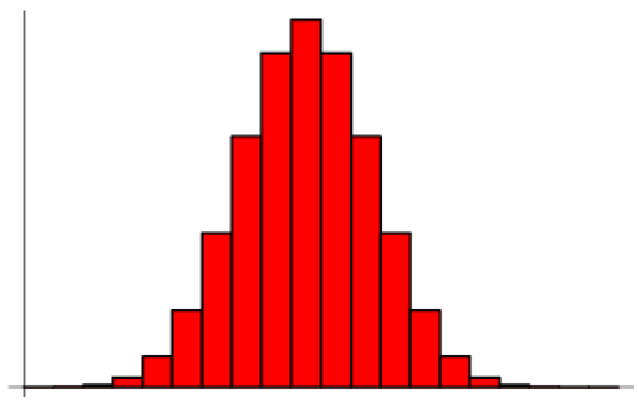
Binomial Distribution: The binomial distribution gives the discrete probability distribution P - p(n|N) of obtaining exactly n successes out of N Bernoulli trials (where the result of each Bernoulli trial is true with probability p and false with probability q=1-p). The binomial distribution is therefore given by

$$\begin{aligned} P_p(n \mid N) &= \binom{N}{n} p^n q^{N-n} \\ &= \frac{N!}{n!\,(N-n)!}\, p^n (1-p)^{N-n}, \end{aligned}$$



where (N; n) is a binomial coefficient. The above plot shows the distribution of n successes out of N=20 trials with p=q=1/2.

$X = \#$ of H from flipping coin 5 times

possible outcomes from 5 flips: $2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 = 2^5 = \boxed{32}$

$$P(X=0) = \frac{1}{32} = \frac{{}_5C_0}{32} \qquad {}_5C_0 = \frac{5!}{\cancel{0!} \cdot (5-0)!} = \frac{5!}{5!} = 1$$

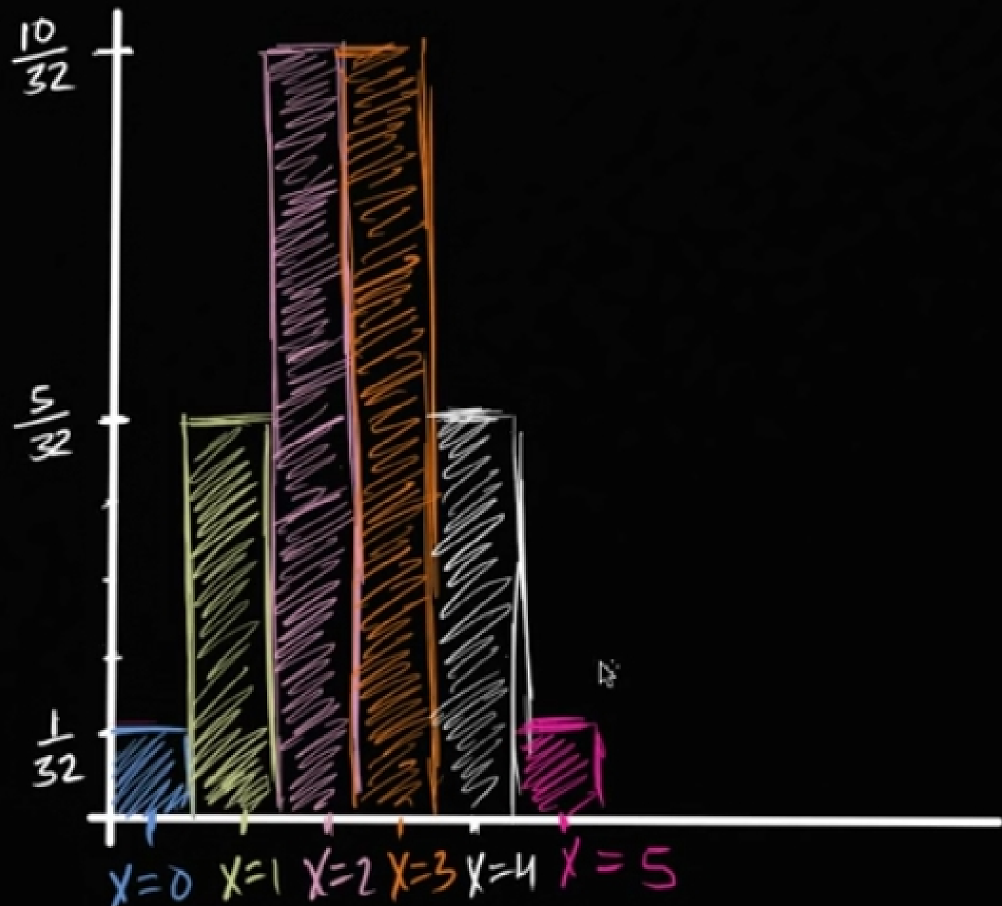
$$P(X=1) = \frac{5}{32} = \frac{{}_5C_1}{32} \qquad {}_5C_1 = \frac{5!}{1!\,(5-1)!} = \frac{5!}{4!} = 5$$

