1342-Notes_Navidi_3-2_measures-of-spread-Empirical-Rule

Wednesday, September 4, 2019 9:18 AM

1342-Notes_Navidi_3-2_measures-of-spread-Empirical-Rule

3.2: Measures of Spread (Measures of Dispersion)

The mean, median, and mode can describe the "middle" of a data set, but none of them can describe how "spread out" the data is.

Range:

The *range* for ungrouped data is the difference between the largest and smallest values. The *range* for grouped data (a frequency distribution) is the difference between the upper boundary of the highest class and the lower boundary of the lowest class.

In other words,

Range = Maximum − Minimum .

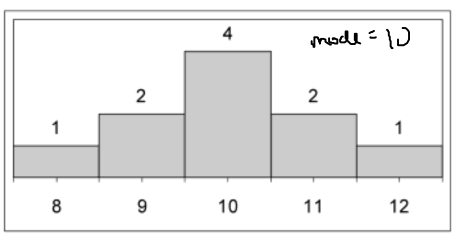
Example 1: Find the range. Range = Max − Min = 1.3 − 0.2 = $\boxed{1.1}$

Commute Times							
0.3	0.7	0.2 min	0.5	0.7	1.2	1.1	0.6
0.6	0.2	1.1	1.1	0.9	0.2	0.4	1.0
1.2	0.9	0.8	0.4	0.6	1.1	0.7	1.2
0.5	1.3 max	0.7	0.6	1.1	0.8	0.4	0.8

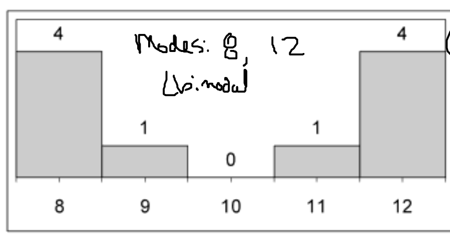
Example 2: Consider these data sets.

$A = \{8, 9, 9, 10, 10, 10, 10, 11, 11, 12\}$, $B = \{8, 8, 8, 8, 9, 11, 12, 12, 12, 12\}$,
$C = \{8, 8, 9, 9, 10, 10, 11, 11, 12, 12\}$

A



B



C

In all:

Mean = 10

Median = 10

Range = Max − Min = 12 − 8 = 4

All are symmetric

Very different-looking data sets

A: 1(8) + 2(9) + 4(10) + 2(11) + 1(12)
= 100

$\mu_A = \frac{100}{10} = 10$

While the range is useful, it is dependent only on the extreme values of the data set. It doesn't tell you whether most of the data points are close to the mean, far from the mean, or evenly distributed. We need something else.

Deviation of a data point:

The deviation of a data point is the difference (i.e., the signed distance) between the data point and the mean.

In other words, the deviation of the *i*th data point, x_i is $x_i - \mu$.
(Note that the deviation is positive if $x_i > \mu$; the deviation is negative if $x_i < \mu$.)

Let's average the deviations for a data set.

Example 3: $A = \{12, 13, 7, 5, 9\}$

$$\mu = \frac{12+13+7+5+9}{5} = 9.2$$

x_i	Deviation $x_i - \mu$
12	$12 - 9.2 = 2.8$
13	$13 - 9.2 = 3.8$
7	$7 - 9.2 = -2.2$
5	$5 - 9.2 = -4.2$
9	$9 - 9.2 = -0.2$
	sum: 0

The deviations always add up to 0.

Variance of a population:

Variance of a population:

If $x_1, x_2, \ldots, x_n$ is a population with mean μ, then the *population variance* σ^2 is given by

know this

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}.$$

In other words, the variance is the average (mean) of the squared deviations.

Alternative formula for the population variance:
(sometimes known as the computational formula, computing formula or shortcut formula)

$$\sigma^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n}$$

(no need to learn this formula)

σ : Greek letter lower case sigma

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

Σ : Upper case sigma

Example 4: Use the ~~alternative (shortcut)~~ formula to find the variance of the population $A = \{12, 13, 7, 5, 9\}$.

