Chapter 2

Methods for Describing Data Sets

class

Class frequency

“Wow

green eyes!” My friend might respond, “Out of how many people?” Twenty people having green eyes

ere were 200 students…

$$\textbf{Relative Frequency} = \frac{Frequency}{n}$$

’s practice with some real data from a recent study conducted by the Pew Research Center.

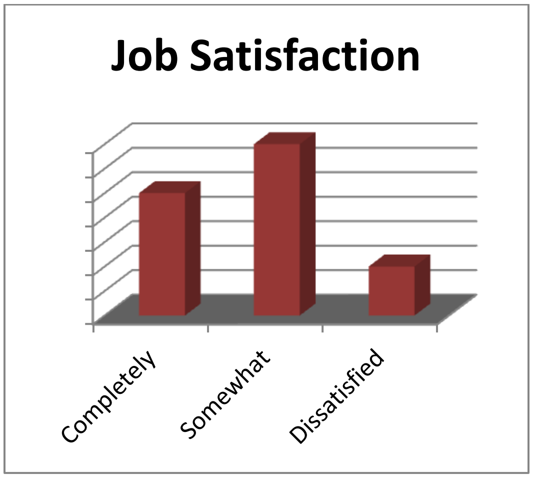
Subject	Job Satisfaction	Subject	Job Satisfaction

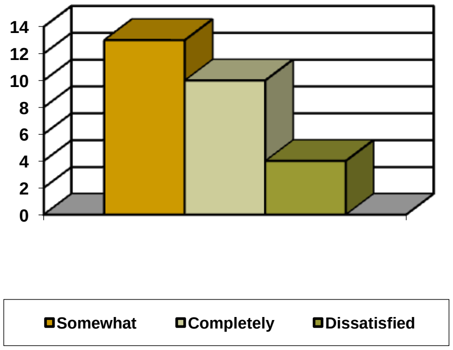
Job Satisfaction	Frequency

Job Satisfaction	Relative Frequency	Class Percentage

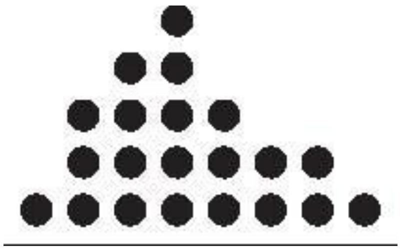
ie charts, bar graphs,…),

Pareto Diagram

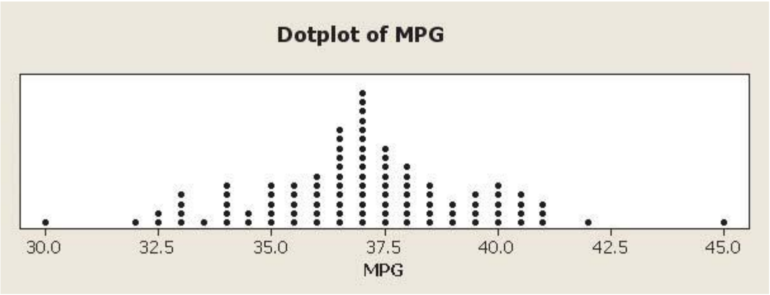




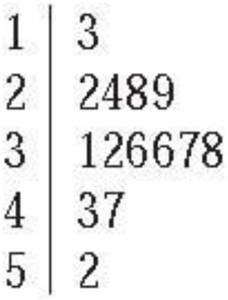
Dot plots



car’s MPG rating from the study:



- **stem-and-leaf display**



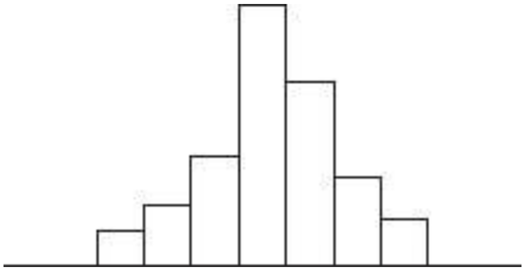
1	3
2	2489
3	126678
4	37
5	2

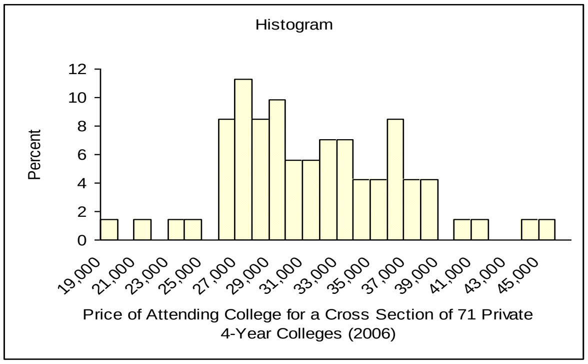
stem unit = 10　　leaf unit =1

Frequency	**Stem**	**Leaf**
9	**3**	**4 6 6 8 8 8 8 9 9**
17	**4**	**0 0 2 2 3 4 4 4 4 5 7 7 8 9 9 9 9**
4	**5**	**2 2 3 3**

