DATA606 - Distributions

Jason Bryer, Ph.D.

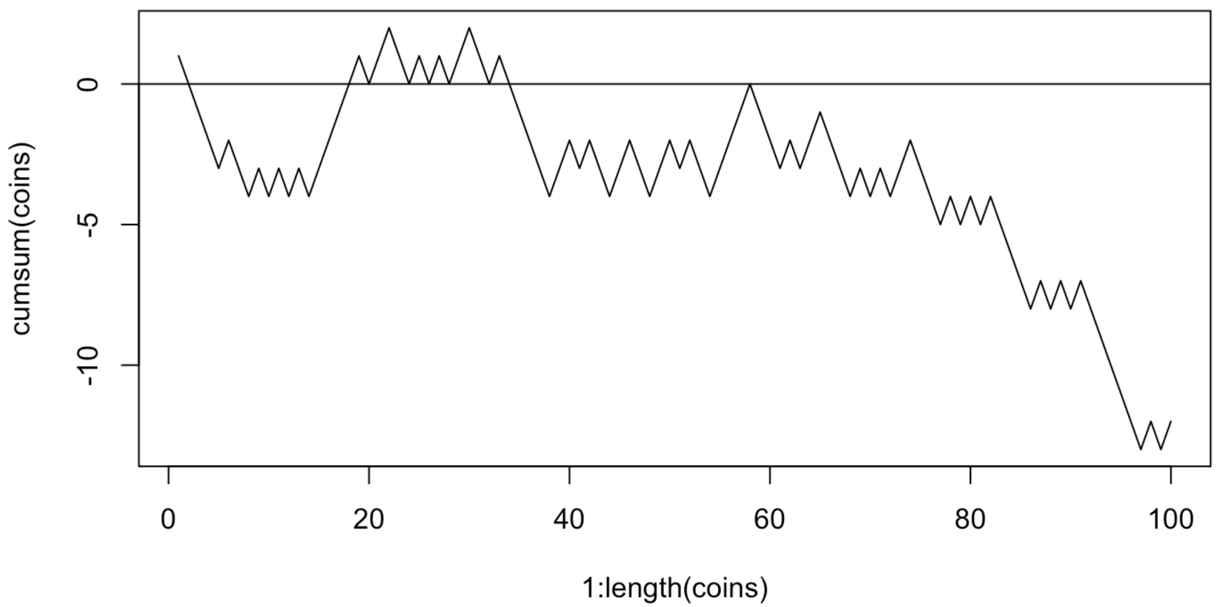
February 16, 2017

Meetup Presentations

- Georgia G. (2.29)
- James Kuruvilla (2.6)

Coin Tosses Revisited

```
coins <- sample(c(-1,1), 100, replace=TRUE)
plot(1:length(coins), cumsum(coins), type='l')
abline(h=0)
```




```
cumsum(coins)[length(coins)]
```

```
## [1] -12
```

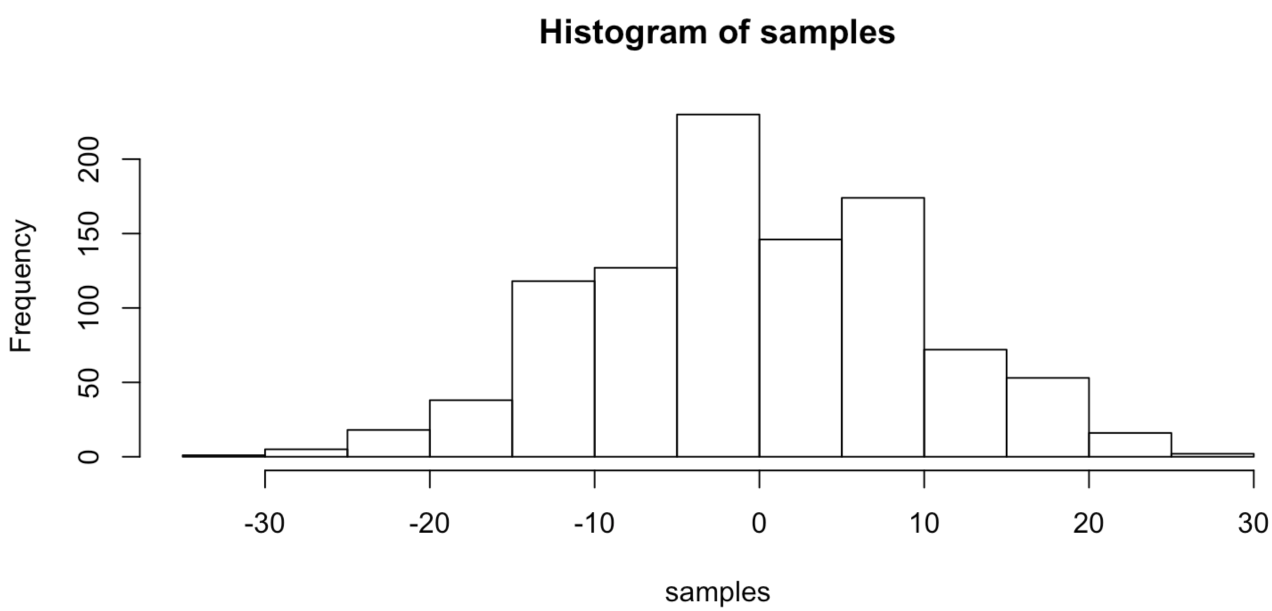
Many Random Samples

```
samples <- rep(NA, 1000)
for(i in seq_along(samples)) {
    coins <- sample(c(-1,1), 100, replace=TRUE)
    samples[i] <- cumsum(coins)[length(coins)]
}
head(samples)
```

```
## [1]  -8   8  -2 -10  -8   6
```

