The Empirical Rule

For <u>only</u> symmetrical bell-shaped frequency distributions, an approximation can be made of the percentage of data within a specified standard deviation from the mean

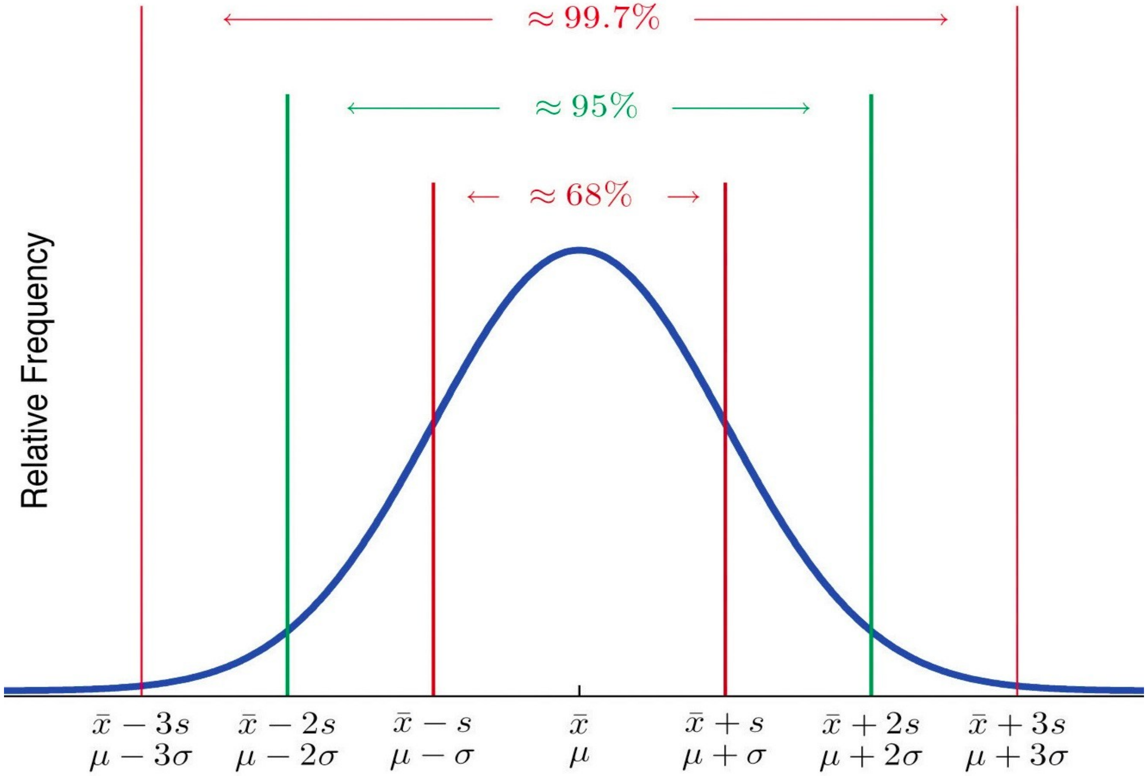
Approximately 68% of the observations lie within ±1s of the mean

About 95% of the observations will lie within ±2s of the mean

Practically all (99.7%) will lie within ±3s of the mean

The Empirical Rule is <u>not</u> valid for non-symmetrical bell shaped distributions

Instead, Chebyshev's theorem is a general rule that applies to other non-symmetrical distributions



Example – Rental Rates

A sample of the rental rates for some apartments approximates a symmetrical, bell-shaped distribution. The $\bar{x}$ = \$600 and the s = \$24. Using the Empirical Rule, answer these questions

(a) About 68% of the rental rates are between what two amounts?

(b) About 95% of the rental rates are between what two amounts?

(c) Almost all of the rental rates are between what two amounts?

