Central Limit Theorem

EPSY 630 - Statistics II

Jason Bryer, Ph.D.

February 16, 2021

Agenda

- Probability (Chatper 3)
- Distributions (Chapter 4)
- **Central Limit Theorem (Chapter 5)**
- Next lab and homework
- One minute papers

Probability

There are two key properties of probability models:

1. P(A) = The probability of event A
2. $0 \leq P(A) \leq 1$

This semester we will examine two interpretations of probabilty:

- **Frequentist interpretation**: The probability of an outcome is the proportion of times the outcome would occur if we observed the random process an infinite number of times.

- **Bayesian interpretation**: A Bayesian interprets probability as a subjective degree of belief: For the same event, two separate people could have different viewpoints and so assign different probabilities. Largely popularized by

Law of Large Numbers

Law of large numbers states that as more observations are collected, the proportion of occurrences with a particular outcome, $\hat{p}_n$, converges to the probability of that outcome, p.

When tossing a fair coin, if heads comes up on each of the first 10 tosses, what do you think the chance is that another head will come up on the next coin toss? 0.5, less 0.5, or greater 0.5?

When tossing a fair coin, if heads comes up on each of the first 10 tosses, what do you think the chance is that another head will come up on the next coin toss? 0.5, less 0.5, or greater 0.5?

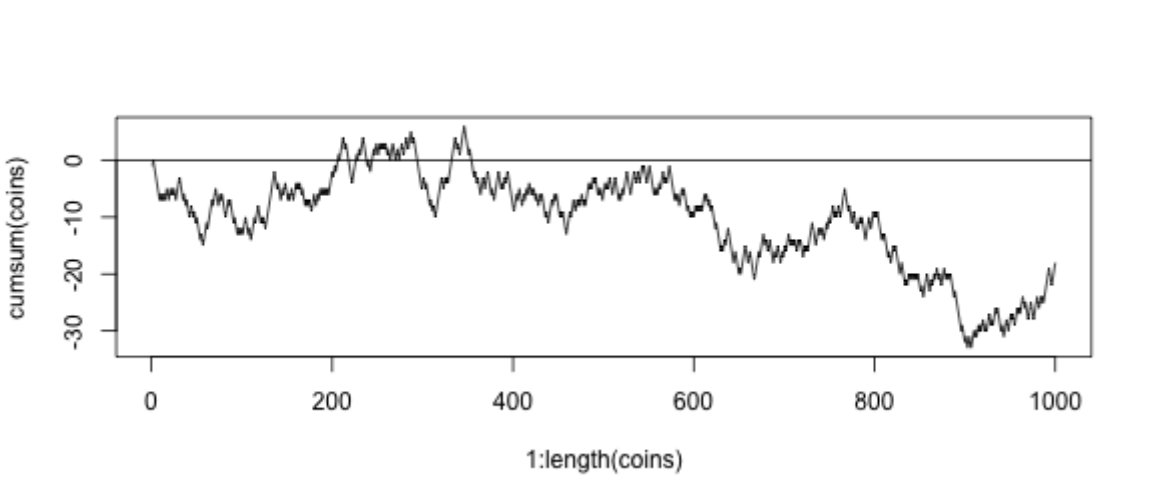
- The probability is still 0.5, or there is still a 50% chance that another head will come up on the next toss.
- The coin is not "due"" for a tail.
- The common misunderstanding of the LLN is that random processes are supposed to compensate for whatever happened in the past; this is just not true and is also called **gambler's fallacy** (or **law of averages**).

Coin Toss Demo

```
library(DATA606)
shiny_demo('gambler')
```

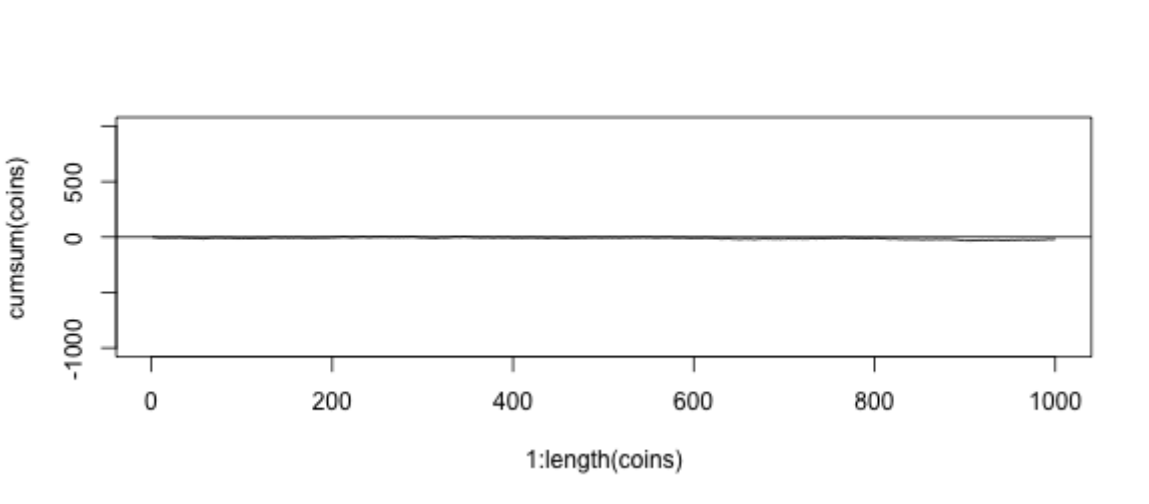
Coin Tosses

```
coins <- sample(c(-1,1), 1000, replace=TRUE)
plot(1:length(coins), cumsum(coins), type='l')
abline(h=0)
```



Coin Tosses (Full Range)

```
plot(1:length(coins), cumsum(coins), type='l', ylim=c(-1000, 1000))
abline(h=0)
```



Disjoint and non-disjoint outcomes

Disjoint (mutually exclusive) outcomes: Cannot happen at the same time.

- The outcome of a single coin toss cannot be a head and a tail. A student both cannot fail and pass a class.
- A single card drawn from a deck cannot be an ace and a queen.

Non-disjoint outcomes: Can happen at the same time.

- A student can get an A in Stats and A in Econ in the same semester.

Probability Distributions

A probability distribution lists all possible events and the probabilities with which they occur.

- The probability distribution for the gender of one kid:

Event	Male	Female
Probabilty	0.5	0.5

Rules for probability distributions:

1. The events listed must be disjoint
2. Each probability must be between 0 and 1
3. The probabilities must total 1

Probabilty Distrubtions (cont.)

The probability distribution for the genders of two kids:

Event	MM	FF	MF	FM
Probability	0.25	0.25	0.25	0.25

Independence

Two processes are independent if knowing the outcome of one provides no useful information about the outcome of the other.

- Knowing that the coin landed on a head on the first toss does not provide any useful information for determining what the coin will land on in the second toss. → Outcomes of two tosses of a coin are independent.
- Knowing that the first card drawn from a deck is an ace does provide useful information for determining the probability of drawing an ace in the second draw. → Outcomes of two draws from a deck of cards (without replacement) are dependent.

Checking for Independence

If P(A occurs, given that B is true) = P(A | B) = P(A), then A and B are independent.

- P(protects citizens) = 0.58
- P(randomly selected NC resident says gun ownership protects citizens, given that the resident is white) = P(protects citizens | White) = 0.67
- P(protects citizens | Black) = 0.28
- P(protects citizens | Hispanic) = 0.64

P(protects citizens) varies by race/ethnicity, therefore opinion on gun ownership and race ethnicity are most likely dependent.