Histogram of Many Random Samples

```
hist(samples)
```



Properties of Distribution

```
(m.sam <- mean(samples))
```

```
## [1] 0.162
```

```
(s.sam <- sd(samples))
```

```
## [1] 9.883088
```

Properties of Distribution (cont.)

```
within1sd <- samples[samples >= m.sam - s.sam & samples <= m.sam + s.sam]
length(within1sd) / length(samples)
```

```
## [1] 0.677
```

```
within2sd <- samples[samples >= m.sam - 2 * s.sam & samples <= m.sam + 2* s.sam]
length(within2sd) / length(samples)
```

```
## [1] 0.951
```

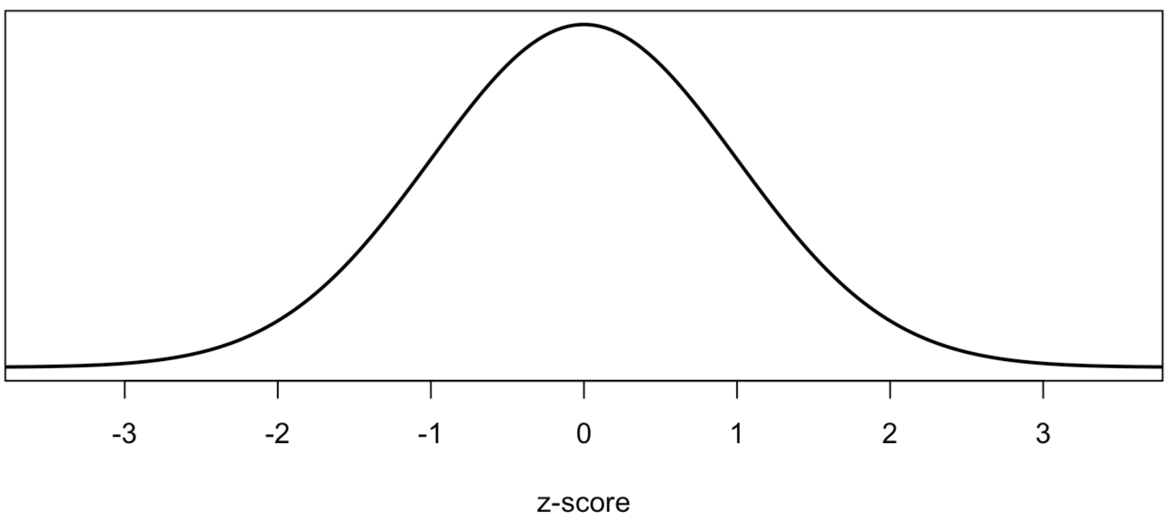
```
within3sd <- samples[samples >= m.sam - 3 * s.sam & samples <= m.sam + 3 * s.sam]
length(within3sd) / length(samples)
```

```
## [1] 0.999
```