(a) About 68% are between ±1s …... ±24
\$576 (\$600 - \$24) and \$624 (\$600 + \$24)

(b) About 95% are between ±2s …... ±48
\$552 (\$600 - 48) and \$648 (\$600 + 48)

(c) Almost all (99.7%) are between ±3s …... ±72
\$528(\$600 - 72) and \$672 (\$600 + 72)

Standard Deviation and Chebyshev's Theorem

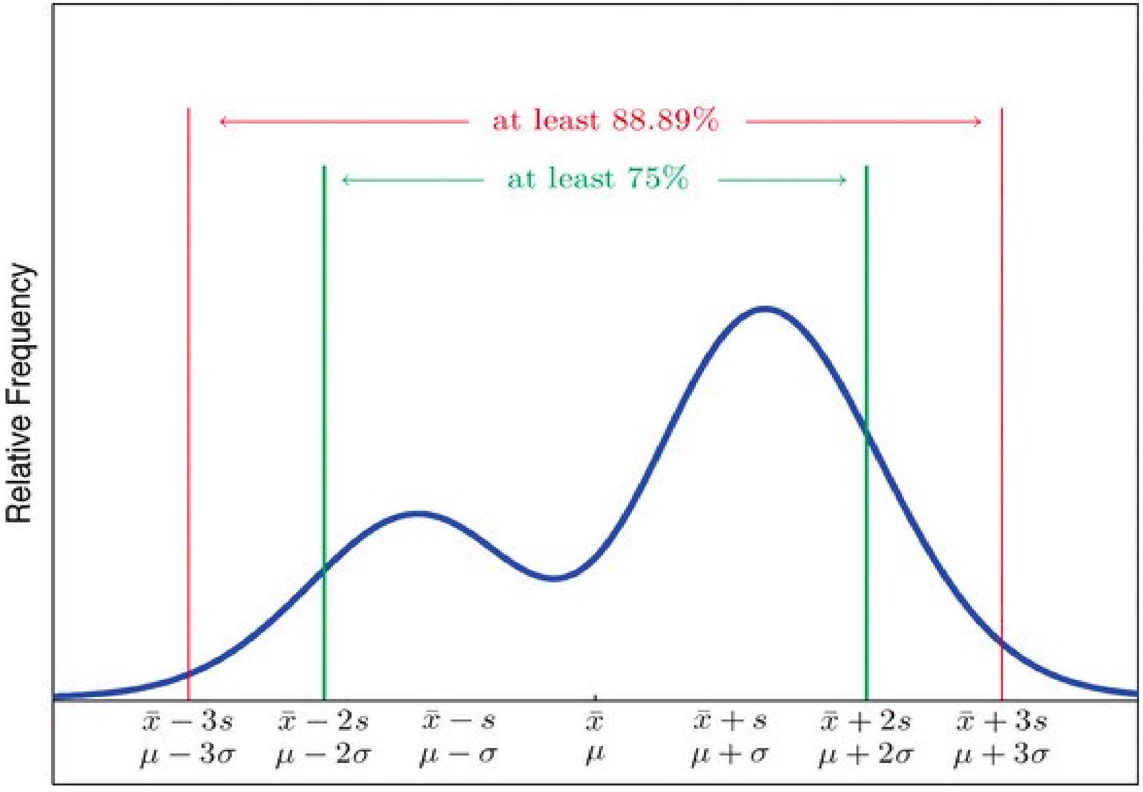
A small standard deviation indicates that the values are located close to the mean. Conversely, a large standard deviation indicates that the values are scattered about the mean.

P. L. Chebyshev (1821-1894) developed a theorem to determine the minimum proportion of values that lie within a specified number of standard deviations from the mean <u>for any data set</u>

3 out of 4 values (75%) lie within ±2s of the mean

8 out of 9 values (88.9%) lie within ±3s of the mean

24 out of 25 values (96%) lie within ±5s of the mean



Chebyshev's Theorem

The theorem applies regardless of the shape of the distribution.

For any set of observations (sample or population), the proportion of the values that lie within k standard deviations of the mean is at least

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

where k is any constant greater than 1. It is usually found in this form with respect to the standard deviation values... ks or kσ.

The change between the mean value ($\bar{x}$ or μ) and the upper or lower boundary can be regarded as Δ = ks or Δ = kσ.

Example – Student Attendance

The daily average number of students who attend a class is $\bar{x}$ = 89.90, and the standard deviation is s = 11.31. At least what percent of students attendance lies within k = 3.5 standard deviations from the mean?

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

$$1-\frac{1}{3.5^2}=1-\frac{1}{12.25}=0.918$$

92% of the observations fall between ±3.5 standard deviations.