Lottery

```
DATA606::shiny_demo('lottery')
```

Random Variables

A random variable is a numeric quantity whose value depends on the outcome of a random event

- We use a capital letter, like X, to denote a random variable
- The values of a random variable are denoted with a lowercase letter, in this case x
- For example, P(X = x)

There are two types of random variables:

- **Discrete random variables** often take only integer values
  Example: Number of credit hours, Difference in number of credit hours this term vs last
- **Continuous random variables** take real (decimal) values
  Example: Cost of books this term, Difference in cost of books this term vs last

Expectation

- We are often interested in the average outcome of a random variable.
- We call this the expected value (mean), and it is a weighted average of the possible outcomes

$$\mu = E(X) = \sum_{i=1}^{k} x_i P(X = x_i)$$

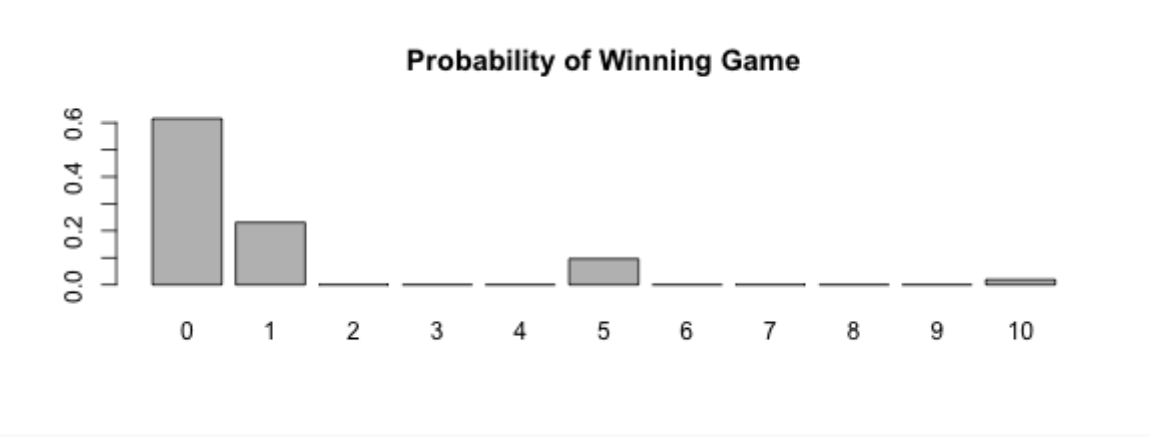
Expected value of a discrete random variable

In a game of cards you win $1 if you draw a heart, $5 if you draw an ace (including the ace of hearts), $10 if you draw the king of spades and nothing for any other card you draw. Write the probability model for your winnings, and calculate your expected winning.

Event	X	P(X)	X P(X)
Heart (not Ace)	1	12/52	12/52
Ace	5	4/52	20/52
King of Spades	10	1/52	10/52
All else	0	35/52	0
Total			$E(X) = \frac{42}{52} \approx 0.81$

Expected value of a discrete random variable

```
cards <- data.frame(Event = c('Heart (not ace)','Ace','King of Spades','All else'),
    X = c(1, 5, 10, 0),    pX = c(12/52, 5/52, 1/52, 32/52) )
cards$XpX <- cards$X * cards$pX
cards2 <- rep(0, 11)
cards2[cards$X + 1] <- cards$pX
names(cards2) <- 0:10
barplot(cards2, main='Probability of Winning Game')
```



Estimating Expected Values with Simulations

```
tickets <- as.data.frame(rbind(
    c(    '$1',    1,     15),
    c(    '$2',    2,     11),
    c(    '$4',    4,     62),
    c(    '$5',    5,    100),
    c(   '$10',   10,    143),
    c(   '$20',   20,    250),
    c(   '$30',   30,    562),
    c(   '$50',   50,   3482),
    c(  '$100',  100,   6681),
    c(  '$500',  500,  49440),
    c('$1500',  1500, 375214),
    c('$2500',  2500, 618000)
), stringsAsFactors=FALSE)
names(tickets) <- c('Winnings', 'Value', 'Odds')
tickets$Value <- as.integer(tickets$Value)
tickets$Odds <- as.integer(tickets$Odds)
```

Estimating Expected Values with Simulations

```
odds <- sample(max(tickets$Odds), 1000, replace=TRUE)
vals <- rep(-1, length(odds))
for(i in 1:nrow(tickets)) {
    vals[odds %% tickets[i,'Odds'] == 0] <- tickets[i,'Value'] - 1
}
head(vals, n=20)
```

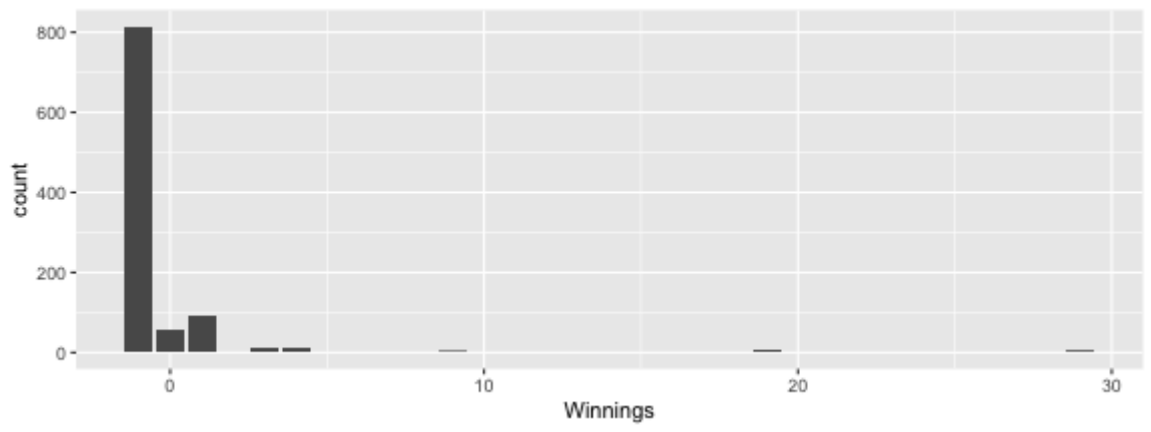
```
##  [1] -1 -1 -1 -1 -1  1 -1 -1 -1  0  0  1  0  0 -1 -1 -1 -1 -1 -1
```

```
mean(vals)
```

```
## [1] -0.396
```

Estimating Expected Values with Simulations

```
ggplot(data.frame(Winnings=vals), aes(x=Winnings)) + geom_bar(binwidth=1)
```



Expected Value of Lottery Example

$$\mu = E(X) = \sum_{i=1}^{k} x_i P(X = x_i)$$

```
tickets
```

```
##    Winnings Value   Odds         xPx
## 1        $1     1     15 0.066666667
## 2        $2     2     11 0.181818182
## 3        $4     4     62 0.064516129
## 4        $5     5    100 0.050000000
## 5       $10    10    143 0.069930070
## 6       $20    20    250 0.080000000
## 7       $30    30    562 0.053380783
## 8       $50    50   3482 0.014359563
## 9      $100   100   6681 0.014967819
## 10     $500   500  49440 0.010113269
## 11    $1500  1500 375214 0.003997719
```

Expected value for one ticket

```
sum(tickets$xPx) - 1
```

```
## [1] -0.3862045
```