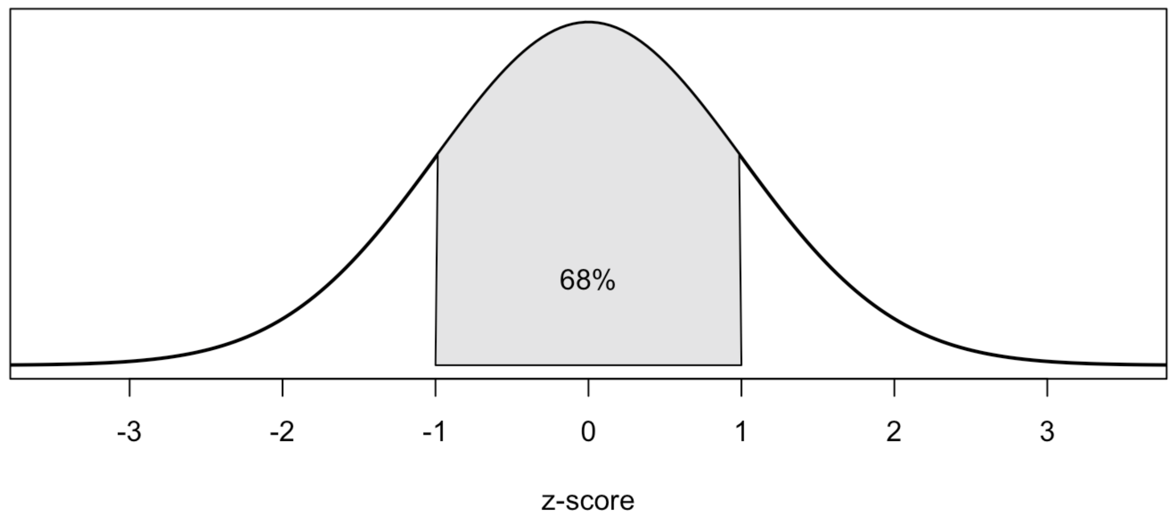
Standard Normal Distribution

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

```
x <- seq(-4,4,length=200); y <- dnorm(x,mean=0, sd=1)
plot(x, y, type = "l", lwd = 2, xlim = c(-3.5,3.5), ylab='', xlab='z-score', yaxt='n')
```

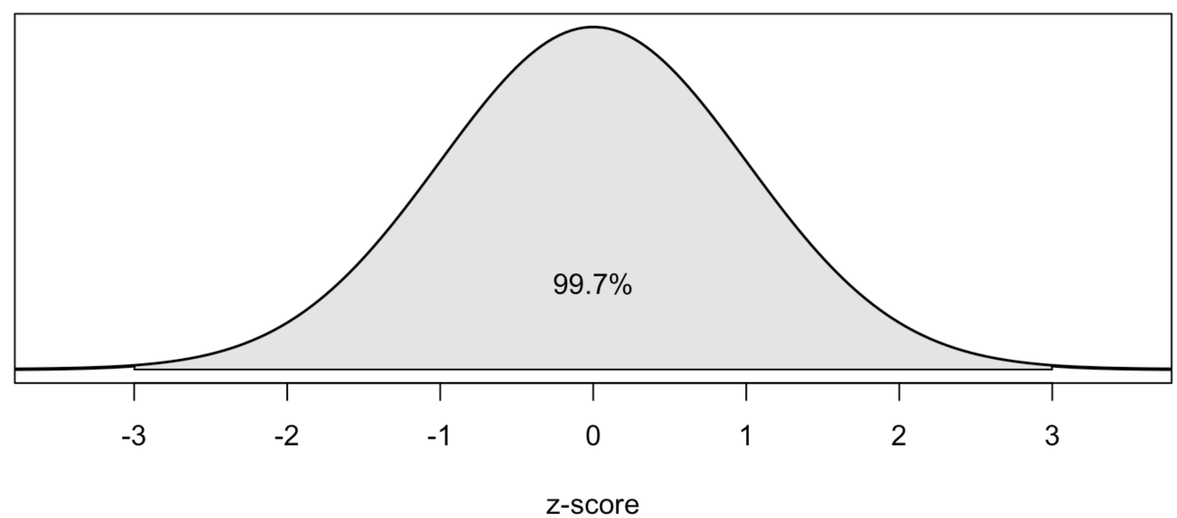


Standard Normal Distribution



Standard Normal Distribution

Standard Normal Distribution



What's the likelihood of ending with 15?

