Distributions

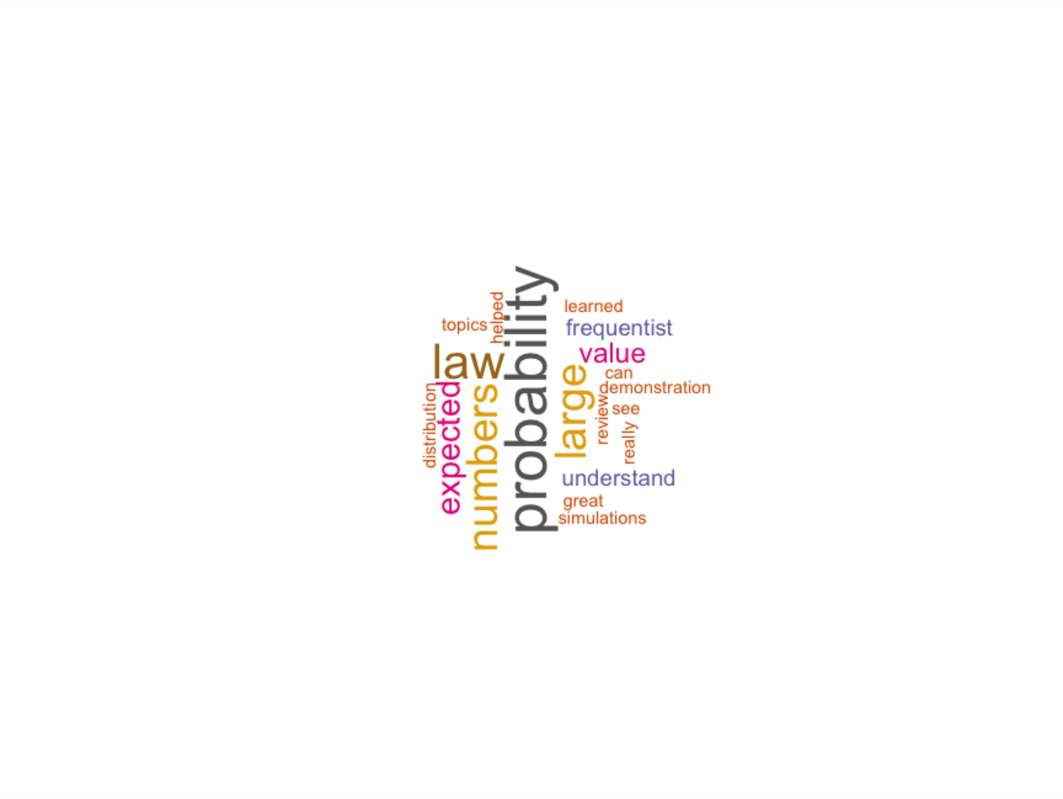
DATA 606 - Statistics & Probability for Data Analytics

Jason Bryer, Ph.D.

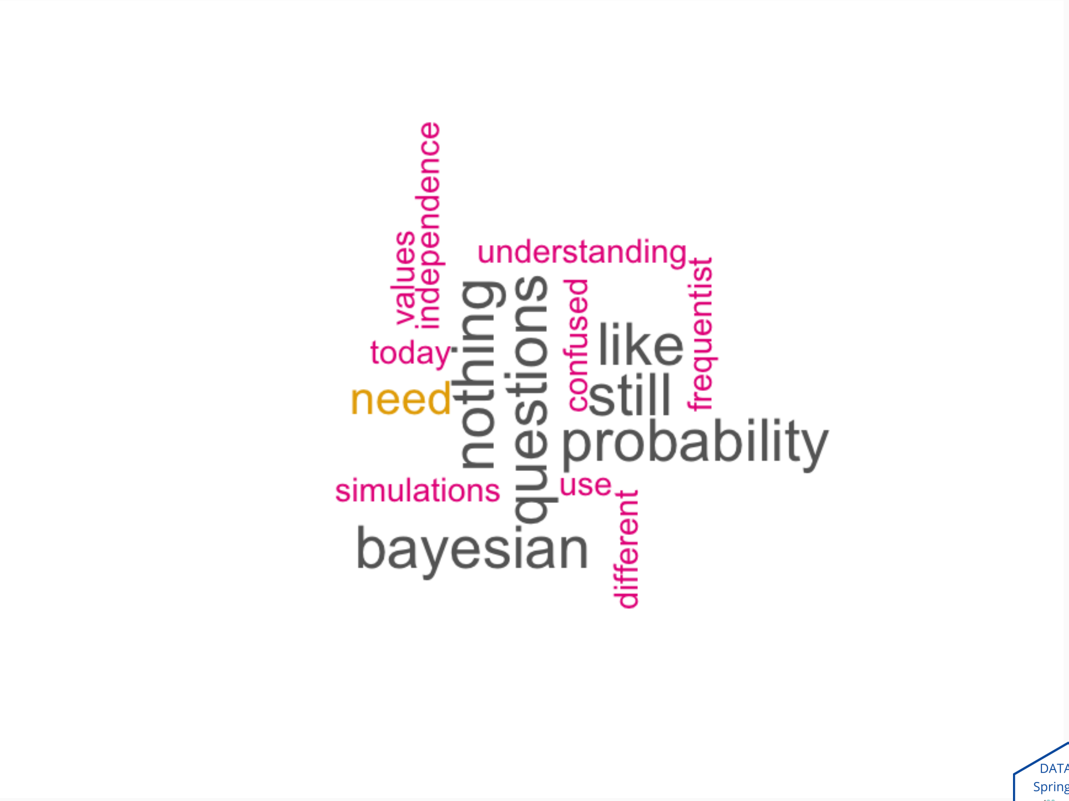
February 24, 2021

One Minute Paper Results

What was the most important thing you learned during this class?



What important question remains unanswered for you?