Expected Value of Lottery Example (cont)

```
sum(tickets$xPx) - 1 # Expected value for one ticket
```

```
## [1] -0.3862045
```

Simulated

```
nGames <- 1
runs <- numeric(10000)
for(j in seq_along(runs)) {
    odds <- sample(max(tickets$Odds), nGames, replace = TRUE)
    vals <- rep(-1, length(odds))
    for(i in 1:nrow(tickets)) {
        vals[odds %% tickets[i,'Odds'] == 0] <- tickets[i,'Value'] - 1
    }
    runs[j] <- cumsum(vals)[nGames]
}
mean(runs)
```

Coin Tosses Revisited

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

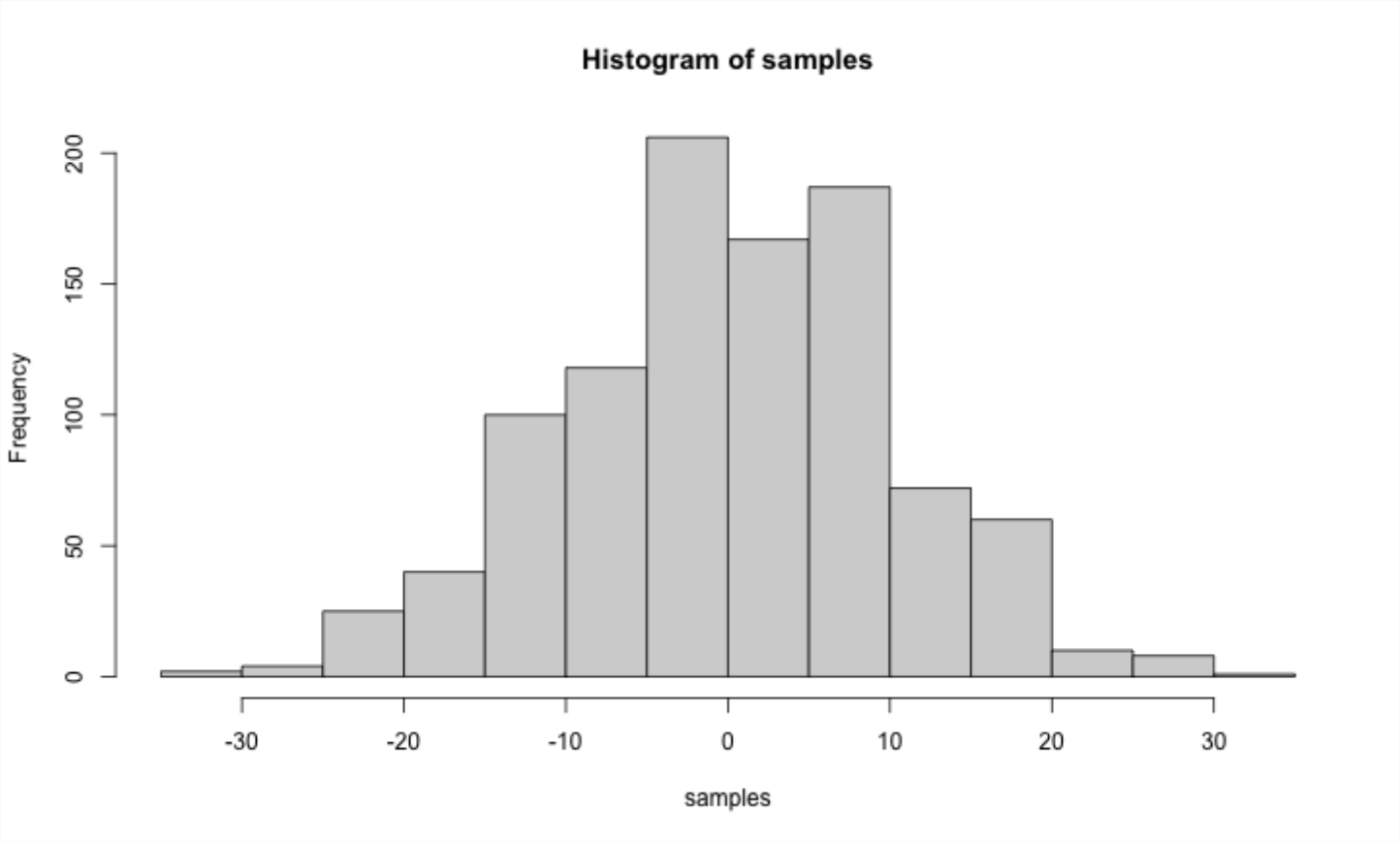
Many Random Samples

```
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 = 10)
```

```
##  [1]   6  -4  -8  -4   4  10  28   0 -14  24
```

Histogram of Many Random Samples

```
hist(samples)
```



Properties of Distribution

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

```
## [1] 0.662
```

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

```
## [1] 10.20813
```

Properties of Distribution (cont.)

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

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

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

```
## [1] 0.95
```

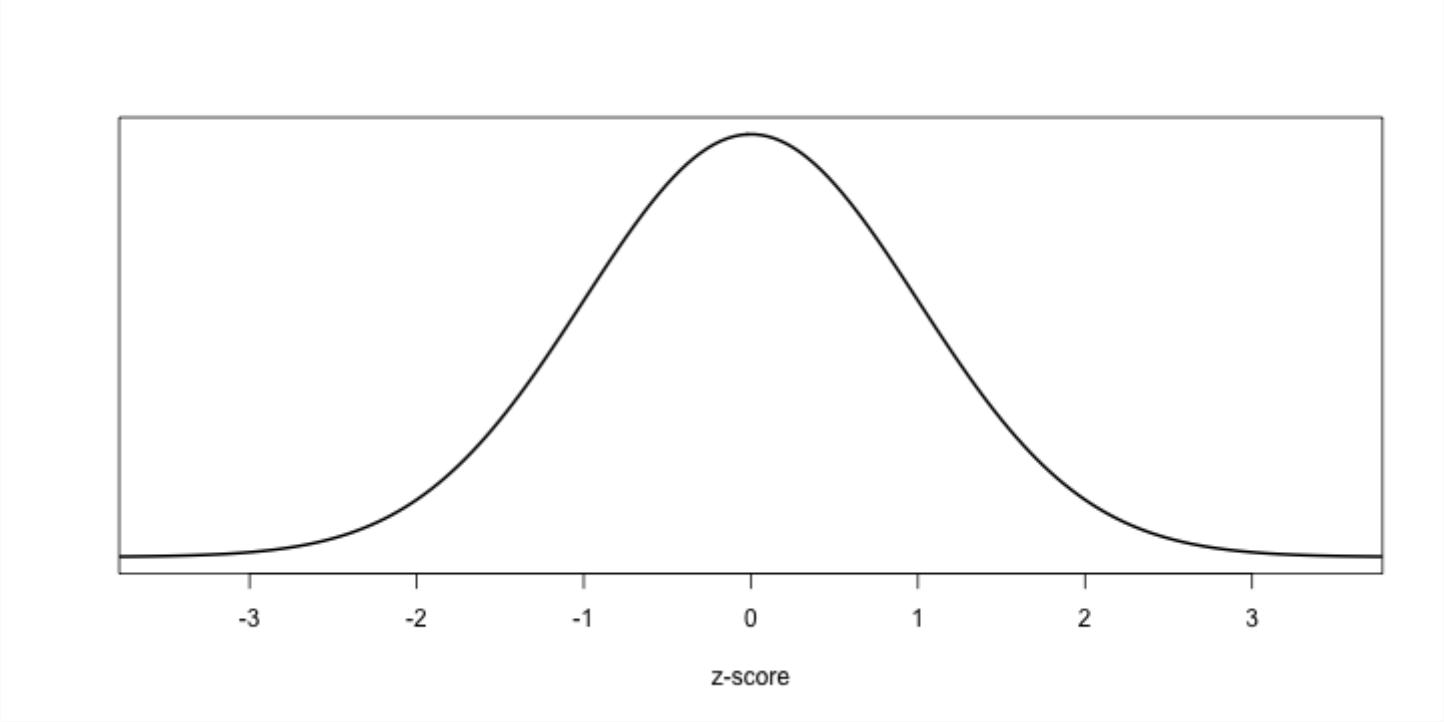
```
within3sd <- samples[samples >= m.sam - 3 * s.sam & samples <= m.sam + 3 * s.sam]
length(within3sd) / length(samples)
```

```
## [1] 0.997
```

