Detecting outliers with Z-scores

$$Z\text{-score} = \frac{X - MEAN}{SD}$$

The 68%–95%–99.7% Rule

68%	have Z-scores between	-1	and	$+1$
95%	have Z-scores between	-2	and	$+2$
99.7%	have Z-scores between	-3	and	$+3$

Z-score criterion for outliers

If the data is approximately normally distributed, then *outliers will have Z-scores less than* -3 *or greater than* $+3$.

We would also label Z-scores in the 2–3 range (positive or negative) as "borderline" outliers.

Example: Accuracy of Nutrition Labels

A laboratory found 40 packaged food items (from canned green beans to cheese curls) and determined the number of calories in each item. They then compared their measurements to the number of calories in the nutrition facts label on each package. The following data lists percent errors,

$$\frac{\text{(calories measured)} - \text{(calories on the box)}}{\text{(calories on the box)}} \times 100\%$$

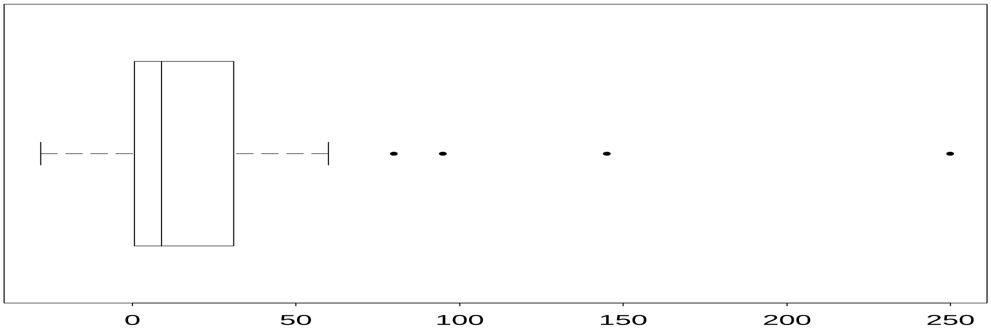
```
  2.0 -28.0  -6.0   8.0   6.0  -1.0  10.0  13.0
 15.0  -4.0  -4.0 -18.0  10.0   5.0   3.0  -7.0
  3.0  -0.5 -10.0   6.0  41.0  46.0   2.0  25.0
 39.0  16.5  17.0  28.0  -3.0  14.0  34.0  42.0
 15.0  60.0 250.0 145.0   6.0  80.0  95.0   3.0
```

Mean = 23.95 and SD = 48.67, so the Z-scores are:

```
-0.45 -1.07 -0.62 -0.33 -0.37 -0.51 -0.29 -0.22
-0.18 -0.57 -0.57 -0.86 -0.29 -0.39 -0.43 -0.64
-0.43 -0.50 -0.70 -0.37  0.35  0.45 -0.45  0.02
 0.31 -0.15 -0.14  0.08 -0.55 -0.20  0.21  0.37
-0.18  0.74  4.64  2.49 -0.37  1.15  1.46 -0.43
```

- The lowest Z-score is -1.07, not an outlier.
- The highest two Z-scores are 2.49 and 4.64, corresponding to lab values of 145 and 250.

*How does this compare to the (1.5)*IQR rule?*



- For the normal distribution, MEAN = Median, and will be about halfway between Q1 and Q3.
- So Q1 is about 0.5 IQR's below the mean; thus

$$Q1 - 1.5 * IQR \approx MEAN - 2 * IQR$$

- For the normal distribution, $2 * IQR = 3 * SD$. Therefore

$$\text{lower fence} \approx MEAN - 3 * SD$$

- This says: Z-score for lower fence ≈ -3
- Similarly: Z-score for upper fence $\approx +3$

So the lower-fence / upper-fence rule for detecting outliers is about the same as the Z-score rule, if the data is normally distributed.

But the boxplot identified four outliers, and the Z-score method only detected one. Why?

