MATH 4720 / MSSC 5720

Instructor: Mehdi Maadooliat

Chapter 3

Department of Mathematical and Statistical Sciences

TOPIC 2 - CHAPTER 3

- Type of Variables
- Frequency distributions
- Histograms
- Mean, median, variance and standard deviation
- Quartiles, interquartile range
- Boxplots
- Correlation

ALL ABOUT VARIABLES

- **Variable:** Any characteristic or quantity to be measured on units in a study
- **Categorical variable:** Places a unit into one of several categories
  - Examples: Gender, race, political party
- **Quantitative variable:** Takes on numerical values for which arithmetic makes sense
  - Examples: SAT score, number of siblings, cost of textbooks
- **Univariate** data has one variable.
- **Bivariate** data has two variables.
- **Multivariate** data has three or more variables.

TYPES OF VARIABLES

Examples:

Variable	Numeric: Discrete	Numeric: Continuous	Categorical
Length		X	
Hours Enrolled	X		
Major			X
Zip Code			X

AUSTRALIAN INSTITUTE OF SPORT DATA

- **Description**
  - Data on 102 male and 100 female athletes collected at the Australian Institute of Sport, courtesy of Richard Telford and Ross Cunningham.
- **Source**
  - Cook and Weisberg (1994), *An Introduction to Regression Graphics*. John Wiley & Sons, New York.

AIS.mjp

Variable	Description
sex	sex
sport	sport
rcc	red cell count
wcc	white cell count
Hc	Hematocrit
Hg	Hemoglobin
Fe	plasma ferritin concentration
bmi	body mass index, weight/(height)
ssf	sum of skin folds
Bfat	body fat percentage
lbm	lean body mass
Ht	height (cm)
Wt	weight (Kg)

SUMMARIZING A SINGLE CATEGORICAL VARIABLE

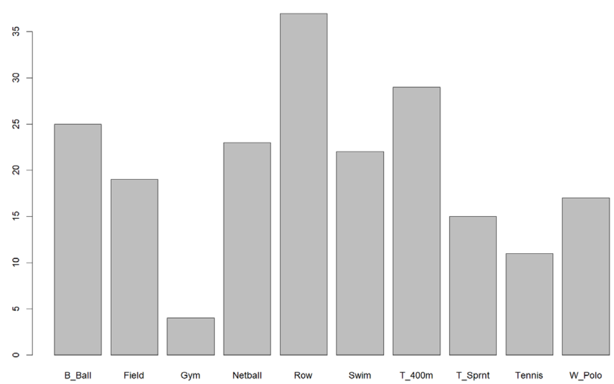
MARQUETTE UNIVERSITY
Be The Difference.

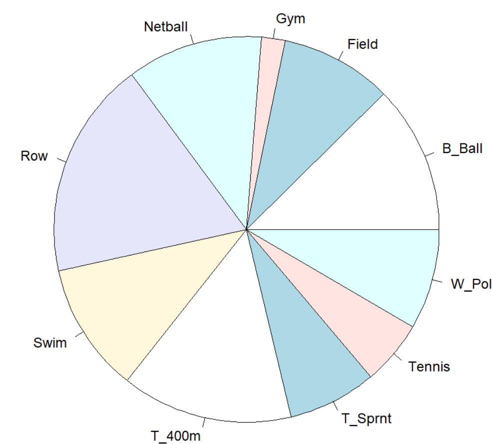
- **Frequency (Count) - number of times the value occurs in the data**
- **Relative frequency (Percent) - proportion of the data with the value**
- **Cumulative Frequency**
- **Cumulative Relative Frequency**
- **ais.csv (D2L/Content/Datasets)**

Tally for Discrete Variables: sport

sport	Count	Percent	CumCnt	CumPct
B_Ball	25	12.38	25	12.38
Field	19	9.41	44	21.78
Gym	4	1.98	48	23.76
Netball	23	11.39	71	35.15
Row	37	18.32	108	53.47
Swim	22	10.89	130	64.36
T_400m	29	14.36	159	78.71
T_Sprnt	15	7.43	174	86.14
Tennis	11	5.45	185	91.58
W_Polo	17	8.42	202	100.00
N=	202			

- **R Code:**
  - library("sn")
  - data("ais")
  - tbl <- table(ais$sport)
  - cumsum(tbl)
  - prop.table(tbl)
  - cumsum(prop.table(tbl))
  - barplot(tbl)
  - pie(tbl)





ANALYZING A SINGLE QUANTITATIVE VARIABLE

- Consider the AIS data which contains 202 athletes.

- What is a **typical** height of athletes?
- How much **spread** is there in their Body fats?

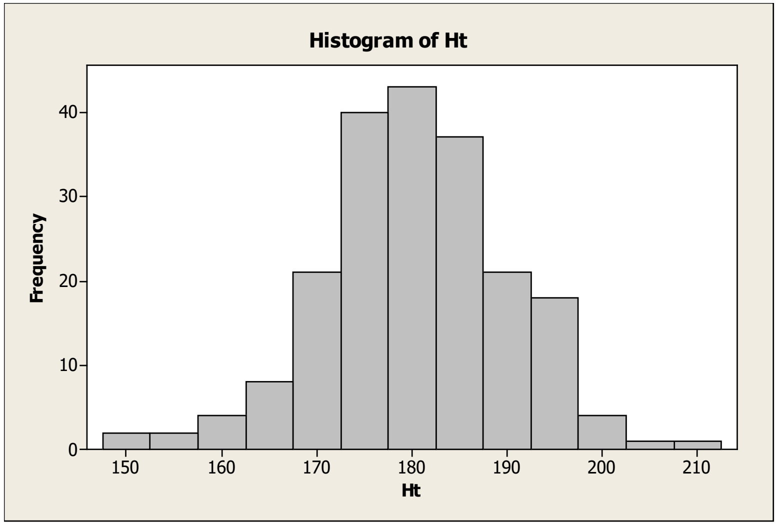
- **Typical** is generally characterized by the **center** of the data