Comparison of the Empirical Rule and Chebyshev's Theorem

	% of Values Found in Intervals Around the Mean	
Interval	**Chebyshev's Theorem (any distribution)**	**Empirical Rule (normal distribution)**
$\mu \pm 1\sigma$, $\bar{x} \pm 1s$	~ 0%	~ 68%
$\mu \pm 2\sigma$, $\bar{x} \pm 2s$	~ 75%	~ 95%
$\mu \pm 3\sigma$, $\bar{x} \pm 3s$	~ 88.89%	~ 99.7%

For the normal distribution, to determine the % for any other non-integer multiple of s or σ...(e.g. 1.3s or 2.8s), we would use a z-table

The z-score

The **z-score** of a data point is the difference between that value and the mean value divided by the standard deviation.

If z-score = 0, then it means that the x value is equal to the mean value

If z-score is very small (negative) value or very large (positive) value, then it means the x value is an **outlier**.

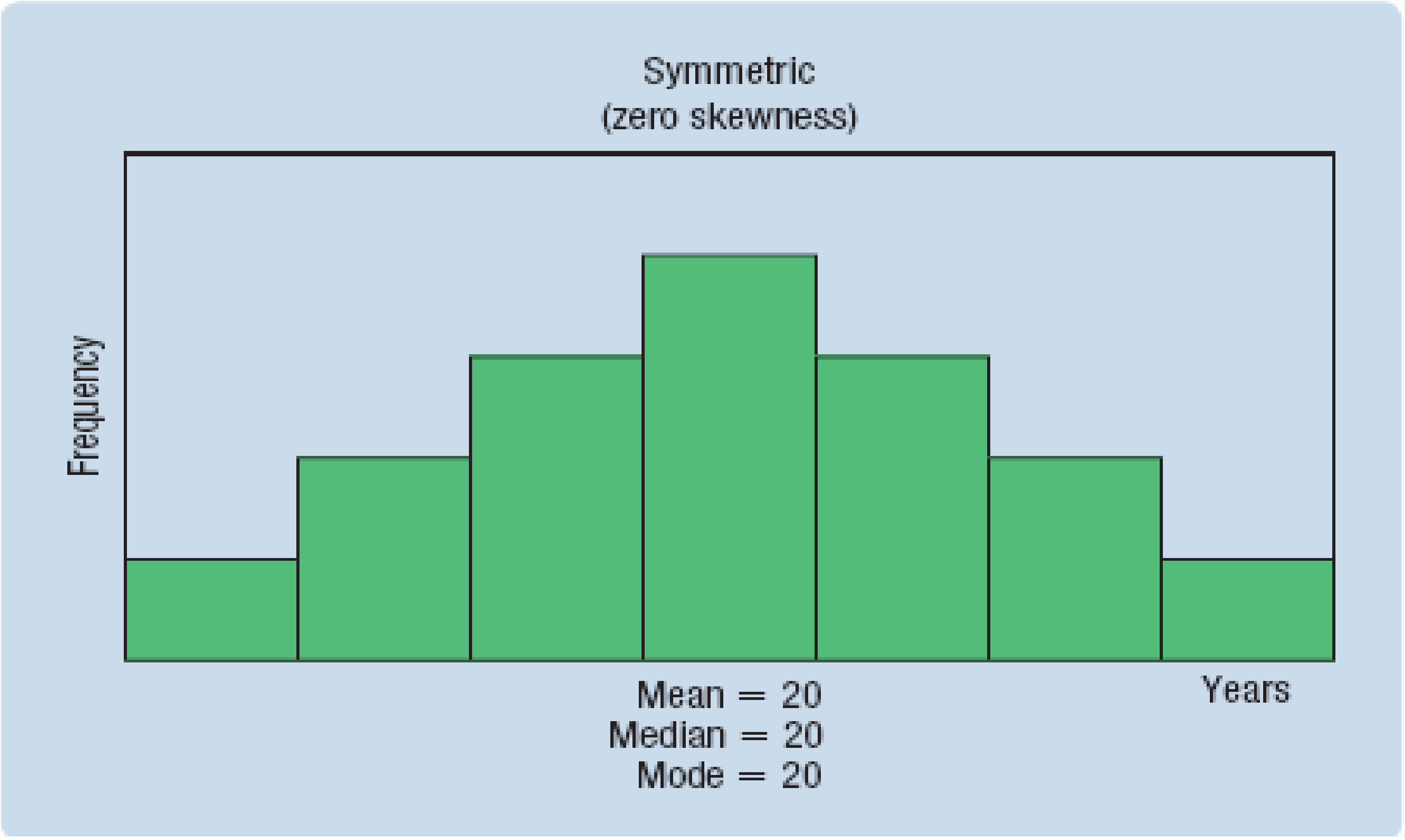
In general, z-scores > +3.0 or z-score < -3.0 indicate an outlier