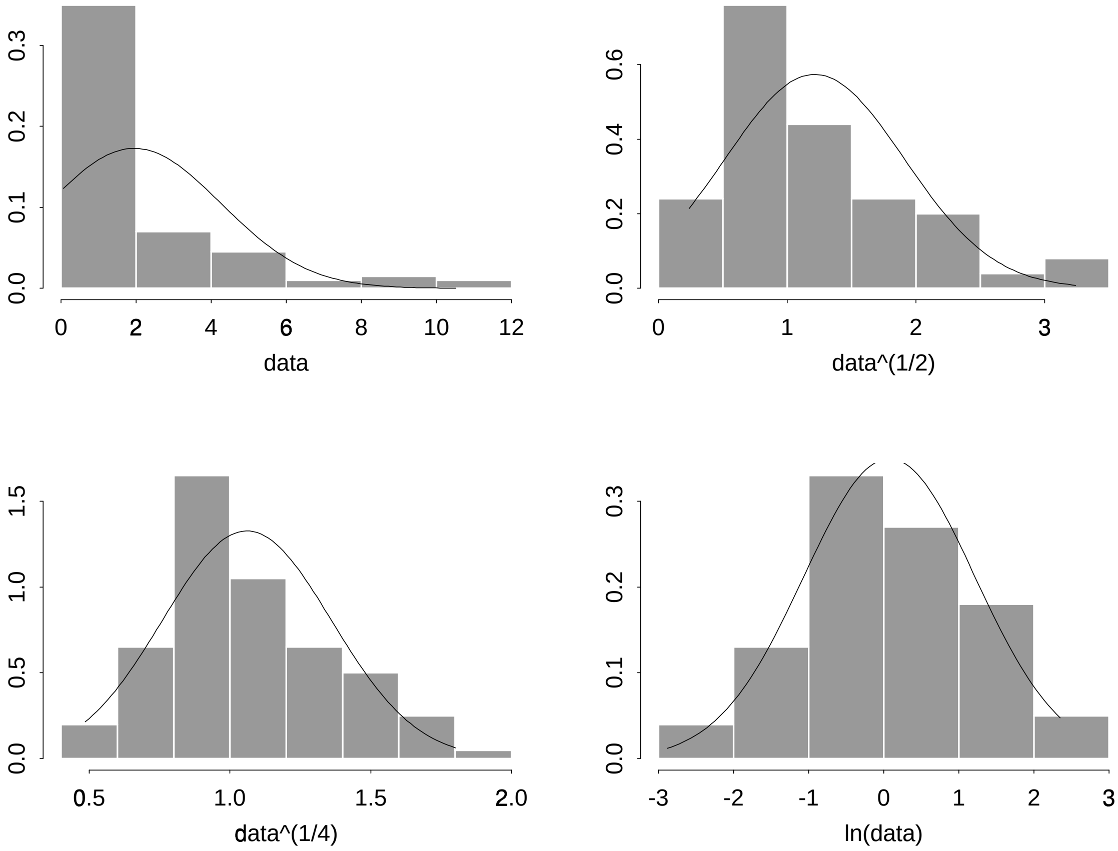
```
Stem-and-leaf plot       N  = 40
Leaf Unit = 10


    1   -0 2
   10   -0 110000000
  (18)   0 000000000011111111
   12    0 2233
    8    0 444
    5    0 6
    4    0 89
    2    1
    2    1
    2    1 4
    1    1
    1    1
    1    2
    1    2
    1    2 5
```

