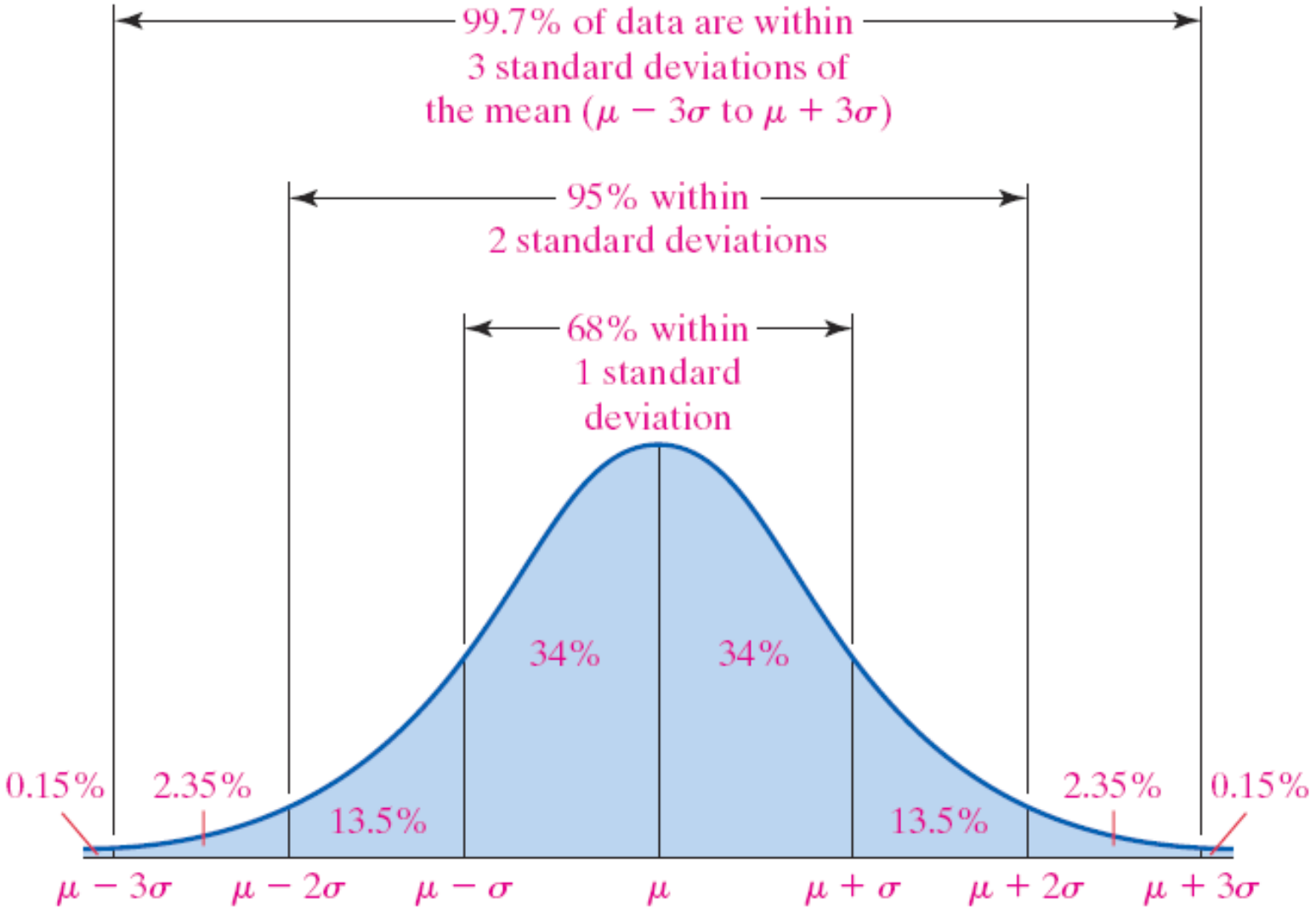
The Empirical Rule

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 $\mu - 1\sigma$ and $\mu + 1\sigma$.
- 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. That is, approximately 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.