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

Shape: Skewness

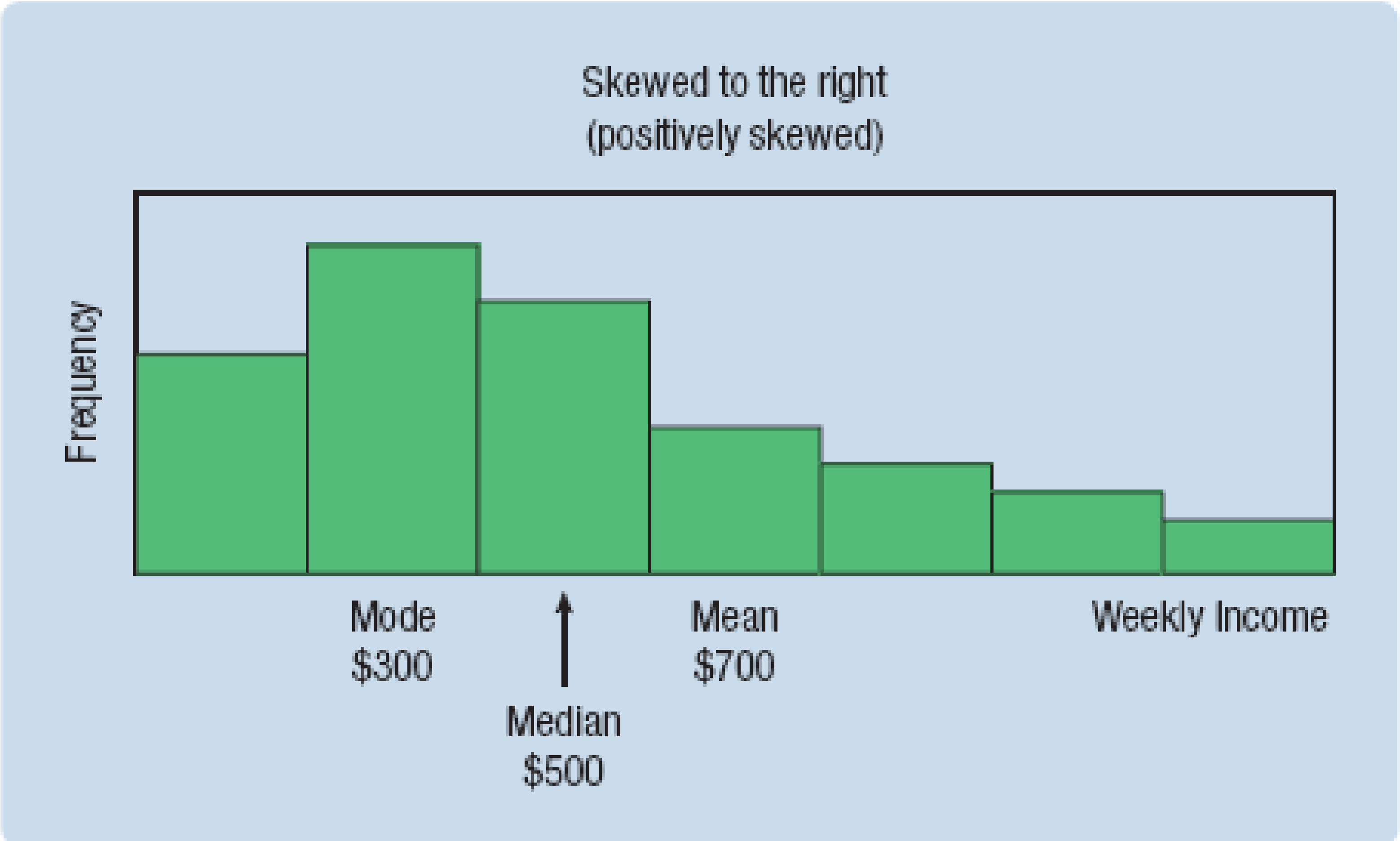
Skewness measures the extent to which the data values are not symmetrical around the mean

For a symmetric unimodal distribution, the mode, median and mean are located at the centre and are always equal



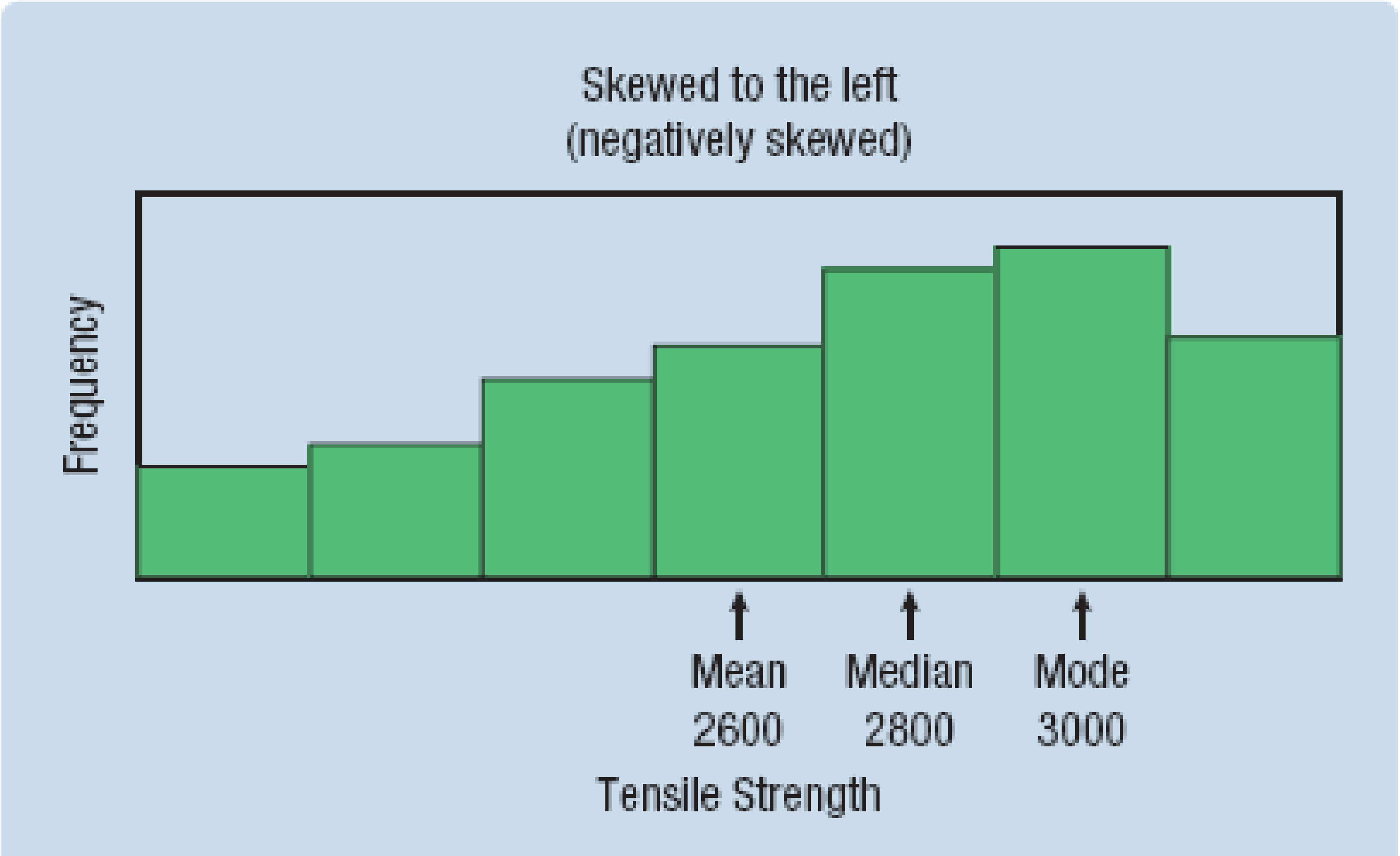
Nonsymmetric Distribution – Positive Skewness

For a **positively skewed (right skewed)** distribution, the mean will be the largest of the measures. The tail of the distribution is to the right. The mode will be the peak.