$$P(X=2) = \frac{{}_5C_2}{32} = \frac{10}{32} \qquad {}_5C_2 = \frac{5!}{2!\,(5-2)!} = \frac{5!}{2! \cdot 3!} = \frac{5 \cdot \cancel{4}^{2} \cdot \cancel{3 \cdot 2}}{\cancel{2} \cdot \cancel{3 \cdot 2}}$$

$$P(X=3) = \frac{{}_5C_3}{32} = \frac{10}{32} \qquad {}_5C_3 = \frac{5!}{3!\,(5-3)!} = \frac{5!}{3! \cdot 2!} = 10$$

$$P(X=4) = \frac{{}_5C_4}{32} = \frac{5}{32} \qquad {}_5C_4 = \frac{5!}{4!\,\cancel{(5-4)}!} = \frac{5!}{4!} = 5$$

$$P(X=5) = \frac{{}_5C_5}{32} = \frac{1}{32} \qquad {}_5C_5 = \frac{5!}{5!\,\cancel{(5-5)}!} = 1$$



X = # of H from flipping co

possible outcomes from 5

$$P(X=0) = \frac{1}{32} = \frac{{}_5C_0}{32}$$

$$P(X=1) = \frac{5}{32} = \frac{{}_5C_1}{32}$$

$$P(X=2) = \frac{{}_5C_2}{32} = \frac{10}{32}$$

$$P(X=3) = \frac{{}_5C_3}{32} = \frac{10}{32}$$

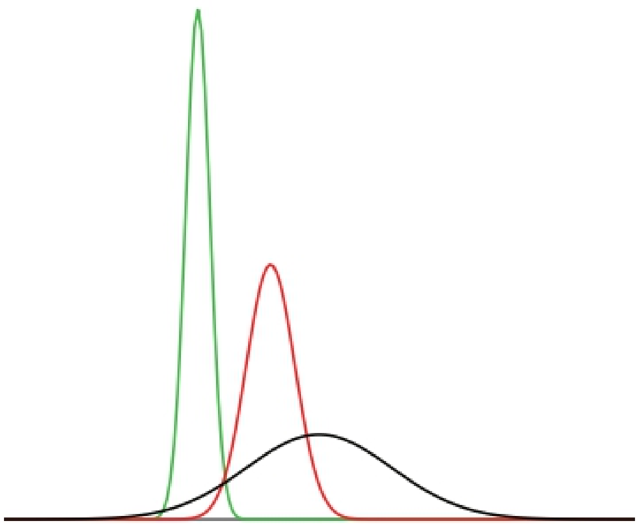
$$P(X=4) = \frac{{}_5C_4}{32} = \frac{5}{32}$$

$$P(X=5) = \frac{{}_5C_5}{32} = \frac{1}{32}$$

finite discrete random variable.

Normal Distribution

The normal distribution is the most important and most widely used distribution in statistics. It is sometimes called the “bell curve,” although the tonal qualities of such a bell would be less than pleasing. It is also called the “Gaussian curve” after the mathematician Karl Friedrich Gauss.