Homework Presentations

- 3.2 Vic Chan
- 3.3 Ethan
- 3.41 MariAlejandra Ginorio
- 3.43 Michael Ippolito

Coin Tosses Revisited

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

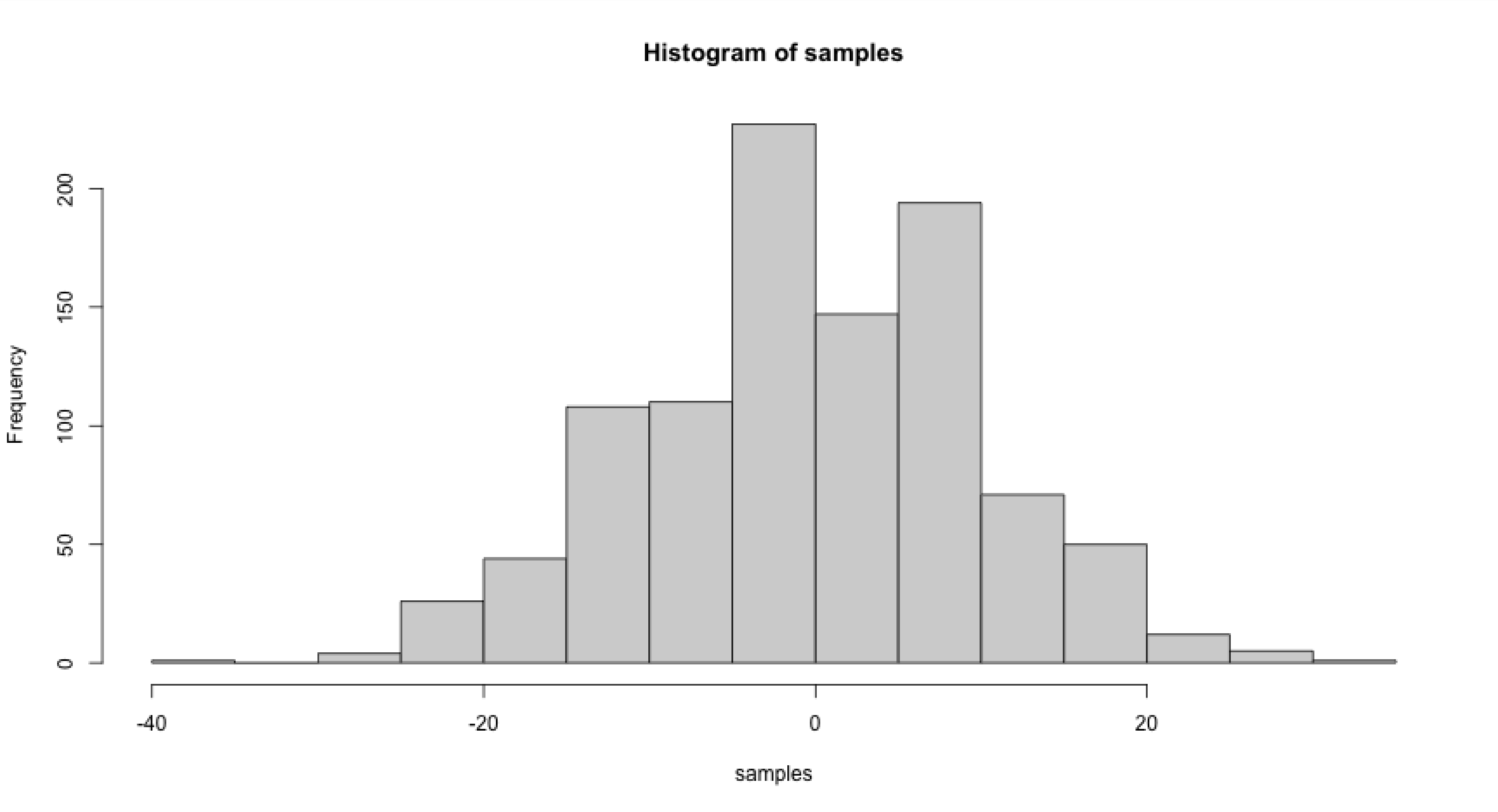
Many Random Samples

```
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, n = 15)
```

```
##  [1]  -4   0  -4   0  18   6   0   2   4  18 -16  -2   8  -2   2
```

Histogram of Many Random Samples

```
hist(samples)
```



Properties of Distribution

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

```
## [1] 0.236
```

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

```
## [1] 10.14096
```

Properties of Distribution (cont.)

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

```
## [1] 0.678
```

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

```
## [1] 0.964
```

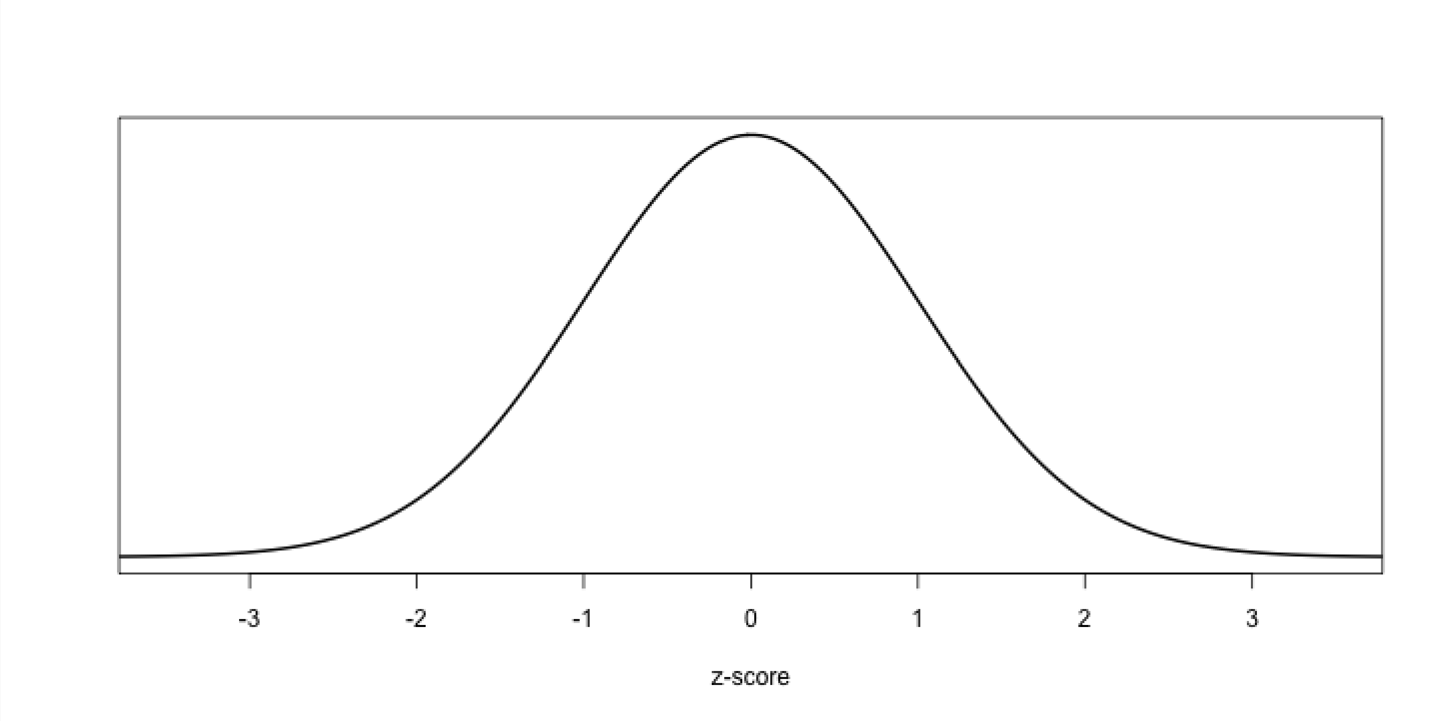
```r
within3sd <- samples[samples >= m.sam - 3 * s.sam & samples <= m.sam + 3 * s.sam]
length(within3sd) / length(samples)
```

```
## [1] 0.998
```