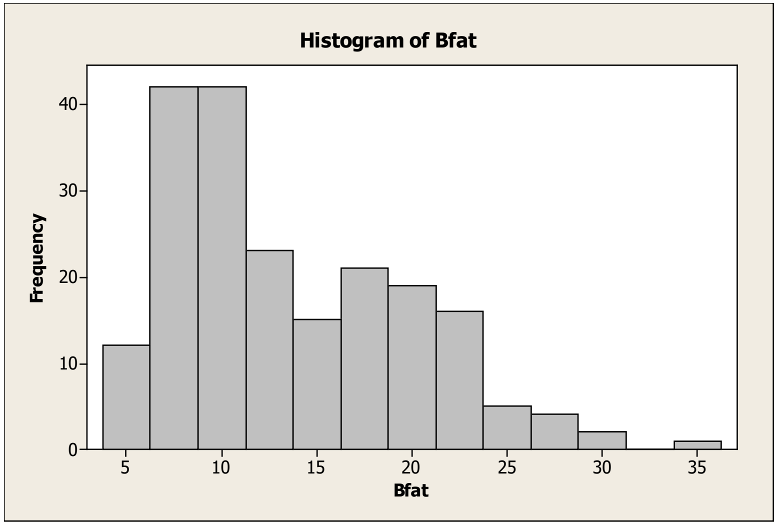
- **Spread** is generally reported as an interval containing most of the data

HISTOGRAMS

- **Histogram** - bar graph of binned data where the height of the bar above each bin denotes the frequency (relative frequency) of values in the bin
- Typical concentration?
- Spread?
- Roughly how many athletes are shorter than 180 cm?

- R Code:
  - hist(ais$Ht)
  - hist(ais$Bfat, breaks = 16)



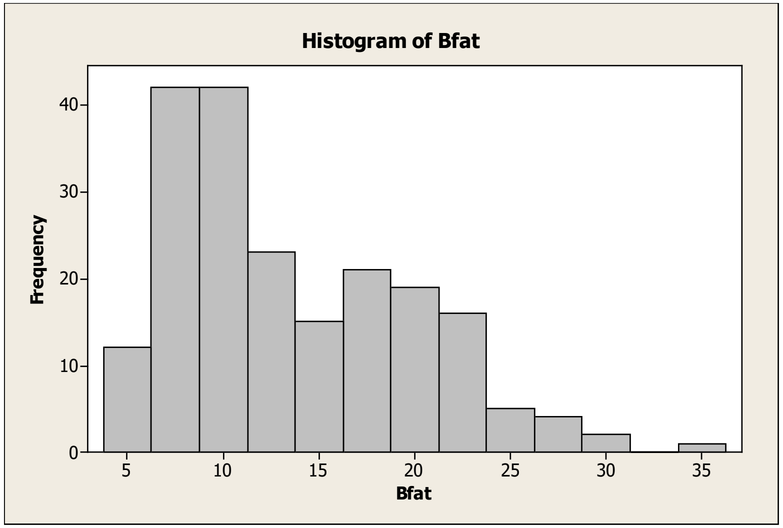


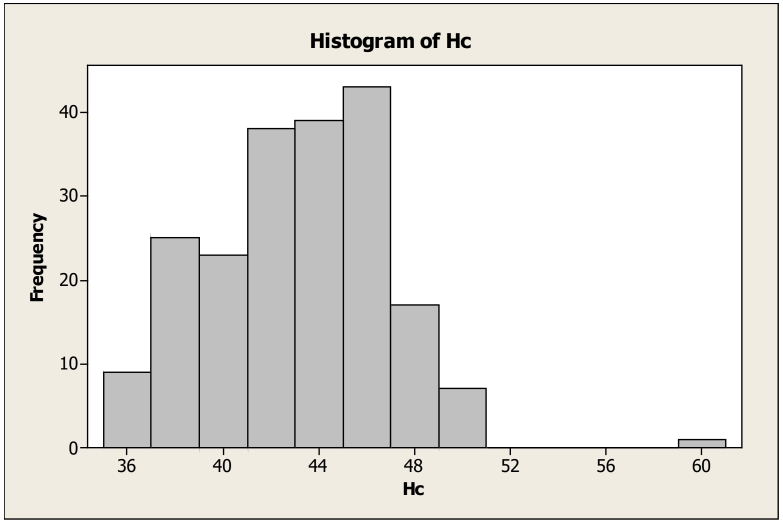
DESCRIBING THE SHAPE OF QUANTITATIVE DATA

- **Symmetric** data has roughly the same mirror image on each side of a center value.
- **Skewed** data has one side (either **right** or **left**) which is much longer than the other relative to the **mode** (peak value).
- The above definitions are most useful when describing data with a single mode.
- **Multimodal** data has more than one mode.
- Beware of **outliers** when describing shape.
- Shape of the AIS Data?

DESCRIBING THE SHAPE (CONT…)

- **Bfat: Body Fat**
  - skewed to the right
  - Bimodal
- **Hematocrit (Hc): Volume percentage (%) of red blood cells in blood**
  - Outlier
- **R Code:**
  - hist(ais$Ht)
  - hist(ais$Bfat, breaks = 16)





STEM AND LEAF PLOTS

- **Separate each value into a *stem* (all but the rightmost digit) and a *leaf* (the rightmost digit)**
- **Write unique sorted stems in a vertical column**
- **Add each leaf to the right of its stem in increasing order**
- **Example from AIS:**
- **rcc (red cell counts)**
  - **Female-Row:**
    - **4.26 4.63 4.36 3.91 4.51 4.37 4.90 4.46 3.95 4.46 5.02 4.26 4.46 4.16 4.49 4.21 4.57 4.87 4.44 4.45 4.41 4.87**
  - **Male-Row:**
    - **4.87 5.04 4.40 4.95 4.78 5.21 5.22 5.18 5.40 4.92 5.24 5.09 4.83 5.22 4.71**
- **R Code:**
  - stem(ais[ais$sport == "Row",]$RCC, scale = 2)

```
Stem-and-leaf of rcc   N  = 37
Leaf Unit = 0.010


 2   39  15
 2   40
 3   41  6
 6   42  166
 8   43  67
 16  44  01456669
 18  45  17
(1)  46  3
 18  47  18
 16  48  3777
 12  49  025
 9   50  249
 6   51  8
 5   52  1224
 1   53
 1   54  0
```

HISTOGRAMS VS. STEM AND LEAF PLOTS

- **Stem and leaf plots (typically) display actual data values whereas histograms do not**
- **Stem and leaf plots are more useful for small data sets (less than 100 values)**
- **Histograms can be constructed for larger data sets**

SUMMARY STATISTICS FOR QUANTITATIVE DATA

- **Measures of center (typical)**
  - The **sample median** is the middle observation if the values are arranged in increasing order.
  - The **sample mean** of *n* observations is the average, the sum of the values divided by *n*.

$$X_1, \dots, X_n \text{ represents } n \text{ data values}$$

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

SUMMARY STATISTICS FOR QUANTITATIVE DATA

- **Measures of spread:**
  - **Interquartile range**, IQR = Q3-Q1, the range of the middle 50% of the data
    - **first quartile (Q1)** is the 25th percentile
    - **third quartile (Q3)** is the 75th percentile
  - **sample variance**, s^2, is the sum of squared deviations from the sample mean divided by n-1

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

  - **sample standard deviation**, s, is the square root of sample variance. Preferred because it has the same units as the data.