99.7% of data are within
3 standard deviations of
the mean ($\mu - 3\sigma$ to $\mu + 3\sigma$)
95% within
2 standard deviations
68% within
1 standard
deviation
34%
34%
0.15%
2.35%
13.5%
13.5%
2.35%
0.15%
$\mu - 3\sigma$
$\mu - 2\sigma$
$\mu - \sigma$
μ
$\mu + \sigma$
$\mu + 2\sigma$
$\mu + 3\sigma$

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

(a) Compute the population mean and standard deviation.

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

(c) Determine the percentage of patients that have serum HDL within 3 standard deviations of the mean according to the Empirical Rule.

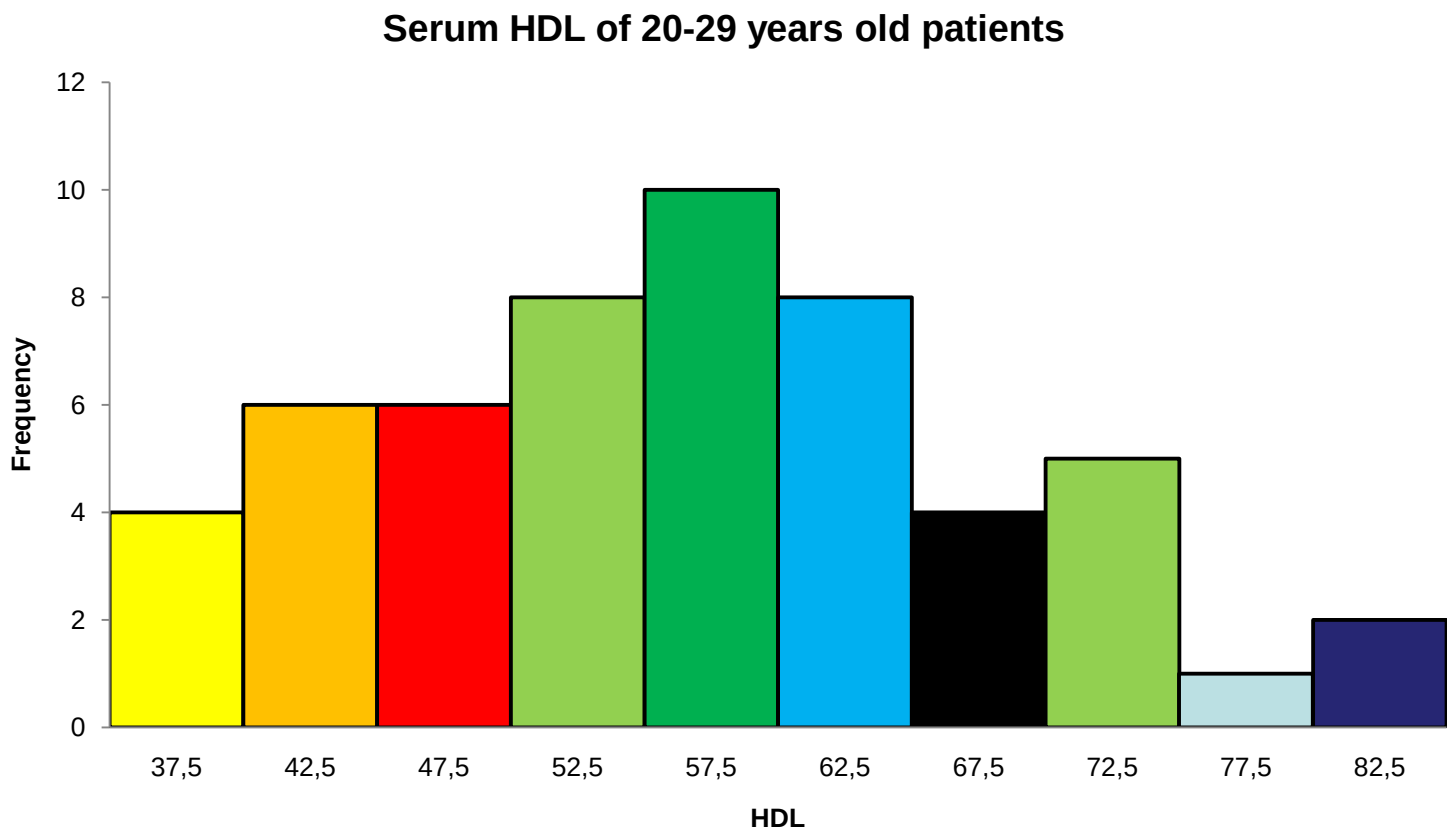
(d) Determine the percentage of patients that have serum HDL between 34 and 69.1 according to the Empirical Rule.

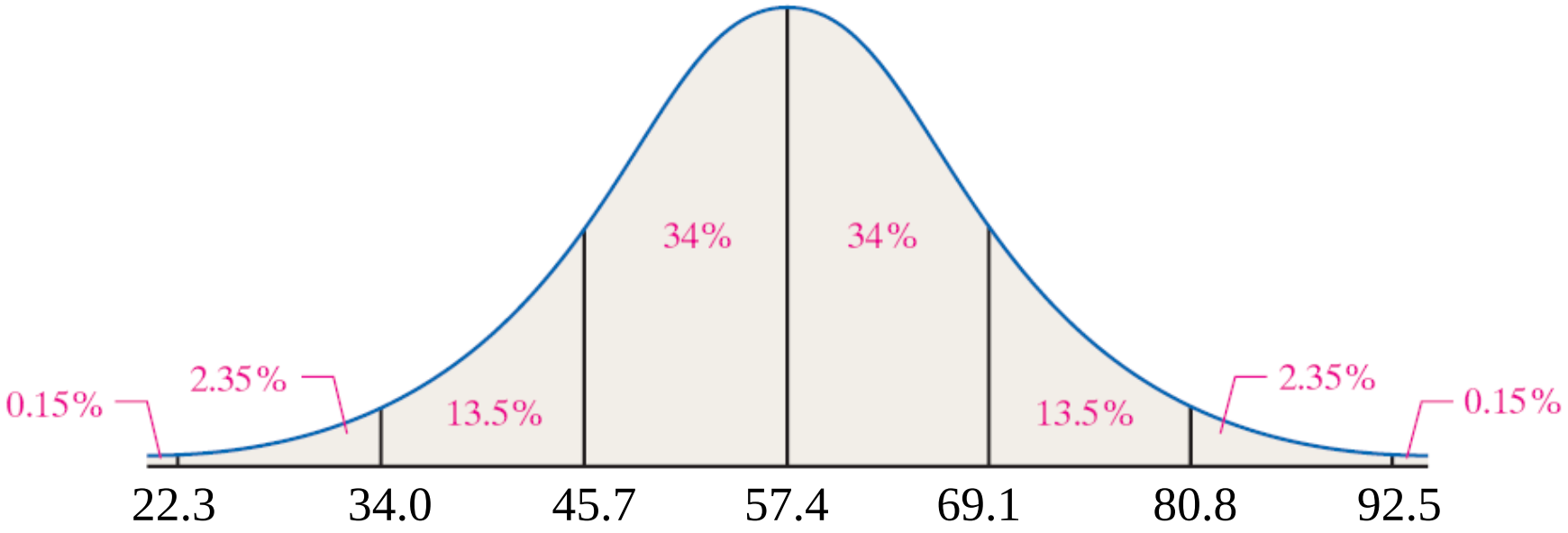
(e) Determine the actual percentage of patients that have serum HDL between 34 and 69.1.

(a) Using MS Excel, we find

$$\mu \approx 57.4 \quad \text{and} \quad \sigma \approx 11.7$$

(b)

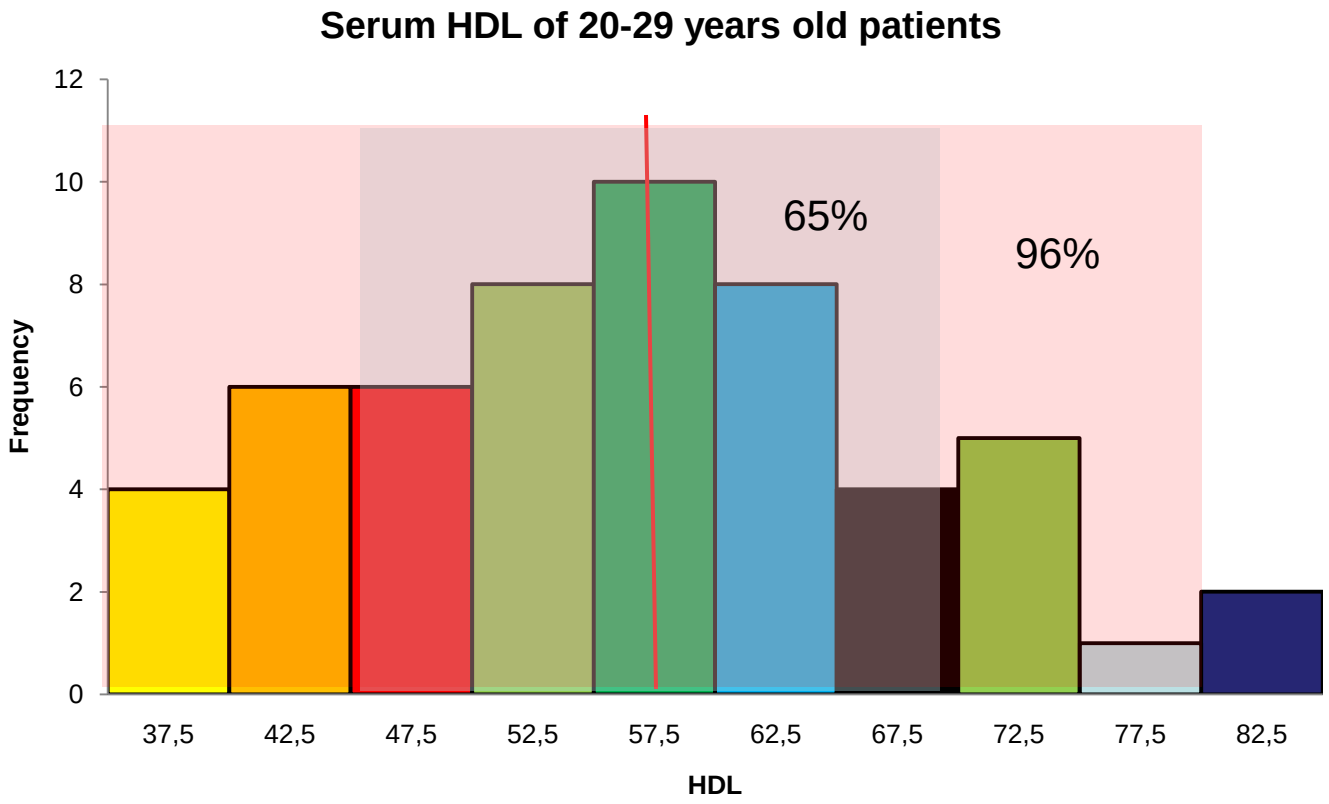




(c) According to the Empirical Rule, 99.7% of the patients that have serum HDL within 3 standard deviations of the mean.

(d) 13.5% + 34% + 34% = 81.5% of 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.


Serum HDL of 20-29 years old patients
12
10
8
6
4
2
0
Frequency
65%
96%
37,5
42,5
47,5
52,5
57,5
62,5
67,5
72,5
77,5
82,5
HDL

- Use Chebyshev's Inequality to Describe Any Set of Data

Chebyshev's Inequality

For any data set, regardless of the shape of the distribution, at least $\left(1 - \frac{1}{k^2}\right)100\%$ of the observations will lie within k standard deviations of the mean, where k is any number greater than 1. That is, at least $\left(1 - \frac{1}{k^2}\right)100\%$ of the data will lie between $\mu - k\sigma$ and $\mu + k\sigma$ for $k > 1$.

Note: We can also use Chebyshev's Inequality based on sample data.

EXAMPLE *Using Chebyshev's Theorem*

Using the data from the previous example, use Chebyshev's Theorem to

(a) determine the percentage of patients that have serum HDL within 3 standard deviations of the mean.

$$\left(1-\frac{1}{3^2}\right)\cdot 100\% = 88.9\%$$

(b) determine the actual percentage of patients that have serum HDL between 34 and 80.8.

$$\left(1-\frac{1}{2^2}\right)\cdot 100\% = 75\%$$