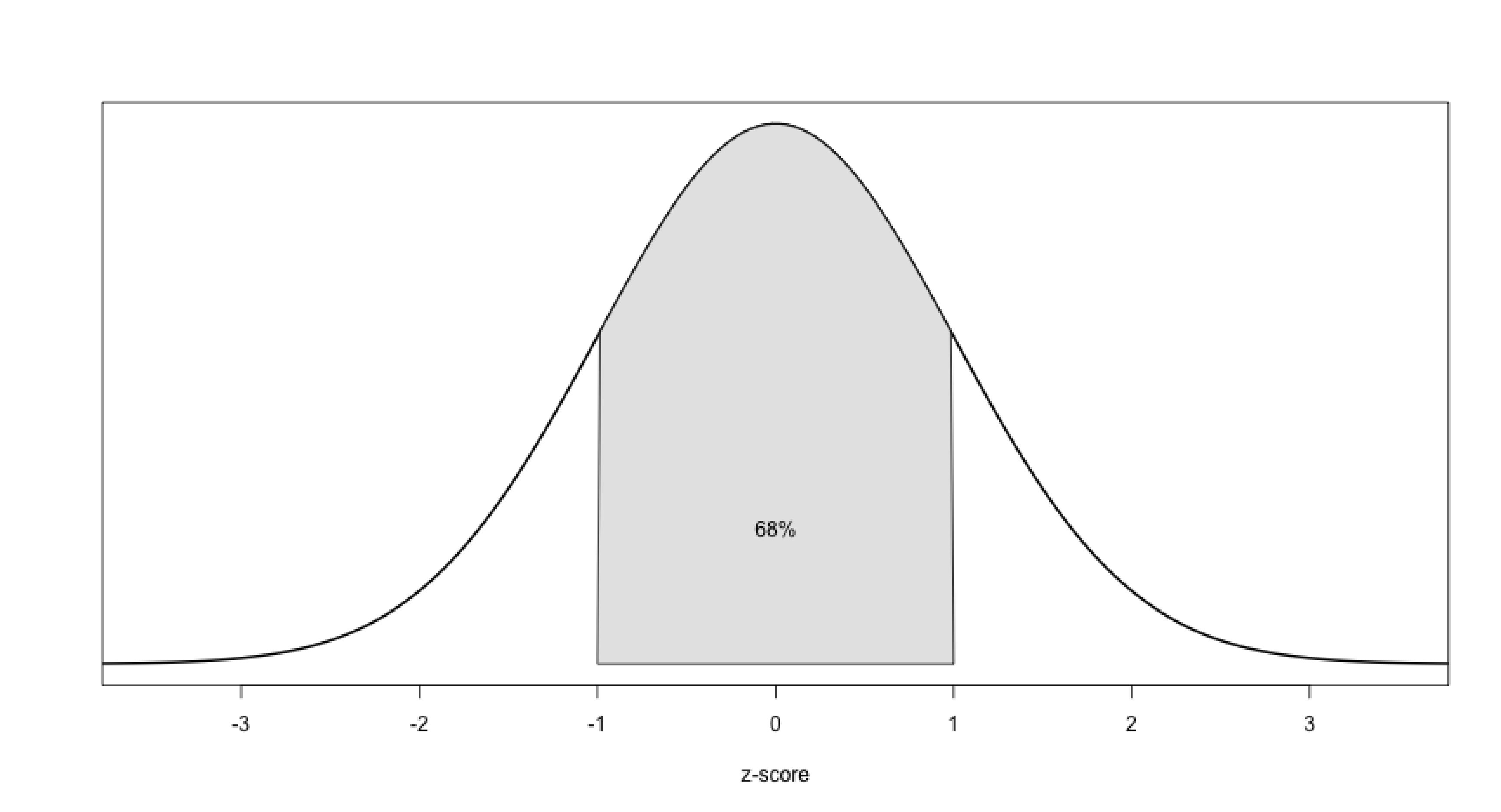
Standard Normal Distribution

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

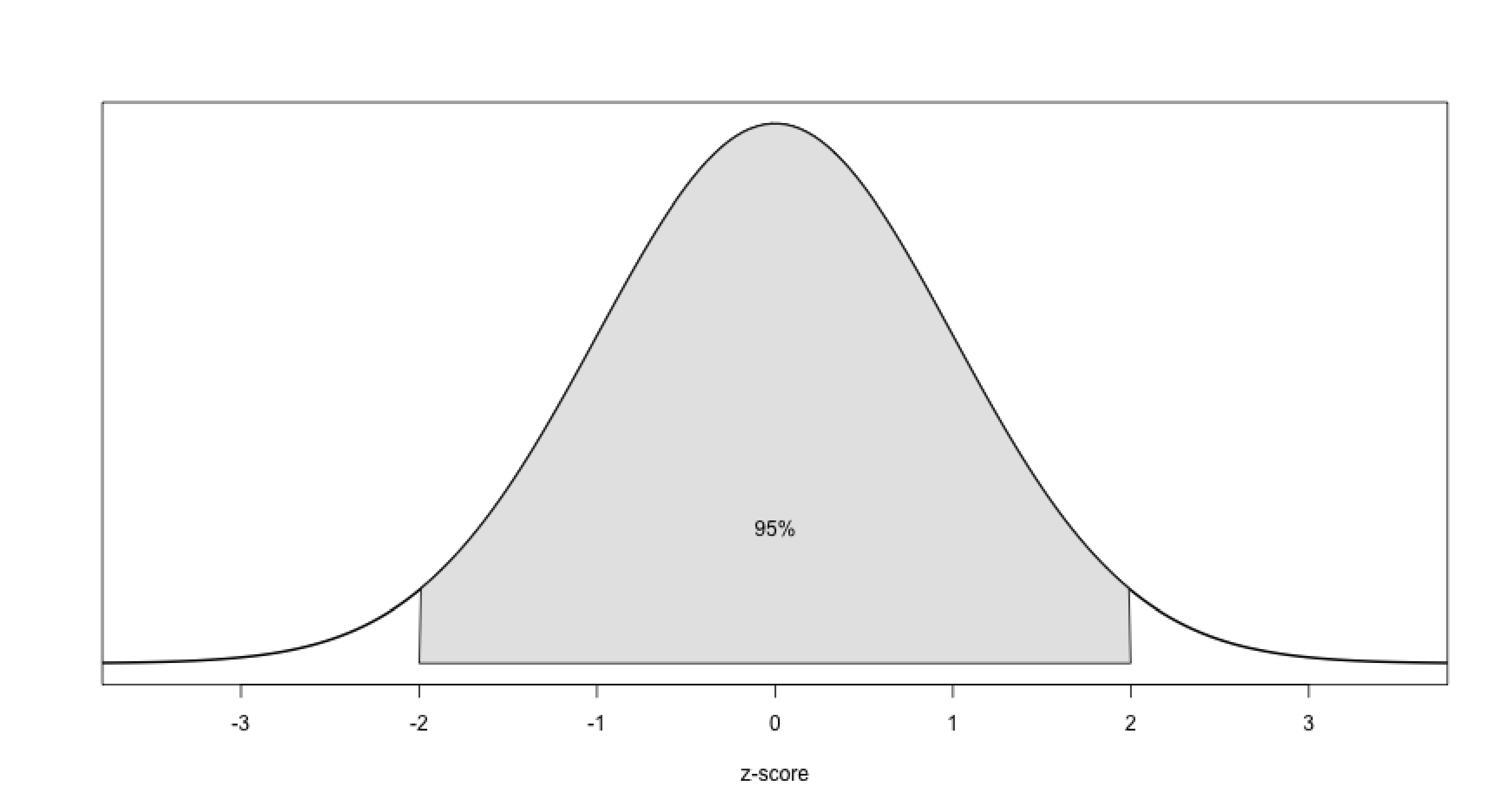
```r
x <- seq(-4,4,length=200); y <- dnorm(x,mean=0, sd=1)
plot(x, y, type = "l", lwd = 2, xlim = c(-3.5,3.5), ylab='', xlab='z-score', yaxt='n')
```



