Standard Deviation and Normality

Chapter 5

Jason Bryer
epsy530.bryer.org

```
## Error: object 'pisana' not found
```

```
## Error: object 'pisausa' not found
```

Comparing Students

- Student A takes the SAT and scores 600 on the math component, 100 points higher than the average student.

- Student B takes the ACT and scores 27 on the math component, 9 points higher than the average student.

- Which student performed better?

How many standard deviations?

	SAT	ACT
Mean	500	18
SD	100	6
Student	600	27

The standard deviation helps compare two different metrics.

- SAT: 600 - 500 = 100 (or 1 standard deviation)

- ACT: 27 - 18 = 9 (or 1.5 standard devations)

The z-Score (or standard score)

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

- The z-score measures the distance of the value from the mean in standard deviations.

- A positive z-score indicates the value is above the mean.

- A negative z-score indicates the value is below the mean.

- A small z-score indicates the value is close to the mean when compared to the rest of the data values.

- A large z-score indicates the value is far from the mean when compared to the rest of the data values.

z-Scores for our Students

	SAT	ACT
Mean	500	18
SD	100	6
Student	600	27

SAT

ACT

$$z = \frac{600 - 500}{100} = 1$$

$$z = \frac{27 - 18}{6} = 1.5$$

Shifting

If the same number is subtracted or added to all data values, then:

- The measures of the spread - standard deviation, range, and IQR - are all unaffected.

- The measures of position - mean, median, and mode - are all changed by that number.

```
ggplot(pisausa, aes(x=Math)) +
geom_histogram(binwidth=20)
```

```
ggplot(pisausa, aes(x=Math - 500)) +
geom_histogram(binwidth=20)
```

```
Error: object 'pisausa' not found
```
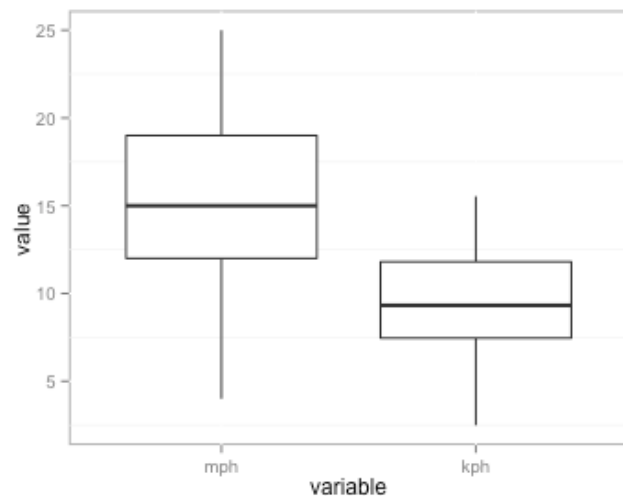
```
Error: object 'pisausa' not found
```

Scaling

If we multiply all data values by the same number, what happens to the position and spread?

When we multiply (or divide) all the data values by a constant, all measures of position and all measures of spread are multiplied (or divided) by that same constant.

```
ggplot(speeds, aes(x=variable, y=value)) +
geom_boxplot()
```

```
ggplot(speeds, aes(x=value)) +
geom_histogram() +
facet_wrap(~ variable, ncol=1)
```
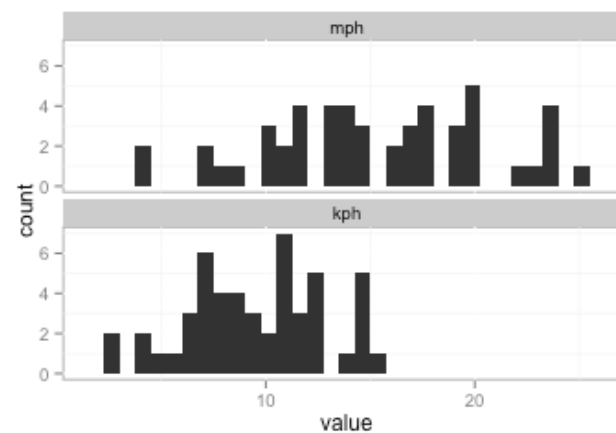
Models

" All models are wrong, but some are useful. "

- George Box

- −1 < z < 1: Not uncommon

- z = ±3: Rare

- z = 6: Shouts out for attention!

The Normal Model

- Bell Shaped: unimodal, symmetric

- A Normal model for every mean and standard deviation.

    - μ (read "mew") represents the population mean.

    - σ (read "sigma") represents the population standard deviation.

    - $N(\mu, \sigma)$ represents a Normal model with mean m and standard deviation s.

$$f(x) \quad = \quad \frac{1}{\sigma\sqrt{2\pi}} \, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Parameters: Numbers that help specify the model (i.e. μ, σ)

- Statistics: Numbers that summarize the data (e.g. $\bar{y}$, s, median, mode)

- $N(0, 1)$ is called the standard Normal model, or the standard Normal distribution.

- The Normal model should only be used if the data is approximately symmetric and unimodal.

The 68-95-99.7 Rule

- 68% of the values fall within 1 standard deviation of the mean.

- 95% of the values fall within 2 standard deviations of the mean.

- 99.7% of the values fall within 3 standard deviations of the mean.

Finding percentiles: http://shiny.albany.edu/stat/stdnormal/
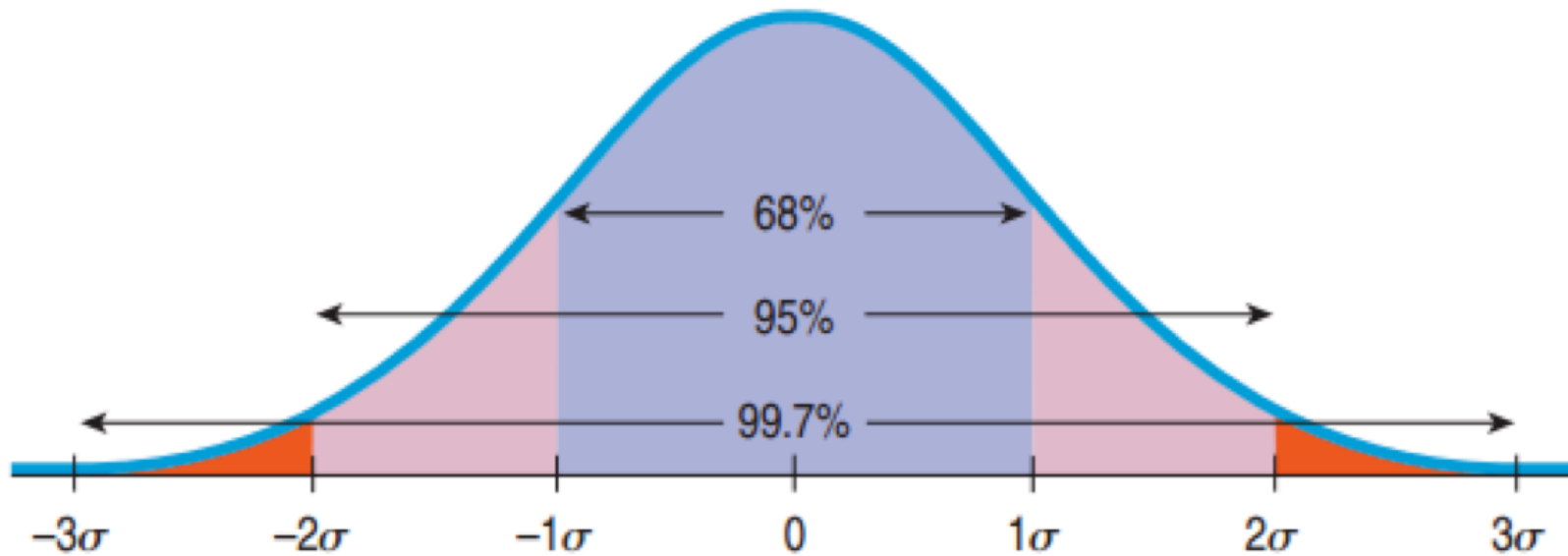
Checking Normality

At a minimum, plot a histogram.

```
ggplot(pisausa, aes(x=Math)) + geom_histogram()
```

```
Error: object 'pisausa' not found
```
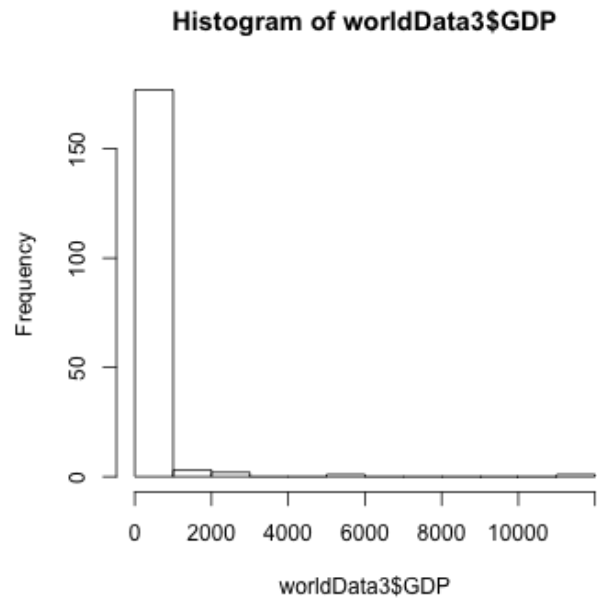
Normal Probability Plots: PISA Math

- Plots each value against the z-score that would be expected had the distribution been perfectly normal.

- If the plot shows a line or is nearly straight, then the Normal model works.

- If the plot strays from being a line, then the Normal model is not a good model.
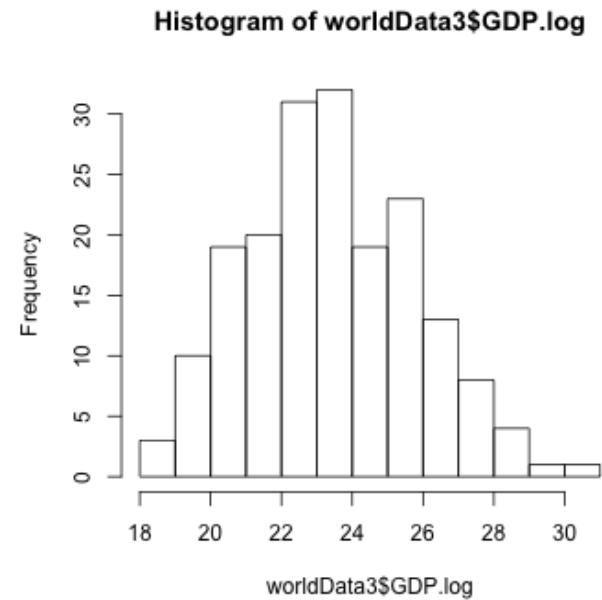
```
qqnorm(pisausa$Math)
```

```
Error: object 'pisausa' not found
```
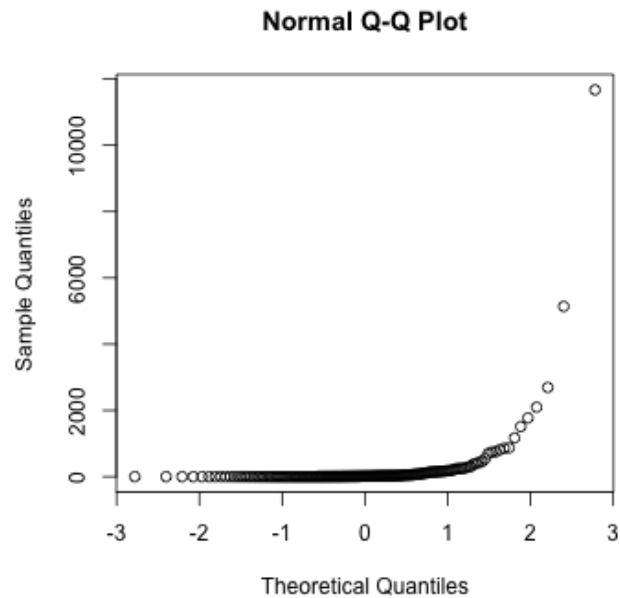
Histograms: GDP

```
hist(worldData3$GDP)
```
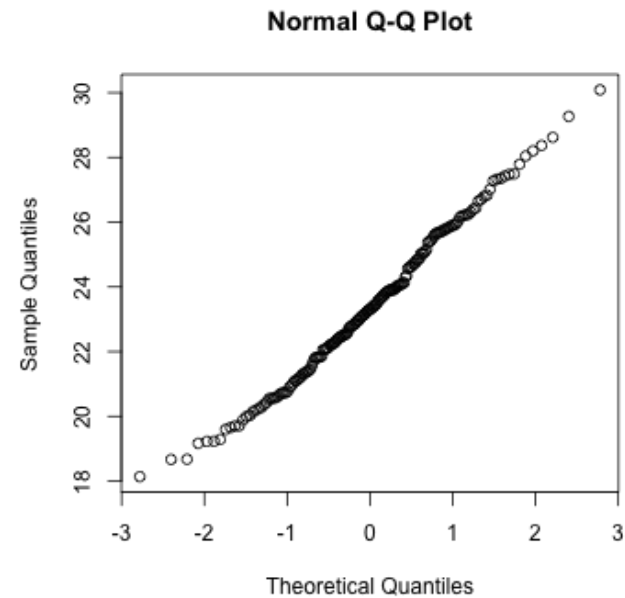
```
hist(worldData3$GDP.log)
```


Histogram of worldData3$GDP

Histogram of worldData3$GDP.log

Normal Probability Plots: GDP

qqnorm(worldData3$GDP)

qqnorm(worldData3$GDP.log)

What can go wrong?

- Don't use the Normal model when the distribution is not unimodal and symmetric.

- Always look at the picture first.

- Don't use the mean and standard deviation when outliers are present.

- Check by making a picture.

- Don't round your results in the middle of the calculation.

- Always wait until the end to round.

- Don't worry about minor differences in results.

- Different rounding can produce slightly different results.