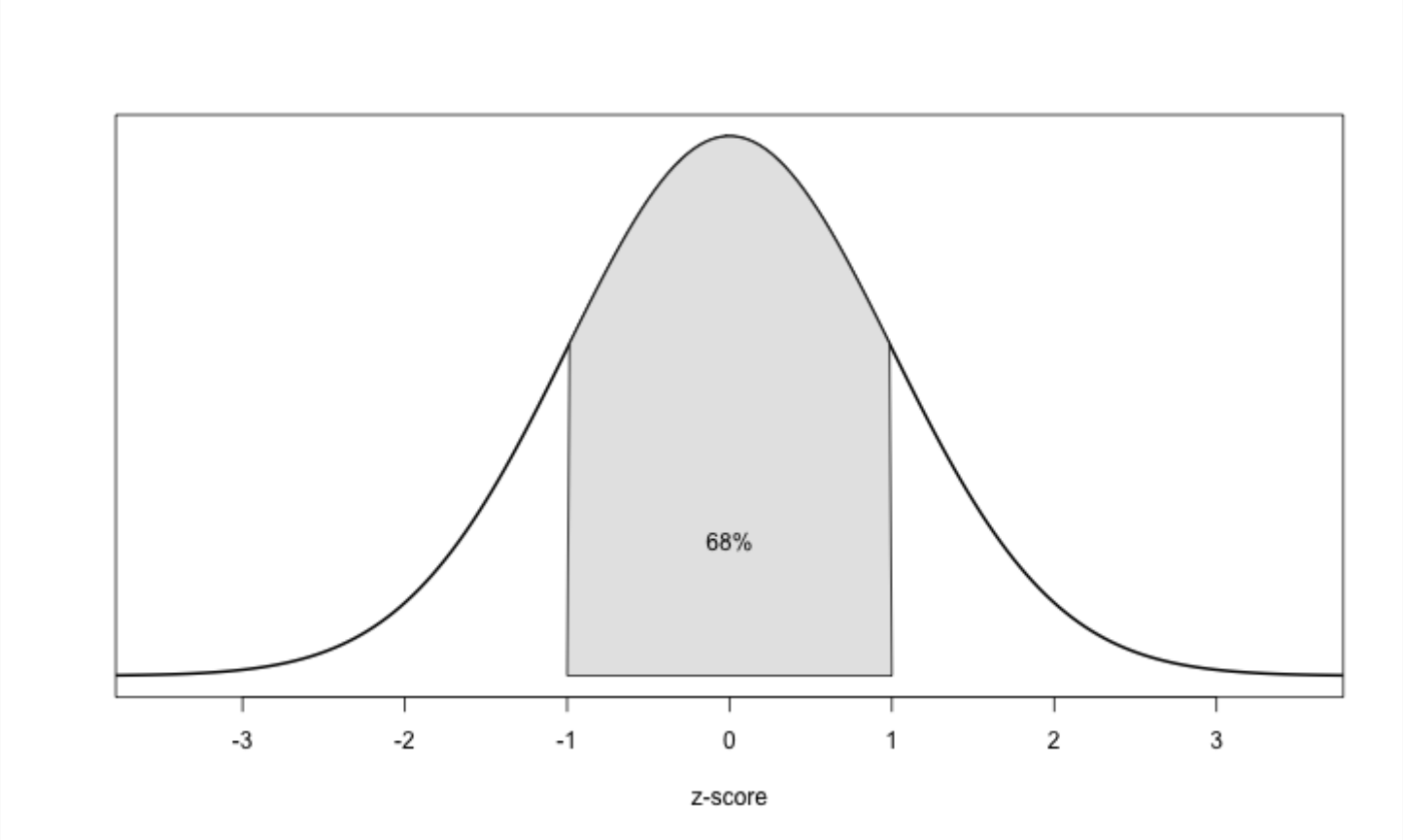
Standard Normal Distribution

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

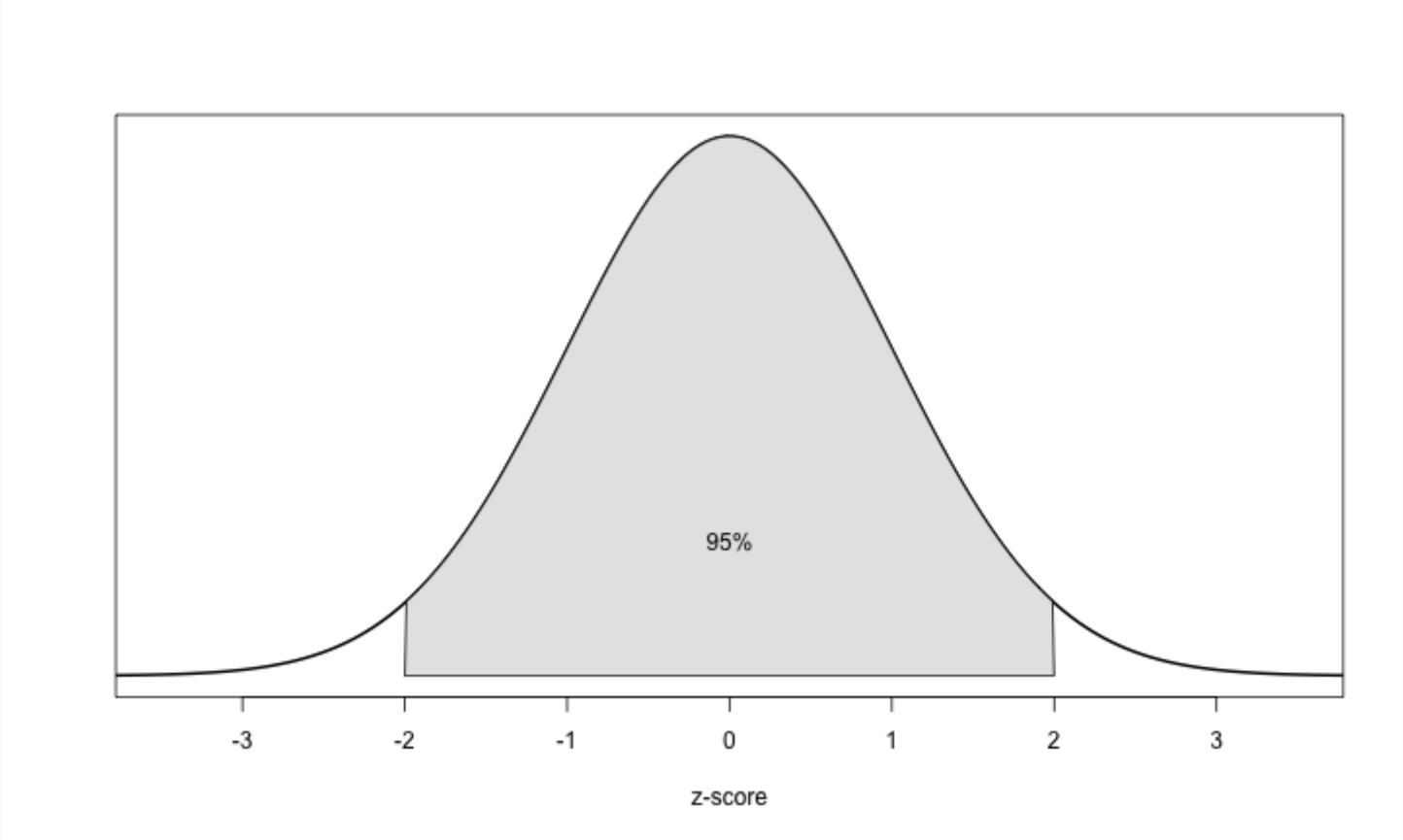
```
x <- seq(-4,4,length=200); y <- dnorm(x,mean=0, sd=1)
plot(x, y, type = "l", lwd = 2, xlim = c(-3.5,3.5), ylab='', xlab='z-score', yaxt='n')
```



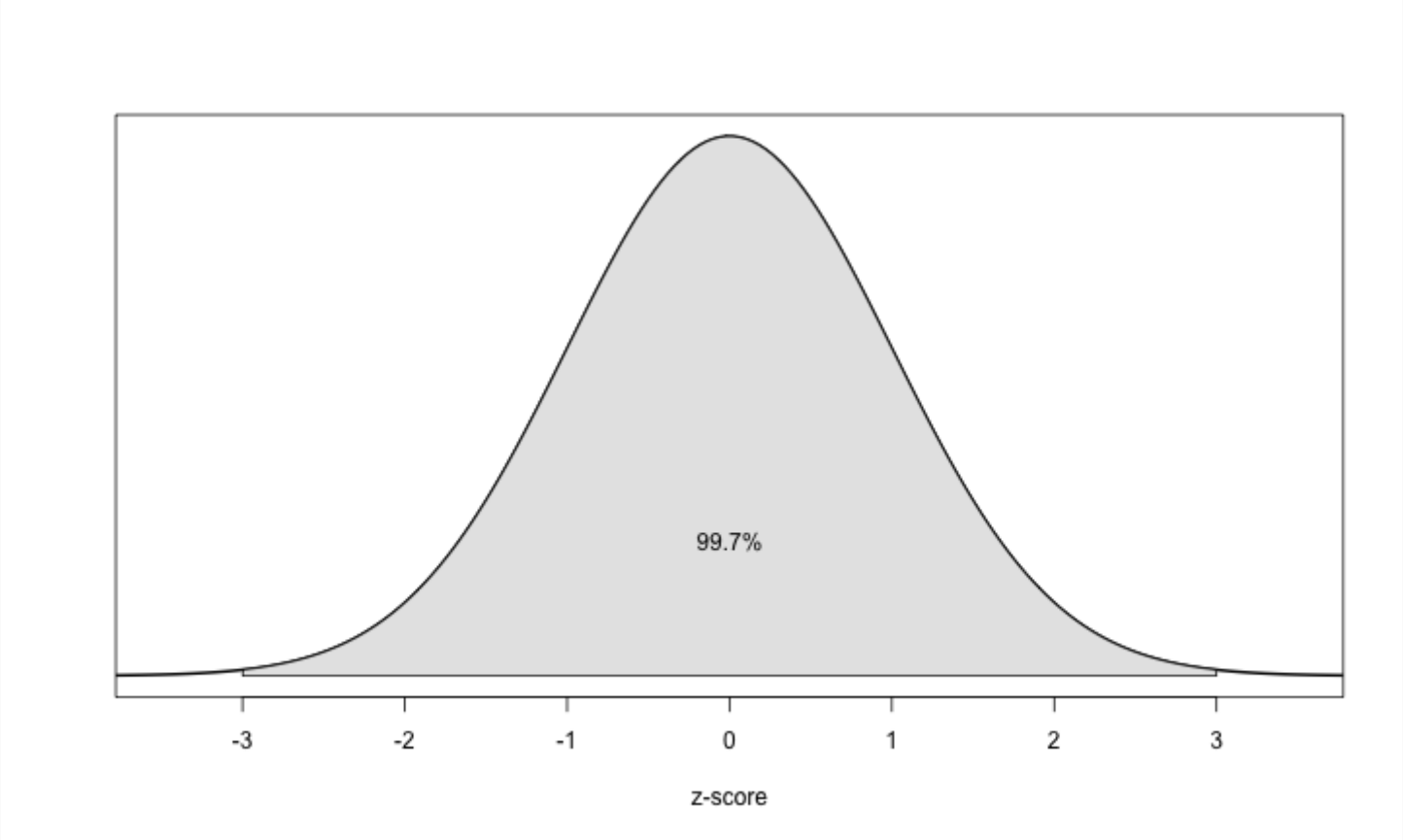
Standard Normal Distribution



Standard Normal Distribution



Standard Normal Distribution



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

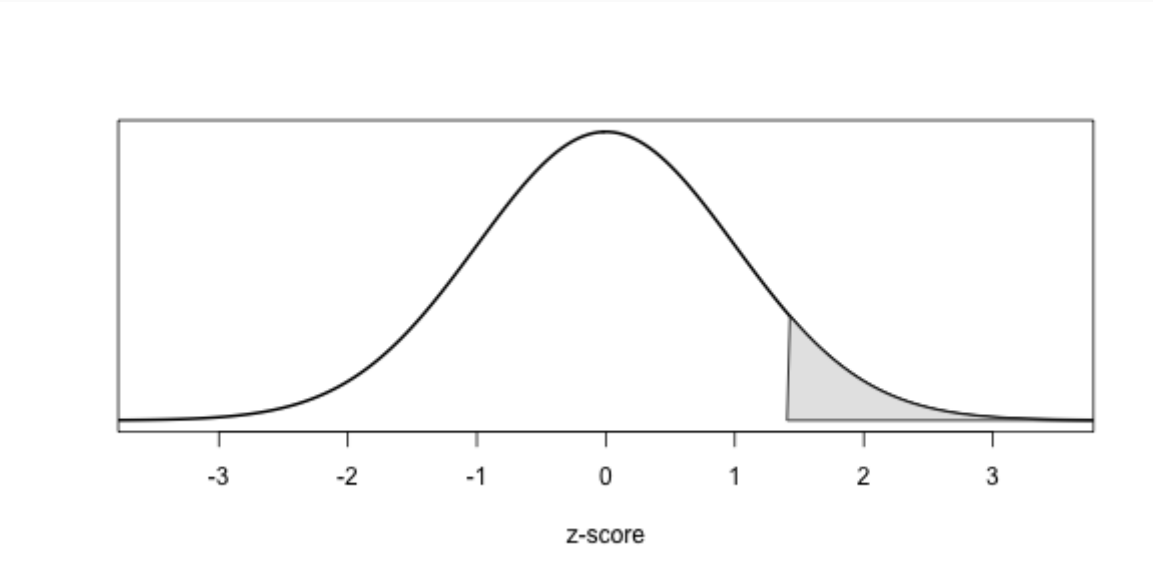
```
pnorm(15, mean=mean(samples), sd=sd(samples))
```

```
## [1] 0.9199249
```

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

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

```
## [1] 0.08007512
```



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:

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

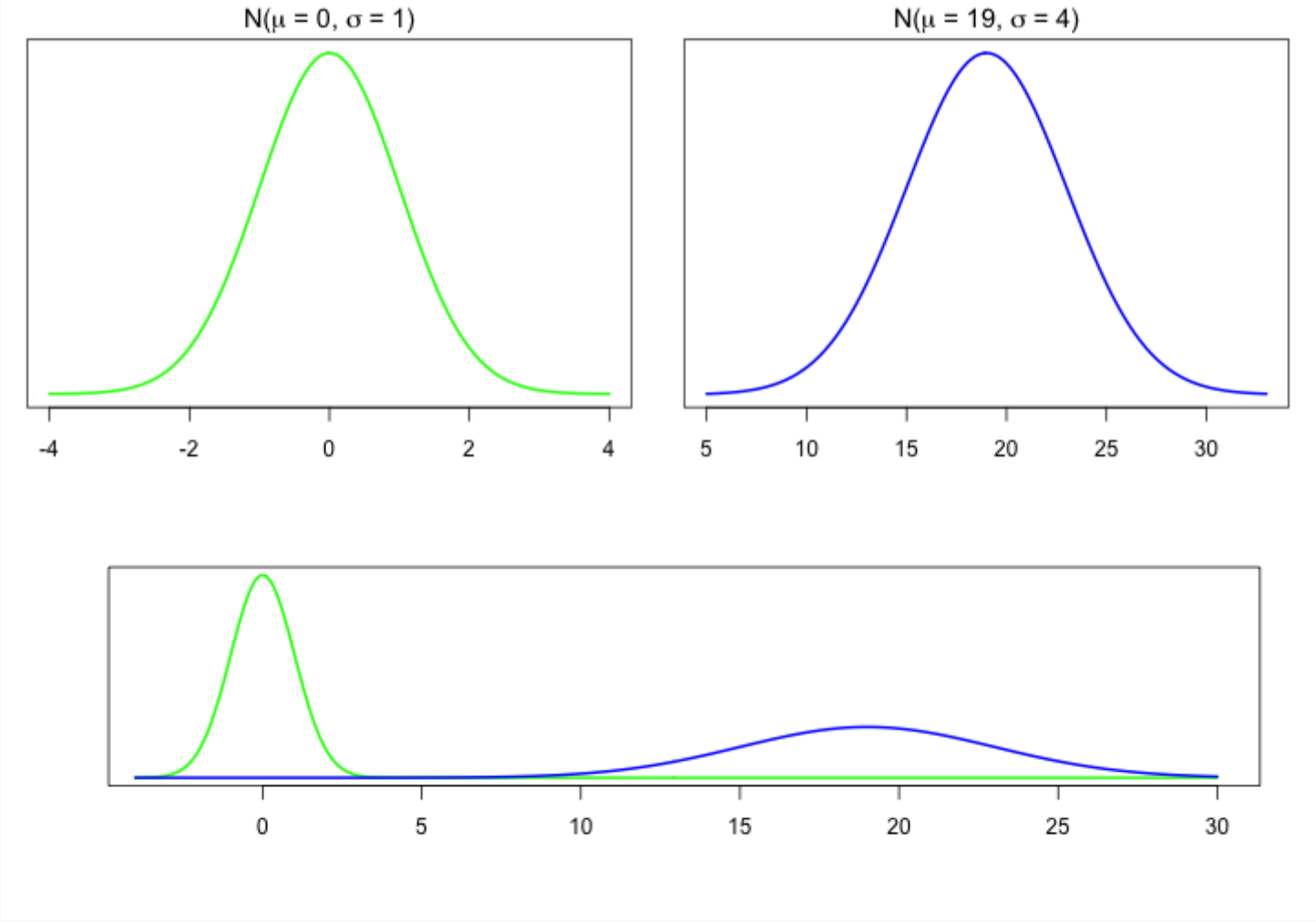
Z-Scores have a mean = 0 and standard deviation = 1.

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



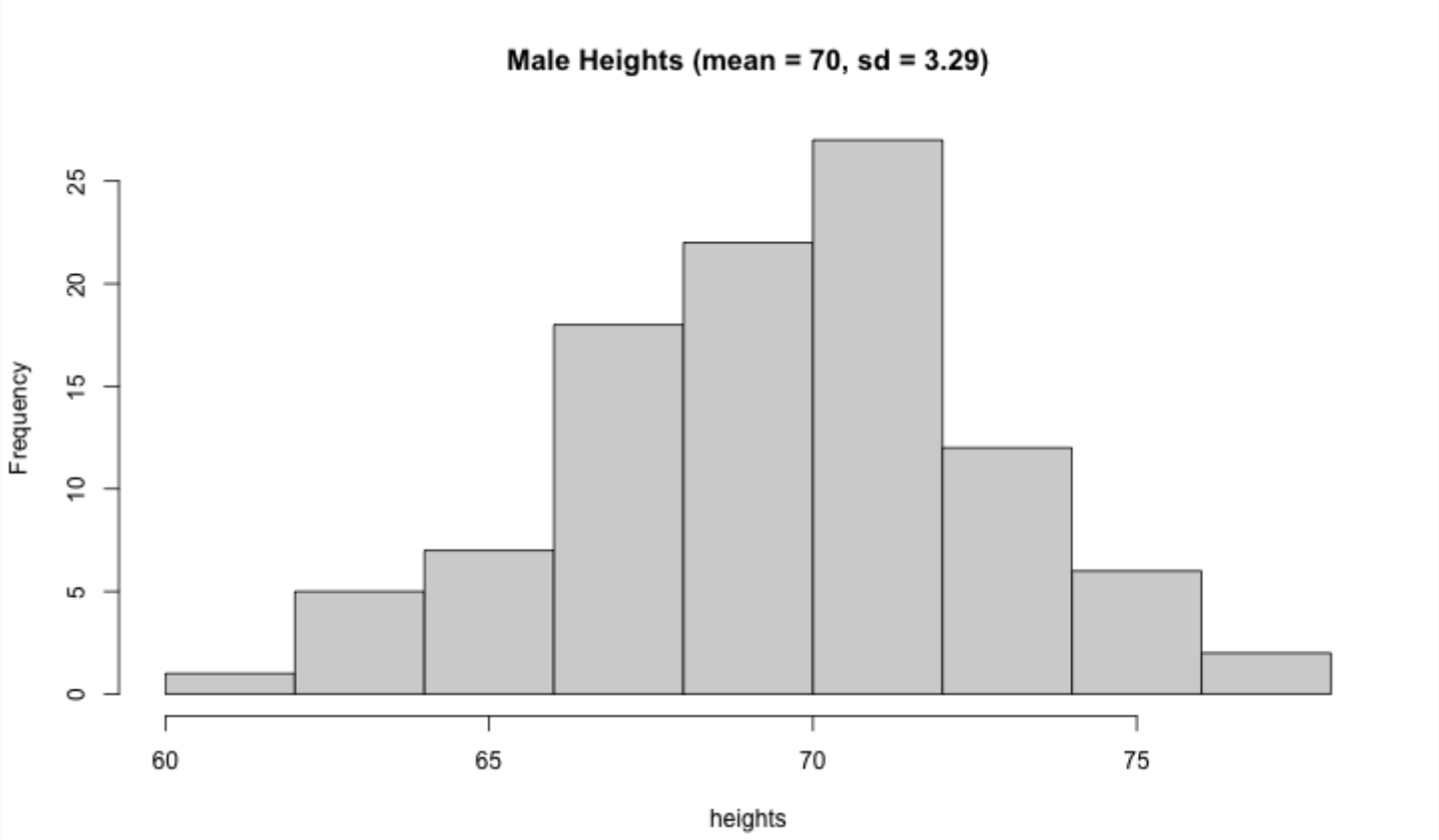
SAT Variability

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.

