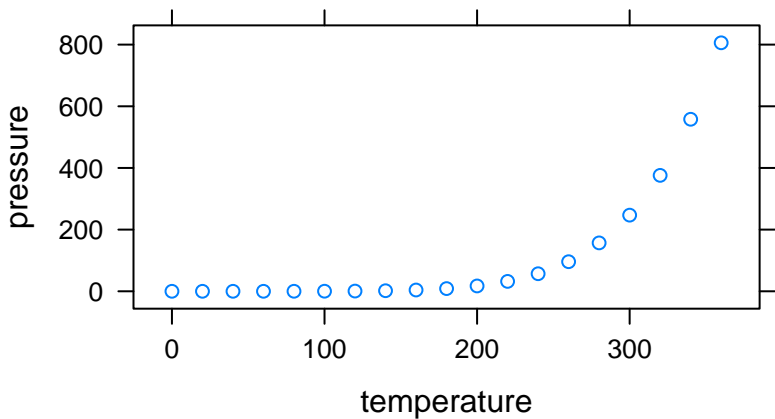


Figure: Scatterplot of iris data

Measuring Linear association

The **sample Pearson product-moment correlation coefficient**:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

- independent of scale
- $-1 \leq r \leq 1$, equality when points lie on straight line

How to do it with R

```
> with(iris, cor(Petal.Width, Petal.Length))
```

```
[1] 0.9628654
```

```
> with(attenu, cor(dist, accel))
```

```
[1] -0.4713809
```

More about linear correlation

- measures *strength* and *direction* of linear association
- Rules of thumb:
 - $0 < |r| < 0.3$, weak linear association
 - $0.3 < |r| < 0.7$, moderate linear association
 - $0.7 < |r| < 1$, strong linear association
- Just because $r \approx 0$ doesn't mean there isn't any association

One quantitative, one categorical

- Break down quantitative var by groups of subjects
 - compare centers and spreads: variation within versus between groups
 - compare clusters and gaps
 - compare outliers and unusual features
 - compare shapes.
- graphical and numerical

Comparison of groups

How to do it with R

```
> stripchart(weight ~ feed, method = "stack",  
+           data = chickwts)  
> library(lattice)  
> histogram(~age | education, data = infert)  
> bwplot(~count | spray, data = InsectSprays)
```