From earlier $\mu = 9.2$

$n = 5$

x_i	$x_i - \mu$	$(x_i - \mu)^2$
12	$12 - 9.2 = 2.8$	$(2.8)^2 = 7.84$
13	$13 - 9.2 = 3.8$	$(3.8)^2 = 14.44$
7	$7 - 9.2 = -2.2$	$(-2.2)^2 = 4.84$
5	$5 - 9.2 = -4.2$	$(-4.2)^2 = 17.64$
9	$9 - 9.2 = -0.2$	$(-0.2)^2 = 0.04$
		44.8

$$\sigma^2 = \frac{44.8}{5} = 8.96$$

$\boxed{\text{Variance: } \sigma^2 = 8.96}$

Note: 5 made biggest contribution to the variance; 9 made smallest contribution

Degrees of freedom:

The quantity known as *degrees of freedom* is the number of scores (data points) in a dataset that are free to vary in the presence of a statistical estimate.

If a sample has n data points and the sample mean $\bar{x}$ is specified, then $n-1$ of the data points can theoretically be anything; the nth data point is forced to be take on whatever value results in the specified mean $\bar{x}$. In other words, the first $n-1$ of the data points are free to vary; the nth data point is not free to vary.

statistical estimate $\bar{x}$

Example 5: Suppose a sample has 5 data points and a mean of 159. Suppose also that the first four data points are 37, 203, 122, and 303. Calculate the fifth data point.

Let x = unknown data point

sample mean: $\bar{x} = 159 = \dfrac{37 + 203 + 122 + 303 + x}{5}$

$5(159) = 37 + 203 + 122 + 303 + x$

$795 = 665 + x$

$130 = x$

$\boxed{\text{The unknown data point is 130}}$

There are 4 degrees of freedom for this sample of 5

$\boxed{df = 4}$

Variance of a sample:

When we calculate the variance of a *sample* (not the whole population), we have no way to calculate the population mean. Therefore, we must use the sample mean (denoted $\bar{x}$) as an estimate of the population mean (denoted μ). Thus, in a sample of n data points, there are $n-1$ degrees of freedom.

When calculating the variance for a sample (not the entire population), we divide by $n-1$ (the degrees of freedom) instead of n. Dividing by n would underestimate the variance, because the points in the sample will be less spread out than those in the population. Using the degrees of freedom, $n-1$, in the denominator provides an unbiased estimate of the population variance.

Variance of a sample:

The *sample variance* s^2 of a set of n sample measurements $x_1, x_2, \ldots, x_n$ with mean $\bar{x}$ is given by

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

Those values are called parameters: μ: population mean; σ^2: population variance

Those values are called statistics: $\bar{x}$: sample mean; s^2: sample variance

Alternative formula for the sample variance:
(sometimes known as the computational formula, computing formula or shortcut formula)

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

(skip)

Example 6: Calculate the variance of this sample: {75, 16, 50, 88, 79, 95, ~~86~~ 81}.

x_i	$(x_i - \bar{x})^2$
75	$(75-69.14)^2 = (75-A)^2 = 34.3061$
16	$(16-A)^2 = 2824.1633$
50	366.4490
88	$(88-A)^2 = 355.5918$
79	$(79-A)^2 = 97.1633$
95	$(95-A)^2 = 668.5918$
80	$(80-A)^2 = 117.8776$
$\Sigma x_i = 483$	4464.1429

sample mean: $\bar{x} = \frac{75+16+50+88+79+95+\cancel{86}\,81}{7}$

$\bar{x} = \frac{484}{7} \approx 69.1428571$

sample mean $\boxed{\bar{x} \approx 69.14}$

store in A

$s^2 = \frac{4464.1429}{7-1} = \frac{4464.1429}{6}$

$= \boxed{744.02 = s^2}$ (sample variance

Standard deviation:

By squaring the deviations, we've changed the units (if there are any). In other words, if we started with "inches", we now have "square inches". This is easily fixed by taking square roots.

std dev:

$s = \sqrt{744.02} \approx \boxed{27.277}$ sample std deviation

Standard Deviation:

The *sample standard deviation* s of a set of n sample measurements $x_1, x_2, \ldots, x_n$ with mean $\bar{x}$ is given by

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

If $x_1, x_2, \ldots, x_n$ is the whole population with mean μ, then the *population standard deviation* σ is given by

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}}.$$

Example 7: Given the following data sample, calculate the standard deviation to two decimal places.

