STATISTICS

INFORMED DECISIONS USING DATA

Fifth Edition

Chapter 3

Numerically Summarizing Data

Pearson

3.2 Measures of Dispersion (Spread)

Learning Objectives

1. Determine the **range** of a variable from raw data
2. Determine the **standard deviation** of a variable from raw data
3. Determine the **variance** of a variable from raw data
4. Use the **Empirical Rule** to describe data that are bell shaped
5. ~~Use Chebyshev's Inequality to describe any data set~~

3.2 Measures of Dispersion

3.2.1 Determine the Range of a Variable from Raw Data (1 of 2)

The **range, *R*,** of a variable is the difference between the largest data value and the smallest data values. That is,

$\boxed{R = max - min}$ Quick & easy way to measure spread.

Range = *R* = Largest Data Value − Smallest Data Value

Units of Range: same as data (ex: 98° − 30° = 68°)

Why study the range? What does it tell us about our data?

Quick → spread

Limitations only uses 2 values of data.

3.2 Measures of Dispersion

3.2.1 Determine the Range of a Variable from Raw Data (2 of 2)

EXAMPLE Finding the Range of a Set of Data

The following data represent the travel times (in minutes) to work for all seven employees of a start-up web development company.

23, 36, 23, 18, 5, 26, 43

min max

Find the range.

$R = 43 - 5 = \boxed{38 \text{ min}}$

UNITS

Pearson

3.2 Measures of Dispersion

3.2.2 Introducing Standard Deviation

Ex: Advil and Motrin IB produce the same headache relief medication with the active ingredient ibuprofen. Each pill should contain 200 mg of ibuprofen. A health agency obtains a sample of ten tablets from both manufacturers and measures how much ibuprofen each pill actually contains.

	Number of milligrams measured									
$\bar{x}_1$ Advil	199.25	198.50	200.10	200.75	201.00	198.00	200.10	199.00	201.10	202.20
$\bar{x}_2$ Motrin IB	205.00	195.80	195.20	203.20	205.80	194.40	204.60	194.60	207.20	194.20

way high way low

Each sample has a mean value of 200 mg. However, based on the given sample values, which company would you prefer to buy from?

$\bar{x}_1 = 200\text{mg} = \bar{x}_2$

which prefer? why?

Advil pills more close to 200mg

why better: · high does can be harm full

· manufacture: can loose money

· low does bad b/c feel cheated!

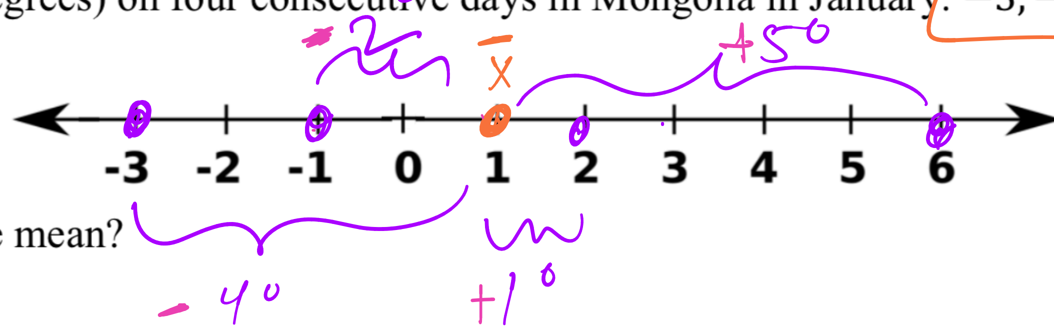
3.2 Measures of Dispersion

3.2.2 Introducing Standard Deviation

Ex: The following are temperatures (in degrees) on four consecutive days in Mongolia in January: $-3, -1, 2, 6$

(a) Find the mean.

(b) How far away is each number from the mean?



a) $\bar{x} = \dfrac{-3-1+2+6}{4} = \dfrac{4}{4} = 1.0$

b) picture

Why do this? take into account how each value compares to mean.

- variation (signed)
- look total variation

problem: total signed variation = 0.

3.2 Measures of Dispersion

3.2.2 Determine the Standard Deviation of a Variable from Raw Data (1 of 16)

" σ "

The **population standard deviation** of a variable is the square root of the sum of squared deviations about the population mean divided by the number of observations in the population, *N*.

That is, it is the square root of the mean of the squared deviations about the population mean.

Notation: population standard deviation is symbolically represented by σ (lowercase Greek "sigma").

Units of Standard Deviation:

same units as data.