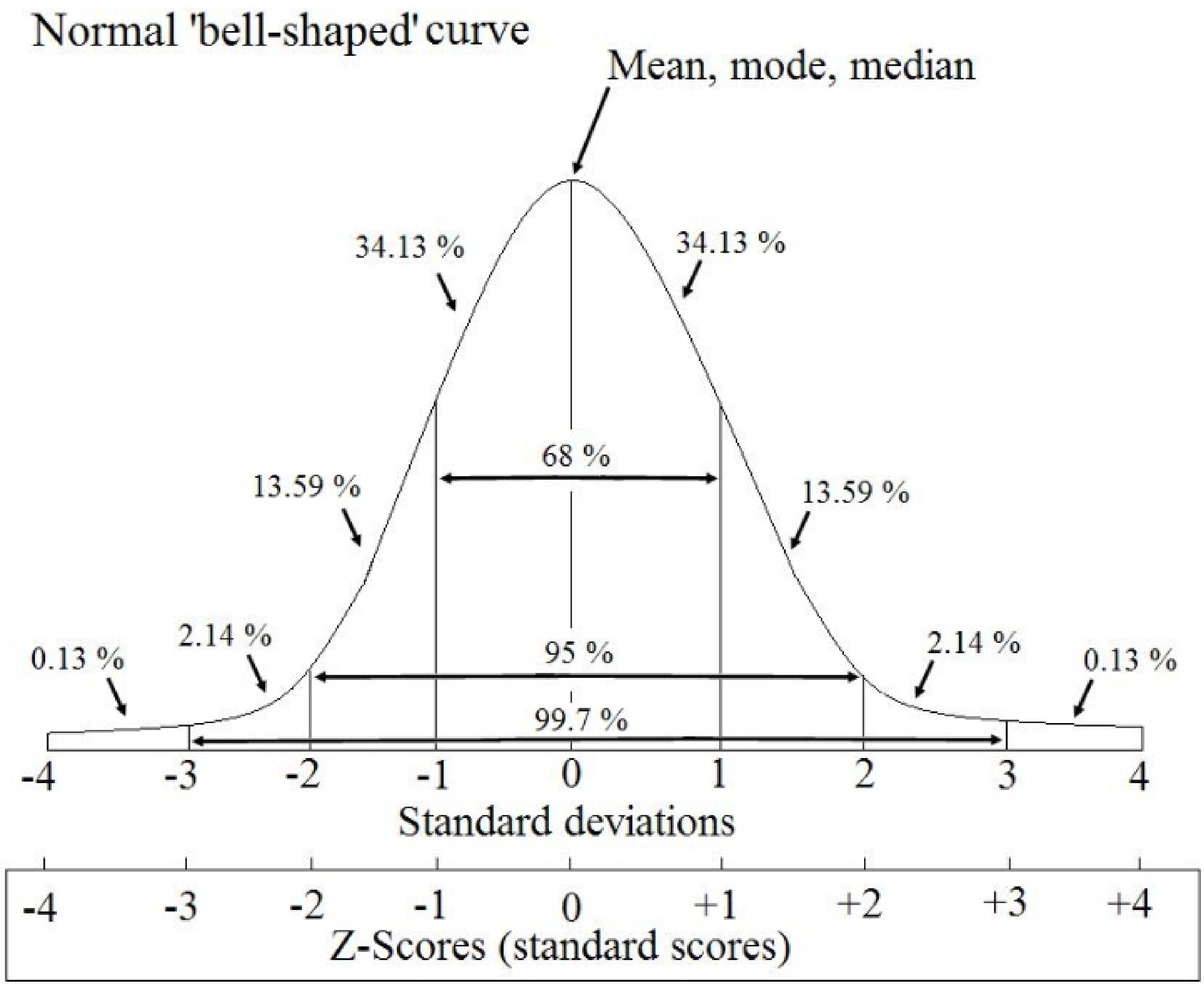


Strictly speaking, it is not correct to talk about “the normal distribution” since there are many normal distributions. Normal distributions can differ in their means and in their standard deviations. Above figure shows three normal distributions. The green (left-most) distribution has a mean of -3 and a standard deviation of 0.5, the distribution in red (the middle distribution) has a mean of 0 and a standard deviation of 1, and the distribution in black (right-most) has a mean of 2 and a standard deviation of 3. These as well as all other normal distributions are symmetric with relatively more values at the center of the distribution and relatively few in the tails.