{70, 39, 54, 84, 68, 93, 75}

$\bar{x} = \frac{\sum x_i}{n} = \frac{483}{7} = 69$

x_i	$(x_i - \bar{x})^2$
70	$(70-69)^2 = 1^2$
39	$(39-69)^2 = (-30)^2 = 900$
54	$(54-69)^2 = (-15)^2 = 225$
84	$(84-69)^2 = (15)^2 = 225$
68	$(68-69)^2 = (-1)^2 = 1$
93	$(93-69)^2 = (24)^2 = 576$
75	$(75-69)^2 = 6^2 = 36$
$\sum x_i = 483$	$\sum (x_i - \bar{x})^2 = 1964$

Variance: $s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$

$= \frac{1964}{7-1} \approx 327.33$

Std. Deviation:

$s = \sqrt{s^2} \approx \sqrt{327.33}$

≈ 18.09

$\boxed{\text{Variance. } s^2 \approx 327.33 \quad \text{Std. Dev. } s \approx 18.09}$

IMPORTANT:

$$\text{Standard Deviation} = \sqrt{\text{Variance}}$$

$$(\text{Standard Deviation})^2 = \text{Variance}$$

Approximating the variance and standard deviation of a frequency distribution:

Variance and standard deviation for grouped data:

Suppose a data set of n sample measurements is grouped into k classes in a frequency table, where x_i is the midpoint and f_i is the frequency of the ith class interval.

Then the *sample variance* s^2 for the grouped data (with mean $\bar{x}$) is given by

$$s^2 = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 f_i}{n-1} \text{ or } s^2 = \frac{\sum_{i=1}^{k} f_i x_i - n\bar{x}^2}{n-1} \text{ (both equivalent)}$$

where $n = \sum_{i=1}^{k} f_i$ = total number of measurements .

Then the *sample standard deviation* s for the grouped data (with mean $\bar{x}$) is given by

$$s = \sqrt{\frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 f_i}{n-1}} \text{ or } s = \sqrt{\frac{\sum_{i=1}^{k} f_i x_i - n\bar{x}^2}{n-1}} \text{ (both equivalent)}$$

where $n = \sum_{i=1}^{k} f_i$ = total number of measurements .

(Assume it's a sample)

Example 1: ~~Find~~ Estimate the variance and standard deviation for the grouped data.

midpoints x_i	Score	Frequency	$(x_i - \bar{x})^2 \cdot f$
$(70+60)/2 = 65$	60–69	12	$(65-80.1163)^2 \cdot 12 = 2742.0227$
75	70–79	7	$(75-80.1163)^2 \cdot 7 = 183.2347$
85	80–89	14	$(85-80.1163)^2 \cdot 14 = 333.9102$
95	90–99	10	$(95-80.1163)^2 \cdot 10 = 2215.2515$
		$n = 43$	$\Sigma = 5474.4186$

$$\bar{x} = \frac{12(65)+7(75)+14(85)+10(95)}{43} = \frac{3445}{43} \approx 80.1163$$

(store in calculator)

Variance: $s^2 = \frac{5474.4186}{43-1} = \frac{5474.4186}{42} \approx 130.3433$

Std deviation: $s = \sqrt{s^2} \approx \sqrt{130.343} \approx 11.4168$

$$\boxed{\begin{array}{l} \text{mean: } \bar{x} \approx 80.17 \\ \text{variance: } s^2 \approx 130.34 \\ \text{std dev: } s \approx 11.42 \end{array}}$$

The Empirical Rule:

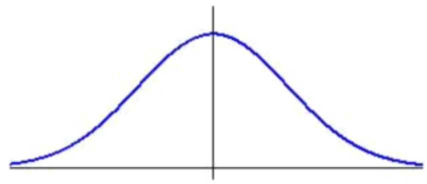
The Empirical Rule:

If a distribution is roughly bell-shaped, then

a) About 68% of the data lie within one standard deviation of the mean.
(or about two-thirds)

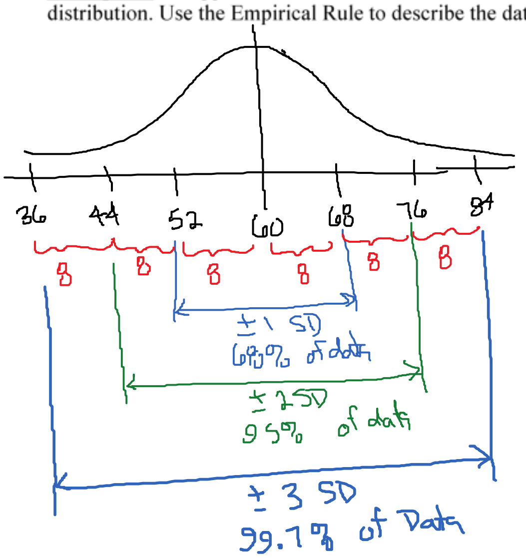
b) About 95% of the data lie within two standard deviations of the mean.

c) About 99.7% of the data lie within three standard deviations of the mean.