3.2 Measures of Dispersion

3.2.2 Determine the Standard Deviation of a Variable from Raw Data (2 of 16)

The **population standard deviation** of a variable is the square root of the mean of the squared deviations about the population mean.

μ – pop. mean

$$\sigma = \sqrt{\frac{(x_1-\mu)^2+(x_2-\mu)^2+\cdots+(x_N-\mu)^2}{N}}$$

$$\boxed{\sigma = \sqrt{\frac{\Sigma(x_i-\mu)^2}{N}}}$$

where $x_1, x_2, \ldots, x_N$ are the N observations in the population and μ is the population mean.

3.2 Measures of Dispersion

3.2.2 Determine the Standard Deviation of a Variable from Raw Data (2 of 16)

Notation:

- **population standard deviation:** σ
- **sample standard deviation:** s (small s)

the actual # not that important

simple <u>is it big</u> or <u>small</u>?

3.2 Measures of Dispersion

3.2.2 Determine the Standard Deviation of a Variable from Raw Data (3 of 16)

IMPORTANT

Why study the standard deviation?

tells us about how spread out our data is

What does it tell us about our data?

Better than range b/c it incorporates

every value of data!

3.2 Measures of Dispersion

3.2.2 Determine the Standard Deviation of a Variable from Raw Data (4 of 16)

EXAMPLE Computing a Population Standard Deviation

The following data represent the travel times (in minutes) to work for all seven employees of a start-up web development company.

23, 36, 23, 18, 5, 26, 43

Compute the population standard deviation of this data.

Mean: $\mu = 24.9$ min per employee

3.2 Measures of Dispersion

$\mu = 24.9$

3.2.2 Determine the Standard Deviation of a Variable from Raw Data (5 of 16)

L1	L2	$\boxed{L3 = L2 - L1}$	$L4 = L3^2$
x_i	μ	$x_i - \mu$	$(x_i - \mu)^2$
23	24.9	23 − 24.9 = −1.9	3.61
36	24.9	36 − 24.9 = 11.1	123.21
23	24.9	23 − 24.9 = −1.9	3.61
18	24.9	18 − 24.9 = −6.9	47.61
5	24.9	5 − 24.9 = −19.9	396.01
26	24.9	26 − 24.9 = 1.1	1.21
43	24.9	43 − 24.9 = 18.1	327.61

$$\boxed{\Sigma(x_i - \mu)^2 = 902.87}$$

$$\sigma = \sqrt{\frac{\Sigma(x_i - \mu)^2}{N}} = \sqrt{\frac{902.87}{7}} \approx 11.4 \text{ minutes}$$

round w/ "stats law"

3.2 Measures of Dispersion

3.2.2 Determine the Standard Deviation of a Variable from Raw Data (5 of 16)

x_i	μ	$x_i - \mu$	$(x_i - \mu)^2$
23	24.85714	–1.85714	3.44898
36	24.85714	11.14286	124.1633
23	24.85714	–1.85714	3.44898
18	24.85714	–6.85714	47.02041
5	24.85714	–19.8571	394.3061
26	24.85714	1.142857	1.306122
43	24.85714	18.14286	329.1633

check w/ "1 VAR STATS"

"σx" =

$$\Sigma(x_i - \mu)^2 = 902.8571$$

$$\sigma = \sqrt{\frac{\Sigma(x_i - \mu)^2}{N}} = \sqrt{\frac{902.8571}{7}} \approx 11.36 \text{ minutes}$$

3.2 Measures of Dispersion

3.2.2 Determine the Standard Deviation of a Variable from Raw Data (7 of 16)

The **sample standard deviation**, *s*, of a variable is the square root of the sum of squared deviations about the sample mean **divided by *n* − 1**, where *n* is the sample size.

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}}$$

$$= \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}}$$

where $x_1, x_2, \ldots, x_n$ are the *n* observations in the sample and $\bar{x}$ is the sample mean.

3.2 Measures of Dispersion

3.2.2 Determine the Standard Deviation of a Variable from Raw Data (9 of 16)

We call $n - 1$ the **degrees of freedom** because the first $n - 1$ observations have freedom to be whatever value they wish, but the n^{th} value has no freedom.

It must be whatever value forces the sum of the deviations about the mean to equal zero.

Why do we divide by the degrees of freedom in the sample standard deviation?

This has to do with the idea of "biased" vs "unbiased" statistic. The standard deviation is biased if we divide by n in a sample, so to correct for this, we divide by n-1 which "unbiases" the sample standard deviation.