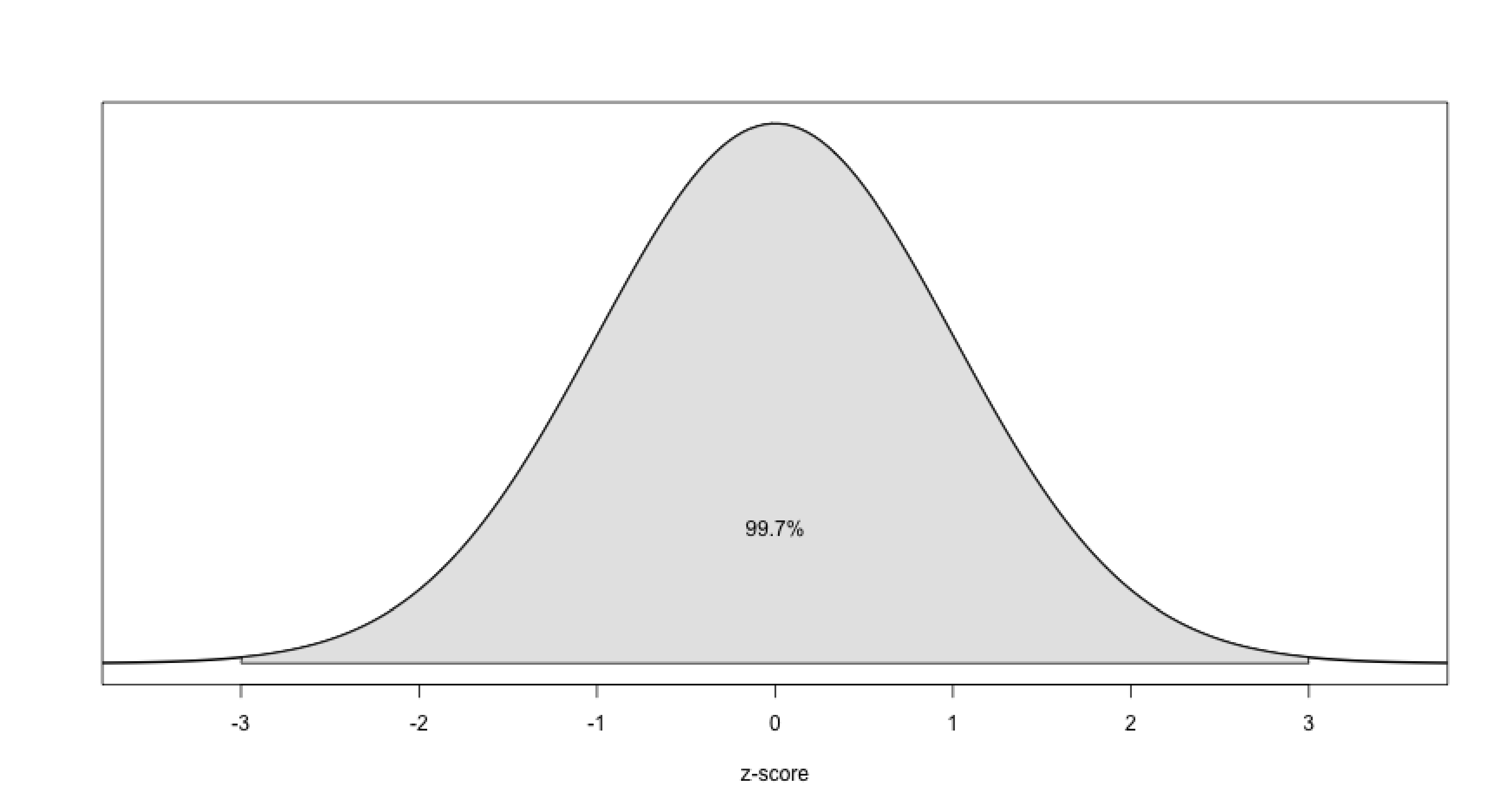
Standard Normal Distribution



Standard Normal Distribution



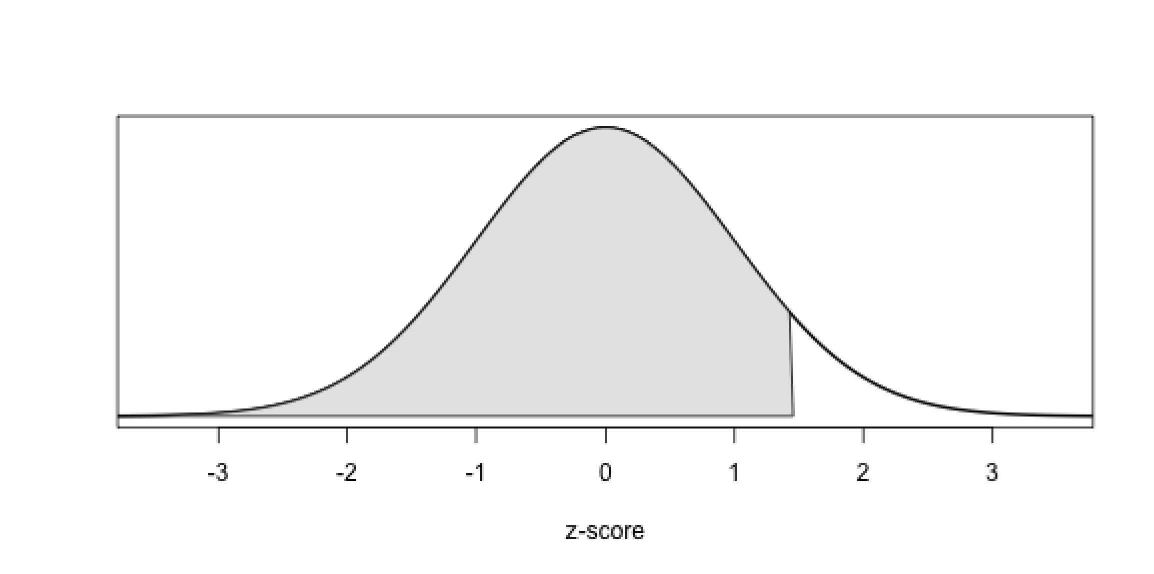
Standard Normal Distribution



What's the likelihood of ending with less than 15?

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

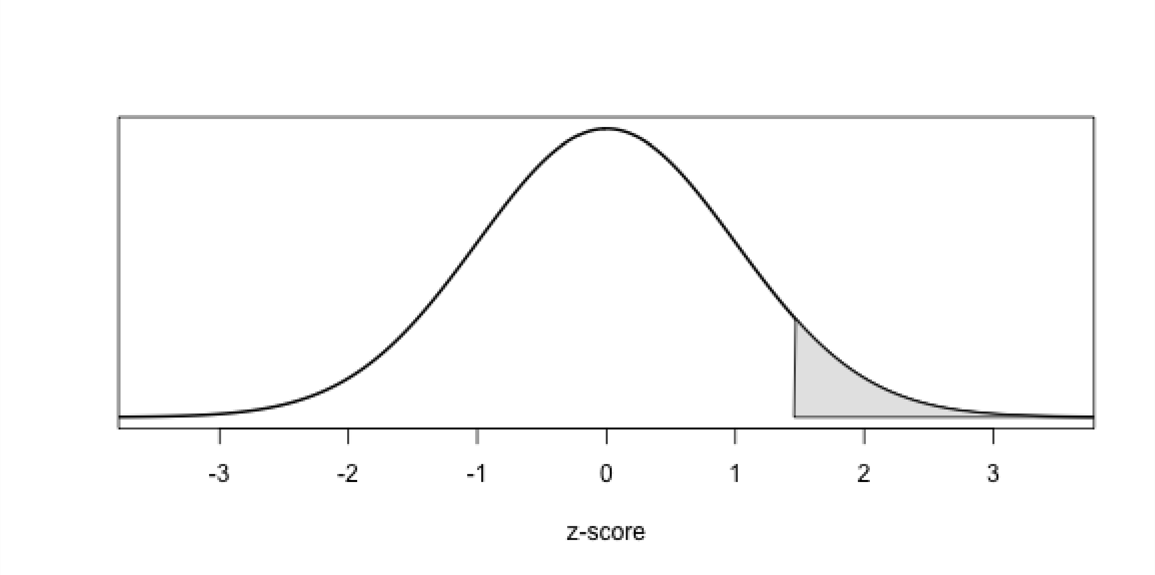
```
## [1] 0.9272867
```



What's the likelihood of ending with more than 15?

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

```
## [1] 0.07271325
```



Comparing Scores on Different Scales

SAT scores are distributed nearly normally with mean 1500 and standard 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:

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

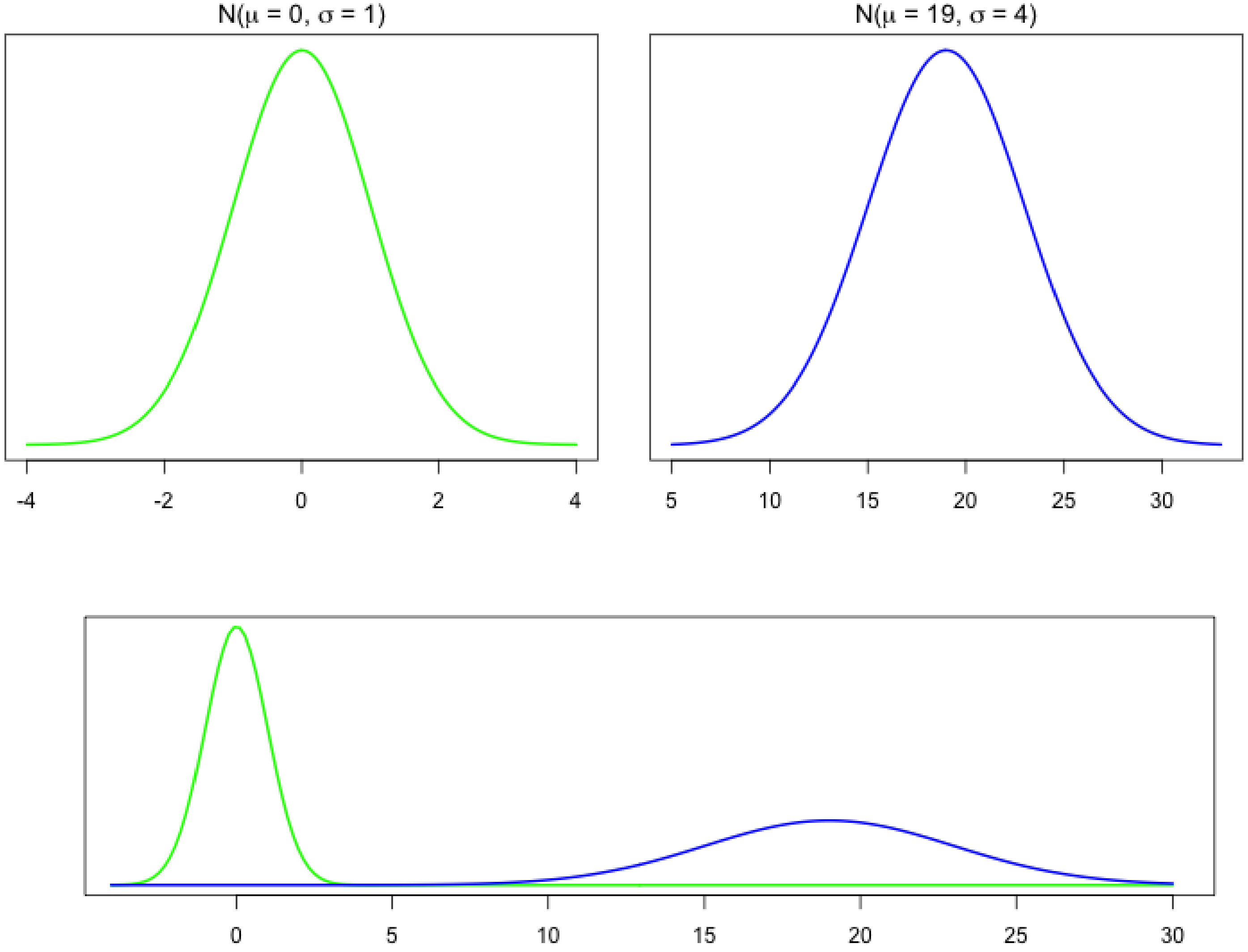
- Z-Scores have a mean = 0 and standard deviation = 1.

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



N(μ = 0, σ = 1)

N(μ = 19, σ = 4)

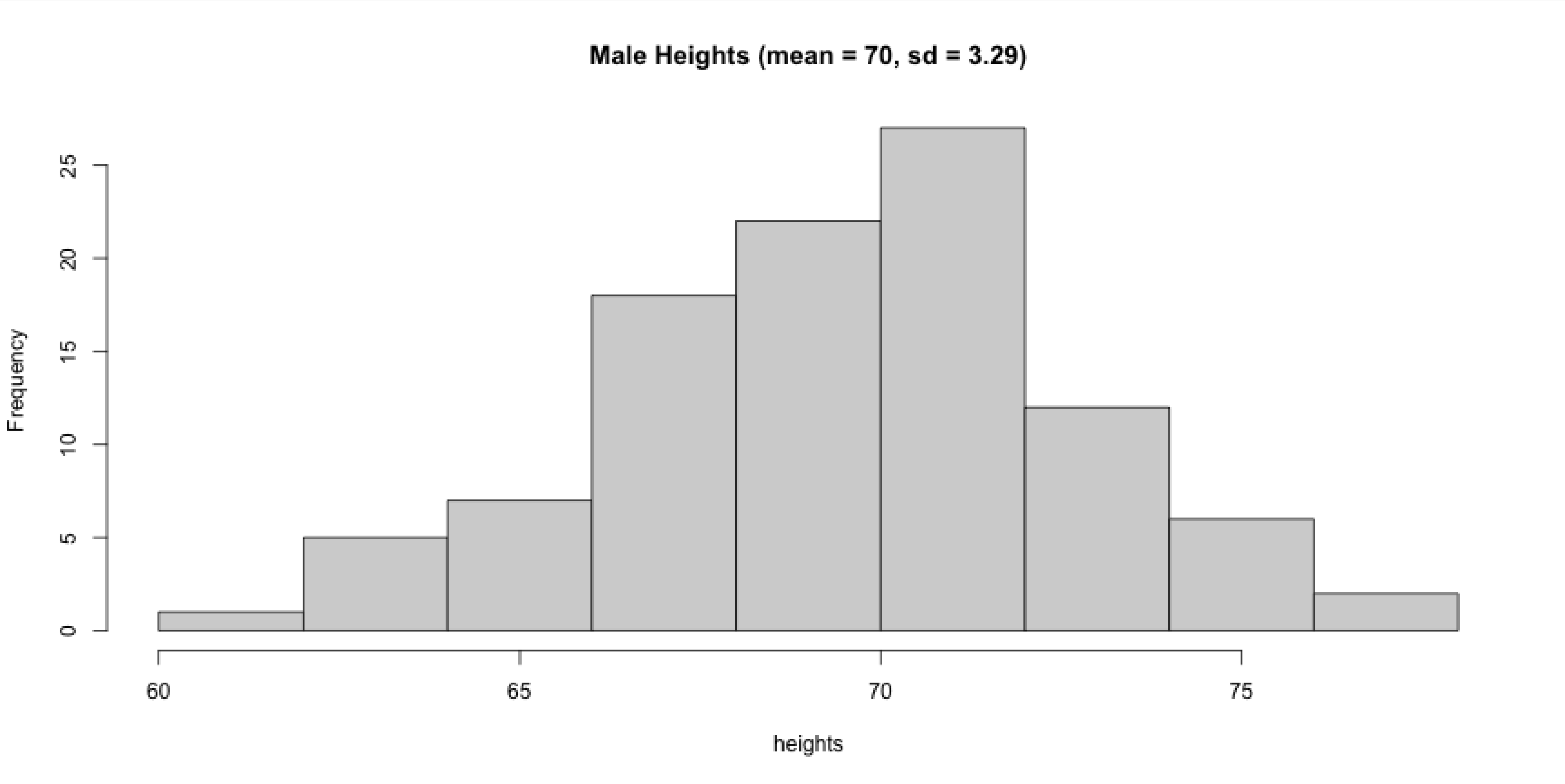
SAT Variability

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.