Standard Deviation is often abbreviated SD.

Example 8: Suppose a data set has mean 60 and standard deviation 8, and has a bell-shaped distribution. Use the Empirical Rule to describe the data set.



About 68% of data is between 52 and 68.

About 95% of data is between 44 and 76.

About 99.7% of data is between 36 and 84.

OR

About 68% in (52, 68).
About 95% in (44, 76)
About 99.7% in (36, 84)

bell-shaped

These 120 numbers were randomly generated from a $\boxed{\text{normal}}$ distribution with mean 60 and standard deviation 8 (same mean and standard deviation as previous example).
(Used Data Analysis ToolPak in Excel)

within 1 SD
80 data points
$\frac{80}{120} \Rightarrow 66.67\%$
(close to 68%)

39.37935	50.39057	54.76075	57.07606	60.01561	63.73369	66.06171	70.43123
42.5313	51.07409	54.76757	57.09698	60.22494	63.89758	66.42436	70.43984
43.05655	51.30639	54.77136	57.31554	60.33393	64.10566	66.62665	70.74113
45.22471	51.5626	54.90775	57.38407	60.57791	64.31159	66.63873	71.55004
45.80155	51.78455	55.13357	57.41827	60.90368	64.40457	66.79456	71.58136
45.91251	52.17896	55.4566	57.51029	61.05229	64.46238	66.79544	73.04499
46.06014	52.4813	55.60325	57.59814	61.07882	64.71196	66.89606	73.29165
46.47654	52.61447	55.80964	58.07242	61.10972	64.76593	66.92538	73.86506
47.10082	52.71231	55.80964	58.12655	61.70779	64.94325	66.99687	74.82269
47.52886	53.2221	55.89434	58.51074	61.95406	65.10725	67.21753	75.07877
47.82743	53.28063	56.37311	59.06982	62.43132	65.40111	68.76018	75.35132
48.44651	53.43097	56.56806	59.31772	62.58109	65.53095	68.88951	75.7777
49.0252	53.81194	56.76762	59.32386	63.5538	65.55196	69.5868	77.55601
49.76189	54.10818	56.94941	59.74017	63.62961	65.6906	69.90556	77.64551
49.77853	54.47837	57.03808	59.79588	63.65133	66.06089	70.21179	79.00524

within 2 SD
$\frac{114}{120} \Rightarrow 95\%$ exactly

How many data points are within 1 standard deviation of the mean?
Thus, how many lie in interval [52, 68]?

80 points so 68% of data lie in [52,68]

How many data points are within 2 standard deviations of the mean?
Thus, how many lie in interval [44, 76]?

114 points so 95% of data

How many data points are within 3 standard deviations of the mean?
Thus, how many lie in interval [36, 84]?

All of them! So 100% of data

This time, I only generated 50 random numbers:

37.48295	53.07996	54.94143	57.29676	60.95421	63.16013	66.48368
44.16628	53.20456	55.31494	57.37955	61.43636	63.3329	67.79641
47.40146	54.25092	55.90412	58.99277	61.91373	63.39574	67.85567
50.54246	54.41687	56.48669	59.10063	62.14886	63.61741	68.12883
51.77209	54.42467	56.64306	59.74812	62.52829	63.79043	68.23415
52.06366	54.87849	56.84053	60.00459	62.58302	63.93691	70.72309
53.05055	54.91449	57.00922	60.59386	63.15417	66.44211	73.3014
						87.41814

within 1 SD

$\frac{40}{50} = 80\%$ of data

within 2 SD

$\frac{48}{50} \Rightarrow 96\%$

within 3 SD

$\frac{49}{50} \Rightarrow 98\%$

How many data points are within 1 standard deviation of the mean?
Thus, how many lie in interval [52, 68]?