Nonsymmetric Distribution – Negative Skewness

For a **negatively skewed (left-skewed)** distribution, the mean will be the smallest of the measures. The tail of the distribution is to the left. The mode will be the peak.



Pearson's Coefficient of Skewness

Another of K. Pearson's (1857-1936) contribution to statistics is a formula to calculate the skewness.

$$sk = \frac{3\left(\bar{x} - median\right)}{s}$$

Accordingly, the coefficient of skewness (sk),

1. can range from –3.00 to +3.00 (negative value means negative skewness and positive value mean positive skewness)

2. a value of 0 indicates a symmetric distribution

Example – Skewness

The following are the earnings per share, in dollars, for a sample of 16 software companies for the year 2008.

$0.08	0.12	0.44	0.52	1.10	1.19	2.49	1.18
4.55	7.36	7.93	8.62	11.15	14.88	17.43	13.13

The mean is $5.76. The standard deviation is $5.85. The median is $3.52. Find the coefficient of skewness using Pearson's estimate.

$$sk = \frac{3\left(\bar{x} - median\right)}{s}$$

$$sk = \frac{3\left(5.76 - 3.52\right)}{5.85}$$

sk = 1.149

sk = 1.149 is a moderate positive skewness

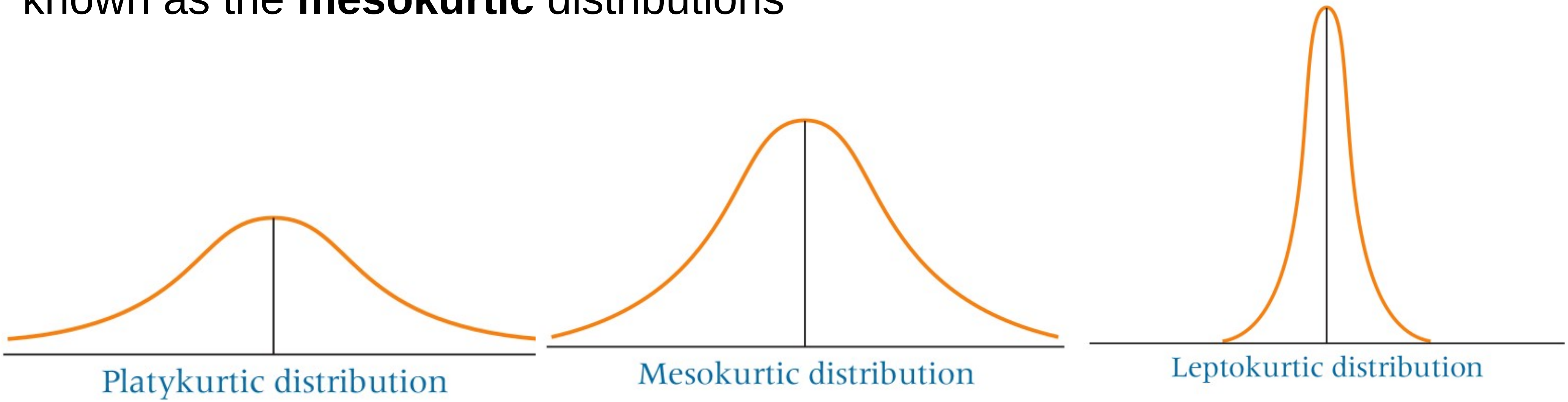
Shape: Kurtosis

Kurtosis describes the amount of peakedness of a distribution

Distributions that are high and thin are called **leptokurtic** distributions

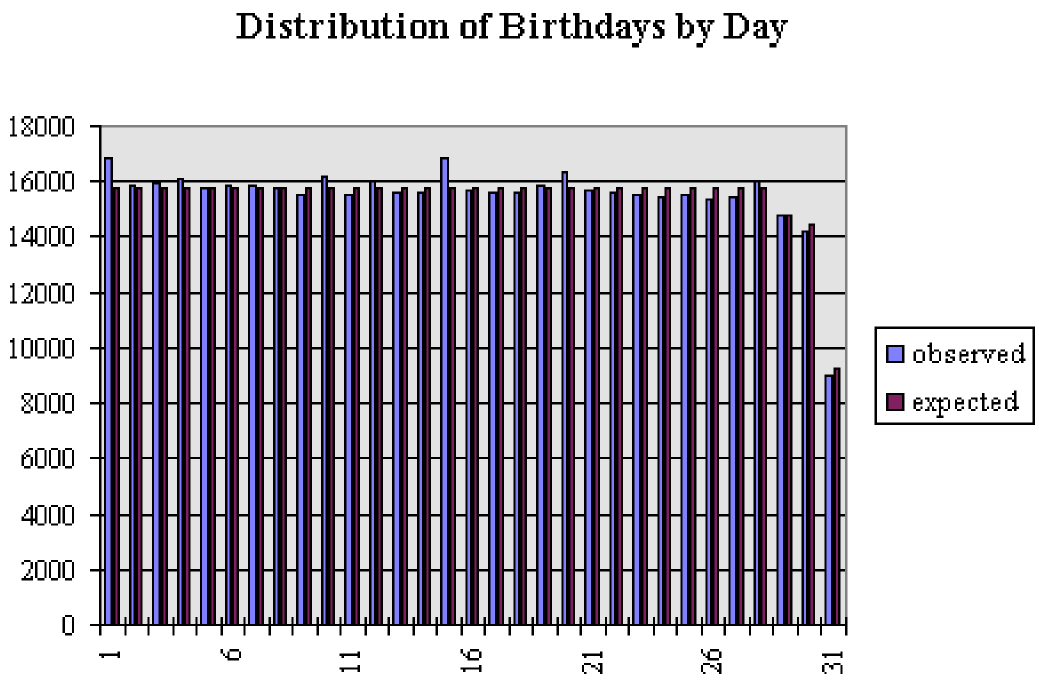
Distributions that are flat and spread out are called **platykurtic** distributions

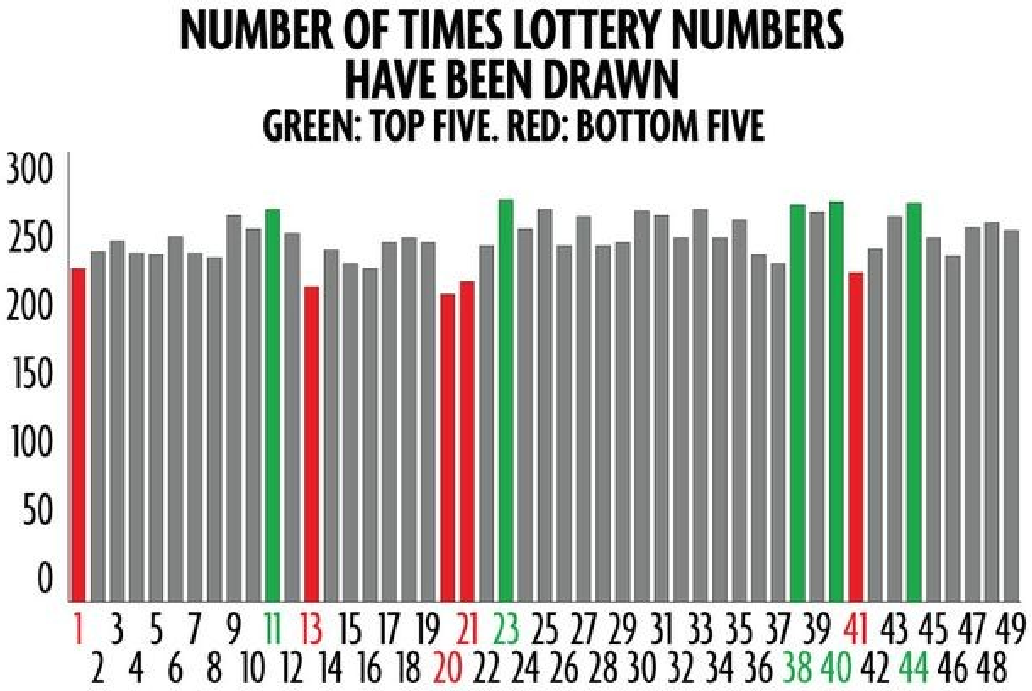
Between the tall and short distributions are the more “normal” in shape, also known as the **mesokurtic** distributions



For identical mean values but varying standard deviations can produce “cousin” distributions with varying heights

Examples of Uniform Distribution





https://www.panix.com/~murphy/bday.html
https://www.mirror.co.uk/news/ampp3d/heres-how-long-its-going-4654080

Review Questions

Review question set 11