EXAMPLE ON HOW TO CALCULATE THE VARIANCE

EXAMPLE 3.9

The time between an electric light stimulus and a bar press to avoid a shock was noted for each of five conditioned rats. Use the given data to compute the sample variance and standard deviation.

Shock avoidance times (seconds): 5, 4, 3, 1, 3

Solution The deviations and the squared deviations are shown in Table 3.11. The sample mean $\bar{y}$ is 3.2.

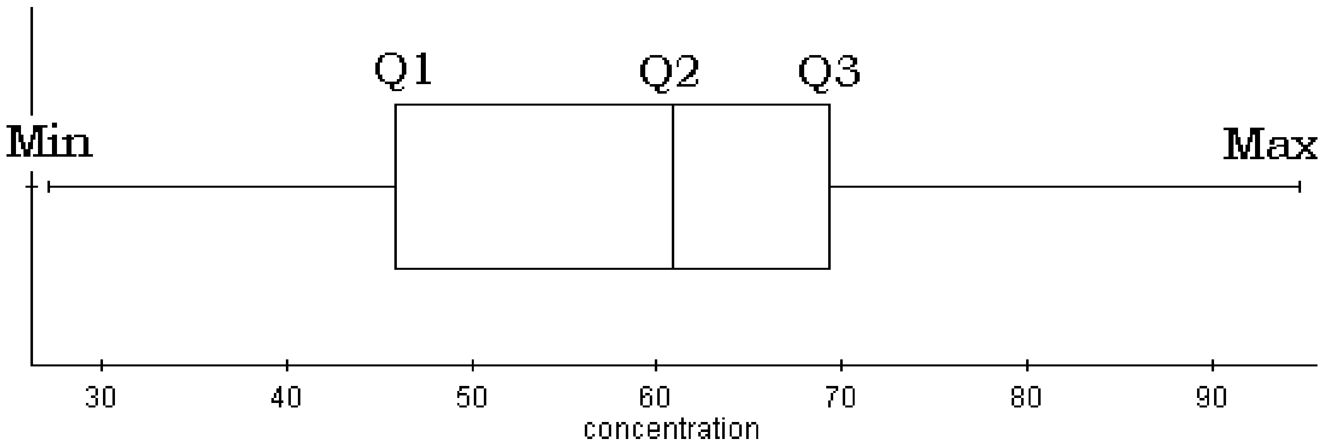
	y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
	5	1.8	3.24
	4	.8	.64
	3	−.2	.04
	1	−2.2	4.84
	3	−.2	.04
Totals	16	0	8.80

Using the total of the squared deviations column, we find the sample variance to be

$$s^2 = \frac{\Sigma_i (y_i - \bar{y})^2}{n - 1} = \frac{8.80}{4} = 2.2$$

SUMMARY STATISTICS FOR QUANTITATIVE DATA

- ***p*th percentile** -the value such that $p \times 100\%$ of values are below it and $(1-p) \times 100\%$ are above it
  - **first quartile (Q1)** is the 25th percentile
  - **second quartile (Q2)** 50th percentile (median)
  - **third quartile (Q3)** is the 75th percentile
- **5-number summary**: Min, Q1, Q2, Q3, Max
  - **Boxplots**: Stacking boxplots can be very useful for comparing multiple groups



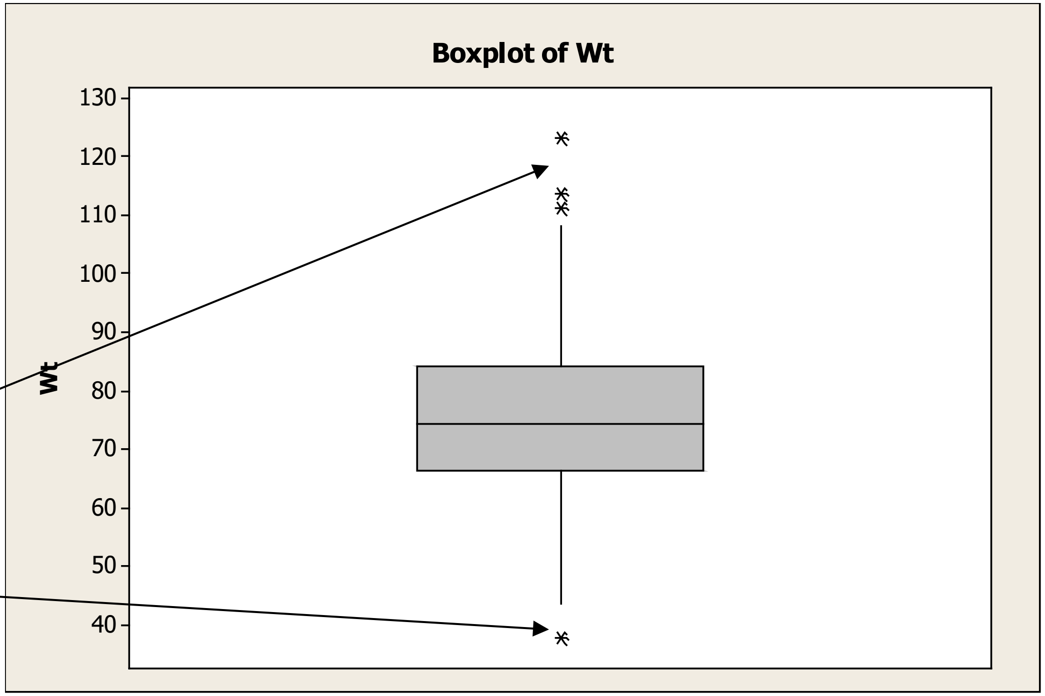
MARQUETTE UNIVERSITY
Be The Difference.