40 data points so 80% of data (expected 68%)

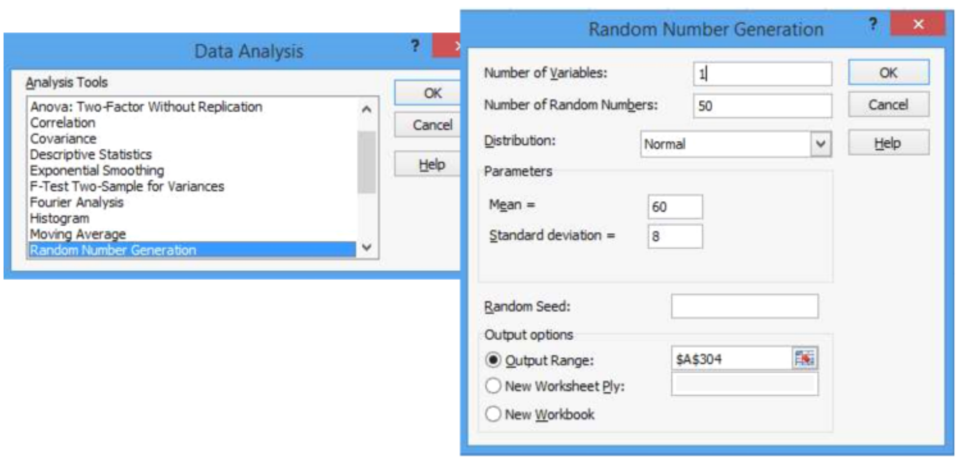
How many data points are within 2 standard deviations of the mean?
Thus, how many lie in interval [44, 76]?

48 data points so 96% of data (expected 95%)

How many data points are within 3 standard deviations of the mean?
Thus, how many lie in interval [36, 84]?

49 points so 98% of data (expected 99.7%)

Some screen shots of the process:



One more time, this time with 300 random numbers:

38.84326	50.86698	54.68468	57.27472	60.58405	63.20981	66.17321	70.96541
42.21232	51.26652	54.72886	57.33431	60.58834	63.21181	66.19631	70.96854
42.98081	51.34825	54.74558	57.35177	60.6086	63.30223	66.33855	70.97636
44.72886	51.37784	54.84689	57.62502	60.67183	63.30689	66.6846	71.418
45.00709	51.4193	54.93544	57.91578	60.68043	63.31289	66.92538	71.52752
45.2281	51.52729	55.12767	57.92022	60.77998	63.36297	66.99418	71.54829
45.74502	51.70324	55.13577	57.92149	60.85565	63.39709	67.01304	71.92155
45.77783	51.7513	55.19161	57.93034	60.94435	63.41921	67.02743	72.24118
46.0039	51.95887	55.21724	58.01314	61.00847	63.58963	67.02743	72.57955
46.27296	51.99328	55.36209	58.17746	61.06586	63.63978	67.02832	72.71404
46.51216	52.16414	55.36352	58.24773	61.08006	63.6826	67.11516	72.97609
46.65451	52.20065	55.40834	58.47087	61.13753	63.72347	67.23043	74.75848
47.0421	52.21638	55.63099	58.49703	61.15298	63.82608	67.43723	74.97872
47.38029	52.25753	55.6459	58.51696	61.20619	63.95696	67.51395	75.81204
47.46869	52.35761	55.68063	58.67048	61.23158	64.05129	67.52156	75.90372
47.58631	52.51833	55.69762	58.8773	61.24397	64.10915	67.68499	76.259
47.64513	52.53727	55.74422	58.96932	61.40466	64.21565	67.68694	77.11451
47.77654	52.56936	55.77662	59.0495	61.45752	64.21987	68.05128	79.49171
47.80597	52.57689	55.8447	59.20525	61.56655	64.37114	68.29572	80.17856
48.32714	52.85754	55.91388	59.22862	61.61336	64.47955	68.3747	81.70971
48.42039	52.90481	55.92015	59.33492	61.72031	64.66111	68.4129	
48.76869	53.05322	55.989	59.39263	61.90177	64.71051	68.54488	
48.8517	53.26581	56.00426	59.43375	61.96163	64.76666	68.67374	
48.92874	53.56296	56.12376	59.49447	61.96414	64.826	68.67927	
49.12779	53.7855	56.20271	59.52696	62.05953	64.83261	68.97397	
49.34563	53.81442	56.24719	59.67464	62.1406	64.93807	69.37314	
49.4033	53.85067	56.26084	59.71445	62.17298	64.94917	69.38531	
49.56304	53.90153	56.26425	59.72853	62.17488	65.08102	69.42562	
49.72711	53.90808	56.3038	59.72914	62.36156	65.14934	69.49211	
49.76327	54.14264	56.47927	59.74996	62.38202	65.22106	69.54926	
49.9019	54.20486	56.50083	59.88402	62.46146	65.28407	69.62202	
49.90597	54.20726	56.57476	59.96481	62.46275	65.39265	69.70709	
49.90734	54.36443	56.64639	60.21147	62.47109	65.39341	69.78032	
49.99756	54.36757	56.67377	60.21637	62.5985	65.42419	69.79844	
50.10628	54.37463	56.73572	60.24942	62.73436	65.52473	69.88977	
50.14039	54.555	56.83193	60.34252	62.75838	65.6011	69.93328	
50.23903	54.55577	56.88944	60.39458	62.81691	65.64185	69.97974	
50.24676	54.58043	56.94217	60.48836	62.86384	65.64969	70.0629	
50.33877	54.59967	57.00134	60.49266	62.95471	65.73562	70.68564	
50.57683	54.66558	57.09895	60.50614	63.09341	65.83657	70.76531	