n = 30

- **Histograms**





Ok, before we begin creating a histogram, let's

relative frequency distribution **<u>frequency table</u>**
__________________ ***"Relative"*** **sample size**
(n). ***"Frequencies"*** **counts.**

$$\textbf{Relative Frequency} = \frac{Frequency}{n}$$

<u>class</u>

<u>Lower class limits</u>

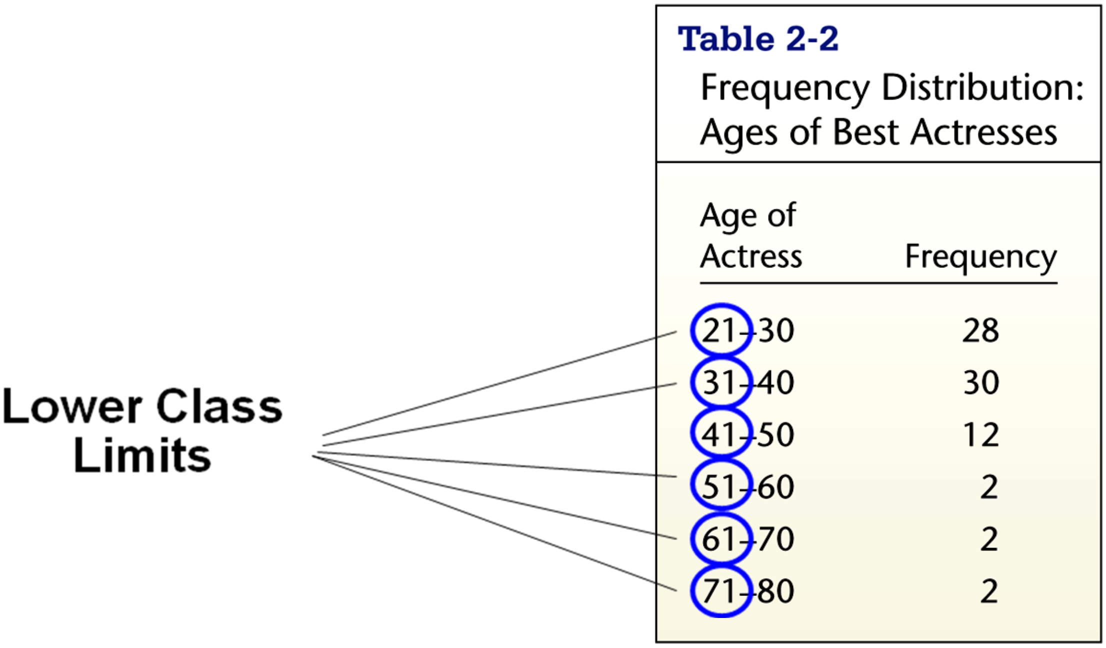

Lower Class Limits

Table 2-2
Frequency Distribution: Ages of Best Actresses

Age of Actress	Frequency
21-30	28
31-40	30
41-50	12
51-60	2
61-70	2
71-80	2

Upper class limits

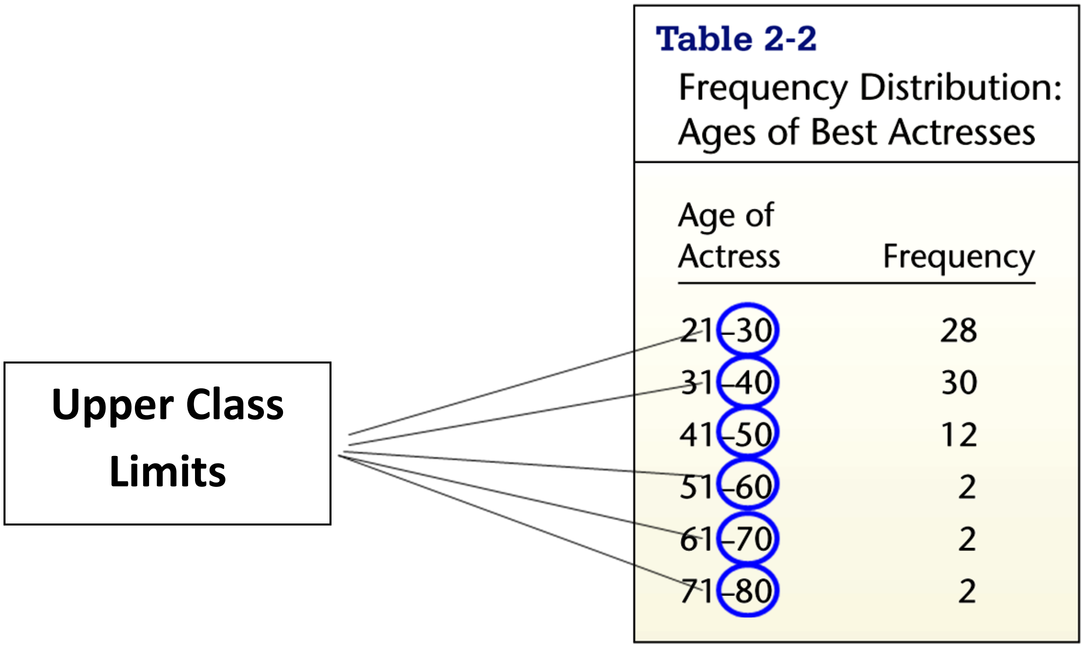

Upper Class
Limits

Table 2-2
Frequency Distribution:
Ages of Best Actresses

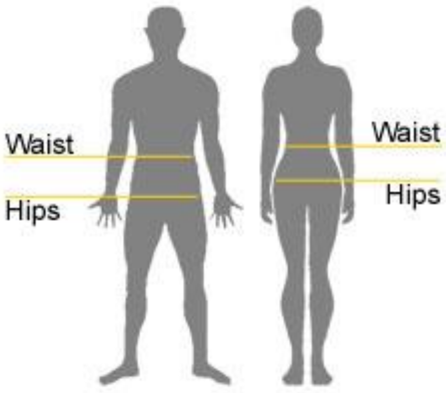
Age of Actress	Frequency
21–30	28
31–40	30
41–50	12
51–60	2
61–70	2
71–80	2

Sample Data –

– **lower class limits**

upper class limits