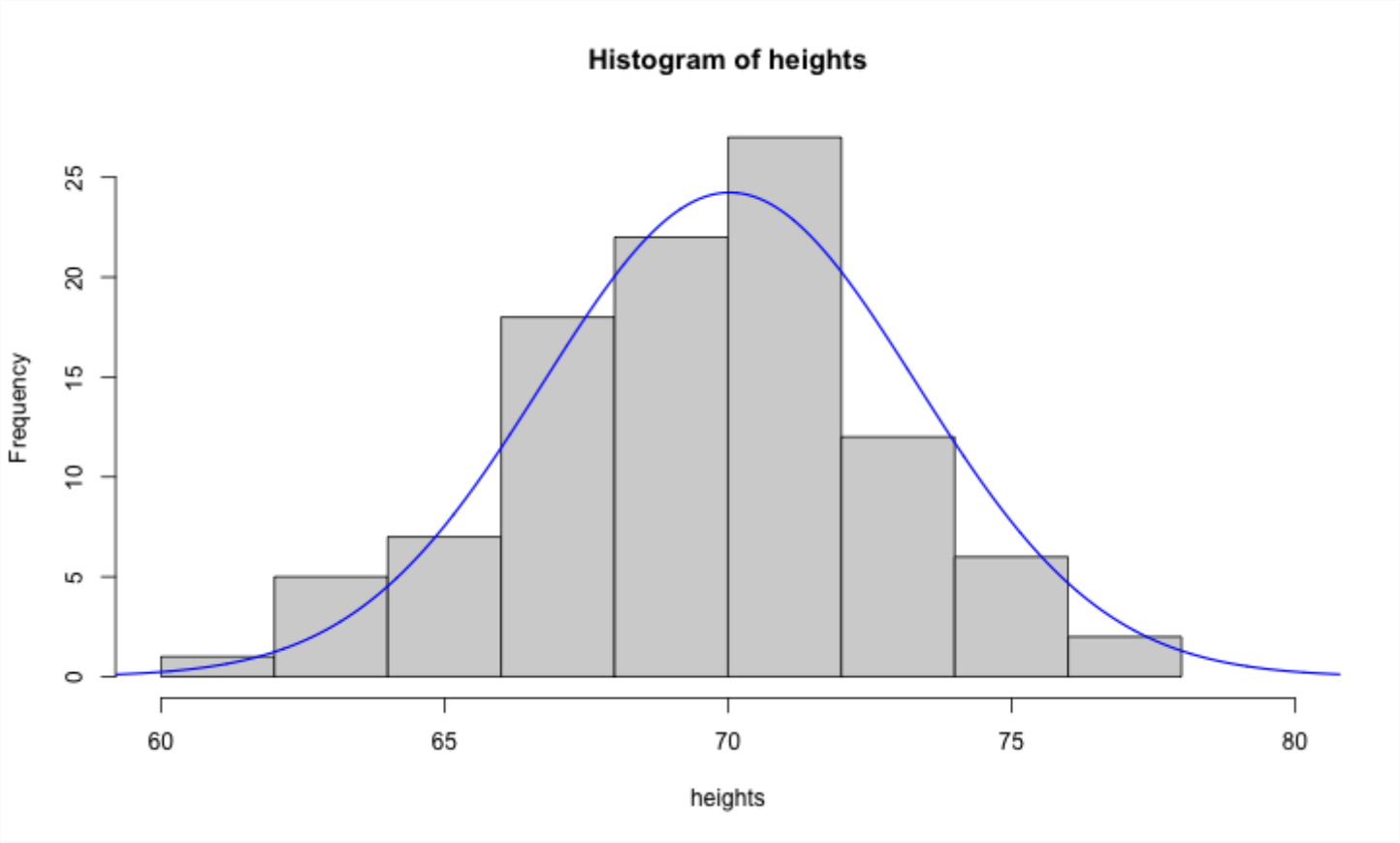
Evaluating Normal Approximation

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).