3.2 Measures of Dispersion

3.2.2 Determine the Standard Deviation of a Variable from Raw Data (10 of 16)

EXAMPLE Computing a Sample Standard Deviation

Here are the results of a random sample taken from the travel times (in minutes) to work for all seven employees of a start-up web development company:

$n = 3$

5, 26, 36

Find the sample standard deviation.

use stats law of rounding

Calc enter into list

~~1 VAR Stats~~

"Sx" sample st. dev

$\boxed{s = 15.8 \text{ min}}$

3.2 Measures of Dispersion

3.2.2 Determine the Standard Deviation of a Variable from Raw Data (11 of 16)

x_i	$\bar{x}$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
5	22.33333	–17.333	300.432889
26	22.33333	3.667	13.446889
36	22.33333	13.667	186.786889

$$\Sigma(x_i - \bar{x})^2 = 500.66667$$

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{500.66667}{2}} \approx 15.82 \text{ minutes}$$

3.2 Measures of Dispersion

3.2.2 Determine the Standard Deviation of a Variable from Raw Data (13 of 16)

IN CLASS ACTIVITY

The Sample Standard Deviation

Using the pulse data we collected from Section 3.1, do the following:

a) Obtain a simple random sample of $n = 4$ students and compute the sample standard deviation.

b) Obtain a second simple random sample of $n = 4$ students and compute the sample standard deviation.

c) Are the sample standard deviations the same? Why?

3.2 Measures of Dispersion

3.2.2 Determine the Standard Deviation of a Variable from Raw Data (14 of 16)

EXAMPLE Comparing Standard Deviations

Determine the standard deviation waiting time for Wendy's and McDonald's. Which is larger? Why?

(Use Calc)

3.2 Measures of Dispersion

3.2.2 Determine the Standard Deviation of a Variable from Raw Data (15 of 16)

σ spread small / values close together

0

more spread

Wait Time at Wendy's

$s = 0.738$

1.50	0.79	1.01	1.66	0.94	0.67
2.53	1.20	1.46	0.89	0.95	0.90
1.88	2.94	1.40	1.33	1.20	0.84
3.99	1.90	1.00	1.54	0.99	0.35
0.90	1.23	0.92	1.09	1.72	2.00

Wait Time at McDonald's

larger!

$s = 1.265$

data is more spread out!

3.50	0.00	0.38	0.43	1.82	3.04
0.00	0.26	0.14	0.60	2.33	2.54
1.97	0.71	2.22	4.54	0.80	0.50
0.00	0.28	0.44	1.38	0.92	1.17
3.08	2.75	0.36	3.10	2.19	0.23

3.2 Measures of Dispersion

3.2.2 Determine the Standard Deviation of a Variable from Raw Data (16 of 16)

EXAMPLE Comparing Standard Deviations

Sample standard deviation for Wendy's:

0.738 minutes

Sample standard deviation for McDonald's:

1.265 minutes

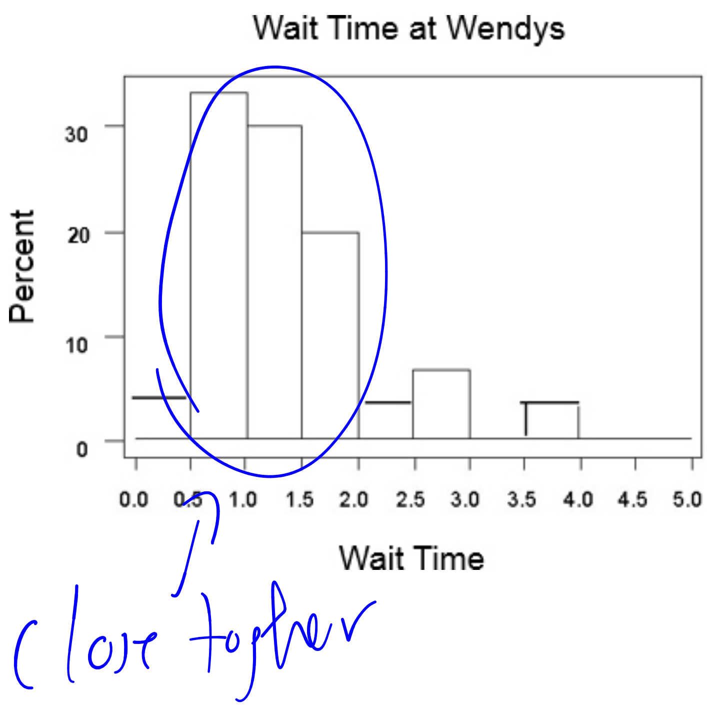
Recall from earlier that the data is more dispersed for McDonald's resulting in a larger standard deviation.

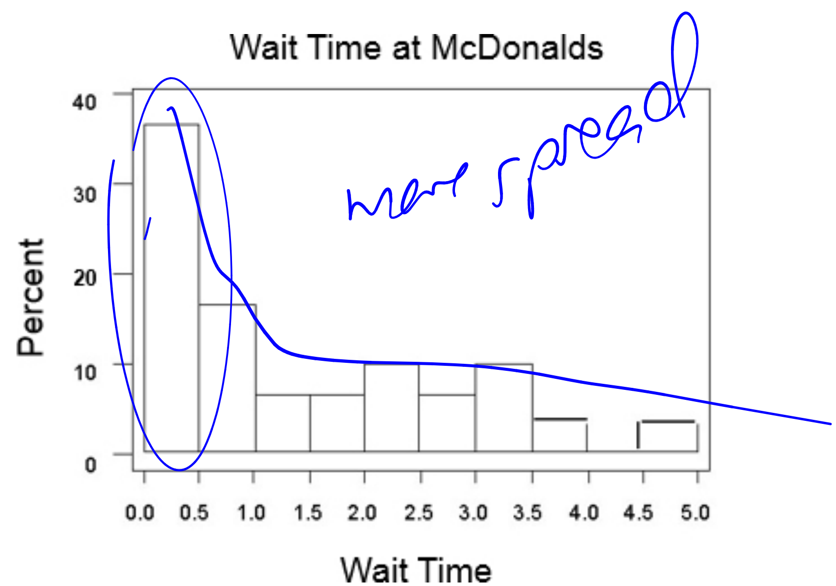
3.2 Measures of Dispersion

3.2.2 Determine the Standard Deviation of a Variable from Raw Data (15 of 16)

The mean wait time in each line is 1.39 minutes.

Histograms for wait time data.





3.2 Measures of Dispersion

3.2.3 Determine the Variance of a Variable from Raw Data (1 of 3)

The **variance** of a variable is the square of the standard deviation.

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N} \qquad \text{vs}$$

"possible test Q"

NOTATION:

- The **population variance** is σ^2
- The **sample variance** is s^2

Units:

- The units of population variance are units squared!
- The units of sample variance are units squared!

3.2 Measures of Dispersion

3.2.3 Determine the Variance of a Variable from Raw Data (2 of 3)

EXAMPLE Computing a Population Variance

The following data represent the travel times (in minutes) to work for all seven employees of a start-up web development company.

23, 36, 23, 18, 5, 26, 43

Compute the population and sample variance of this data.

Note: calculator doesn't compute variance. You need to compute it from the standard deviation by squaring it.

3.2 Measures of Dispersion

3.2.3 Determine the Variance of a Variable from Raw Data (3 of 3)

EXAMPLE Computing a Population Variance

Recall that the population standard deviation (from previous slide) is $\boxed{\sigma = 11.36}$

so the population variance is…

$\sigma^2 = 11.36^2 = 129.0496 = \boxed{129.1 \text{ minutes squared}}$

Recall that the sample standard deviation is $s = 15.82$,

so the sample variance is…

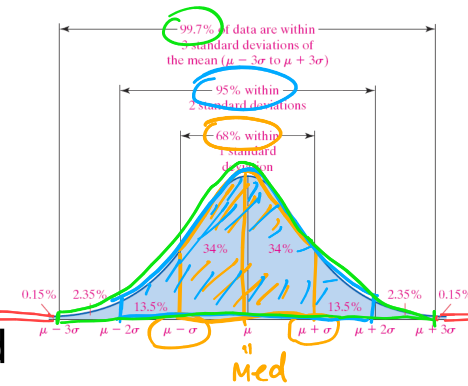
3.2 Measures of Dispersion

3.2.4 Use the Empirical Rule to Describe Data that are Bell Shaped (1 of 7)

68-95-99.7% Rule

The Empirical Rule

If a distribution is roughly bell shaped, then

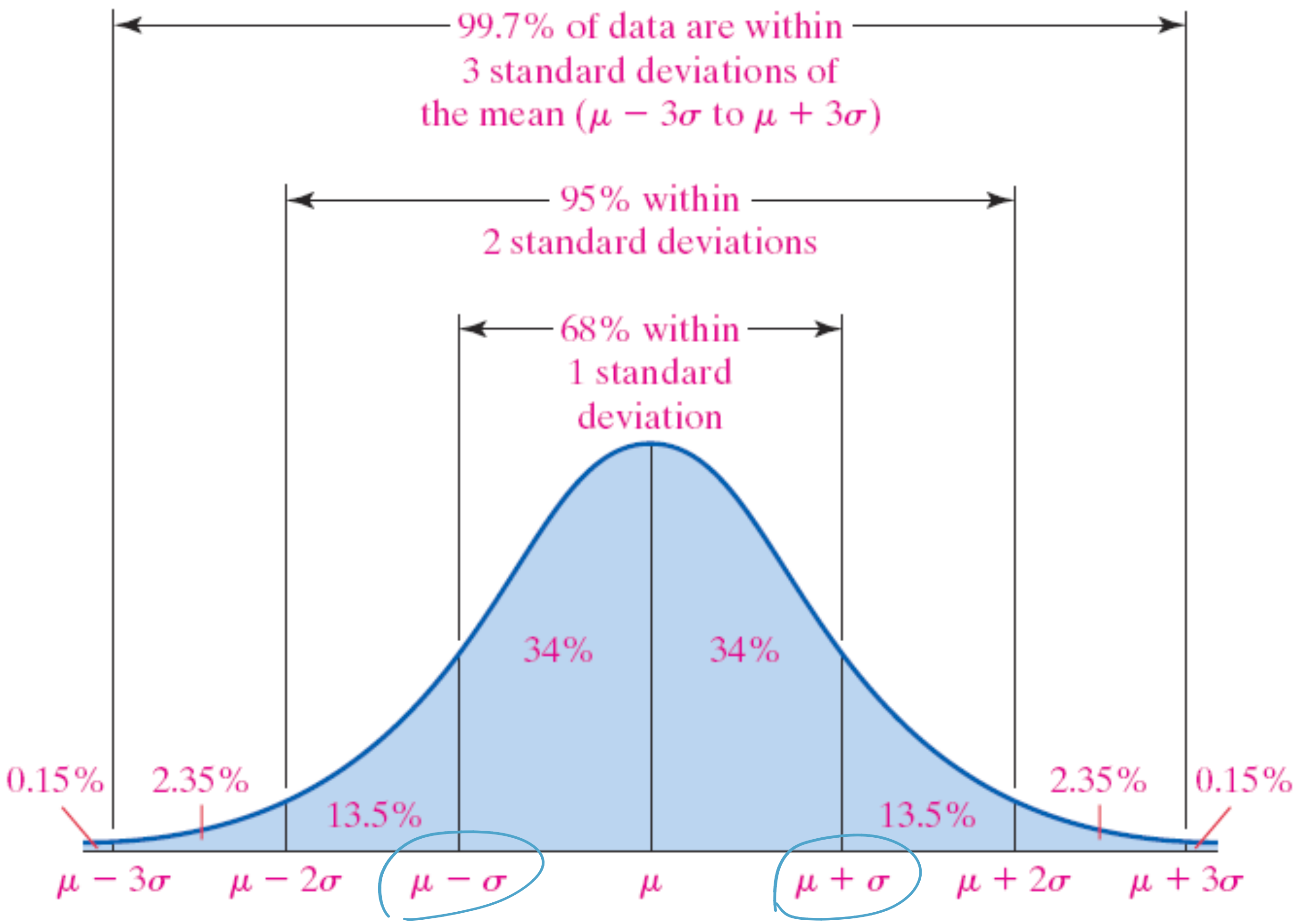


- **Approximately 68%** of the data will lie within 1 standard deviation of the mean. That is, approximately 68% of the data lie between $\mu - \sigma$ and $\mu + \sigma$
- **Approximately 95%** of the data will lie within 2 standard deviations of the mean. That is, approximately 95% of the data lie between $\mu - 2\sigma$ and $\mu + 2\sigma$
- **Approximately 99.7%** of the data will lie within 3 standard deviations of the mean. So, approx. 99.7% of the data lie between $\mu - 3\sigma$ and $\mu + 3\sigma$

Note: We can also use the Empirical Rule based on sample data with $\bar{x}$ used in place of μ and s used in place of σ.

3.2 Measures of Dispersion

3.2.4 Use the Empirical Rule to Describe Data that are Bell Shaped (3 of 7)



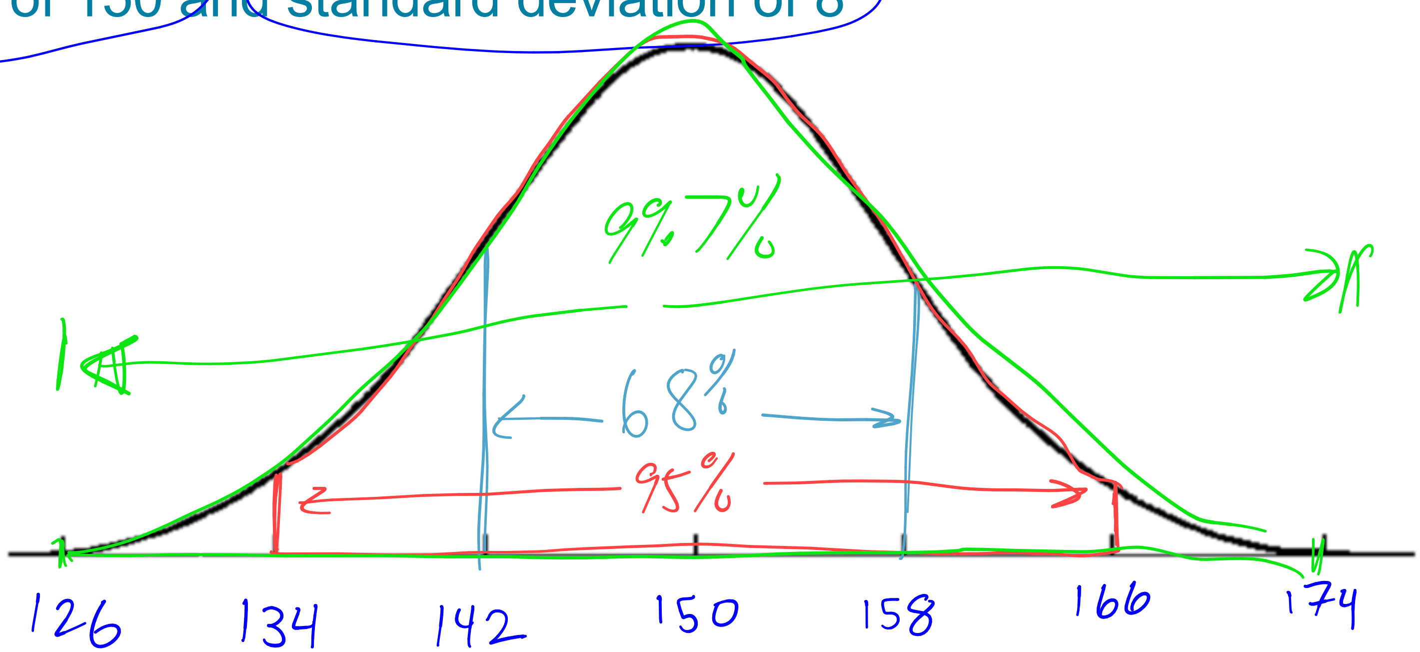
3.2 Measures of Dispersion

3.2.4 Use the Empirical Rule to Describe Data that are Bell Shaped (3 of 7)

$\mu = 150$
$\sigma = 8$

Use the Empirical Rule to fill out the normal distribution:

mean of 150 and standard deviation of 8



3.2 Measures of Dispersion

3.2.4 Use the Empirical Rule to Describe Data that are Bell Shaped (4 of 7)

population

EXAMPLE Using the Empirical Rule

The following data represent the serum HDL cholesterol of the 54 female patients of a family doctor.

41	48	43	38	35	37	44	44	44
62	75	77	58	82	39	85	55	54
67	69	69	70	65	72	74	74	74
60	60	60	61	62	63	64	64	64
54	54	55	56	56	56	57	58	59
45	47	47	48	48	50	52	52	53

The population mean and the standard deviation is:

$$\boxed{\mu = 57.4} \text{ and } \boxed{\sigma = 11.8}$$

3.2 Measures of Dispersion

3.2.4 Use the Empirical Rule to Describe Data that are Bell Shaped (5 of 7)

a) Draw a histogram to verify the data is approximately bell-shaped.

b) Determine the percentage of all patients that have serum HDL within 3 standard deviations of the mean according to the Empirical Rule. Ans 99.7%

c) Determine the percentage of all patients that have serum HDL between 34 and 69.1 according to the Empirical Rule.
Ans $68 + 13.5 = \boxed{81.5\%}$

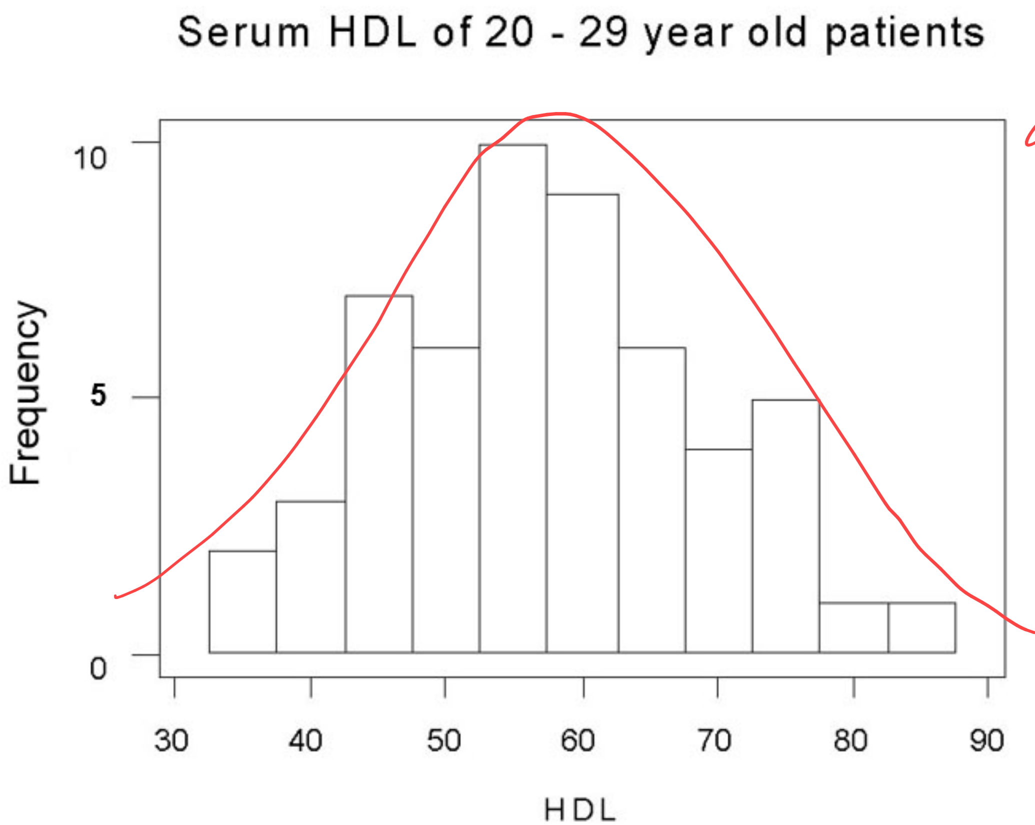
d) Determine the actual ~~percentage~~ # of patients that have serum HDL between 34 and 69.1.

54 patients $* 0.815 = 44.01 \rightsquigarrow$ 45 female patients have HDL levels between 34 & 69.1

Pearson

3.2 Measures of Dispersion

3.2.4 Use the Empirical Rule to Describe Data that are Bell Shaped (6 of 7)



approximately bell shaped

can use Emp. Rule.

3.2 Measures of Dispersion

3.2.4 Use the Empirical Rule to Describe Data that are Bell Shaped (6 of 7)

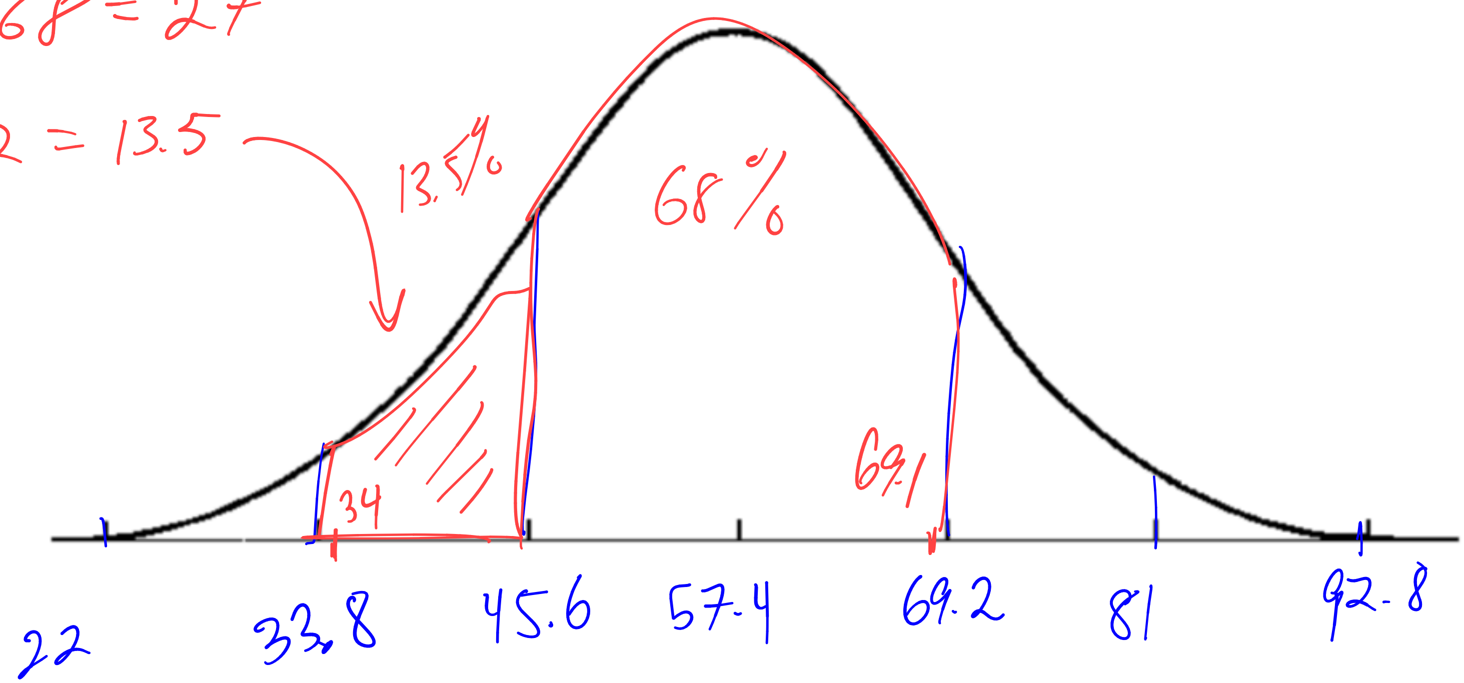
Use the Empirical Rule to fill out the normal distribution:

$\mu = 57.4$

$\sigma = 11.8$

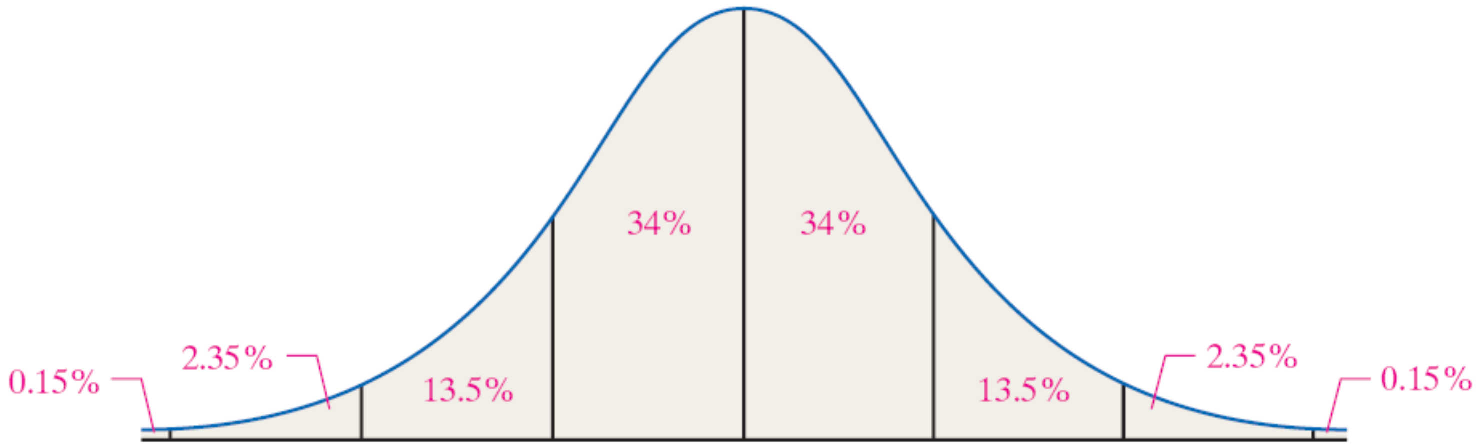
$95 - 68 = 27$

$27/2 = 13.5$



3.2 Measures of Dispersion

3.2.4 Use the Empirical Rule to Describe Data that are Bell Shaped (7 of 7)



(c) According to the Empirical Rule, 99.7% of the all patients that have serum HDL within 3 standard deviations of the mean.

(d) 13.5% + 34% + 34% = 81.5% of all patients will have a serum HDL between 34.0 and 69.1 according to the Empirical Rule.

(e) 45 out of the 54 or 83.3% of the patients have a serum HDL between 34.0 and 69.1.

3.2 Measures of Dispersion

3.2.4 Use the Empirical Rule to Describe Data that are Bell Shaped (7 of 7)

Rounding Rules:

- Stats Law of Rounding
  - When rounding statistics based on data, round (final answers) to one more significant figure than the original data.
  - Example: mean, median, standard deviation, etc
- Miscellaneous
  - When rounding people, always round UP
    - Ex: if we estimate that 23.2 people, then round up to 24 people