Minimum, Q_1, Median, Q_3, and Maximum of AIS-weight

- **R Code:**
  - boxplot(ais$Wt)

- **These five numbers are called the**
  Five Number Summary

- **What are these points?**
  Outliers



- **Interquartile Range (IQR):**
  Distance between the first quartile (Q_1) and the third quartile (Q_3). IQR = $Q_3 - Q_1$

BOX PLOT (CONT…)

- Outliers: observations that are unusually far from the bulk of the data.

- What are some possible explanations for outliers?
  - The data point was recorded wrong.
  - The data point wasn't actually a member of the population we were trying to sample.
  - We just happened to get an extreme value in our sample.

- The 1.5 x IQR Criterion for Outliers: Designate an observation a suspected outlier if it falls more than 1.5 x IQR below the first quartile or above the third quartile.

1.5*IQR CRITERION EXAMPLE

- Suppose you had the following data set:

  -2, 15, 3, 7, 10, 21, 1, 5, 12, 8, 1, 35, 10

List data from smallest to largest:

Find Q_1, Median, Q_3, Min, and Max:

$IQR = Q_3 - Q_1 =$ ________

$1.5 * IQR =$ __________

$Q_1 - 1.5 * IQR =$ ___________ If less than this number, then outlier

$Q_3 + 1.5 * IQR =$ ___________ If more than this number, then outlier

Are there any outliers in this data set?

1.5*IQR CRITERION EXAMPLE

- Suppose you had the following data set:

  -2, 15, 3, 7, 10, 21, 1, 5, 12, 8, 1, 35, 10

List data from smallest to largest:

-2, 1, 1, 3, 5, 7, 8, 10, 10, 12, 15, 21, 35

Find Q_1, Median, and Q_3:

$Q_1 = (1+3)/2 = 2$ Median = 8 $Q_3 = (12 + 15)/2 = 13.5$

IQR = $Q_3 - Q_1 = 11.5$

1.5*IQR = 17.25

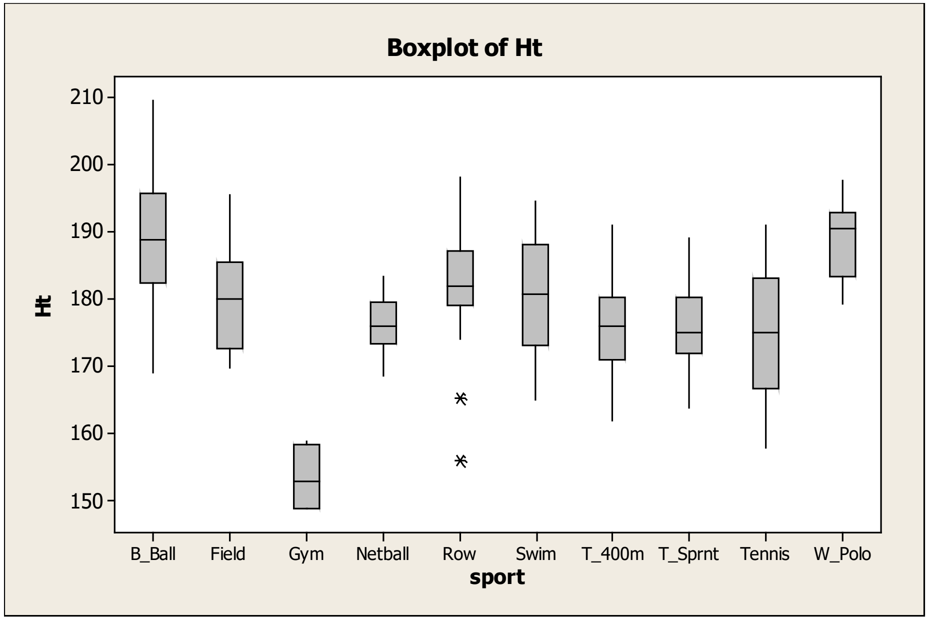
$Q_1 - 1.5*IQR = -15.25$ If less than this number, then outlier

$Q_3 + 1.5*IQR = 30.75$ If more than this number, then outlier

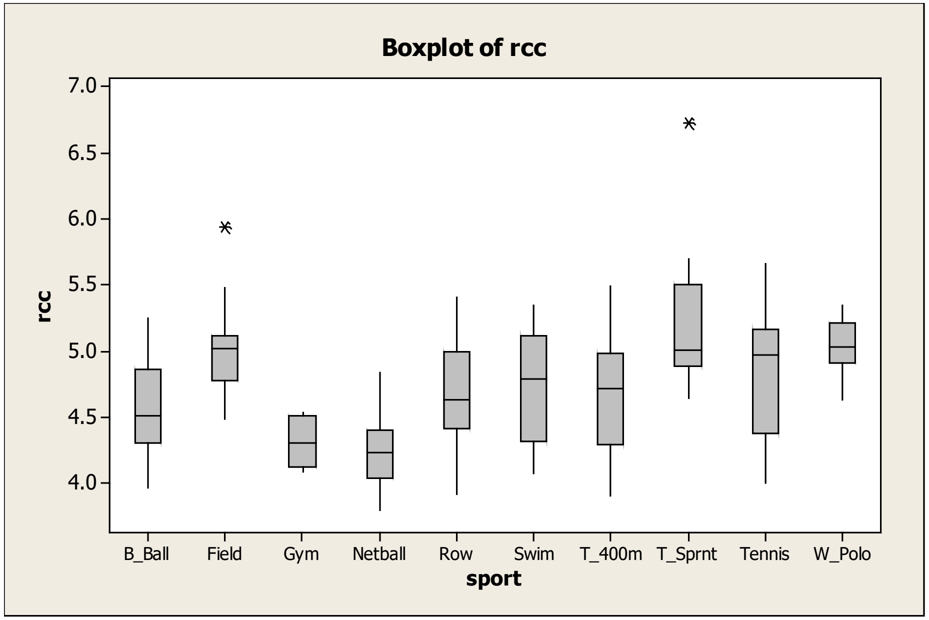
Are there any outliers in this data set? Yes, 35

SIDE-BY-SIDE BOX PLOT

- **R Code:**
  - boxplot(Ht ~ sport, data=ais)

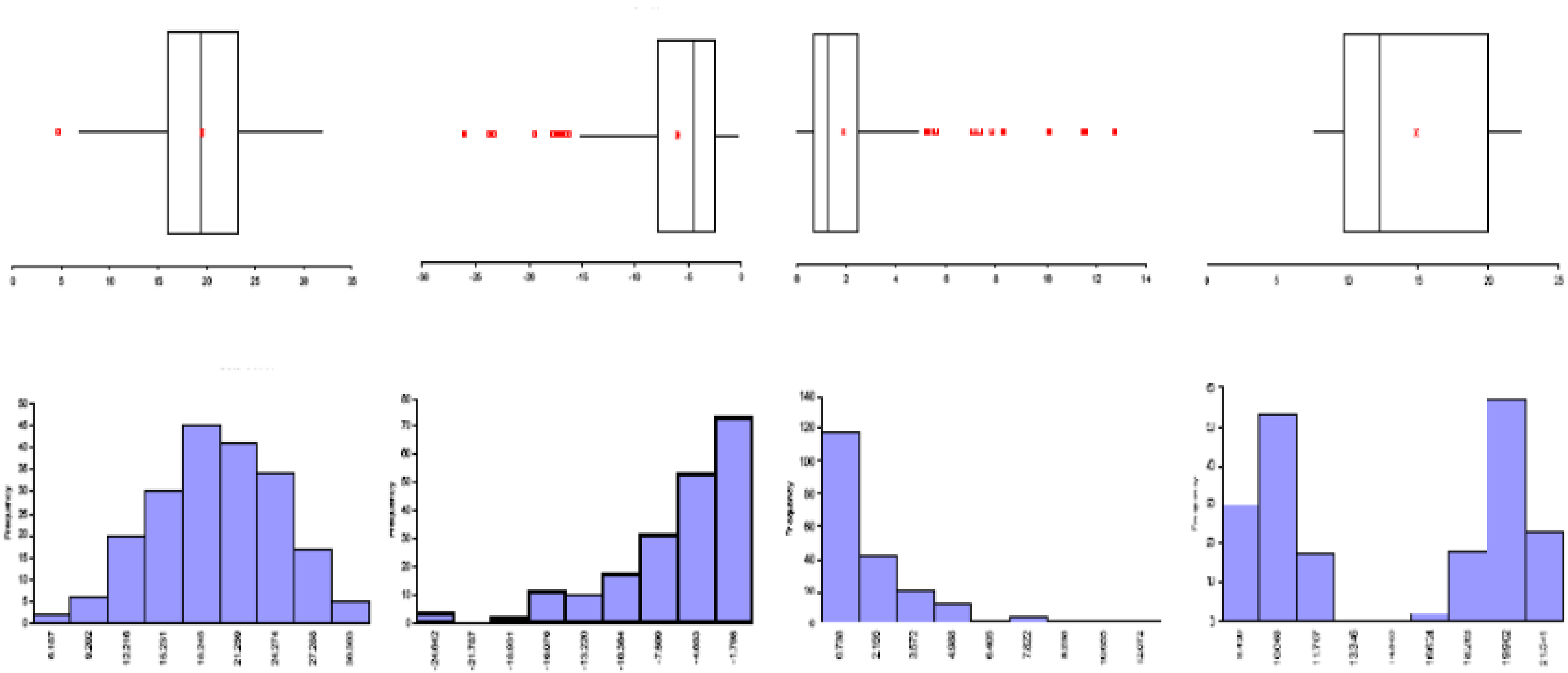


  - boxplot(RCC ~ sport, data=ais)



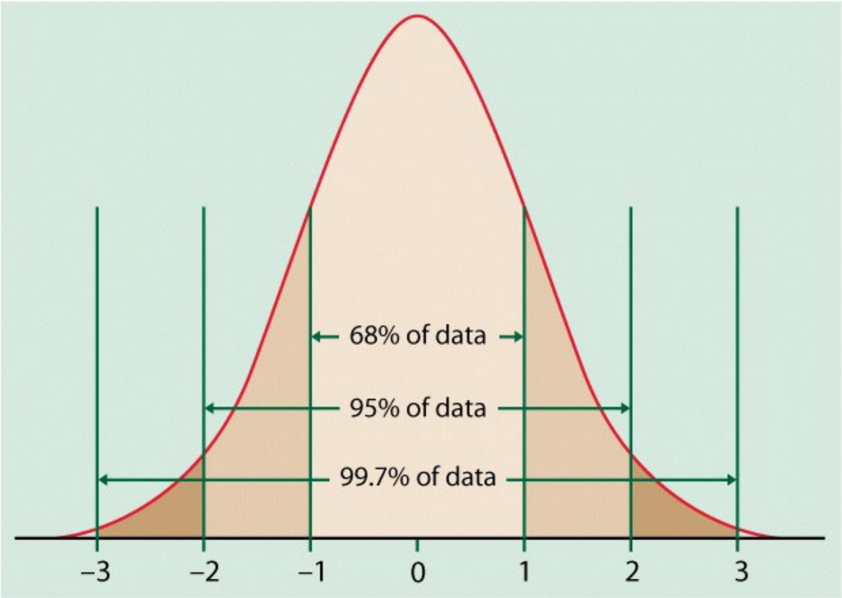
COMPARING HISTOGRAMS AND CORRESPONDING BOXPLOTS

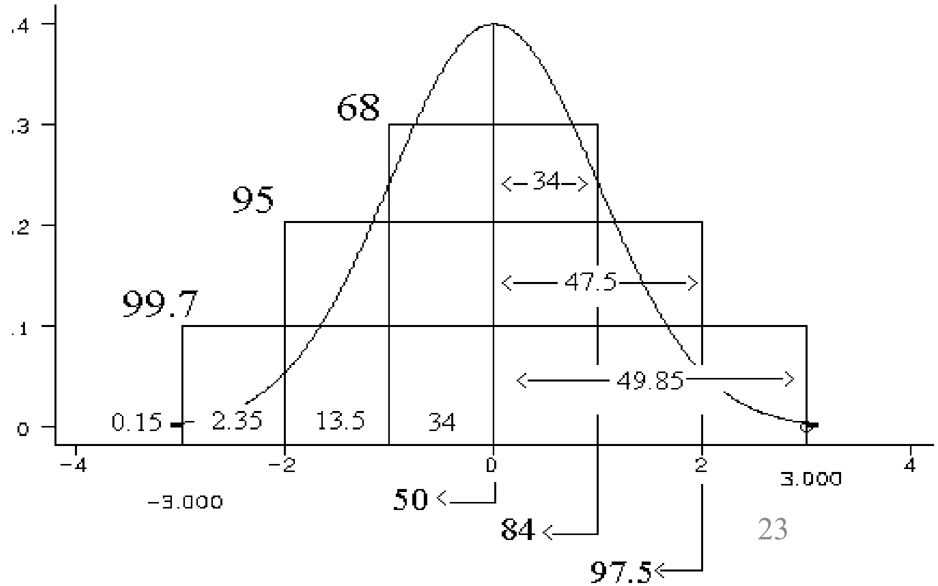
Symmetric | Left Skewed | Right Skewed | Bimodal



EMPIRICAL RULE (THE 68-95-99.7 RULE)

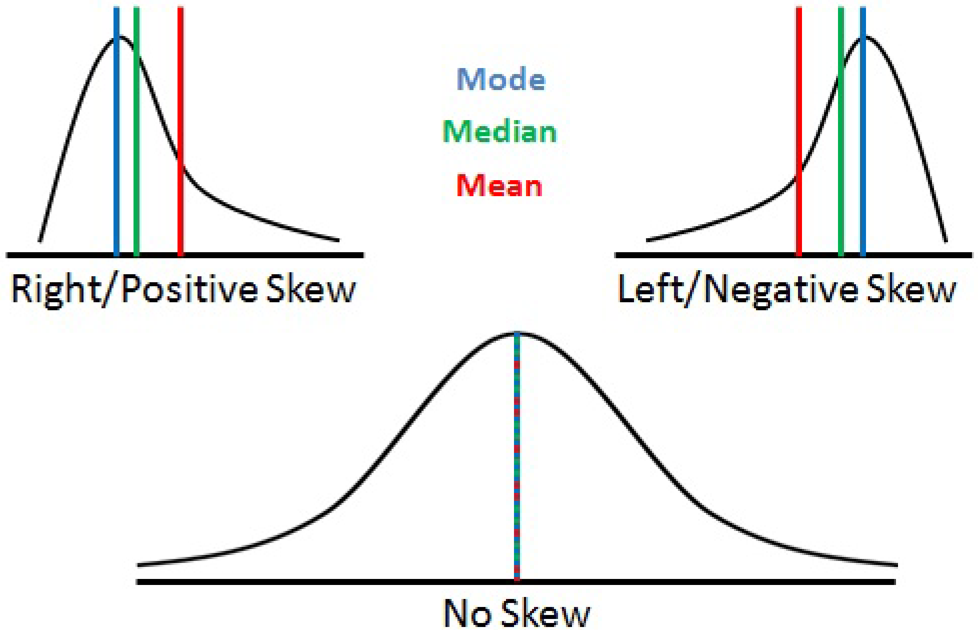
- If the distribution is <u>mound-shaped</u>, then
  - Approximately 68% of the data falls within one standard deviation of the mean
  - Approximately 95% of the data falls within two standard deviations of the mean
    - Approximate value of $s = \frac{\text{range}}{4}$
  - Approximately 99.7% of the data falls within three standard deviations of the mean





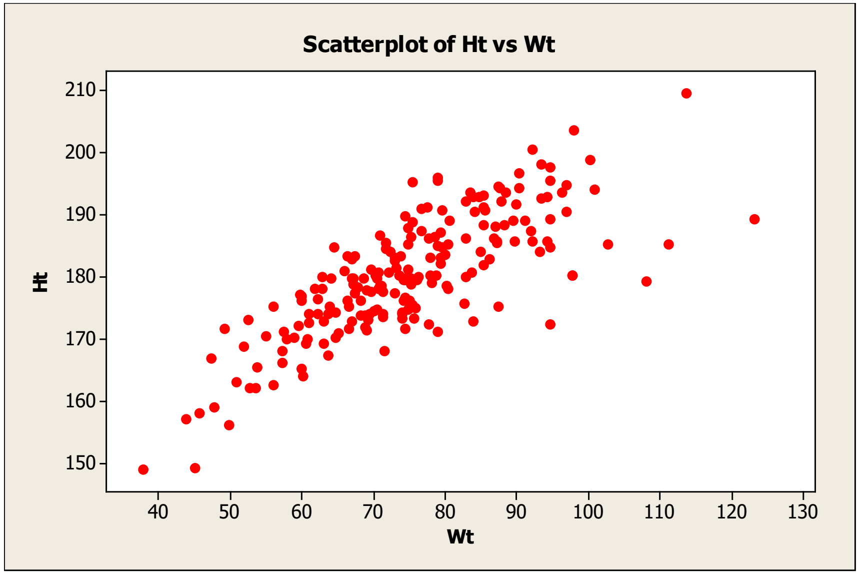
COMPARING MEASURES OF CENTER AND SPREAD

- The **sample mean** and the **sample standard deviation** are good measures of center and spread, respectively, for **symmetric** data
- If the data set is **skewed** or has **outliers**, the **sample median** and the **interquartile range** are more commonly used
- Mean versus median



RELATIONSHIPS BETWEEN 2 NUMERIC VARIABLES

- Depending on the situation, one of the variables is the explanatory variable and the other is the response variable.
- There is not always an explanatory-response relationship.
- Examples:
  - Height and Weight
  - Income and Age
  - SAT scores on math exam and on verbal exam
  - Amount of time spent studying for an exam and exam score



- **R Code:**
  - plot(Ht ~ Wt, data=ais)

RELATIONSHIPS BETWEEN 2 NUMERIC VARIABLES

- **Scatterplots**
  - Look for overall pattern and any striking deviations from that pattern.
  - Look for outliers, values falling outside the overall pattern of the relationship
  - You can describe the overall pattern of a scatterplot by the form, direction, and strength of the relationship.
    - Form: Linear or clusters
    - Direction
      - Two variables are positively associated when above-average values of one tend to accompany above-average values of the other and likewise below-average values also tend to occur together.
      - Two variables are negatively associated when above-average values of one variable accompany below-average values of the other variable, and vice-versa.
    - Strength-how close the points lie to a line