Transformations, Part II

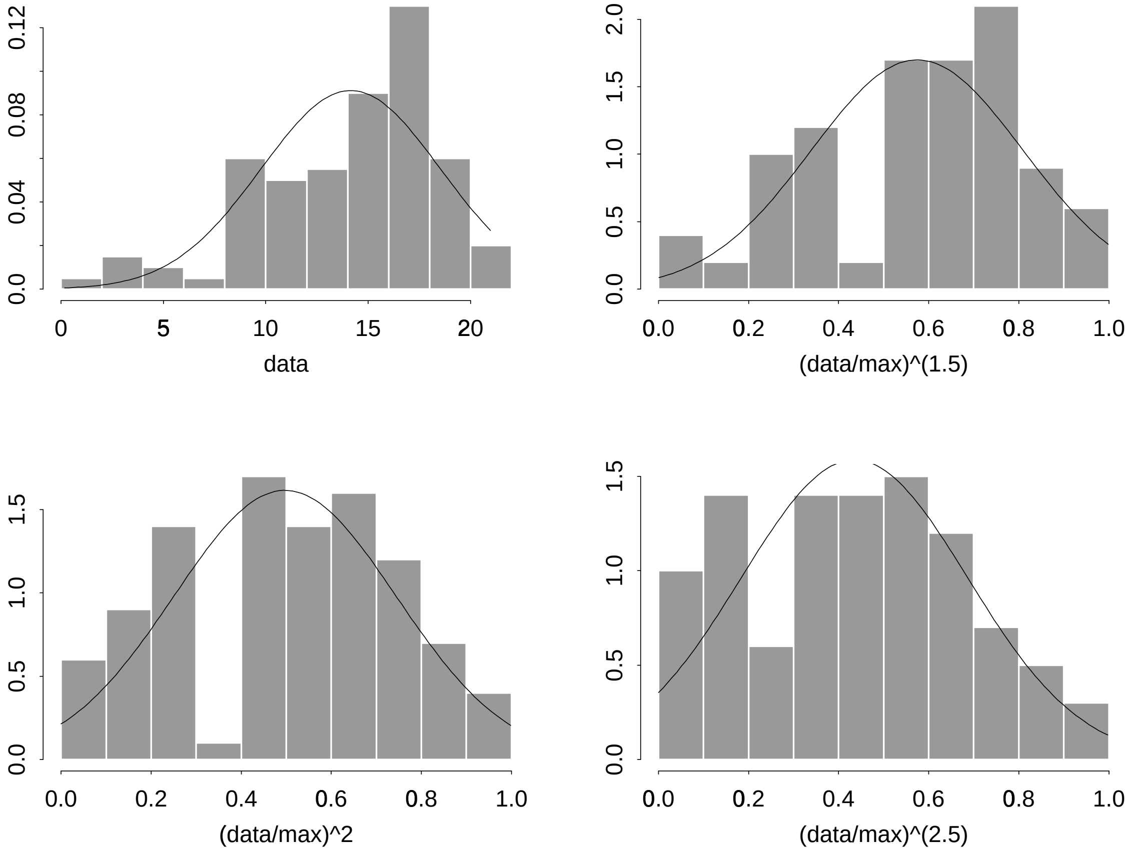
If the normal distribution is so great, how can we get it?

When data is skewed, power transformations ($y = x^p$) or logarithms ($y = \ln(x)$ or $y = \log(x)$) can "move" the data back to the Normal bell-shape.

- Skewed right: Try $x^{2/3}, x^{1/2}, x^{1/3}, x^{1/4}$, etc. (powers less than one), or try $\ln(x)$.



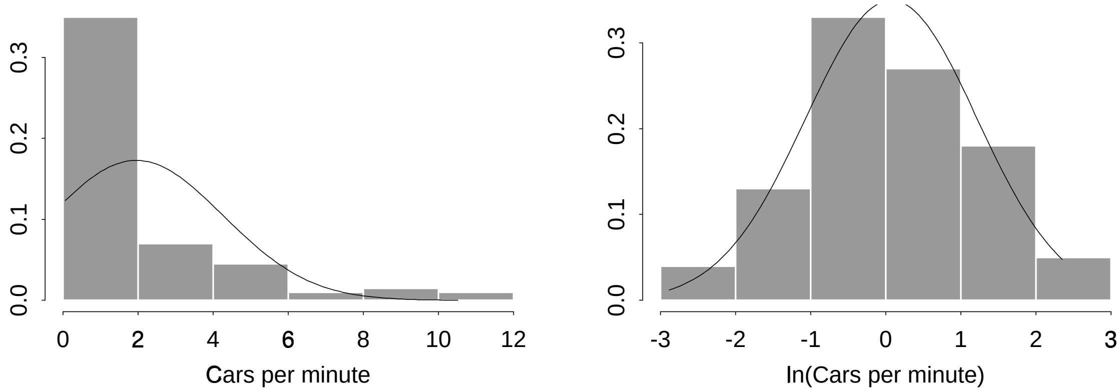
- Skewed left: Try x^2, x^3, etc. (powers greater than one).



- If x can be negative, add a constant (the MIN of the data) to all the x's first so they are all positive before you get started.

- If the numbers are getting out of hand, divide by a constant (the MAX of the data) first, so they are all between 0 and 1.

Example: Traffic Patterns

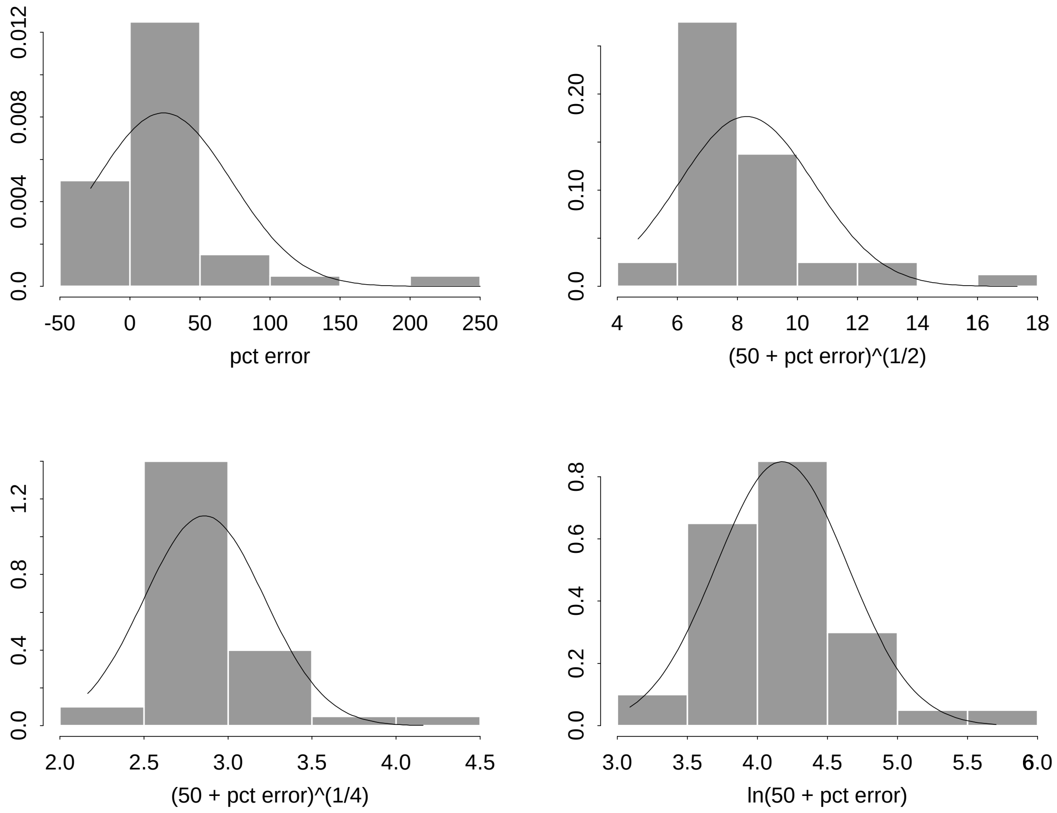


The histogram on the left represents the number of cars per minute passing through a certain street intersection. Use the 68–95–99.7 rule to answer the following questions:

(a) About what percent of the time are there 3 or fewer cars per minute going through the intersection?

(b) In the busiest 2.5% of the time, about how many cars were going through the intersection?

Example: Can we fix the Calorie Data?



If you are looking at a list of data:

```
  2.0 -28.0  -6.0   8.0   6.0  -1.0  10.0  13.0
 15.0  -4.0  -4.0 -18.0  10.0   5.0   3.0  -7.0
  3.0  -0.5 -10.0   6.0  41.0  46.0   2.0  25.0
 39.0  16.5  17.0  28.0  -3.0  14.0  34.0  42.0
 15.0  60.0 250.0 145.0   6.0  80.0  95.0   3.0
```

- *Order of magnitude:* the power of 10 (number of digits to the left or right of the decimal point) where the number begins.
- Different orders of magnitude ⇒ do a transformation.