Probability Density Function of a normal distribution is

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$$

1. Normal distributions are symmetric around their mean.
2. The mean, median, and mode of a normal distribution are equal.
3. The area under the normal curve is equal to 1.0.
4. Normal distributions are denser in the center and less dense in the tails.
5. Normal distributions are defined by two parameters, the mean (μ) and the standard deviation (σ).
6. 68% of the area of a normal distribution is within one standard deviation of the mean.
7. Approximately 95% of the area of a normal distribution is within two standard deviations of the mean.


Normal 'bell-shaped' curve
Mean, mode, median
34.13 %
34.13 %
68 %
13.59 %
13.59 %
2.14 %
2.14 %
0.13 %
0.13 %
95 %
99.7 %
-4 -3 -2 -1 0 1 2 3 4
Standard deviations
-4 -3 -2 -1 0 +1 +2 +3 +4
Z-Scores (standard scores)

Data Visualization

Bar Chart: Subjects and Teachers likes

3 – Maths
4 – Physics
1 – English
2 – Social Sciences
X- axis – subject labels
Y- axis – counts of teacher likes

* *A bar graph is a nice way to display categorical data.*

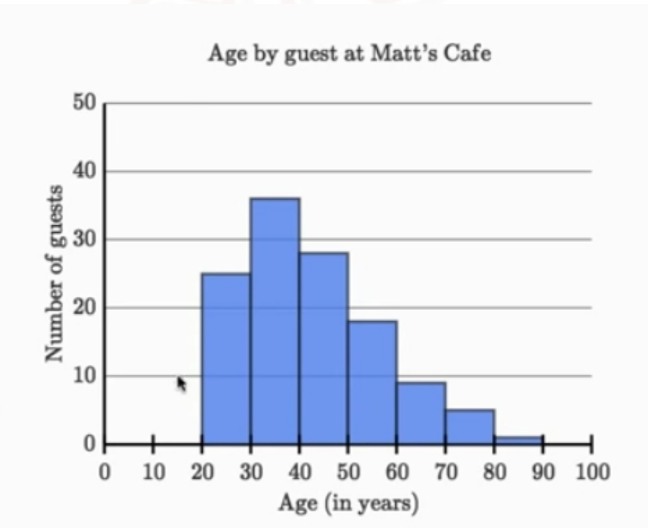
Histogram: When you have just a list of numbers, then you make some buckets of the data and plot a bar chart.

Example: ages – 1, 3, 27, 32, 5, 63, 26, 25, 18, 16, 4, 45, 29, 19, 51, 9, 42, 6

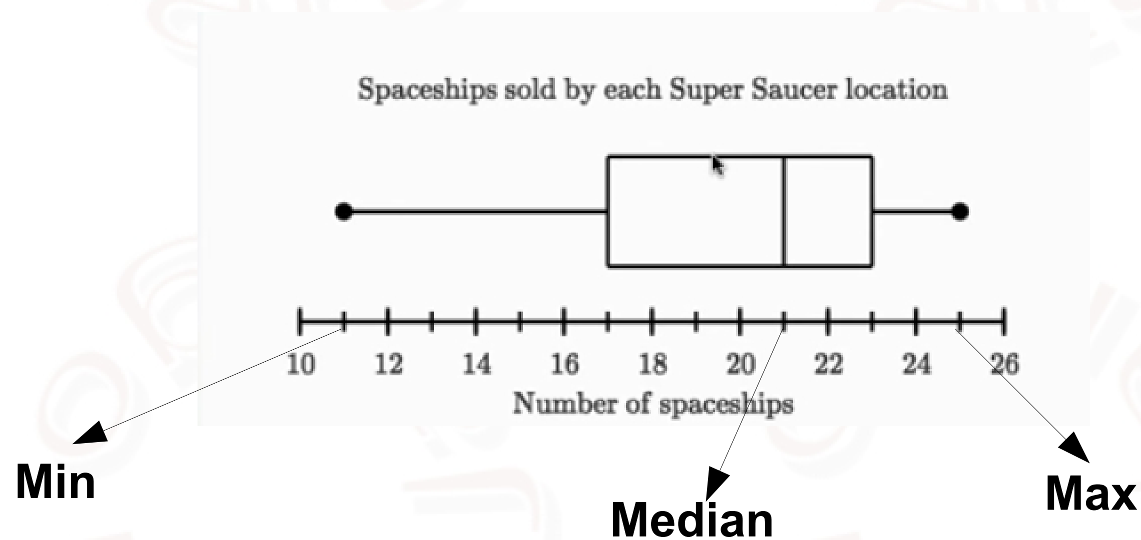
Bucket	# of people
0-9	6
10-19	3
20-29	5
30-39	1
40-49	2
50-59	2
60-69	1

Shapes of Distributions:

- If the bar graph is right tailed – it says data right tailed (right skewed), example below.



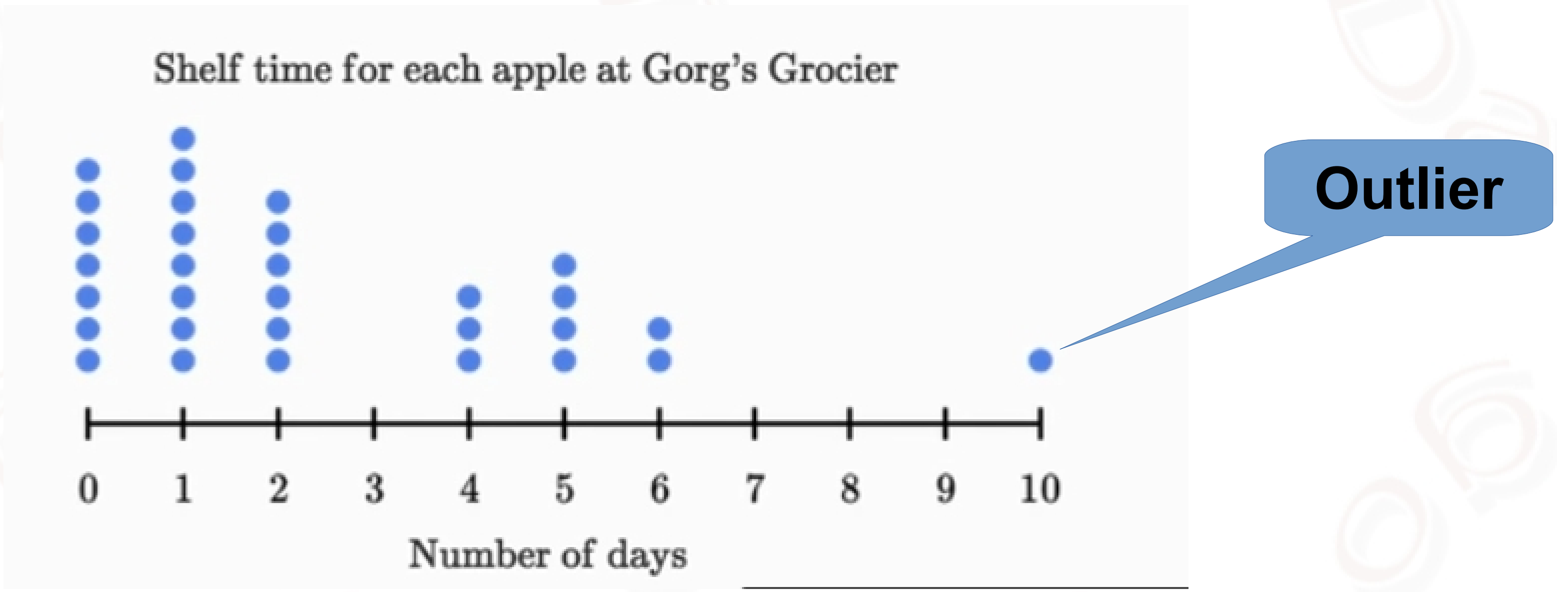
- Box whisker plot of the left tailed data
  (the data is left skewed)



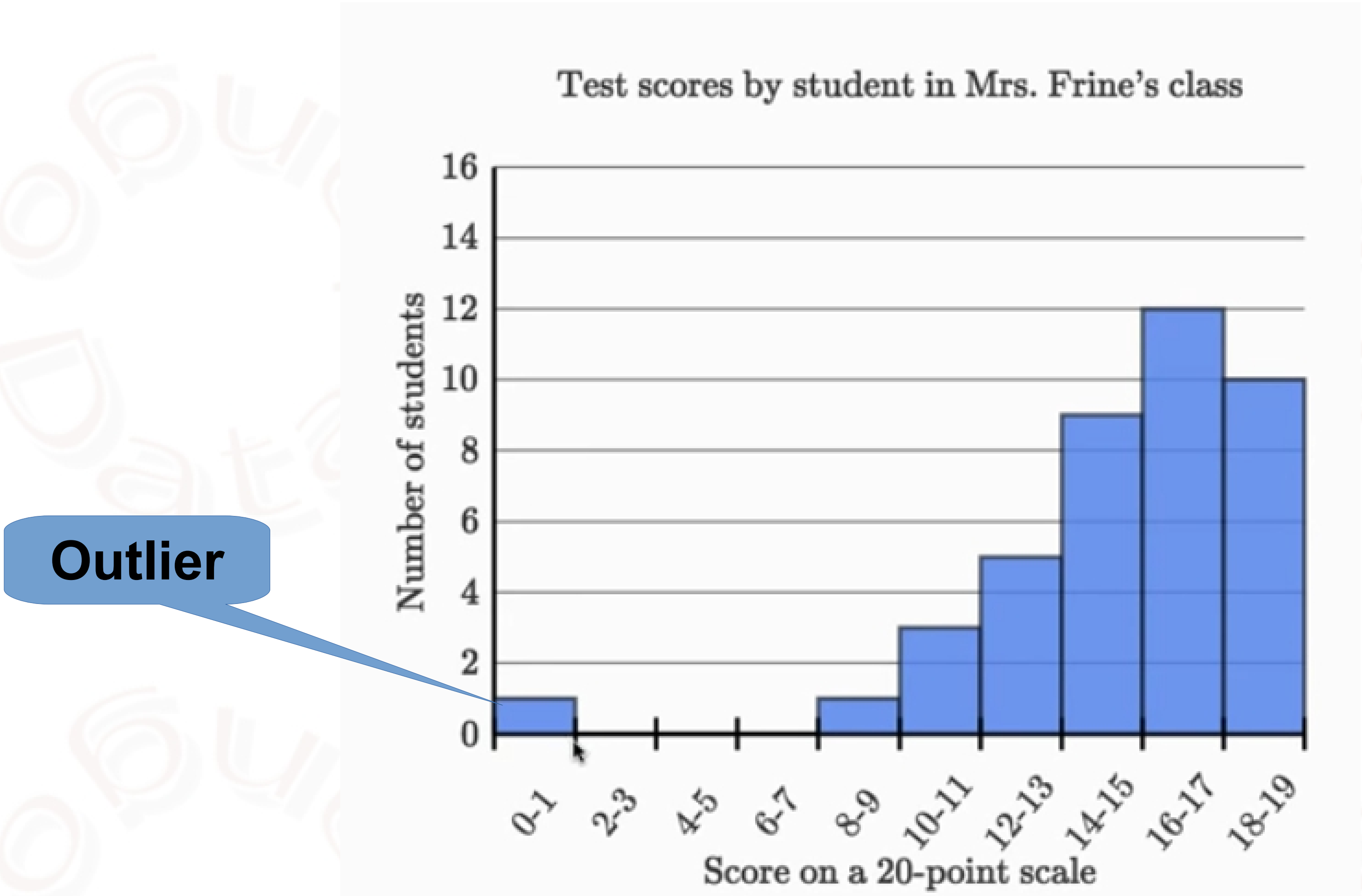
Outliers are extreme values that might be errors in measurement and recording, or might be accurate reports of rare events.