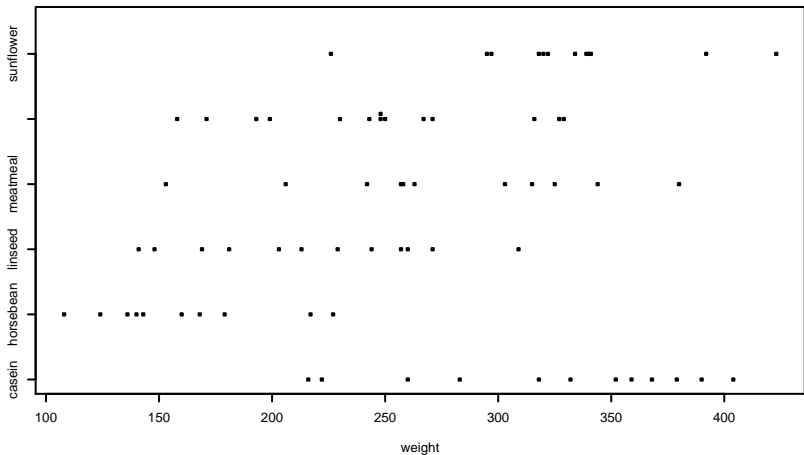



Figure: Stripcharts of chickwts data

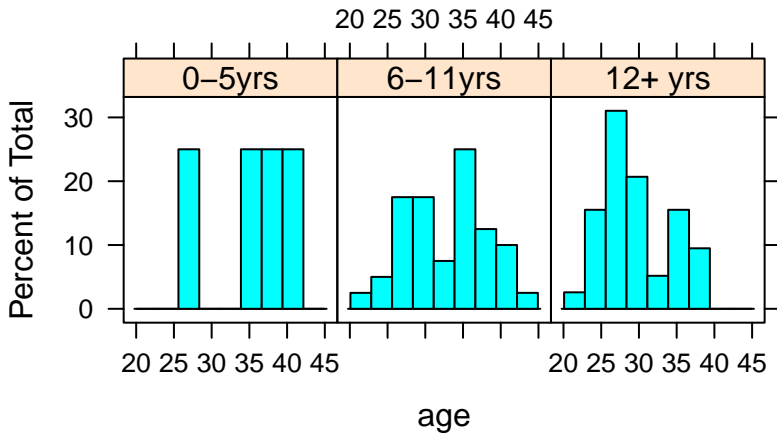


Figure: Histograms of infert data

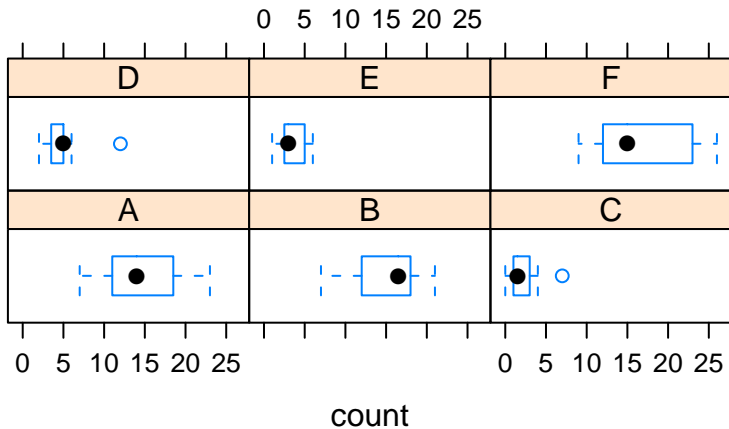


Figure: Boxplots of InsectSprays data

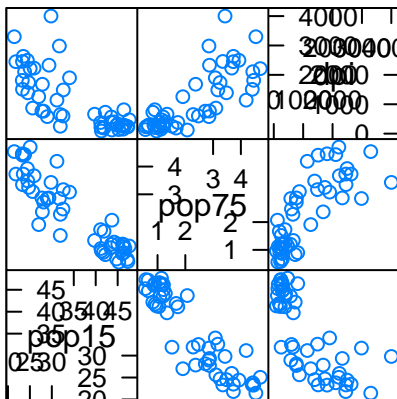
Multiple variables

With more variables, complexity increases

- multi-way contingency tables (bunch of categorical vars)
 - mosaic plots, dotcharts
- sample variance-covariance matrices
 - scatterplot matrices
- comparing groups: coplots

How to do it with R

```
> splom(~cbind(Murder, Assault, Rape),  
+       data = USArrests)  
> `?`(dotchart)  
> `?`(xyplot)  
> `?`(mosaicplot)
```



Scatter Plot Matrix

Figure: Scatterplot matrix of LifeCycleSavings data

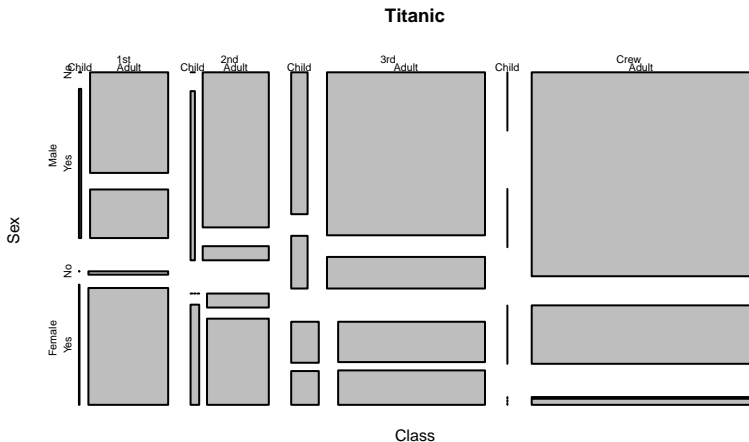


Figure: Mosaic plot of Titanic data

Given : depth