–	
–	
–	
–	
–	
–	
–	



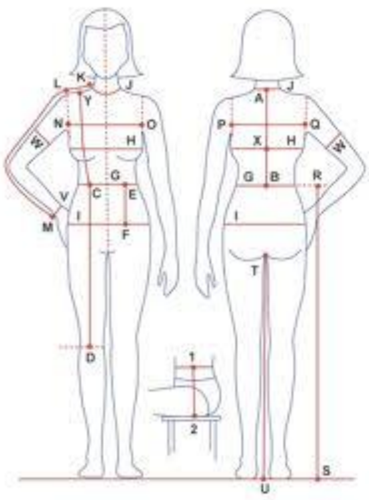
Class boundaries

Class boundaries are obtained as follows:

Step 1:

Step 2:

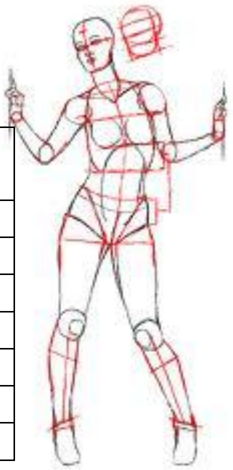
Step 3:



–		
–		
–		
–		
–		
–		
–		

<u>Class midpoints</u>

–	–		
–	–		
–	–		
–	–		
–	–		
–	–		
–	–		



<u>class width</u>

– –

Guidelines for creating a relative frequency table:

Determine the number of classes to use:

Sturges’ K = 1 + 3.3219 * log n, where K is the number of classes, and n is the number of values in the data set.)

Calculate the Range: –

Determine the class width:

$$\geq \frac{Range}{numberofclasses}$$

Select the lower limit of the first class:

Use the class width to obtain the other lower class limits:

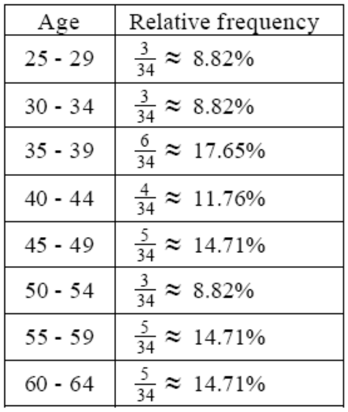
Determine the upper class limits:

Determine the frequencies:

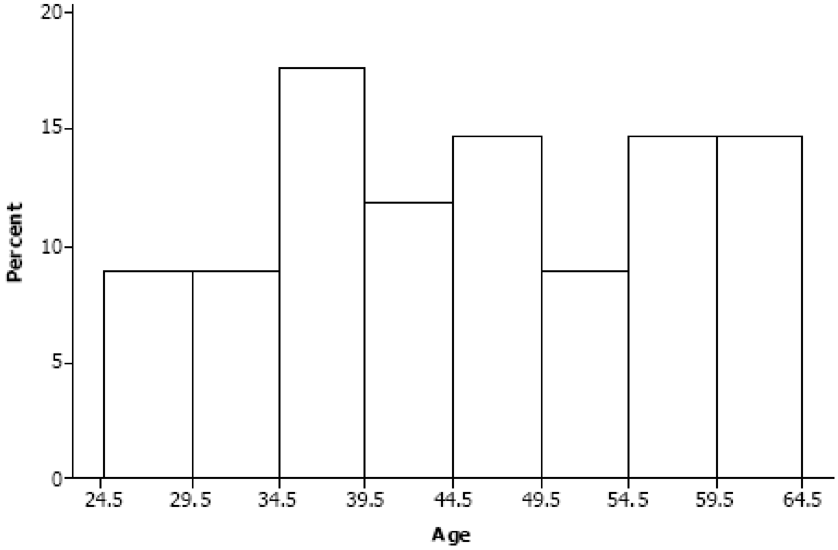
Calculate the relative frequencies:

<u>relative frequency histogram</u>

<u>classes</u>

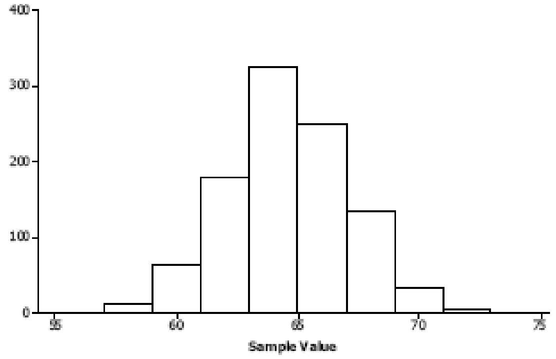


Age	Relative frequency
25 - 29	$\frac{3}{34} \approx$ 8.82%
30 - 34	$\frac{3}{34} \approx$ 8.82%
35 - 39	$\frac{6}{34} \approx$ 17.65%
40 - 44	$\frac{4}{34} \approx$ 11.76%
45 - 49	$\frac{5}{34} \approx$ 14.71%
50 - 54	$\frac{3}{34} \approx$ 8.82%
55 - 59	$\frac{5}{34} \approx$ 14.71%
60 - 64	$\frac{5}{34} \approx$ 14.71%

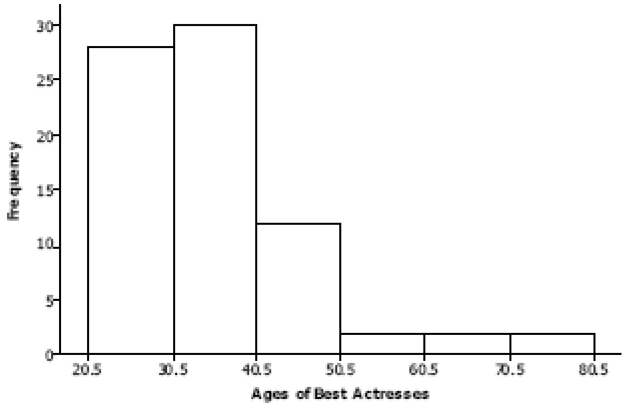


Left End Point Convention for Continuous Data:

–	
–	
–	
–	
–	



[redacted]



[redacted]

- **Individual observations in a data set are denoted**

  x_1, x_2, x_3, x_4, … x_n.

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \ldots + x_n$$

$$\sum_{i=1}^{n} x_i = 1 + 2 + 3 + 4 = 10$$

-

$$\sum_{i=1}^{n} x_i^2 = x_1^2 + x_2^2 + x_3^2 + \ldots + x_n^2$$

-

$$\left(\sum_{i=1}^{n} x_i\right)^2 = \left(x_1 + x_2 + x_3 + \ldots + x_n\right)^2$$

$$\sum_{i=1}^{5} X_i$$

$$\left(\sum_{i=1}^{5} X_i\right)^2$$

$$\sum_{i=1}^{5} X_i^2$$

$$\sum_{i=1}^{5}\left(x_i-3\right)^2$$

■

○

○

Central tendency

Variability **ariability**

central tendency:

Mean

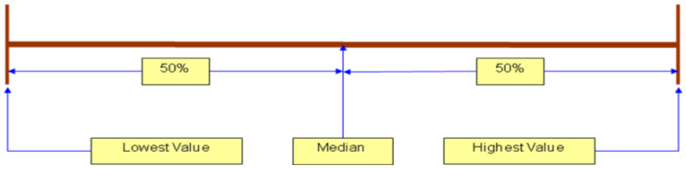
$$\bar{x} = \frac{\sum_{i=1}^{n} X_i}{n}$$

■ *sample* *population mean*
µ (pronounced mew)

$$\mu = \text{population mean} \quad \bar{x} = \text{sample mean}$$

$$\overline{x} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{21+2+1+3+24+120+36+1+1+1}{10} = 21$$

Median



Notice the median only looks at one or two numbers in the center of the data set. Doesn’t that seem

isn’t unduly affected by really big or small numbers in the data set. For example, what

artificially high, and would not truly capture the typical American’s personal

-
-

$\tilde{x}$

η

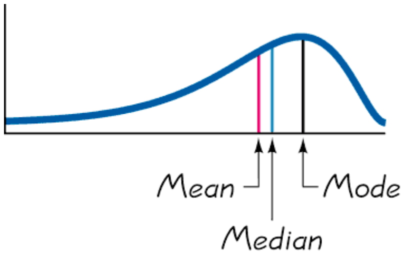
Mode

—

color in the USA, correct? You can't add and divide eye colors to find an average, nor could you put

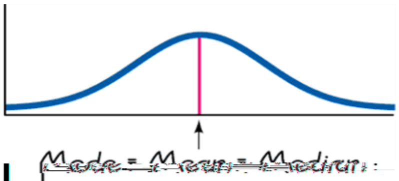
Skewed

■



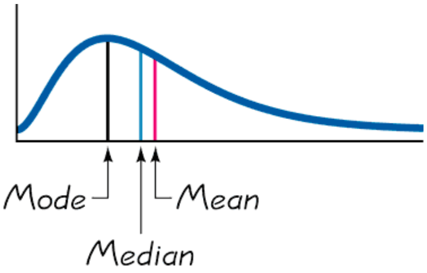
(a) Skewed to the Left
(Negatively)

■

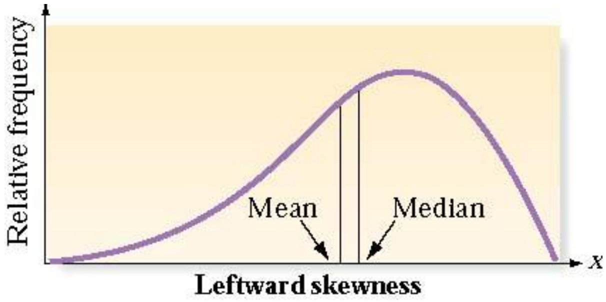


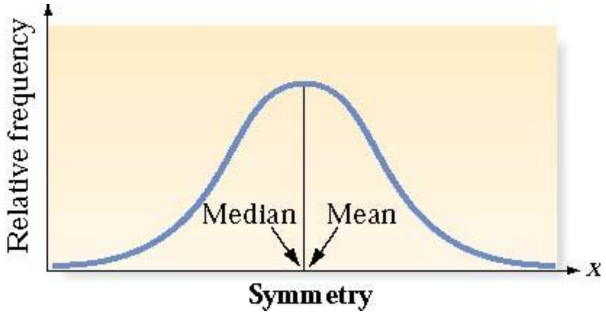
(b) Symmetric

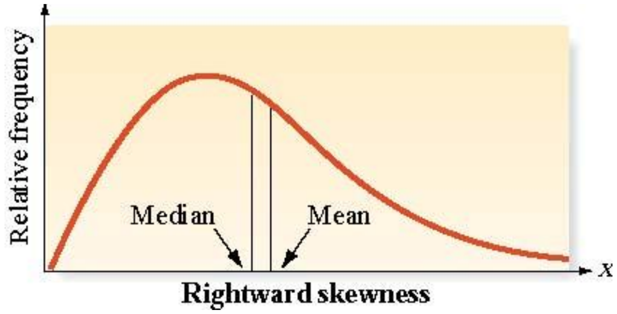
■



(c) Skewed to the Right (Positively)




Relative frequency
Median
Mean
x
Symmetry


Relative frequency
Median
Mean
x
Rightward skewness

measures of variability

Variability **ariability**

You know what the word ‘vary’ means. If a population’s data values do not vary much, then the