The best way to handle outliers depends on “domain knowledge”; that is, information about where the data come from and what they mean. And it depends on what analysis you are planning to perform.

Finding outliers on a dot-plot

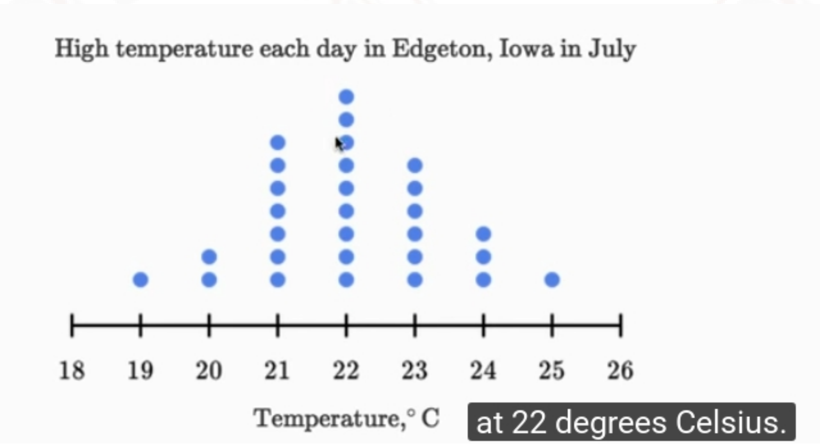


Finding outliers on a Histogram

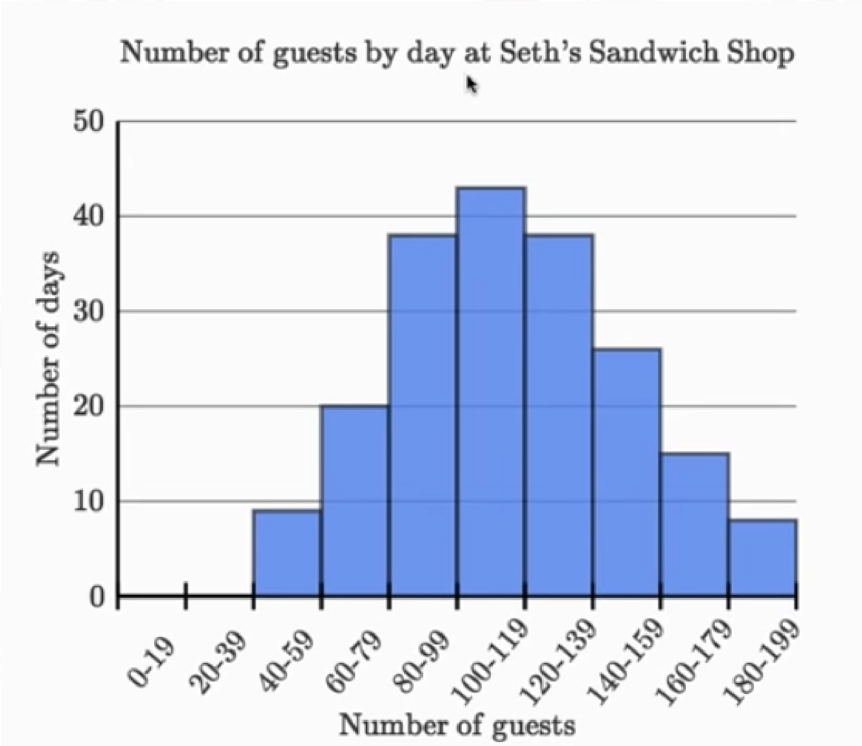


Examples for no outlier data

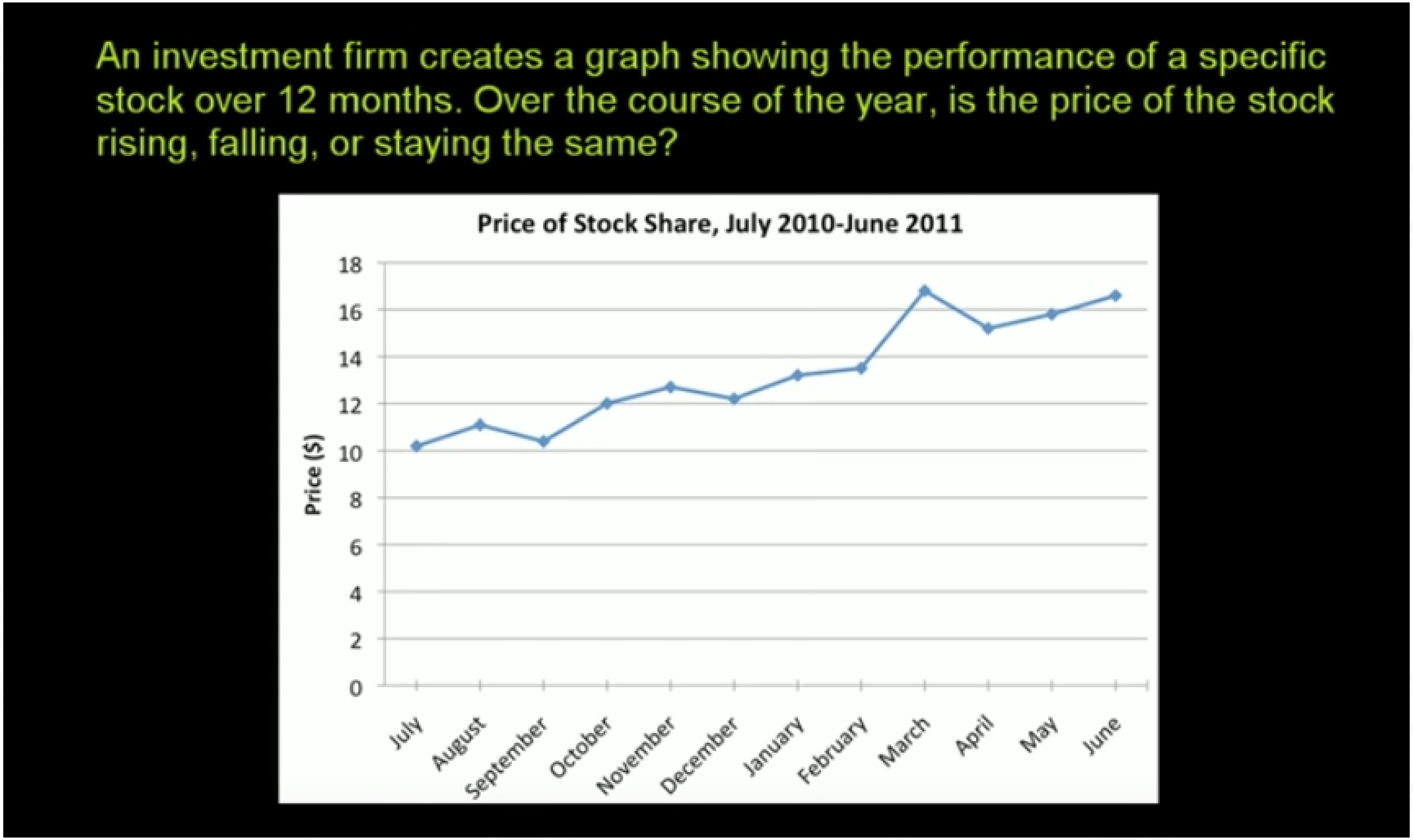
Dot-plot



Histogram



Line plot : This shows the trend of the data. The scale is very important when comparing two or more line plots.



Importance of the Scale: Comparing two line graphs with out being misled.

Thrill Soda hired a marketing company to help them promote their brand against Yummy Cola. The company gathered the following data about consumers' preference of soda:

Year	% of respondents who prefer Yummy Cola	% of respondents who prefer Thrill Soda	% of respondents who have no preference
2006	80%	12%	8%
2007	76%	19%	5%
2008	77%	19%	4%
2009	73%	20%	7%
2010	73%	21%	6%
2011	68%	25%	7%