How many data points are within 1 standard deviation of the mean?
Thus, how many lie in interval [52, 68]?

How many data points are within 2 standard deviations of the mean?
Thus, how many lie in interval [44, 76]?

How many data points are within 3 standard deviations of the mean?
Thus, how many lie in interval [36, 84]?

You shouldn't expect that any sample will match the Empirical Rule exactly. However, it should be close, especially with a large sample.

Example 9: The mean value from a sample of used cars is \$2400, with a standard deviation of \$450. Between what two values should about 95% of the data lie? Assume the data is approximately bell-shaped.

Bell shaped, so about 95% lie within 2 SDs of mean

$\bar{x} \pm s = \$2400 \pm \$450$

$2400 + 450 = 2850$

$2400 - 450 = 2350$

Chebyshev's Inequality:

> Chebyshev's Rule (Chebyshev's Inequality):
>
> For any data set or distribution, at least $1-\frac{1}{k^2}$ of the data points lie within k standard deviations of the mean, where k is any number greater than 1.
> (In other words, at least $1-\frac{1}{k^2}$ of the observations lie in the interval $[\bar{x}-ks, \bar{x}+ks]$.
>
> Note: Chebyshev's Inequality is true even when the distribution is not bell-shaped.

Example 2: What does Chebyshev's Inequality tell us for $k=1$, $k=2$, $k=3$, $k=4$?

$k=2 \Rightarrow 1-\frac{1}{k^2} = 1-\frac{1}{2^2} = 1-\frac{1}{4} = \frac{3}{4} \Rightarrow 75\%$

At least 75% of data lie within 2 SDs of the mean

$k=3 \Rightarrow 1-\frac{1}{k^2} = 1-\frac{1}{3^2} = 1-\frac{1}{9} = \frac{8}{9} \Rightarrow 88.9\%$

At least 88.9% lie within 3 SDs of the mean

$k=4 \Rightarrow 1-\frac{1}{k^2} = 1-\frac{1}{4^2} = 1-\frac{1}{16} = \frac{15}{16} \Rightarrow 93.75\%$

At least 93.75% are within 4 SDs of the mean

Example 10: Suppose the mean time for women's 200 meter track athletes is 57.07 seconds with a standard deviation of 1.05. The shape of the data distribution is unknown. Find the interval that contains at least 75% of the data.

at least 75% of data lie within 2 SDs.

$57.07 + 2(1.05) = 59.17$

$57.07 - 2(1.05) = 54.97$

At least 75% of data lie within $[54.97, 59.17]$

Example 11: Suppose a data set includes 120 observations. At least how many observations lie within three standard deviations of the mean?

Example 12: Suppose a data set includes 68 observations and has mean 55 and standard deviation 7.5. At least how many observations lie between 40 and 70?

The coefficient of variation:

The coefficient of variation (CV) describes how large the standard deviation is, expressed as a proportion of the mean. This lets us compare the spreads of data sets that have different means.

For example, a CV of 0.2 means the standard deviation is 20% of the mean. A standard deviation of 0.34 means the standard deviation is 34% of the mean.

Coefficient of Variation (CV):

To find the coefficient of variation, divide the standard deviation by the mean.

$$CV = \frac{\sigma}{\mu}$$

Example 13: