Calculating Descriptive Statistics

→ Measures of variation

{Determining the dispersion of the data}

→ Range, Interquartile range

Outliers, variance, Standard deviation

& visualizing distributions

Range How wide is your data

Range = highest value − Smallest value

→ always a postive no.

Interquartile Range : finding the middle 50 percent of the data

$$IQR = Q_3 - Q_1$$

make sure data is sorted from least to greatest, so that the 3rd highest value is actually in the third position

Outliers Separating the good data from the bad

→ An extremely high or low data value, as compared to the rest of the data

Lower limit for outliers $= Q_1 - 1.5\,IQR$ & Upper limit for outliers $= Q_3 + 1.5\,IQR$

Variance

$X = (x_1, x_2, \ldots, x_n)$

$$Var(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

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

$$= \frac{1}{n}\sum x_i^2 + \frac{1}{n}\cdot n\bar{x}^2 - \frac{2}{n}\bar{x}\sum x_i$$

$$= \frac{1}{n}\sum x_i^2 + \bar{x}^2 - 2\bar{x}^2$$

$$= \frac{1}{n}\sum x_i^2 - \bar{x}^2$$

$$= \frac{1}{n}\left[\sum x_i^2 - \frac{1}{n}(\sum x_i)^2\right]$$

$$= \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n}$$

Short-cut Method

Requires fewer Computations

Variance denote

Population — σ^2

Sample — s^2

denominator — n

denominator — $(n-1)$

Variance	Formula
Population	$\frac{1}{n}\sum(x_i-\bar{x})^2$ or $\frac{1}{n}\sum x_i^2 - \bar{x}^2$ or $\frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n}$
Sample	$\frac{1}{n-1}\sum(x_i-\bar{x})^2$ or $\frac{1}{n-1}\sum x_i^2 - \frac{n}{n-1}\bar{x}^2$ or $\frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}$

Grouped data

$$Var(x) = \frac{1}{N}\sum f_i (x_i - \bar{x})^2$$

$$= \frac{1}{N}\sum f_i (x_i^2 + \bar{x}^2 - 2x_i\bar{x})$$

$$= \frac{1}{N}\sum f_i x_i^2 + \bar{x}^2 \frac{1}{N}\sum f_i - \frac{2\bar{x}}{N}\sum f_i x_i$$

$$= \frac{1}{N}\sum f_i x_i^2 + \bar{x}^2 - 2\bar{x}^2$$

$$= \frac{1}{N}\sum f_i x_i^2 - \bar{x}^2$$

$$= \frac{1}{N}\left[\sum f_i x_i^2 - \frac{1}{N}\left(\sum f_i x_i\right)^2\right]$$

$$= \frac{\sum f_i x_i^2 - \frac{\left(\sum f_i x_i\right)^2}{N}}{N}$$

In classes. Use m_i — midpoints instead of x_i

Let $Y = X + a$

$y_1 = x_1 + a$

$y_2 = x_2 + a$

$y_n = x_n + a$

Then $Var(Y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$

$= \frac{1}{n}\sum_{i=1}^{n}(x_i + a - \bar{x} - a)^2$

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

$= Var(X)$

$Var(X \pm a) = Var(X)$

$Var(aX) = a^2 Var(X)$

$Var\left(\frac{X}{a}\right) = \frac{1}{a^2} Var(X)$

always ≥ 0

Let $Y = aX$

Then $Var(Y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$

$= \frac{1}{n}\sum_{i=1}^{n}(ax_i - a\bar{x})^2$

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

$= a^2 Var(X)$

Standard deviation : always ≥ 0

$$Sd(X) = \sqrt{Var(X)}$$

$$Sd(X \pm a) = Sd(X)$$

$$Sd(aX) = a\,Sd(X)$$

Coefficient of Variation

$$CV = \frac{Sd}{mean} \times (100\%)$$

→ Useful when you are comparing two data sets that aren't exactly alike, especially if the different data sets aren't measured using the same units.

Visualizing distributions

→ Box-and-whisker plot

↳ Consists of five data points

Smallest data value

Q_1, Q_2, Q_3

Largest data value

Data: number of children per day who attend an after school program over an 11-day period

20 22 29 37 49 56 64 70 70 87 92

Smallest value

Largest value

$Q_1 = \frac{11}{4} = 2.75^{th}$ observation $= 29$

$Q_2 = \frac{11+1}{2} = 6^{th}$ observation $= 56$

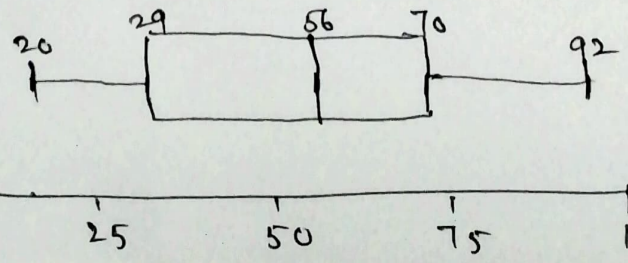
$Q_3 = \frac{3 \times 11}{4} = 8.25^{th}$ observation $= 70$

Outliers $IQR = Q_3 - Q_1 = 70 - 29 = 41$

Lower limit $= Q_1 - 1.5\,IQR = 29 - 1.5 \times 41 = -32.5$

Upper limit $= Q_3 + 1.5\,IQR = 70 + 1.5 \times 41 = 131.5$

There are no outliers in this data set.



Stem-and-Leaf plot

↳ displays the distribution of a data set by separating each value into a stem and a leaf.

Stem — first digit (or digits) of the number

Leaf — last digit

Leaves with common stem are grouped together in ascending order.

Data: 28, 34, 42, 47, 49

2	8
3	4
4	2 7 9

Data: total annual snowfall (in inches) for 30 cities

11	12	14	17	20	20	22	25	25	26
26	28	30	32	32	34	35	35	38	39
39	41	41	43	45	46	48	49	50	56

Stem of each data value is its 10 digit

1	1 2 4 7
2	0 0 2 5 5 6 6 8
3	0 2 2 4 5 5 8 9 9
4	1 1 3 5 6 8 9
5	0 6

Leaf is one-digit

Data:

117	104	102	98	97	96
95	91	90	90	89	88
87	87	86	85	83	82
81	78	77	77	76	75
74	71	71	71	70	70

7	0	0	1	1	1	4	5	6	7	7	8
8	1	2	3	5	6	7	7	8	9		
9	0	0	1	5	6	7	8				
10	2	4									
11	7										

7(0)	0	0	1	1	4	
7(5)	5	6	7	7	8	
8(0)	1	2	3			
8(5)	5	6	7	7	8	9
9(0)	0	0	1			
9(5)	5	6	7	8		
10(0)	2	4				
10(5)						
11(0)						
11(5)	7					

Splitting the stems in half is a good idea when some stems have a lot of leaves and others don't have as many.

You are trying to see how spread out the data is, and sometimes you need to spread out the stems to do that

The following two tables list the numbers of home runs hit by the leaders in this category in the National League and the American League for the 2008 Major League Baseball season. Construct a back-to-back stem-and-leaf diagram comparing the two leagues. What conclusions can you draw based on this diagram?

Sorted National League Home Run Leaders

48	40	38	37	37	37	36	34	33	33
33	33	32	32	29	29	29	28	28	27
27	27	26	26	25	25	25	25	25	25

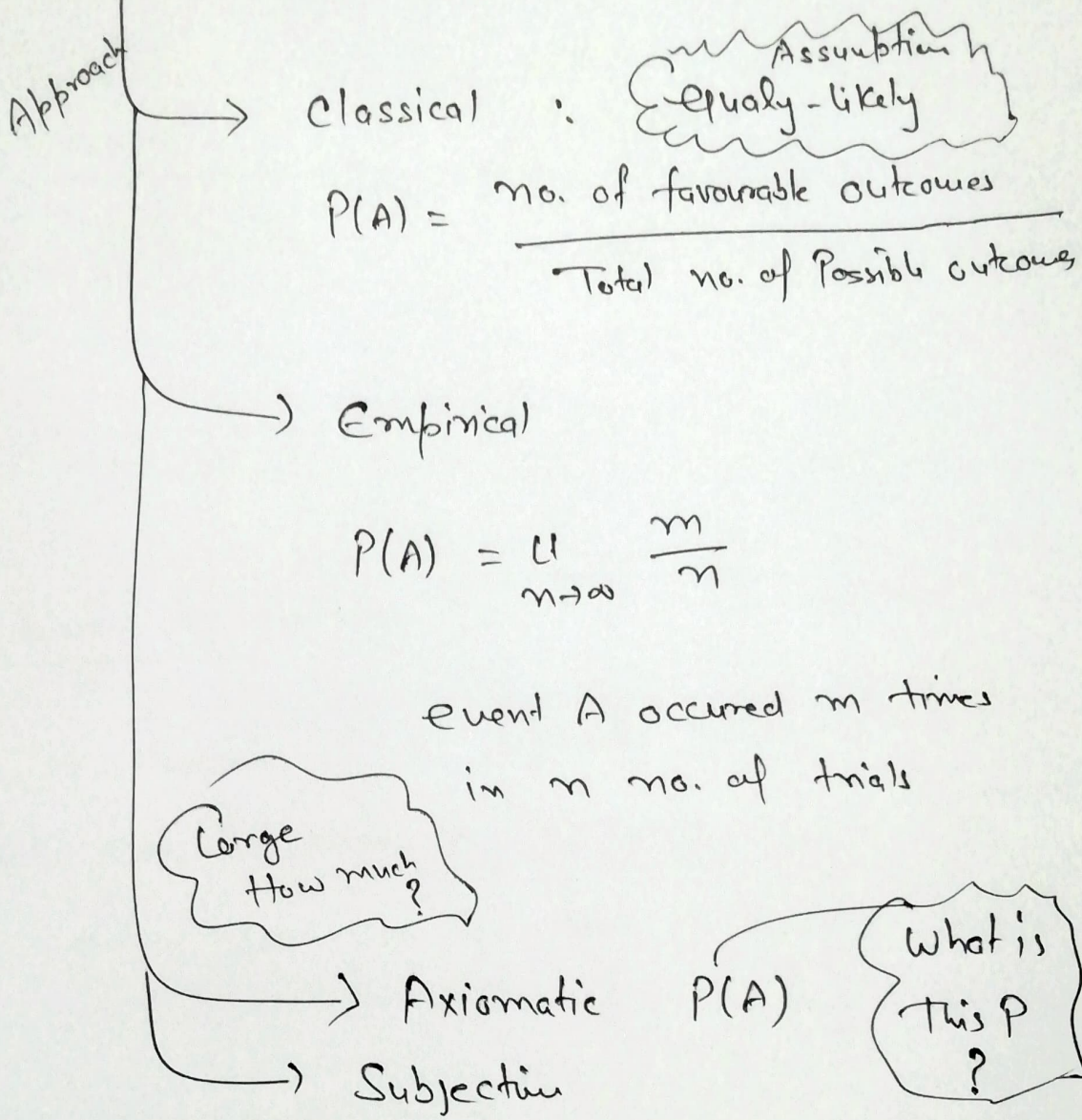
Sorted American League Home Run Leaders

37	36	35	34	34	33	33	32	32	32
31	29	27	27	25	25	24	23	23	23
23	23	23	22	22	22	27	21	21	21

National League →		American League →
	2(0)	1 1 1 1 2 2 2 3 3 3 3 3 3 4
8 8 7 7 7 6 6 5 5 5 5 5	2(5)	5 5 7 7 9
9 9 9		
4 3 3 3 3 2 2	3(0)	1 2 2 2 3 3 4 4
8 7 7 7 6	3(5)	5 6 7
0	4(0)	
8	4(5)	

The majority of National League's batters hit between 25 and 29 home runs. Most of the American League leaders hit between 20 & 24 home runs, and None of them hit 40 or more

Introduction to Probability : What are the chances?



Sample space — Set of all possible outcomes of a particular experiment

Event — any collection of possible outcomes of an experiment
any subset of S (including S itself)

Disjoint $A \& B$: $A \cap B = \phi$

Pairwise disjoint $A_i \cap A_j = \phi$ for all $i \neq j$

Partition: A_i if (i) $\bigcup A_i = S$ & (ii) $A_i \cap A_j = \phi$ for $i \neq j$

No. of possible arrangements of size r from n objects

	Without Replacement	With Replacement
Ordered	${}^{n}P_r = \frac{n!}{(n-r)!}$	n^r
Unordered	${}^{n}C_r = \frac{n!}{(n-r)!\,r!}$	${}^{n+r-1}C_r$

Fundamental Theorem of Counting

If a job consists of k separate tasks, the ith of which can be done in n_i ways, $i = 1, 2, \dots k$

Then the entire job can be done in $n_1 \times n_2 \times \dots \times n_k$ ways.