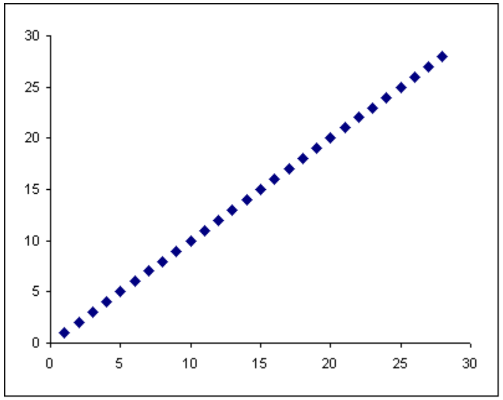
RELATIONSHIPS BETWEEN 2 NUMERIC VARIABLES

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$

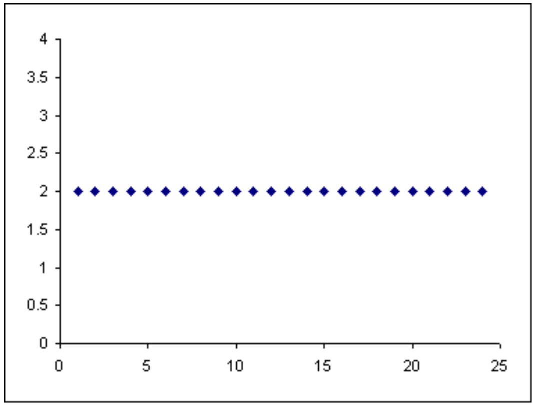
$$= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

- Examples of extreme cases

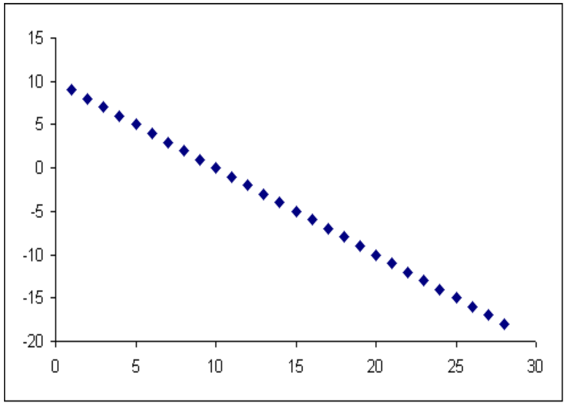
$r = 1$



$r = 0$



$r = -1$



EXAMPLE FOR CORRELATION

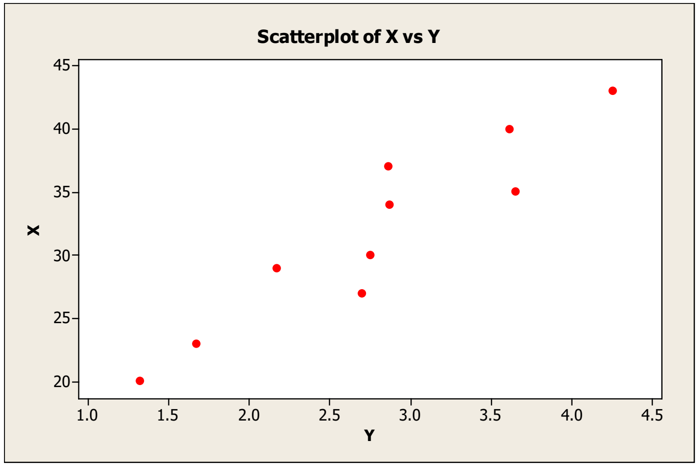
EXAMPLE 3.16

For the data in Table 3.17, compute the value of the correlation coefficient.

$$\bar{x} = 31.80 \text{ and } \bar{y} = 2.785$$

$$x - \bar{x} = 20 - 31.8 = -11.8, \qquad y - \bar{y} = 1.32 - 2.785 = -1.465,$$
$$(x - \bar{x})(y - \bar{y}) = (-11.8)(-1.465) = 17.287,$$
$$(x - \bar{x})^2 = (-11.8)^2 = 139.24, \qquad (y - \bar{y})^2 = (-1.465)^2 = 2.14623$$

	x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$	$(y - \bar{y})^2$
	20	1.32	−11.8	−1.465	17.287		
	23	1.67	−8.8	−1.115	9.812		
	29	2.17	−2.8	−0.615	1.722		
	27	2.70	−4.8	−0.085	0.408		
	30	2.75	−1.8	−0.035	0.063		
	34	2.87	2.2	0.085	0.187		
	35	3.65	3.2	0.865	2.768		
	37	2.86	5.2	0.075	0.390		
	40	3.61	8.2	0.825	6.765		
	43	4.25	11.2	1.465	16.408		
Total	318	27.85	0	0	55.810	485.60	7.3641
Mean	31.80	2.785					



A form of r that is somewhat more direct in its calculation is given by

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{55.810}{\sqrt{(485.6)(7.3641)}} = .933$$

RELATIONSHIPS BETWEEN 2 NUMERIC VARIABLES

- **Correlation** or r : measures the direction and strength of the **linear** relationship between two numeric variables
  - General Properties
    - It must be between -1 and 1, or ($-1 \leq r \leq 1$).
    - If r is negative, the relationship is negative.
    - If $r = -1$, there is a perfect negative linear relationship (extreme case).
    - If r is positive, the relationship is positive.
    - If $r = 1$, there is a perfect positive linear relationship (extreme case).
    - If r is 0, there is no **linear** relationship.
    - r measures the strength of the **linear** relationship.
    - If explanatory and response are switched, r remains the same.
    - r has no units of measurement associated with it
    - Scale changes do not affect r
- Correlation Applet