Range

$$Range = Max - Min$$

Variance

–

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

$$s^2 = \frac{n\sum x^2 - \left(\sum x\right)^2}{n(n-1)}$$

Standard Deviation

$$s=\sqrt{s^2}=\sqrt{\frac{n\sum x^2-\left(\sum x\right)^2}{n(n-1)}}$$

■

σ^2 s^2

σ s

ft., inches, miles, yrs, dollars,…) for

range rule of thumb. We will see that at the end of the next section, but first let's find a way

where the data will lie relative to the mean

2.6 Chebyshev'

Chebyshev's Rule:

$$1-\frac{1}{k^2}$$

$[\mu - k\sigma, \mu + k\sigma]$ $1-\frac{1}{k^2}$

Chebyshev's Rule

○ *any*

○ *k* $1-\frac{1}{k^2}$ *k*

first

Chebyshev’s

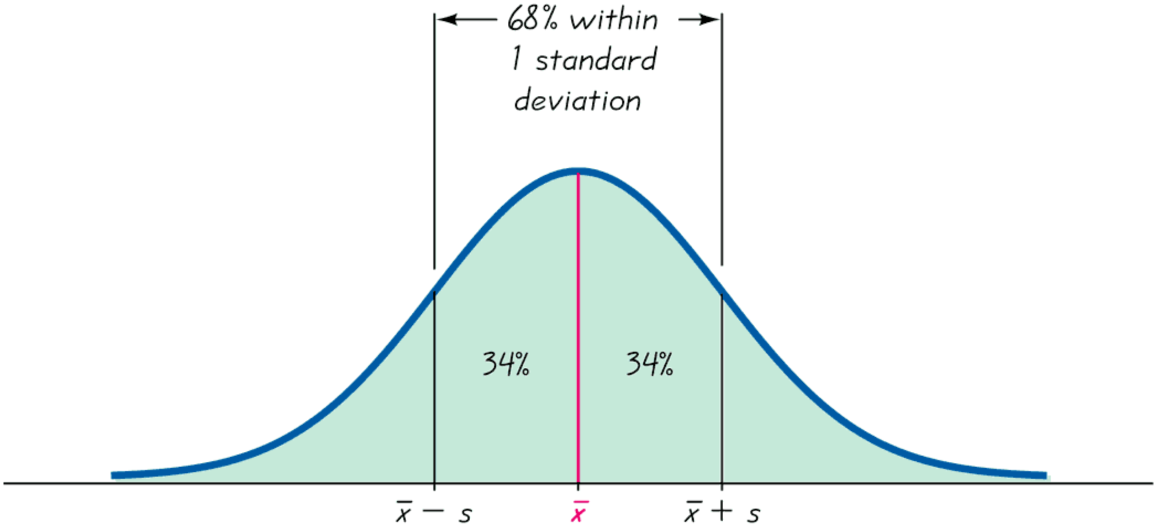
- **The Empirical Rule**
  - **Useful for mound-shaped, symmetrical distributions**
  - **~68% will be within the range** $(\overline{x}-s, \overline{x}+s)$
  - **~95% will be within the range** $(\overline{x}-2s, \overline{x}+2s)$
  - **~99.7% will be within the range** $(\overline{x}-3s, \overline{x}+3s)$

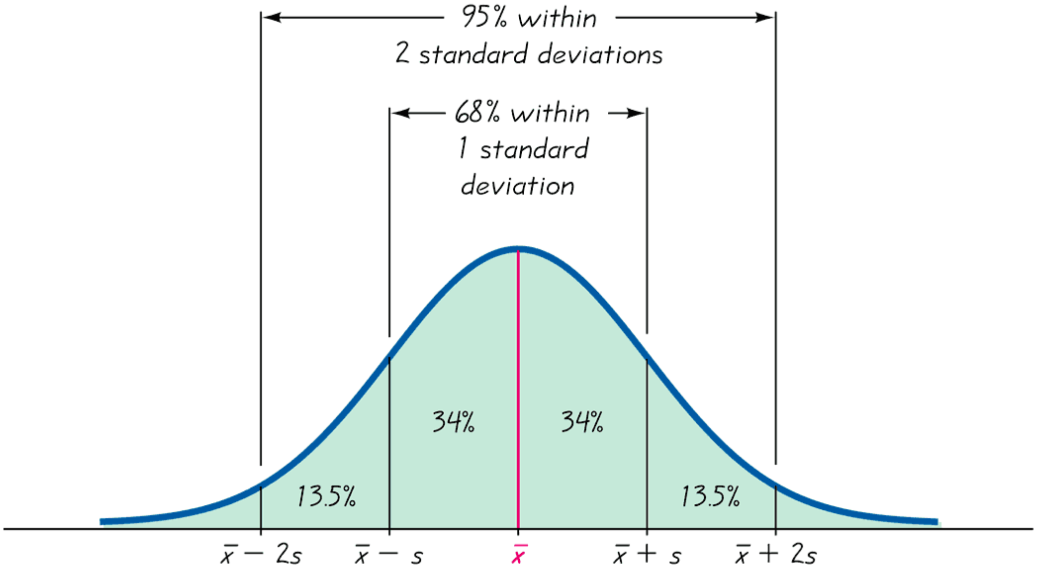
Empirical Rule:

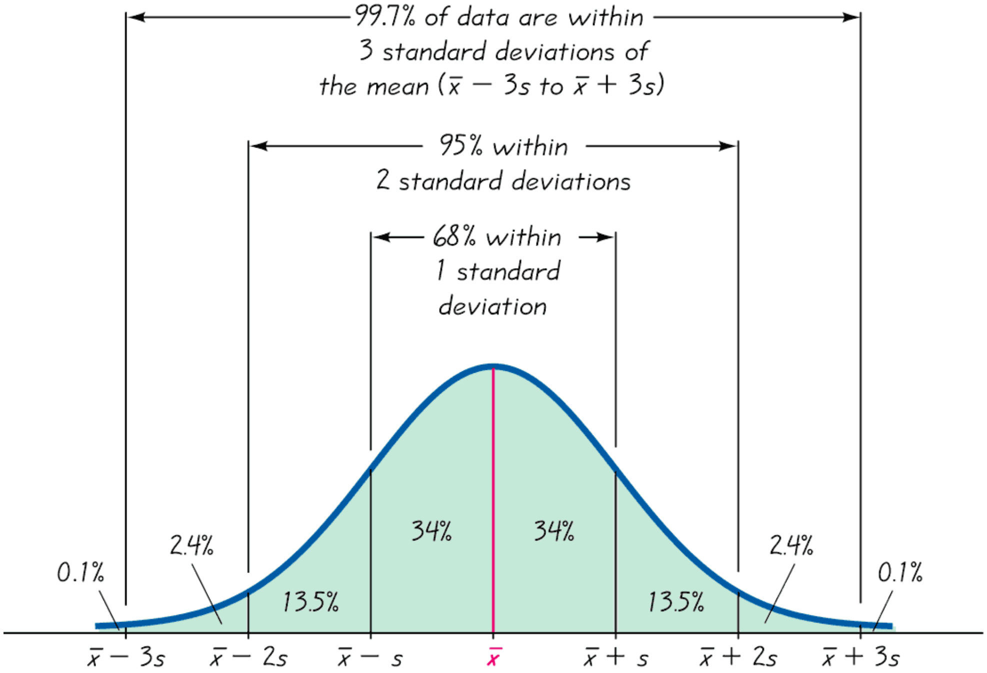
σ

σ’s of the mean

σ’s of the mean


68% within 1 standard deviation
34%
34%
$\bar{x} - s$
$\bar{x}$
$\bar{x} + s$


95% within 2 standard deviations
68% within 1 standard deviation
34%
34%
13.5%
13.5%
$\bar{x} - 2s$
$\bar{x} - s$
$\bar{x}$
$\bar{x} + s$
$\bar{x} + 2s$


99.7% of data are within 3 standard deviations of the mean ($\bar{x} - 3s$ to $\bar{x} + 3s$)
95% within 2 standard deviations
68% within 1 standard deviation
0.1%
2.4%
13.5%
34%
34%
13.5%
2.4%
0.1%
$\bar{x} - 3s$
$\bar{x} - 2s$
$\bar{x} - s$
$\bar{x}$
$\bar{x} + s$
$\bar{x} + 2s$
$\bar{x} + 3s$

○

○

○

Range Rule of Thumb

s

$\left[\frac{R}{6}, \frac{R}{4}\right]$ –

$\frac{R}{6}$ $\frac{R}{6}$ $\frac{R}{4}$

$\bar{x} - 3s$ $\bar{x} + 3s$

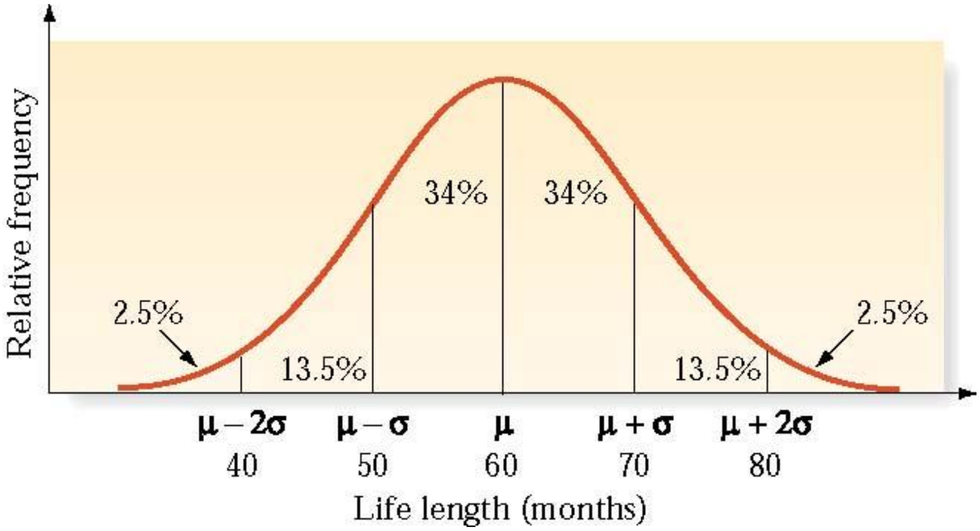
$\bar{x} - 3s$

$\bar{x} + 3s$ –

$(\bar{x} + 3s) - (\bar{x} - 3s)$ $(\bar{x} + 3s - \bar{x} + 3s) = 6s$

$s \approx \frac{R}{6}$ $\bar{x} - 3s$ $\bar{x} + 3s$

$\frac{R}{4}$



-

- *below* *above*

❖

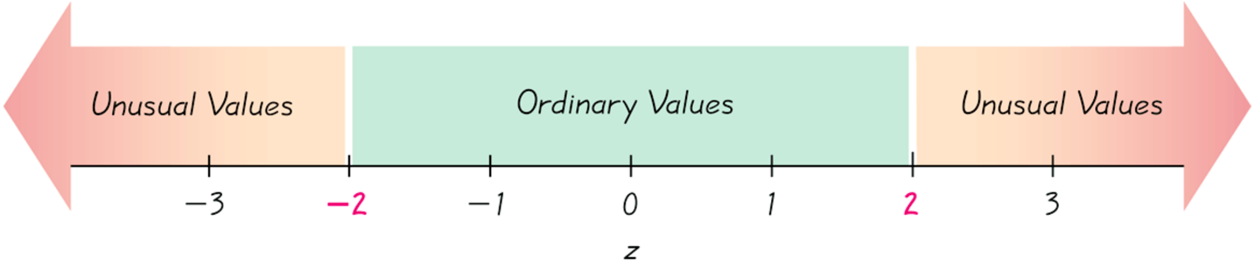
z -

x

$$z = \frac{x - \bar{x}}{s}$$

$$z = \frac{x - \mu}{\sigma}$$

<u>**z-score**</u>



Z-scores and the Empirical rule:

○

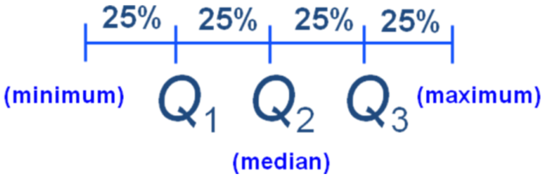
○

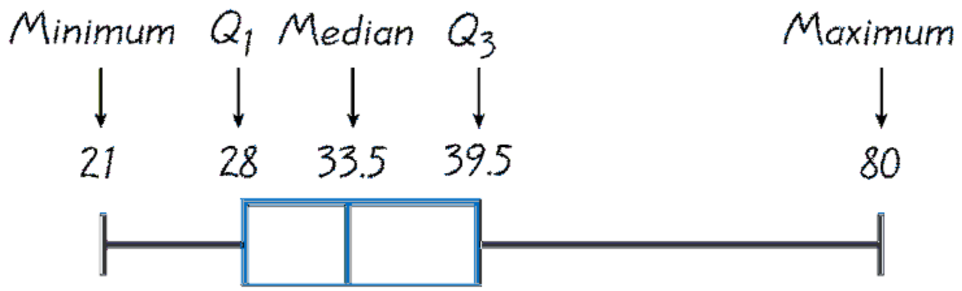
○

Percentiles

pth percentile

***Q_1*, *Q_2*, *Q_3* divide ranked scores into four equal parts:**





Guidelines for finding the approximate kth – percentile:

$$L = \left(\frac{K}{100}\right)n$$

Another approach is to use the following formula:

$$L_k = (n+1)\frac{k}{100}$$

*What to do if L_k is a decimal: if the locator ended up being 14.35 you would add 0.35 (the decimal part of the locator) times the difference between the 14th and 15th value to the 14th value (the whole number part of the locator). For example, if the locator was 14.35, the 14th value was 80, and the 15th value was 83, we would perform the following calculation: 80 + (83 - 80)*0.35 = 81.05.*

Guidelines for finding the approximate percentile of a given number:
