Distributions

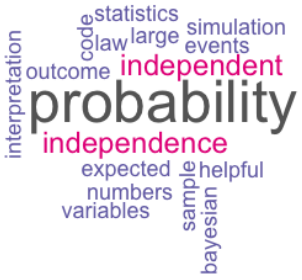
DATA 606 - Statistics & Probability for Data Analytics

Jason Bryer, Ph.D. and Angela Lui, Ph.D.

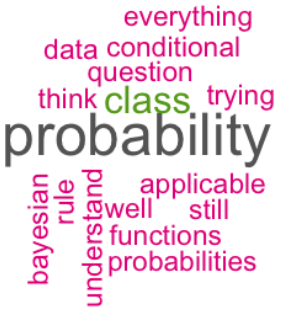
February 23, 2022

One Minute Paper Results

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



What important question remains unanswered for you?



Announcements

- **Next week's meetup (March 2nd) will be at 3pm!**
- We have pushed the midterm back one week. Will no be March 16th through March 22nd.
- The solution to simulating the card game is located here: https://github.com/jbryer/DATA606Spring2022/blob/main/R/probability_card_game_simulation.R
- Reminder: I will be giving a talk on R Package Development on March 1st at 7pm. More info here: https://www.meetup.com/Albany-R-Users-Group/events/282413860/

Coin Tosses Revisited

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

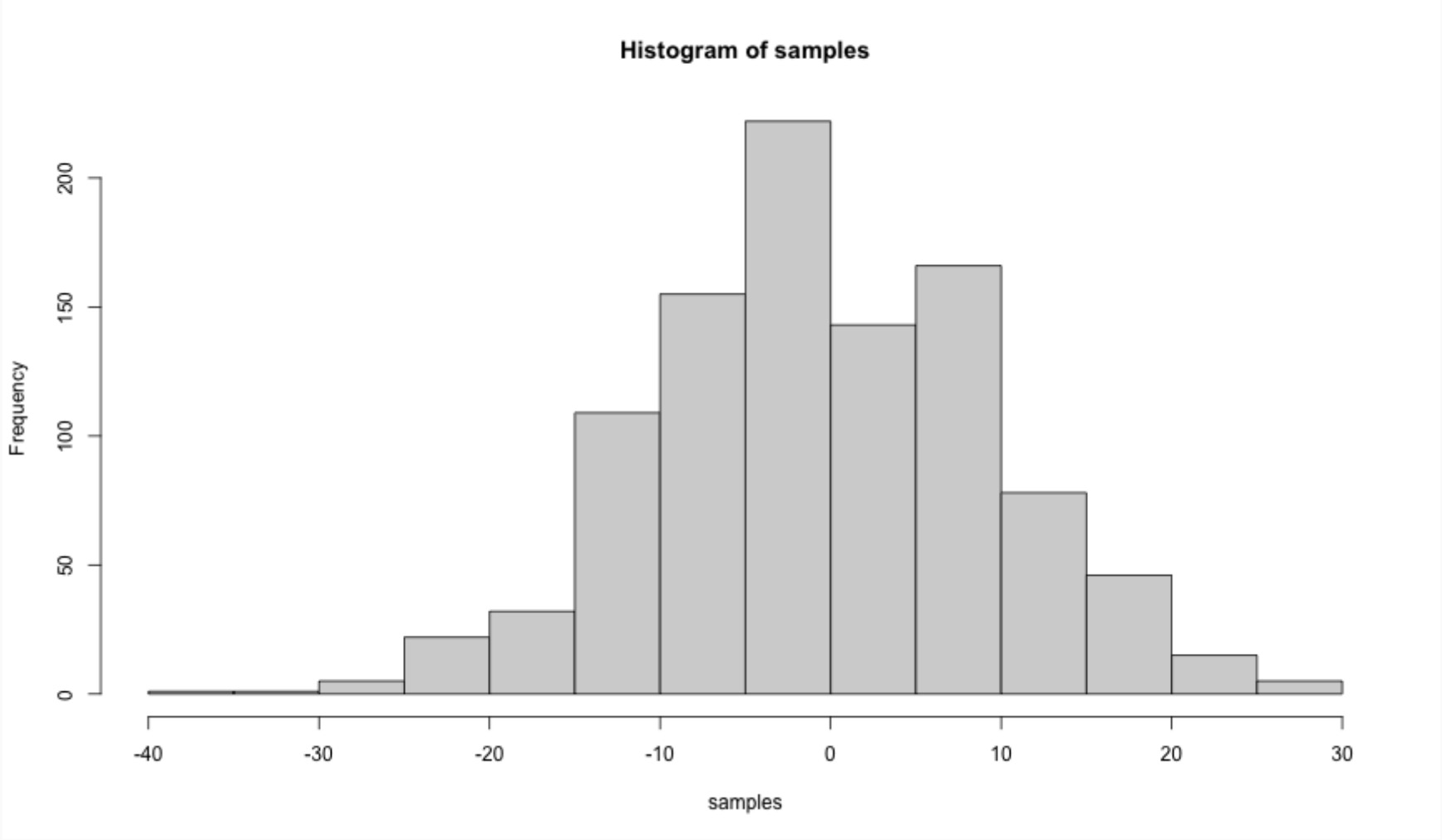
Many Random Samples

```
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]  -8   6 -14  10 -14 -14 -10   2  24  26  -2   2  10   4  -4
```

Histogram of Many Random Samples

```
hist(samples)
```



Properties of Distribution

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

```
## [1] 0.066
```

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

```
## [1] 9.989767
```

Properties of Distribution (cont.)

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

```
## [1] 0.686
```

```
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)
```

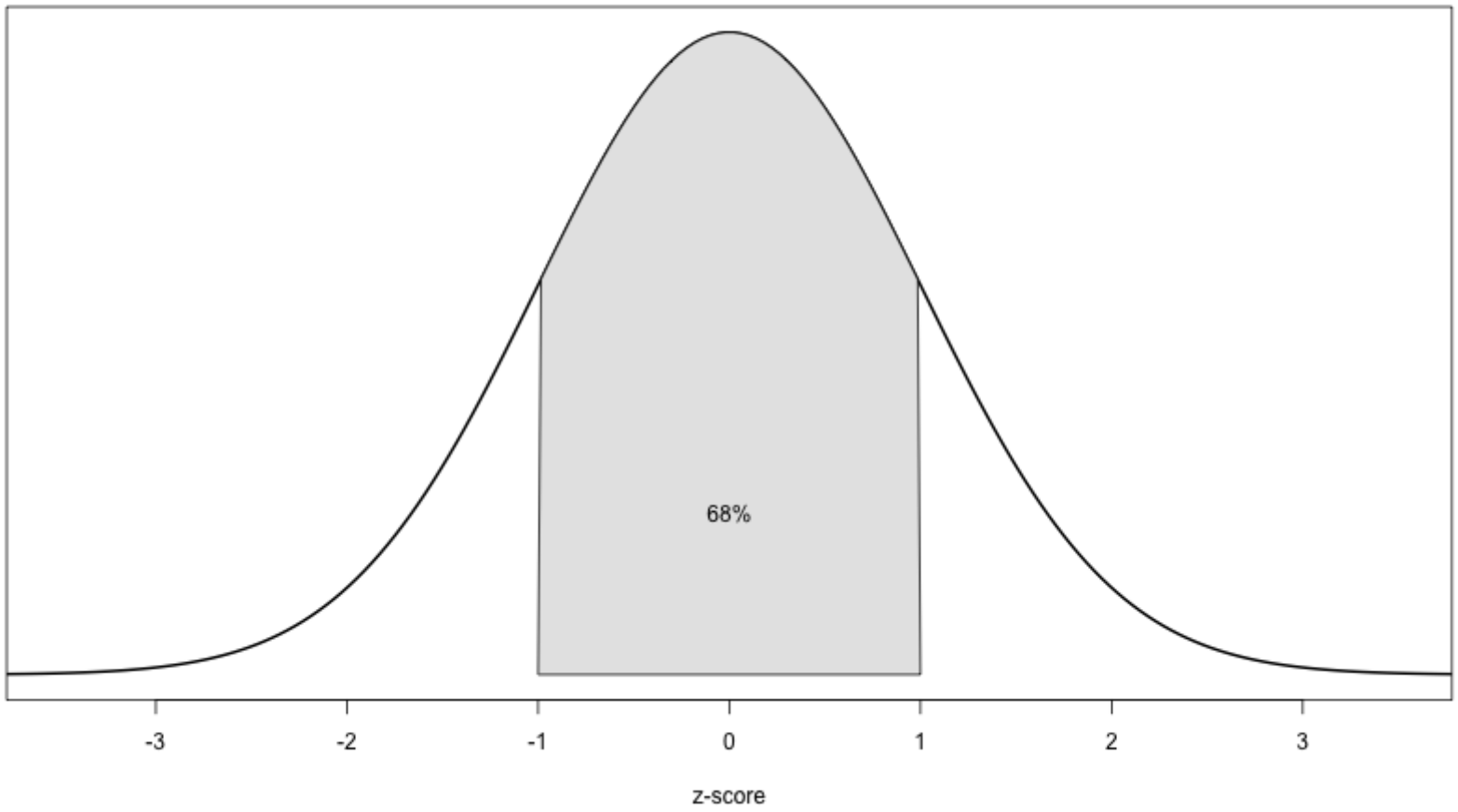
```
## [1] 0.998
```

Standard Normal Distribution

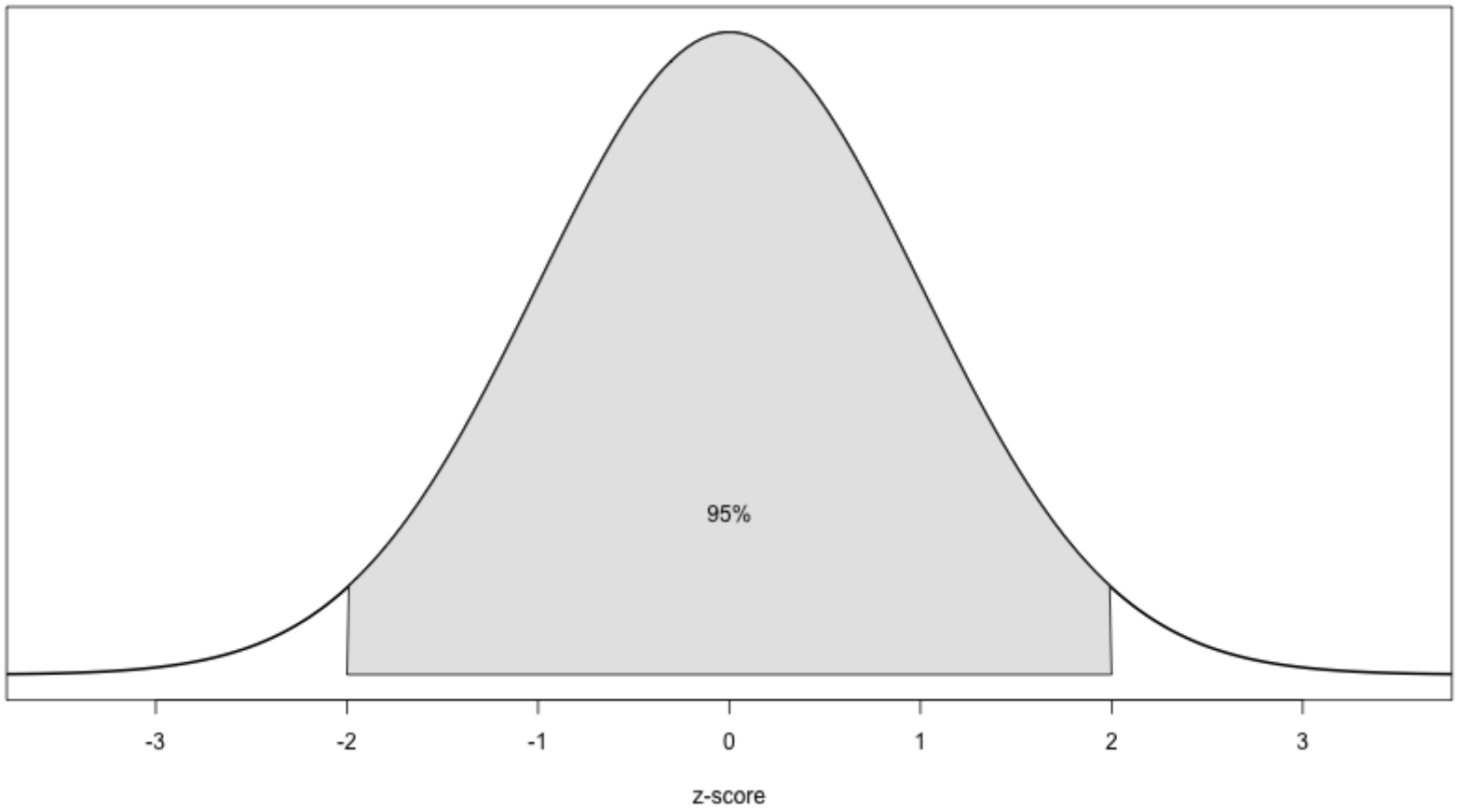
$$f\left(x|\mu,\sigma\right)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

```
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')
```

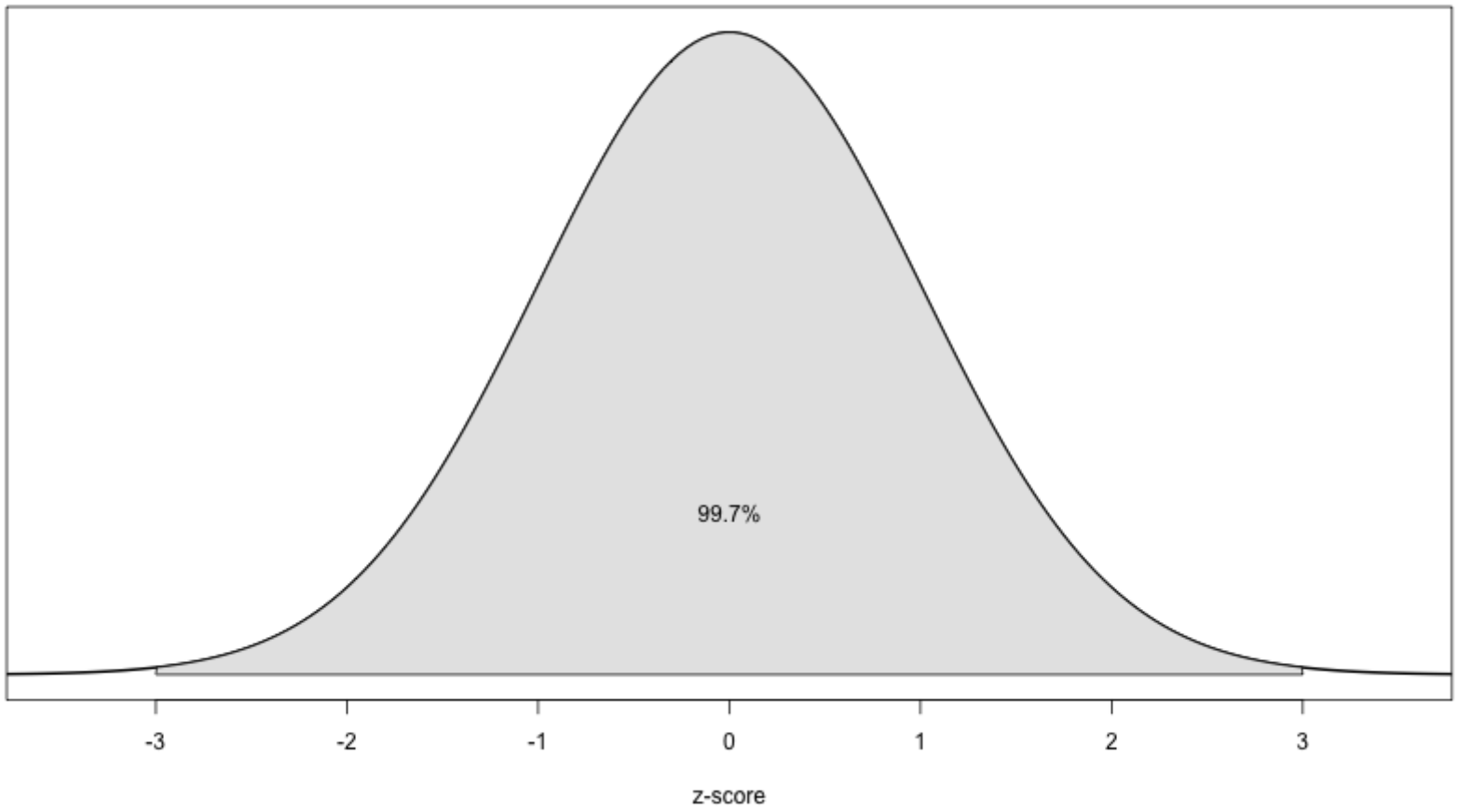
Standard Normal Distribution



Standard Normal Distribution



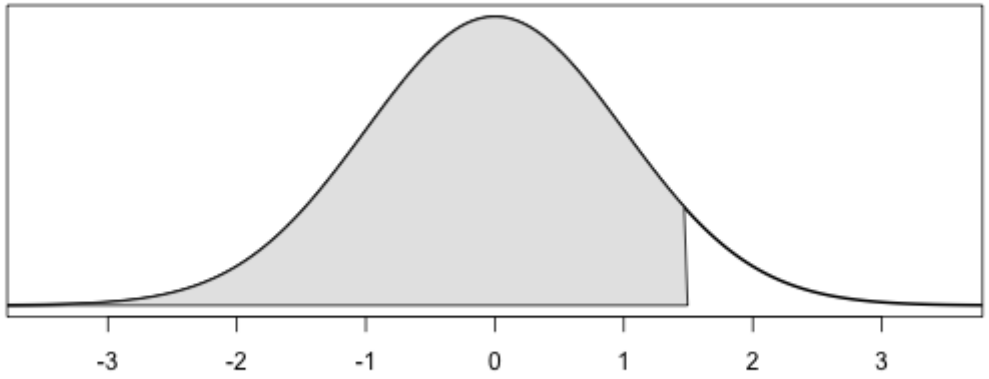
Standard Normal Distribution



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

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

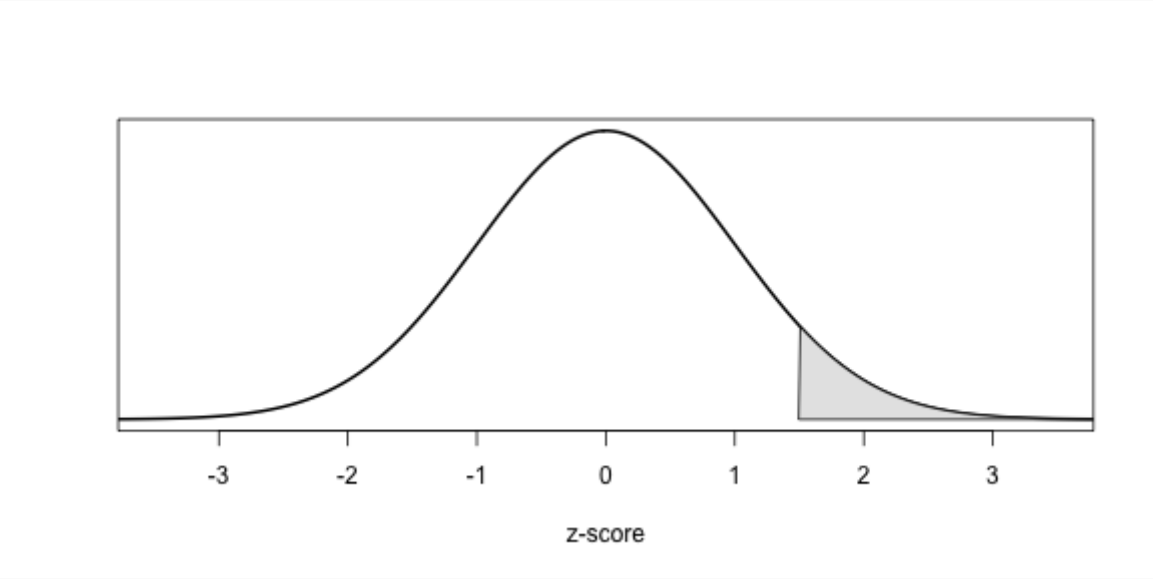
```
## [1] 0.9325336
```



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

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

```
## [1] 0.06746638
```



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:

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

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

Dual Scales

Some problems[1]:

- The designer has to make choices about scales and this can have a big impact on the viewer
- "Cross-over points" where one series cross another are results of the design choices, not intrinsic to the data, and viewers (particularly unsophisticated viewers)
- They make it easier to lazily associate correlation with causation, not taking into account autocorrelation and other time-series issues
- Because of the issues above, in malicious hands they make it possible to deliberately mislead

This example looks at the relationship between NZ dollar exchange rate and trade weighted index.

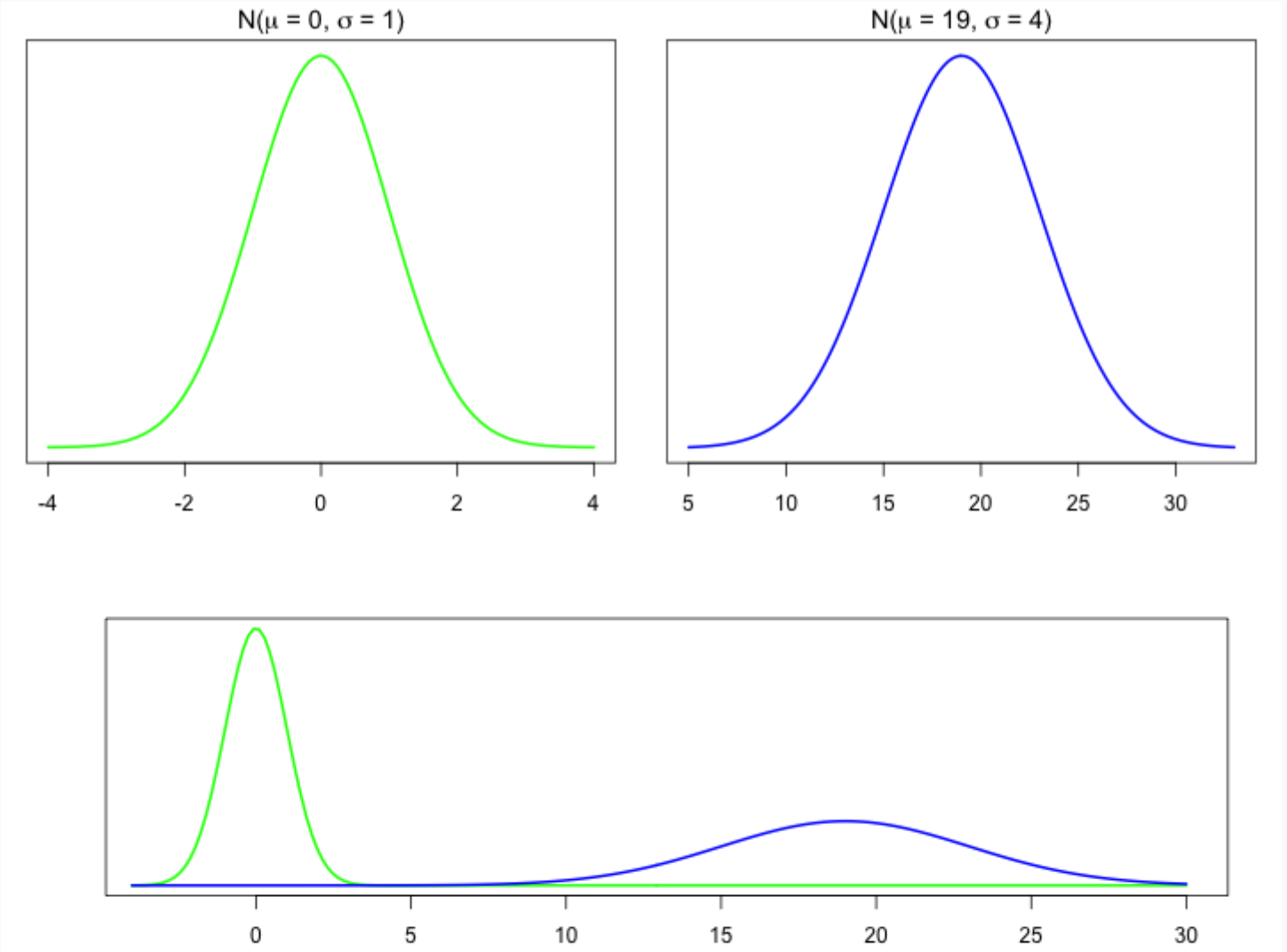
```
DATA606::shiny_demo('DualScales', package='DATA606')
```

My advise:

- Avoid using them. You can usually do better with other plot types.
- When necessary (or compelled) to use them, rescale (using z-scores, we'll discuss this in a few weeks)

[1] http://blog.revolutionanalytics.com/2016/08/dual-axis-time-series.html
[2] http://ellisp.github.io/blog/2016/08/18/dualaxes

Standard Normal Parameters



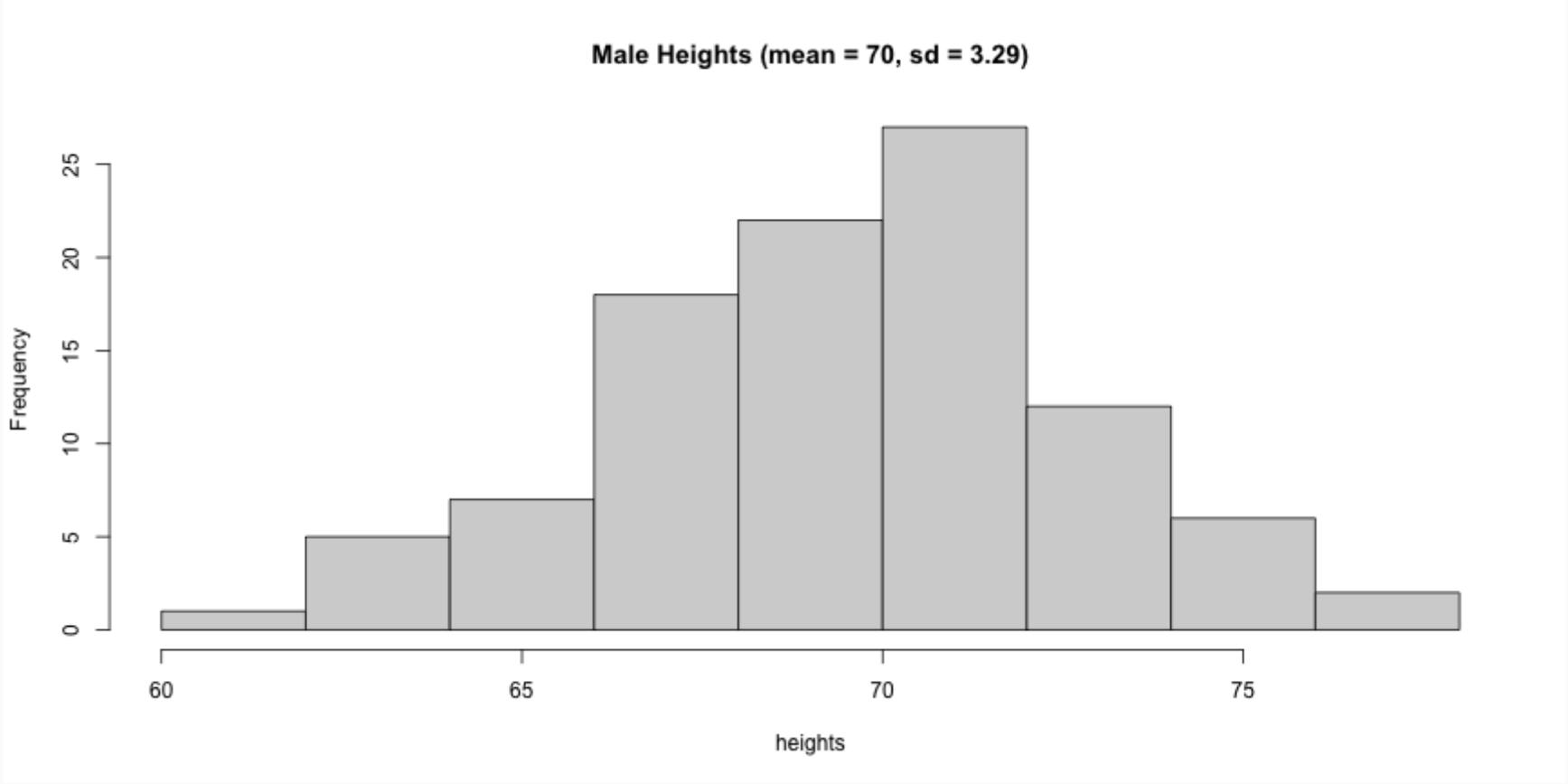
SAT Variability

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.

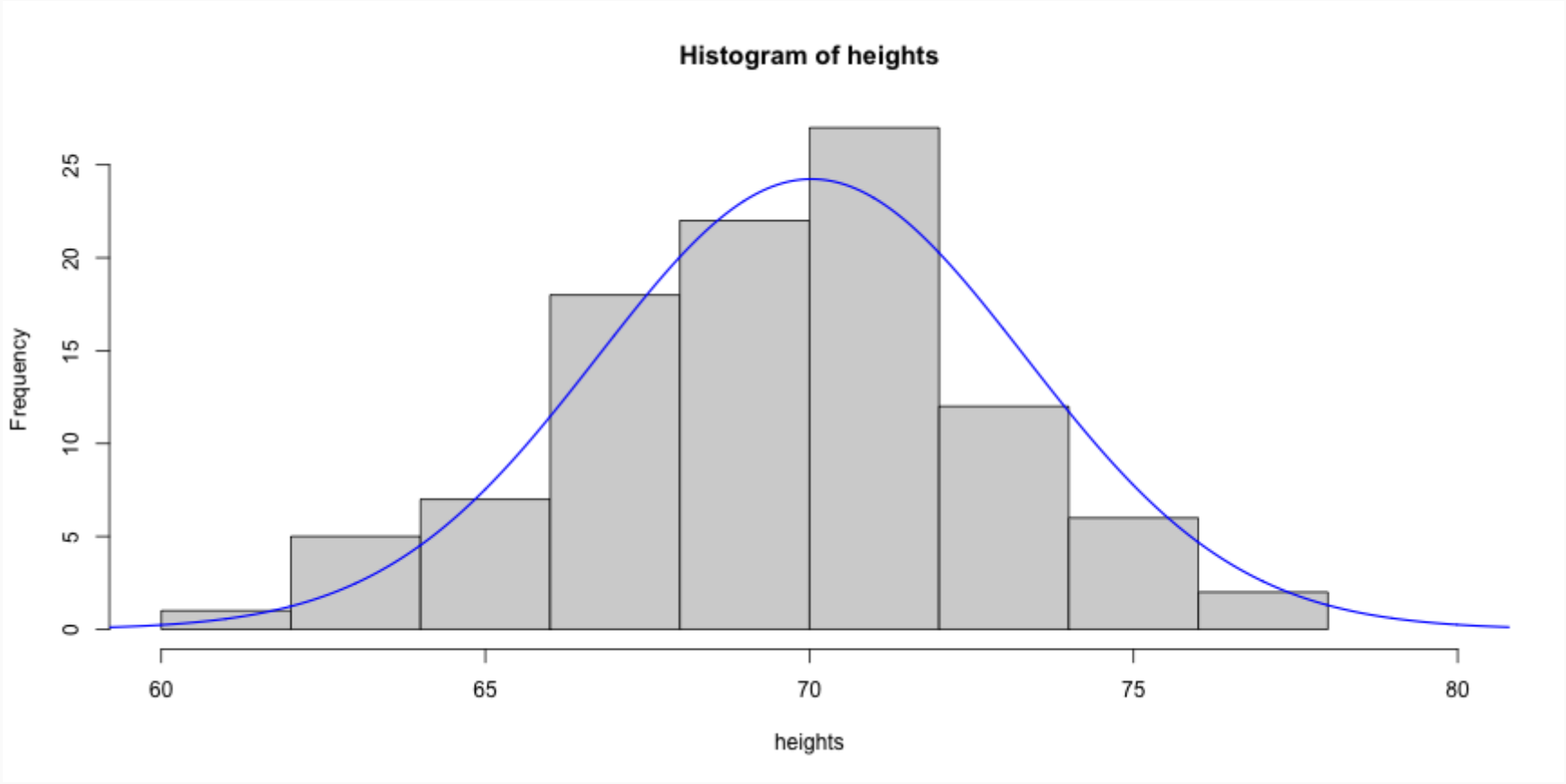
Evaluating Normal Approximation

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).

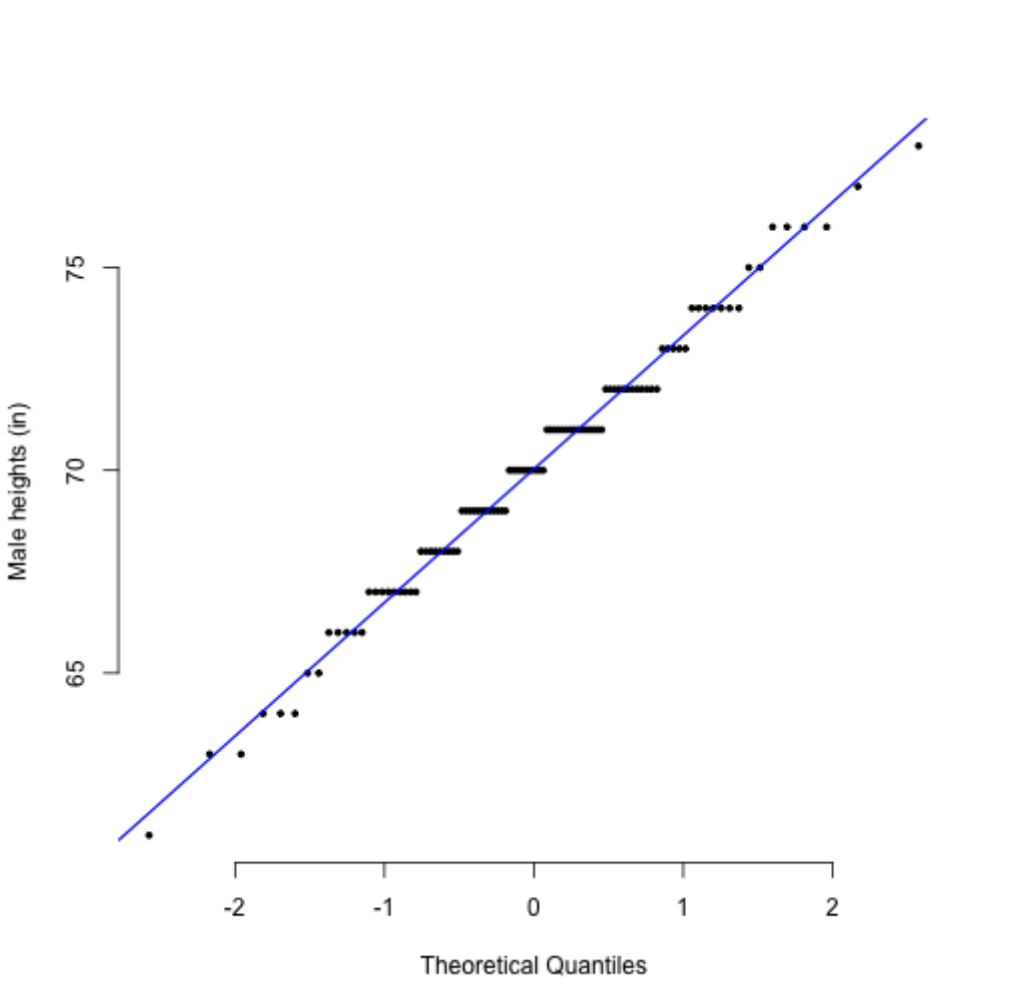


Evaluating Normal Approximation

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

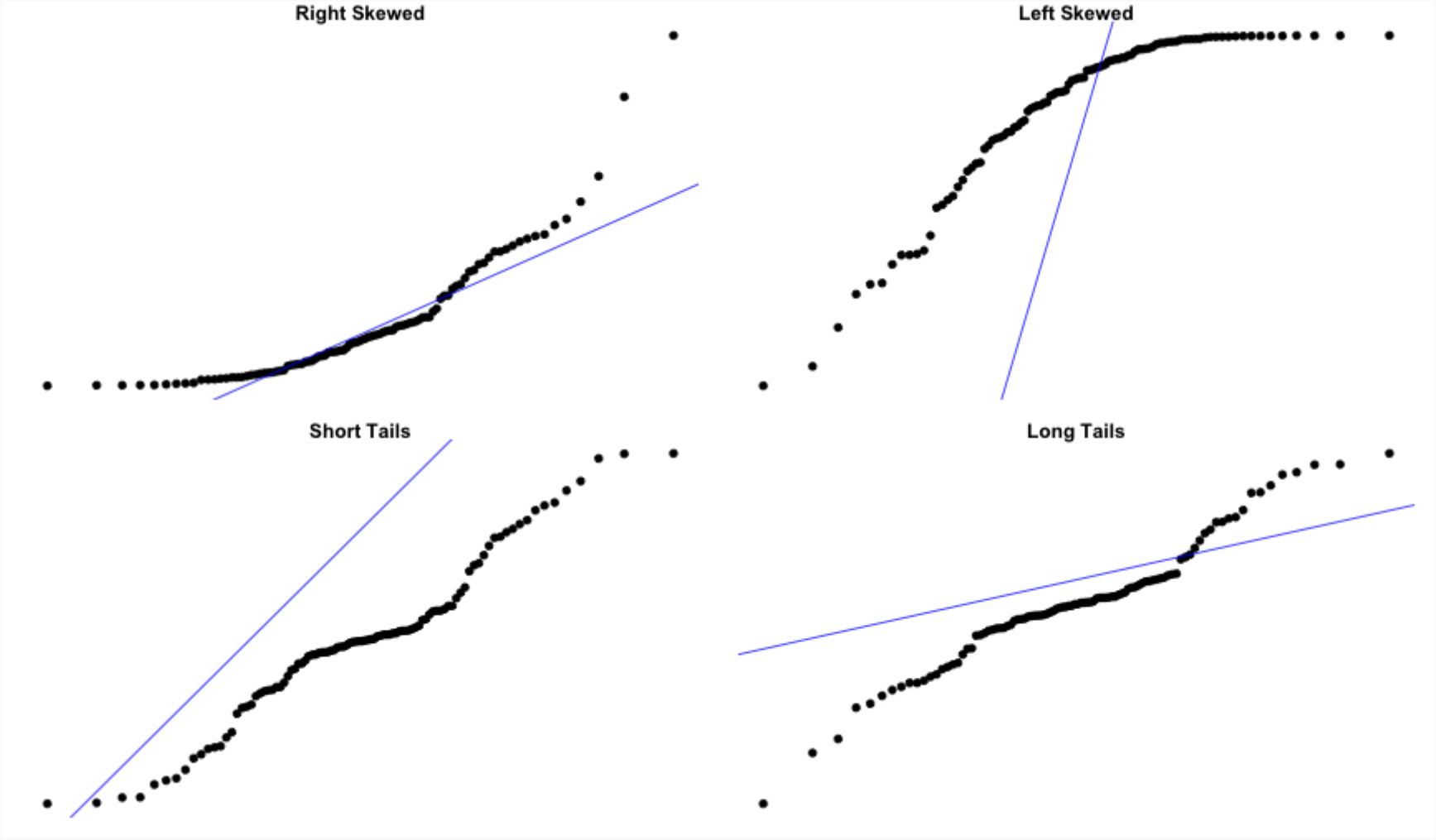


Normal Q-Q Plot



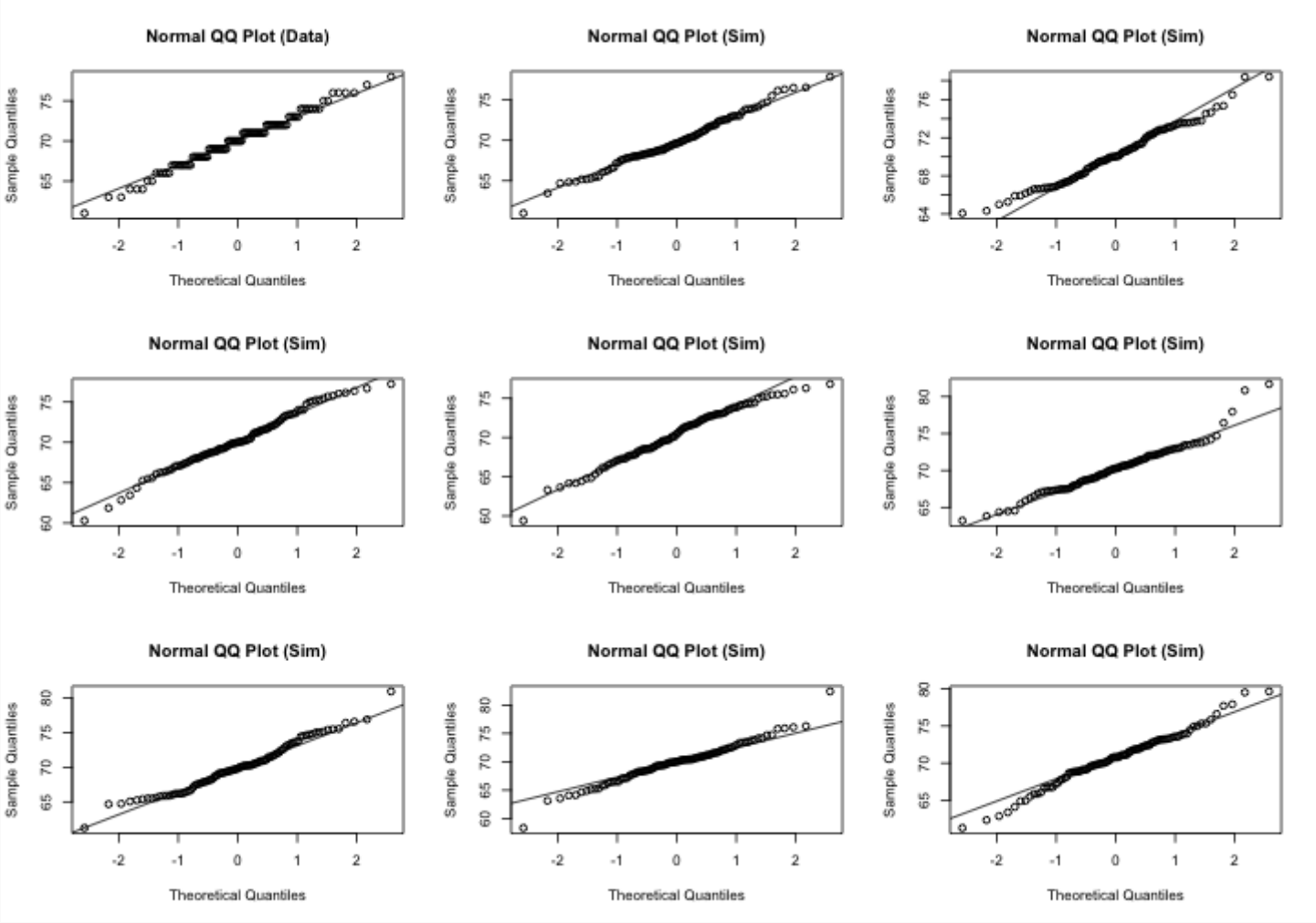
- Data are plotted on the y-axis of a normal probability plot, and theoretical quantiles (following a normal distribution) on the x-axis.
- If there is a linear relationship in the plot, then the data follow a nearly normal distribution.
- Constructing a normal probability plot requires calculating percentiles and corresponding z-scores for each observation, which is tedious. Therefore we generally rely on software when making these plots.

Skewness



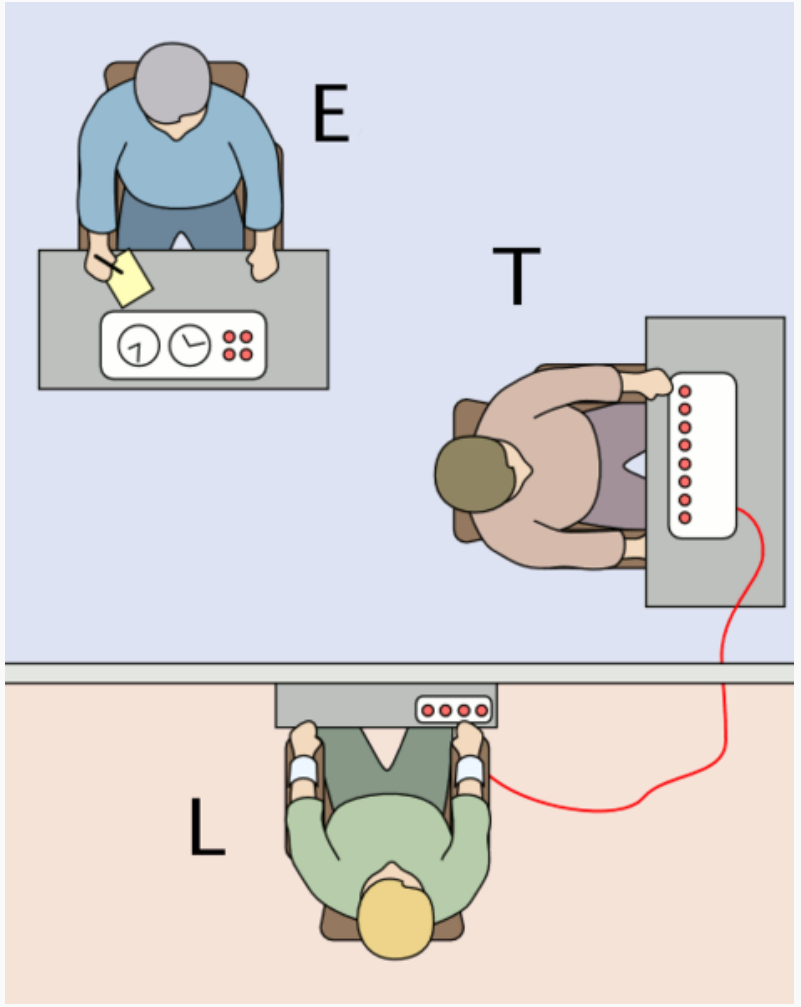
Simulated Normal Q-Q Plots

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



Milgram Experiment

- Stanley Milgram conducted a series of experiments on obedience to authority starting in 1963.
- Experimenter (E) orders the teacher (T), the subject of the experiment, to give severe electric shocks to a learner (L) each time the learner answers a question incorrectly.



Milgram Experiment (cont.)

- The learner is actually an actor, and the electric shocks are not real, but a prerecorded sound is played each time the teacher administers an electric shock.
- These experiments measured the willingness of study participants to obey an authority figure who instructed them to perform acts that conflicted with their personal conscience.
- Milgram found that about 65% of people would obey authority and give such shocks.
- Over the years, additional research suggested this number is approximately consistent across communities and time.

Bernoulli Sequences

- Each person in Milgram's experiment can be thought of as a trial.
- A person is labeled a success if she refuses to administer a severe shock, and failure if she administers such shock.
- Since only 35% of people refused to administer a shock, probability of success is p = 0.35.
- When an individual trial has only two possible outcomes, it is called a **Bernoulli** random variable.

A random variable X has a *Bernoulli distribution* with parameter p if

$$P(X = 1) = p \quad and \quad P(X = 0) = 1 - p$$

for $0 < p < 1$

Geometric distribution

Dr. Smith wants to repeat Milgrams experiments but she only wants to sample people until she finds someone who will not inflict a severe shock. What is the probability that she stops after the first person?

$$P(1^{st} \quad person \quad refuses) = 0.35$$

the third person?

$$P(1^{st} and 2^{nd} shock, 3^{rd} refuses) = \frac{S}{0.65} \times \frac{S}{0.65} \times \frac{R}{0.35} = 0.65^2 \times 0.35 \approx 0.15$$

the tenth person?

Geometric distribution (cont.)

Geometric distribution describes the waiting time until a success for *independent and identically distributed* (iid) Bernouilli random variables.

- independence: outcomes of trials don't affect each other
- identical: the probability of success is the same for each trial

Geometric probabilities

If p represents probability of success, $(1-p)$ represents probability of failure, and n represents number of independent trials

$$P(success \quad on \quad the \quad n^{th} \quad trial) = (1-p)^{n-1}p$$

Expected value

How many people is Dr. Smith expected to test before finding the first one that refuses to administer the shock?

The expected value, or the mean, of a geometric distribution is defined as $\frac{1}{p}$.

$$\mu = \frac{1}{p} = \frac{1}{0.35} = 2.86$$

She is expected to test 2.86 people before finding the first one that refuses to administer the shock.

But how can she test a non-whole number of people?

Expected value and its variability

$$\mu = \frac{1}{p} \qquad \sigma = \sqrt{\frac{1-p}{p^2}}$$

Going back to Dr. Smith's experiment:

$$\sigma = \sqrt{\frac{1-p}{p^2}} = \sqrt{\frac{1-0.35}{0.35^2}} = 2.3$$

Dr. Smith is expected to test 2.86 people before finding the first one that refuses to administer the shock, give or take 2.3 people.

These values only make sense in the context of repeating the experiment many many times.

Milgram Part 2

Suppose we randomly select four individuals to participate in this experiment. What is the probability that exactly 1 of them will refuse to administer the shock

Let's call these people Allen (A), Brittany (B), Caroline (C), and Damian (D). Each one of the four scenarios below will satisfy the condition of "exactly 1 of them refuses to administer the shock":

$$\text{Scenario 1:} \quad \underbrace{0.35}_{\text{(A) } \textit{refuse}} \times \underbrace{0.65}_{\text{(B) shock}} \times \underbrace{0.65}_{\text{(C) shock}} \times \underbrace{0.65}_{\text{(D) shock}} = 0.0961$$

$$\text{Scenario 2:} \quad \underbrace{0.65}_{\text{(A) shock}} \times \underbrace{0.35}_{\text{(B) } \textit{refuse}} \times \underbrace{0.65}_{\text{(C) shock}} \times \underbrace{0.65}_{\text{(D) shock}} = 0.0961$$

$$\text{Scenario 3:} \quad \underbrace{0.65}_{\text{(A) shock}} \times \underbrace{0.65}_{\text{(B) shock}} \times \underbrace{0.35}_{\text{(C) } \textit{refuse}} \times \underbrace{0.65}_{\text{(D) shock}} = 0.0961$$

$$\text{Scenario 4:} \quad \underbrace{0.65}_{\text{(A) shock}} \times \underbrace{0.65}_{\text{(B) shock}} \times \underbrace{0.65}_{\text{(C) shock}} \times \underbrace{0.35}_{\text{(D) } \textit{refuse}} = 0.0961$$

The probability of exactly one 1 of 4 people refusing to administer the shock is the sum of all of these probabilities.

$0.0961 + 0.0961 + 0.0961 + 0.0961 = 4 \times 0.0961 = 0.3844$

Binomial distribution

The question from the prior slide asked for the probability of given number of successes, k, in a given number of trials, n, (k = 1 success in n = 4 trials), and we calculated this probability as

$$\boxed{\text{\# of scenarios} \times \text{P(single scenario)}}$$

Number of scenarios: there is a less tedious way to figure this out, we'll get to that shortly...

$$P(single \quad scenario) = p^k(1-p)^{(n-k)}$$

The *Binomial* distribution describes the probability of having exactly k successes in n independent Bernouilli trials with probability of success p.

Choose Function

The choose function is useful for calculating the number of ways to choose k successes in n trials.

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

For example, :

$$\binom{9}{2} = \frac{9!}{2!(9-2)!} = \frac{9 \times 8 \times 7!}{2 \times 1 \times 7!} = \frac{72}{2} = 36$$

```
choose(9,2)
```

```
## [1] 36
```

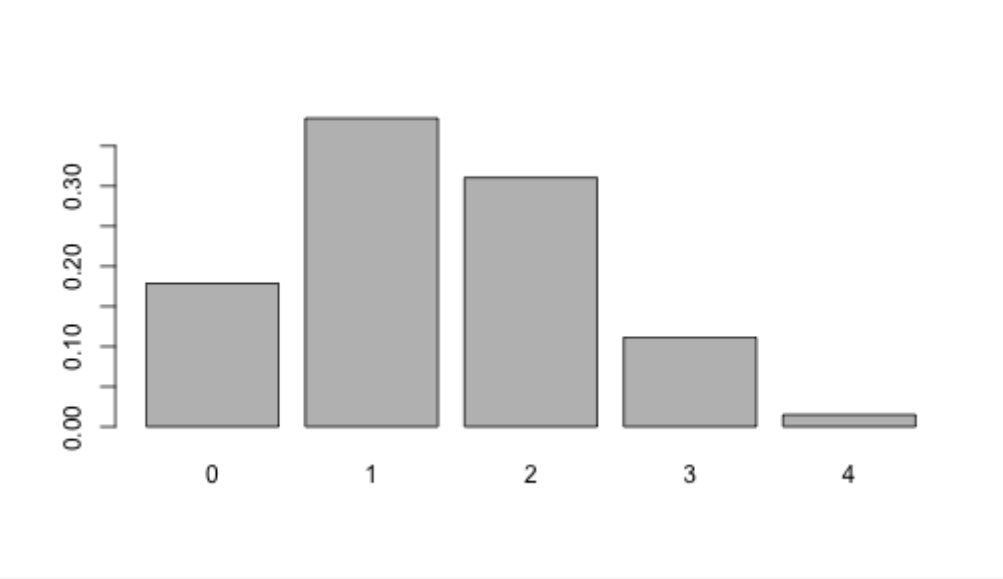
Binomial distribution

If p represents probability of success, (1 - p) represents probability of failure, n represents number of independent trials, and k represents number of successes

$$P(k \quad successes \quad in \quad n \quad trials) = \binom{n}{k} p^k (1-p)^{(n-k)}$$

Binomial distribution

```
n <- 4
p <- 0.35
barplot(dbinom(0:n, n, p), names.arg=0:n)
```




```
dbinom(1, 4, p)
```

```
## [1] 0.384475
```

One Minute Paper

Complete the one minute paper:

https://forms.gle/qxRnsCyydx1nf8sXA

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