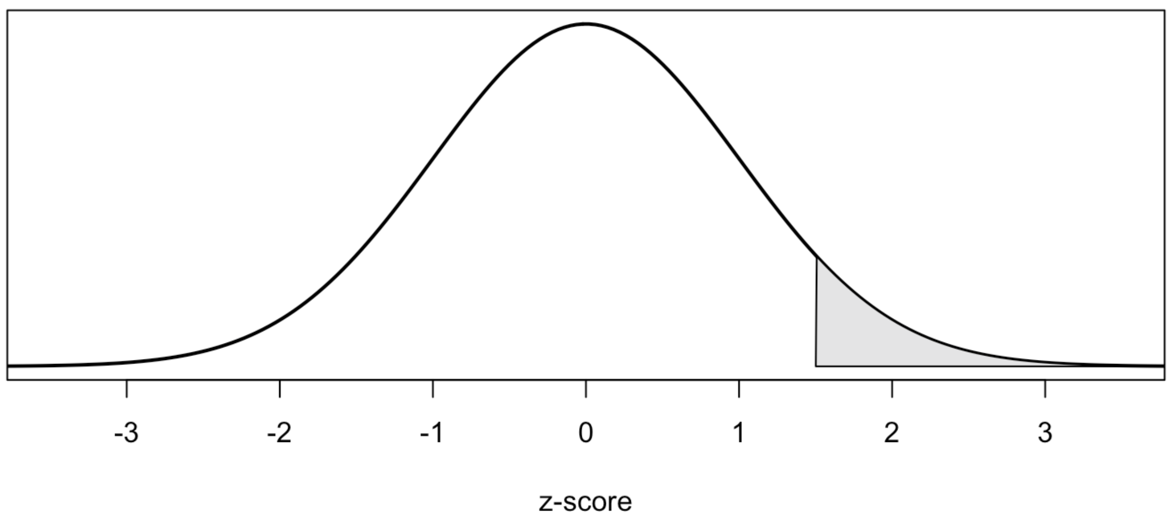
```
pnorm(15, mean=mean(samples), sd=sd(samples))
```

```
## [1] 0.9333678
```

What's the likelihood of ending with 15?

```
1 - pnorm(15, mean=mean(samples), sd=sd(samples))
```

```
## [1] 0.06663219
```



Comparing Scores on Different Scales

SAT scores are distributed nearly normally with mean 1500 and stan- dard deviation 300. ACT scores are distributed nearly normally with mean 21 and standard deviation 5. A college admissions officer wants to determine which of the two applicants scored better on their standardized test with respect to the other test takers: Pam, who earned an 1800 on her SAT, or Jim, who scored a 24 on his ACT?

Z-Scores

- Z-scores are often called standard scores:

$$Z = \frac{observation - mean}{SD}$$

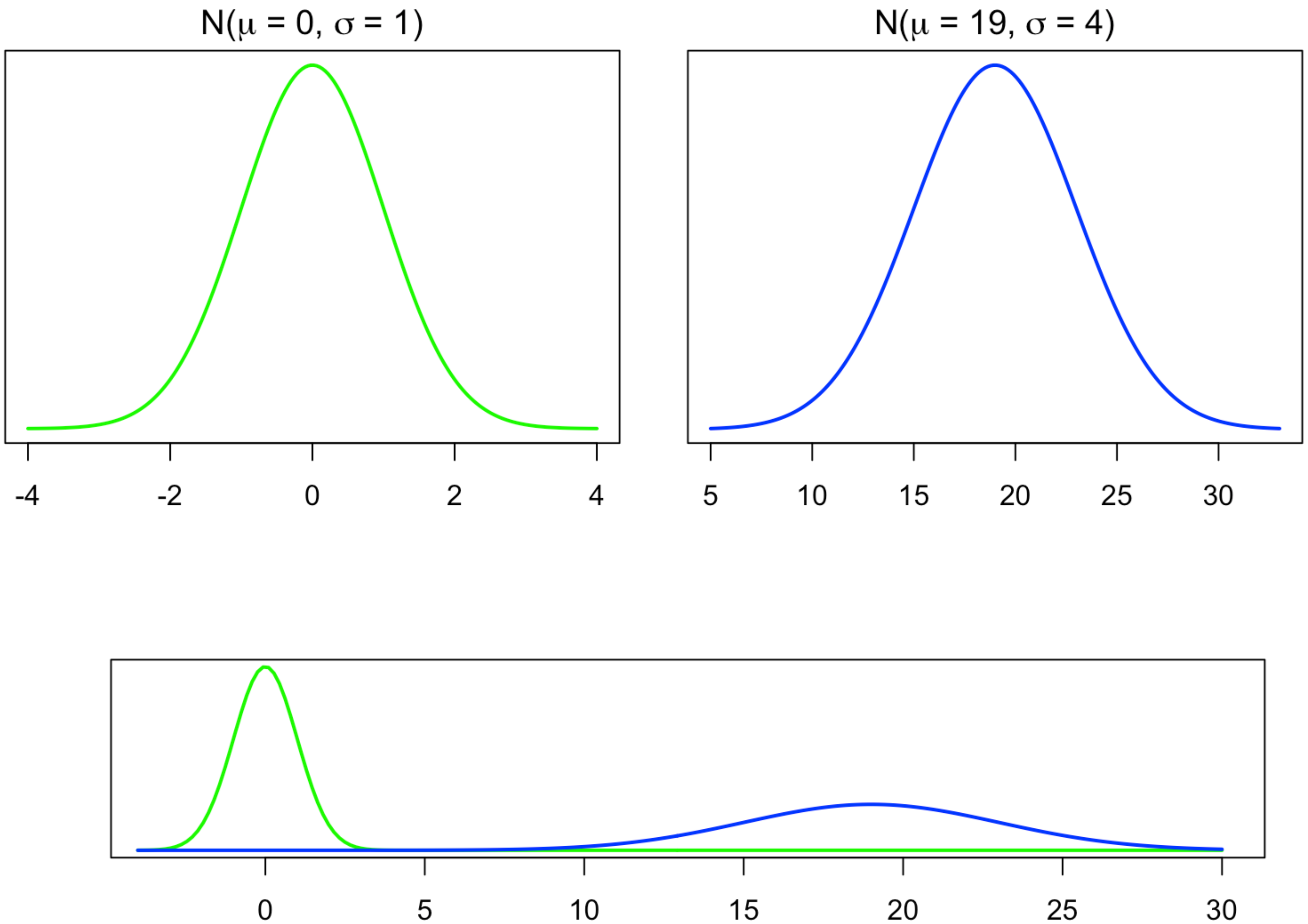
- Z-Scores have a mean = 0 and standard deviation = 1.

Converting Pam and Jim's scores to z-scores:

$$Z_{Pam} = \frac{1800 - 1500}{300} = 1$$

$$Z_{Jim} = \frac{24 - 21}{5} = 0.6$$

Standard Normal Parameters



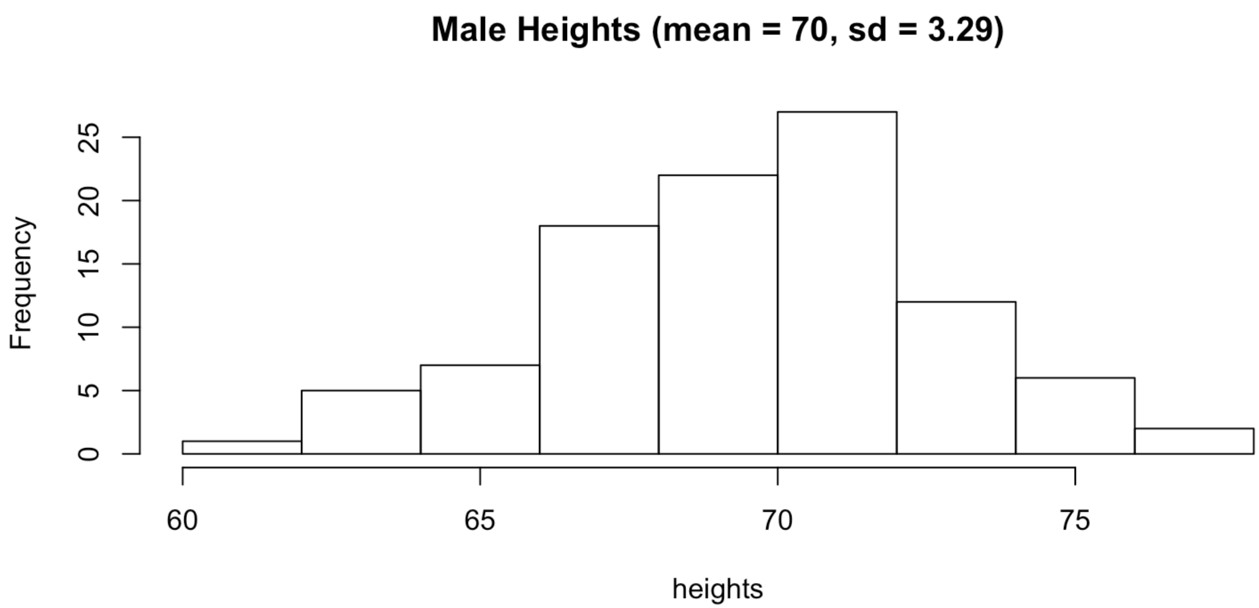
SAT Variability

SAT scores are distributed nearly normally with mean 1500 and standard deviation 300.

- ~68% of students score between 1200 and 1800 on the SAT.
- ~95% of students score between 900 and 2100 on the SAT.
- ~99.7% of students score between 600 and 2400 on the SAT.

Evaluating Normal Approximation

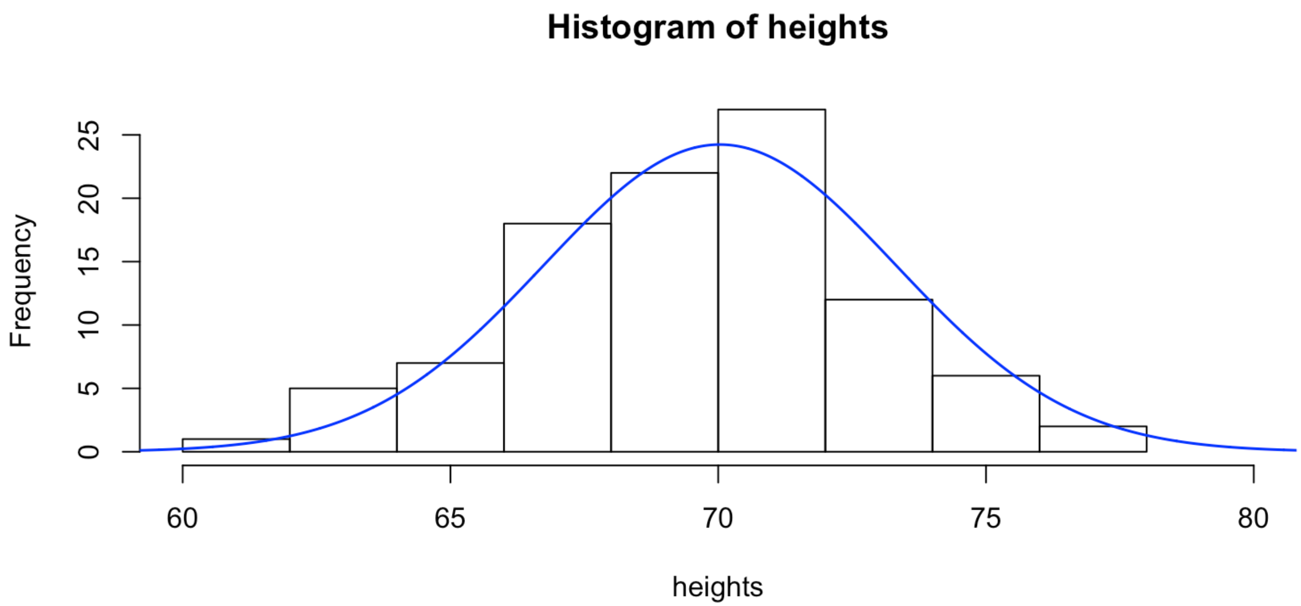
To use the 68-95-99 rule, we must verify the normality assumption. We will want to do this also later when we talk about various (parametric) modeling. Consider a sample of 100 male heights (in inches).