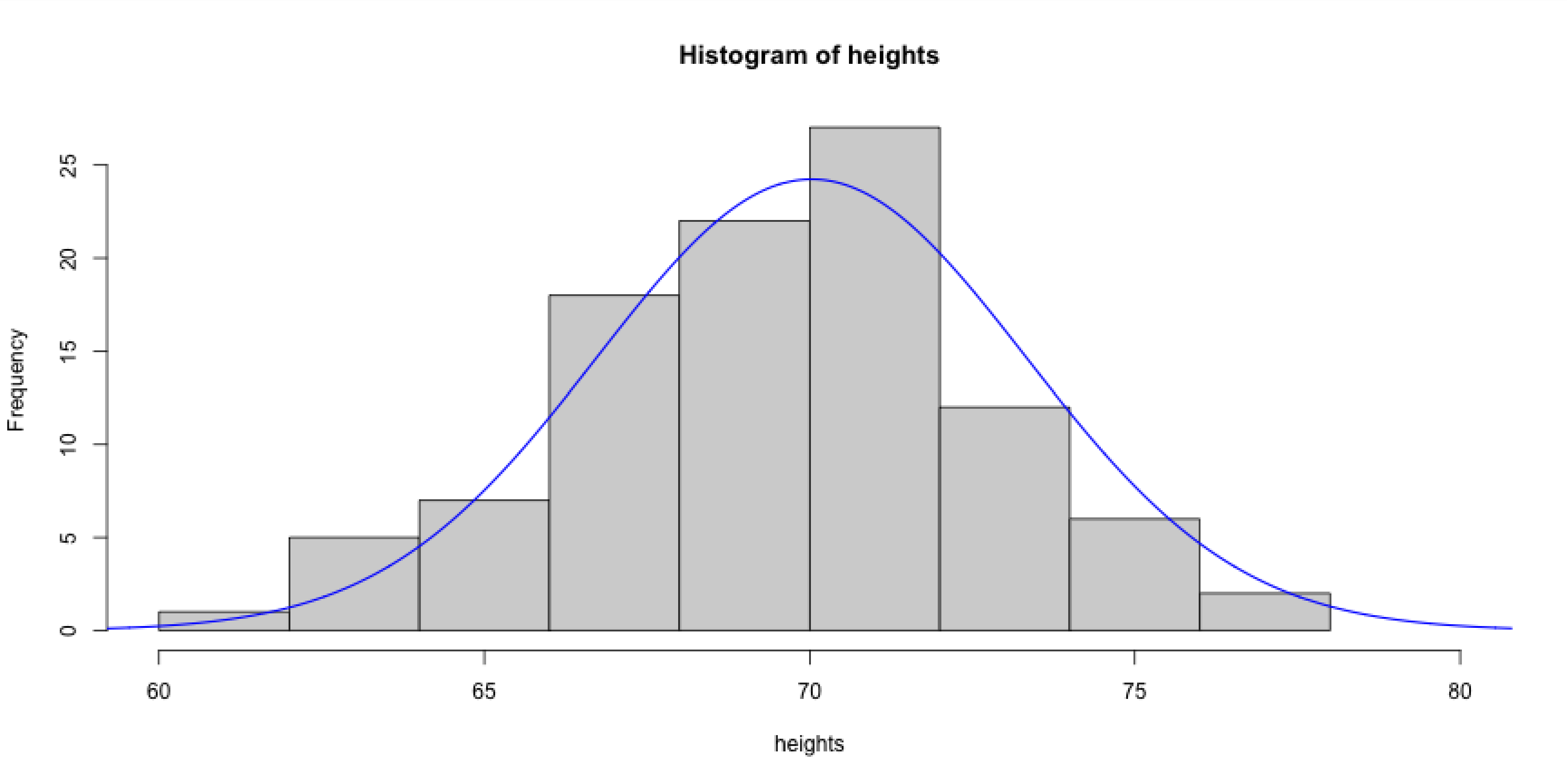
Evaluating Normal Approximation

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).