Correlation $r = 0$

Correlation $r = -0.3$

Correlation $r = 0.5$

Correlation $r = -0.7$

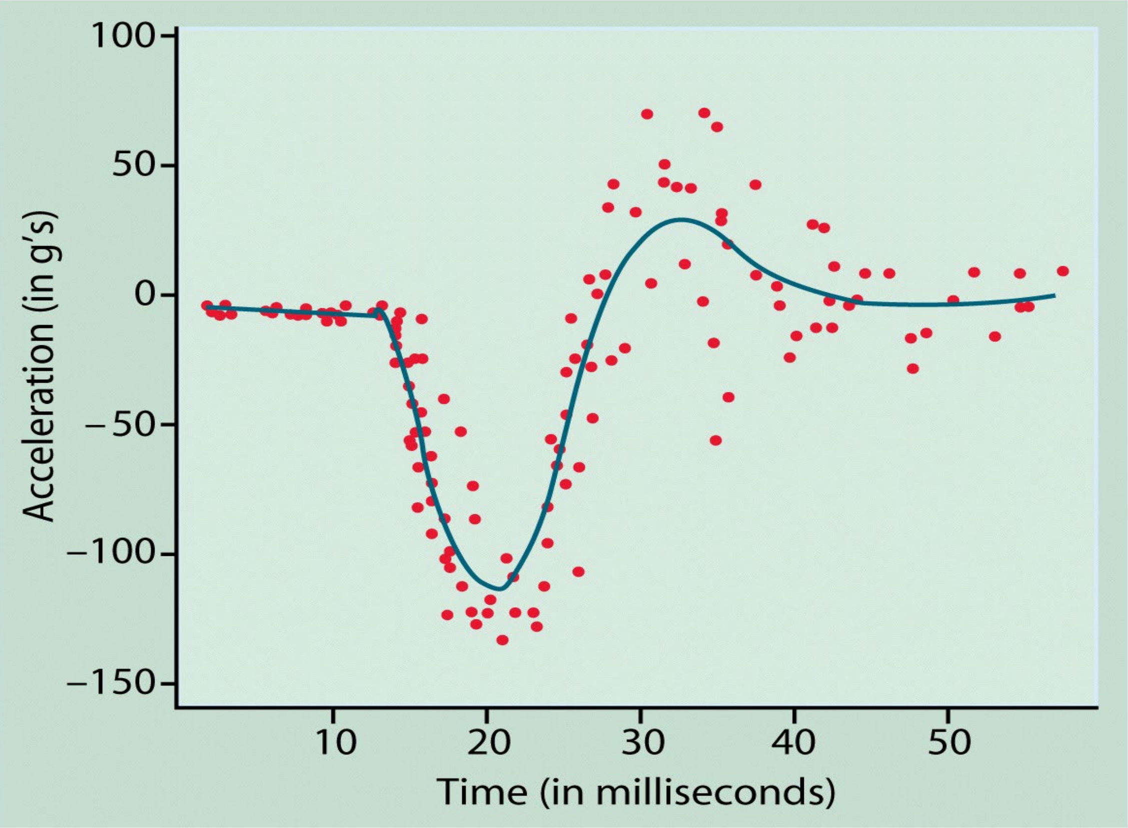
Correlation $r = 0.9$

Correlation $r = -0.99$

RELATIONSHIPS BETWEEN 2 NUMERIC VARIABLES

It is possible for there to be a strong relationship between two variables and still have $r \approx 0$.

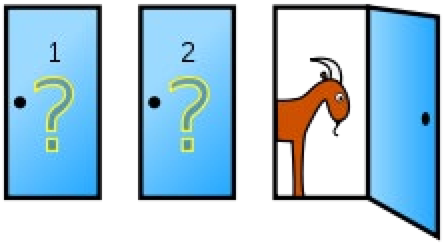
EX.



MARQUETTE UNIVERSITY
Be The Difference.

LET'S MAKE A DEAL

- Let's Make a Deal (Monty Hall problem)
  - http://en.wikipedia.org/wiki/Monty_Hall_problem

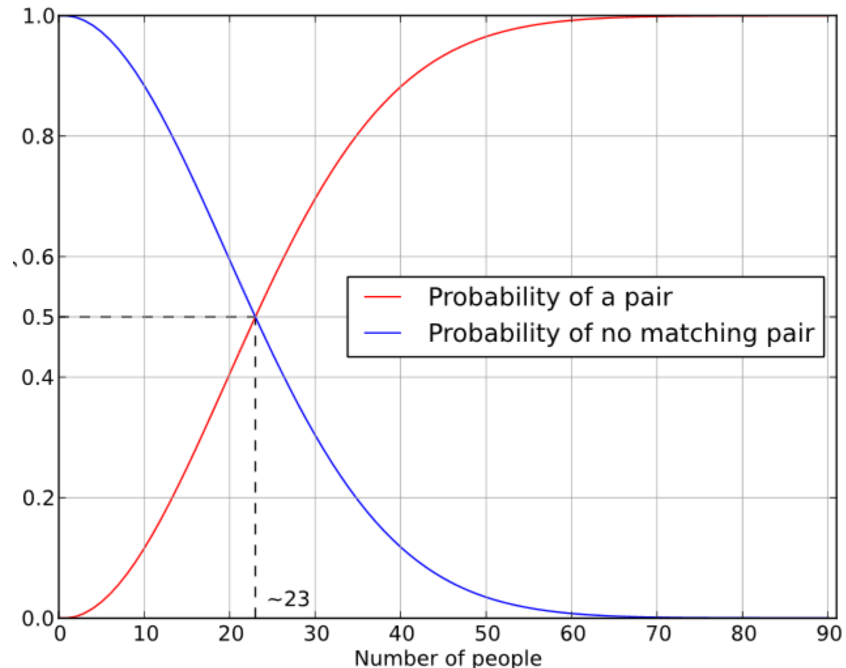


- This is motivation to study probability.

- Should you switch or should you stay with your original choice?

BIRTHDAY PARADOX

- **What's the chances that two people in our class have the same birthday?**



- **R Code:**

```
p <- function(n) {
    p <- NA
    for (i in 1:length(n)) {
        p[i] <- prod(365:(365 - (n[i] - 1))) / 365^n[i]
    }
    return(p)
}
plot(n,p(n), col="blue", type="l", lwd=2, xlab= "Number of people")
points(n,1-p(n), col="red", lwd=2, type="l")
abline(v=23, lty=2)
abline(h=0.5, lty=2)
legend("right",c("Probability of a pair","Probability of no maching pair"),
    lty=1, lwd=2, col=c("red","blue"), cex=2)
```