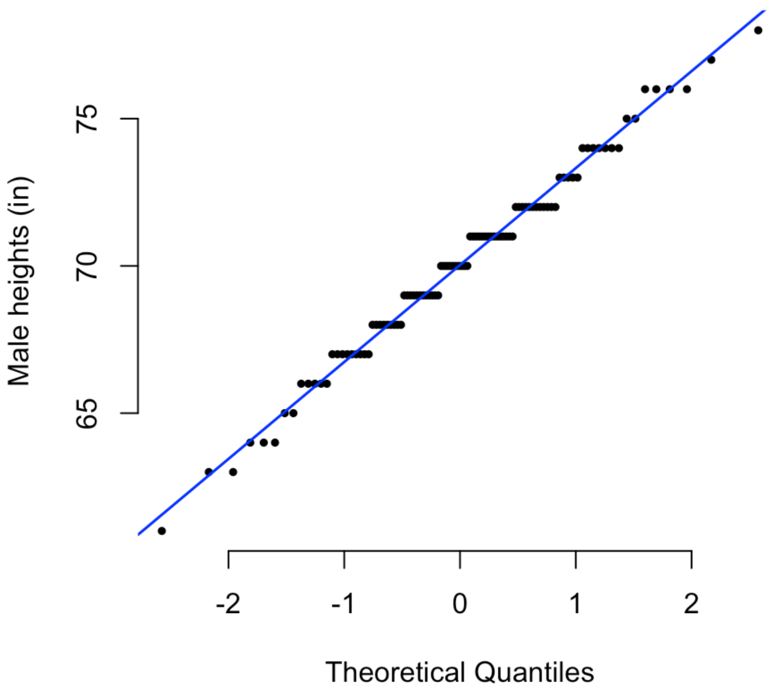
18/22

Evaluating Normal Approximation

Histogram looks normal, but we can overlay a standard normal curve to help evaluation.

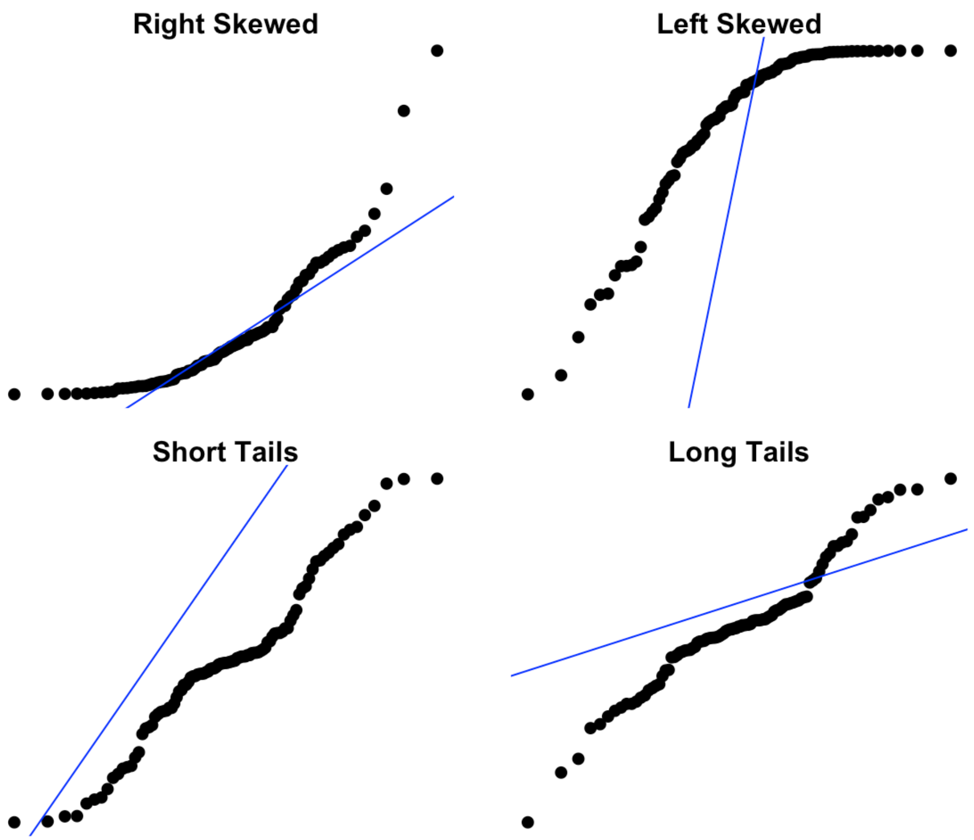


Normal Q-Q Plot



- Data are plotted on the y-axis of a normal probability plot, and theoretical quantiles (following a normal distribution) on the x-axis.
- If there is a linear relationship in the plot, then the data follow a nearly normal distribution.
- Constructing a normal probability plot requires calculating percentiles and corresponding z-

Skewness



Simulated Normal Q-Q Plots

```
DATA606::qqnormsim(heights)
```