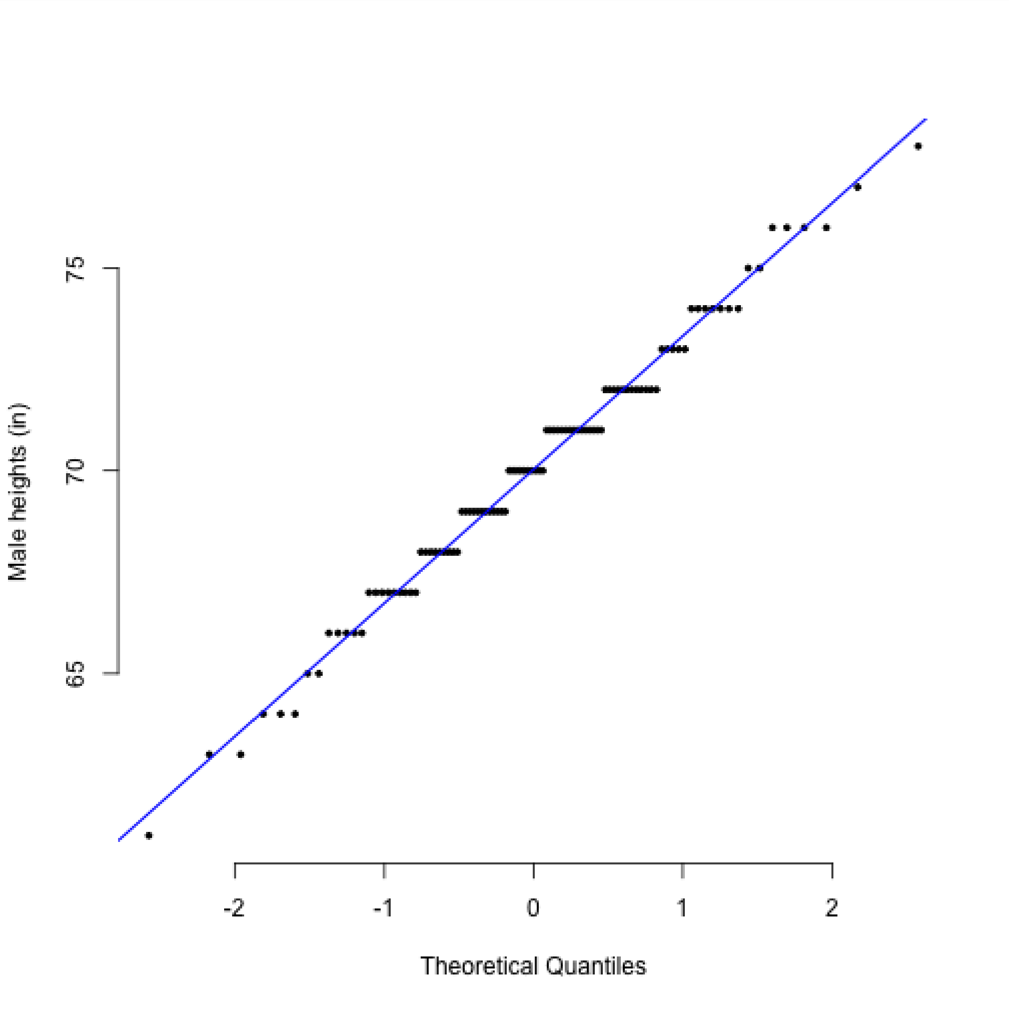


Evaluating Normal Approximation

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



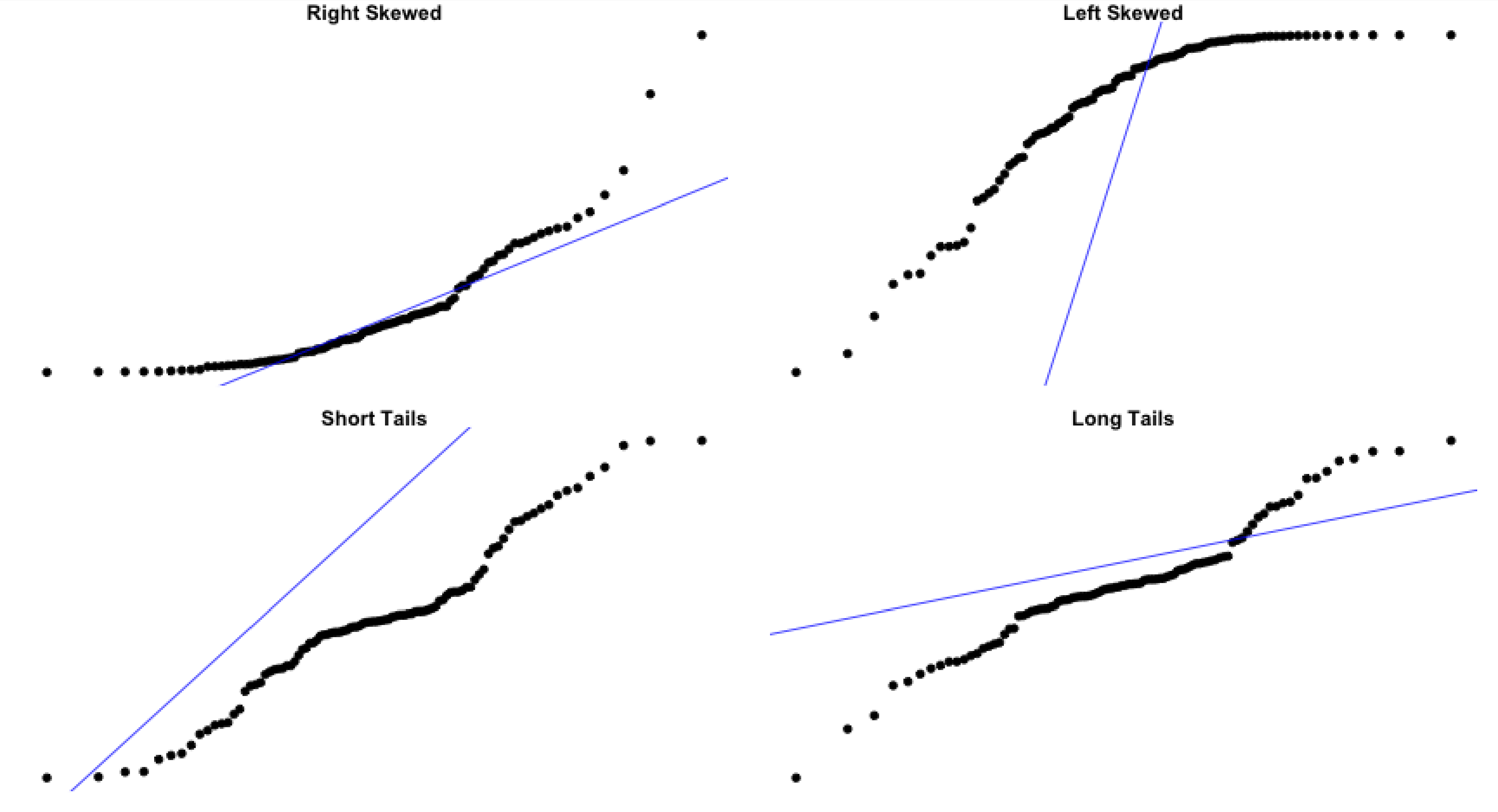
Normal Q-Q Plot



- 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-scores for each observation, which is tedious. Therefore we generally rely on software when making these plots.

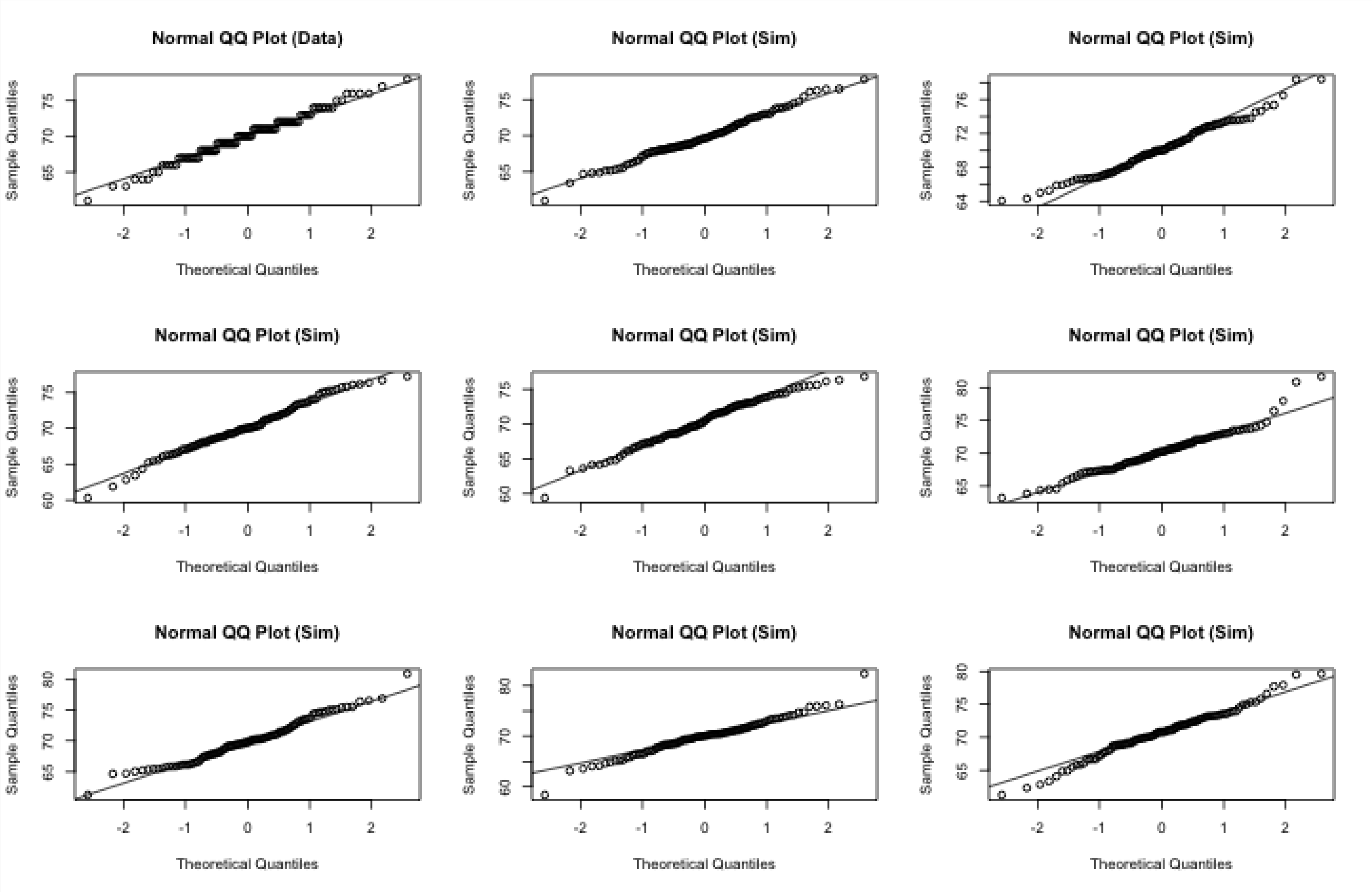
DATA 606
Spring 2021

Skewness



Simulated Normal Q-Q Plots

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



One Minute Paper

Complete the one minute paper:

https://forms.gle/gY9SeBCPggHEtZYw6

1. What was the most important thing you learned during this class?
2. What important question remains unanswered for you?