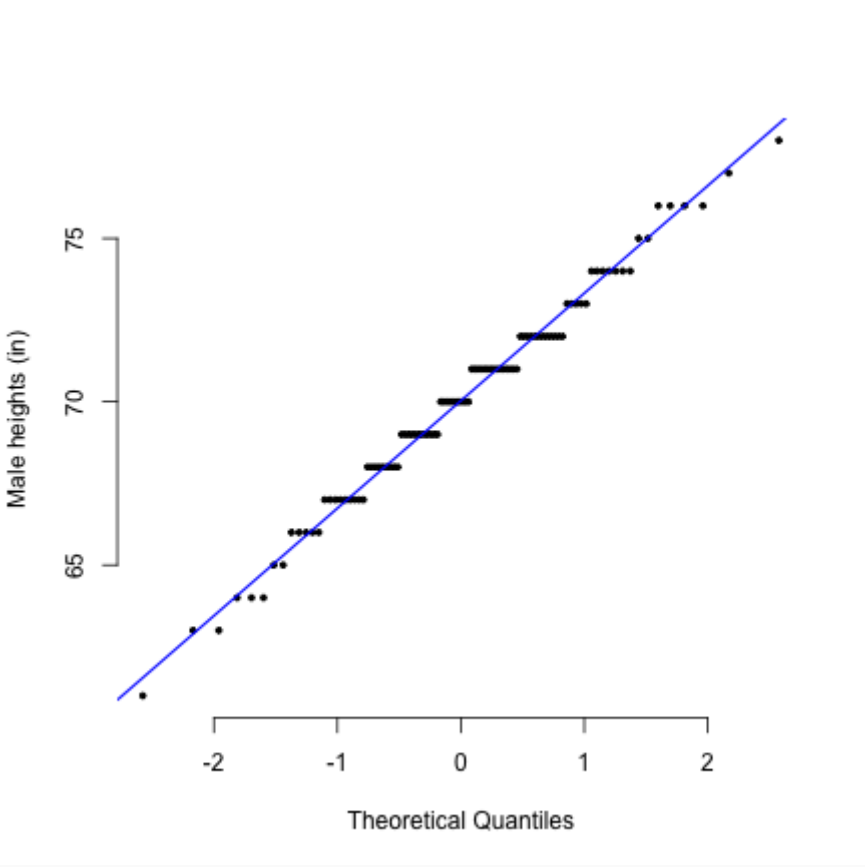
Evaluating Normal Approximation

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

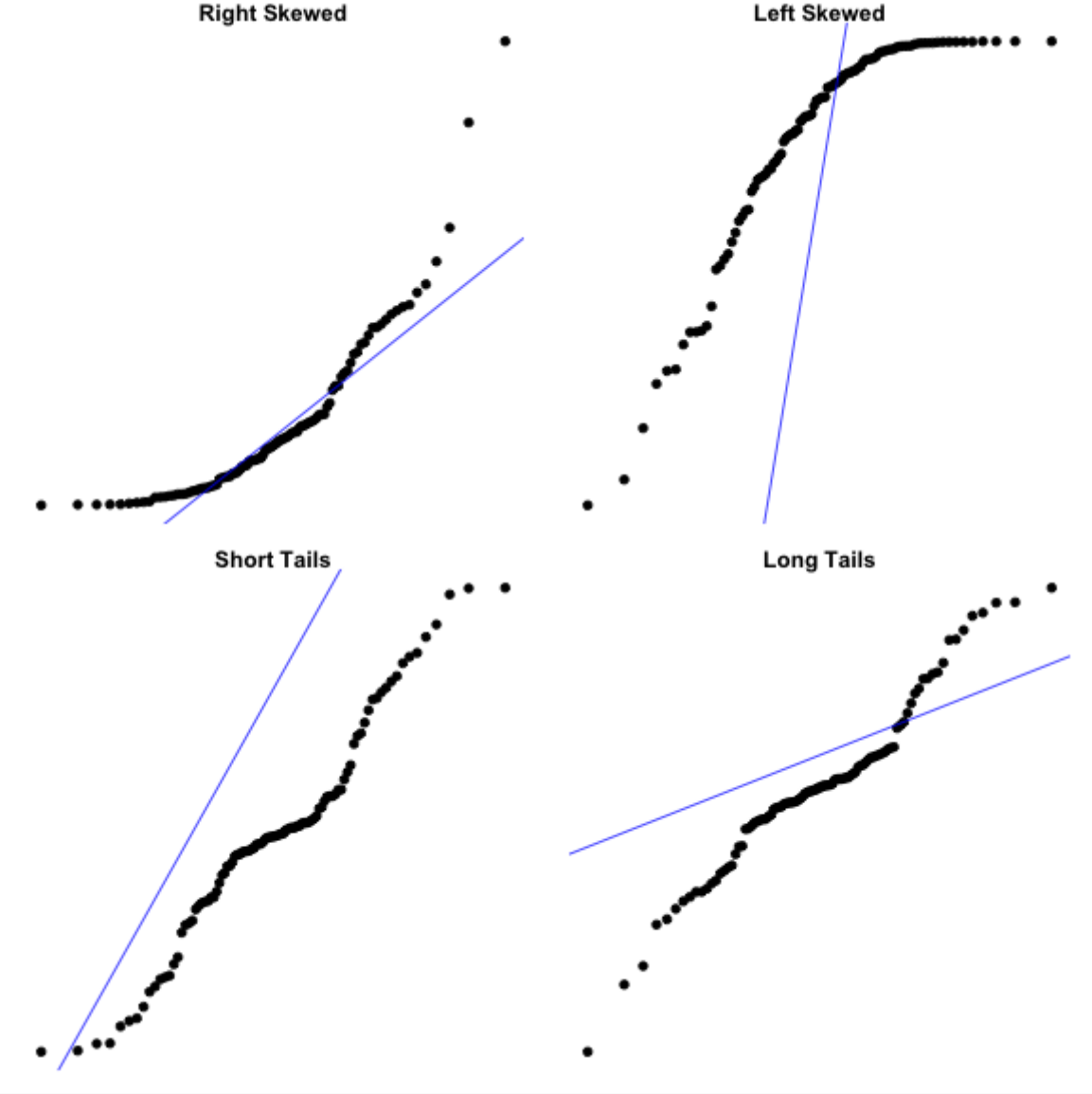


Normal Q-Q Plot

- 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



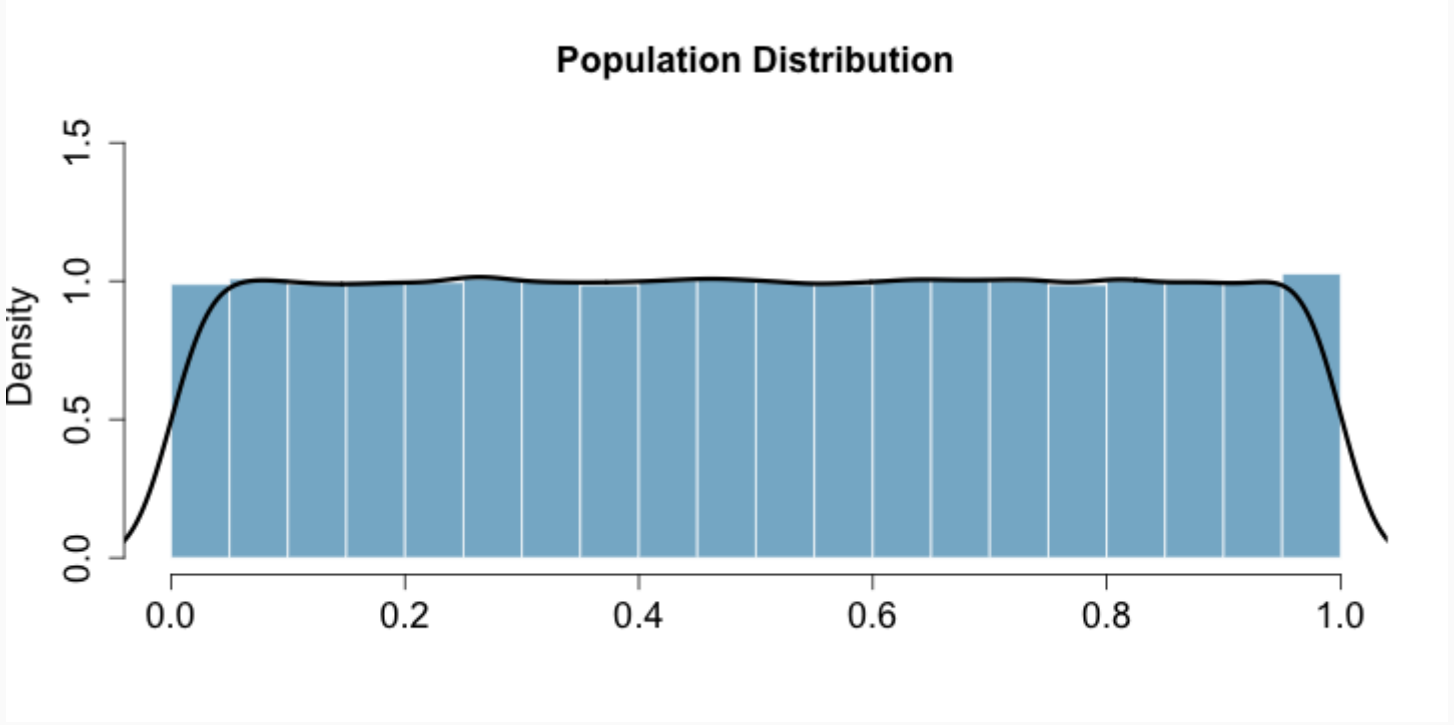
Simulated Normal Q-Q Plots

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

Population Distribution (Uniform)

```
n <- 1e5
pop <- runif(n, 0, 1)
mean(pop)
```

```
## [1] 0.5008915
```



Random Sample (n=10)

```
samp1 <- sample(pop, size=10)
mean(samp1)
```

```
## [1] 0.5462923
```

```
hist(samp1)
```

Random Sample (n=30)

```
samp2 <- sample(pop, size=30)
mean(samp2)
```

```
## [1] 0.5130467
```

```
hist(samp2)
```

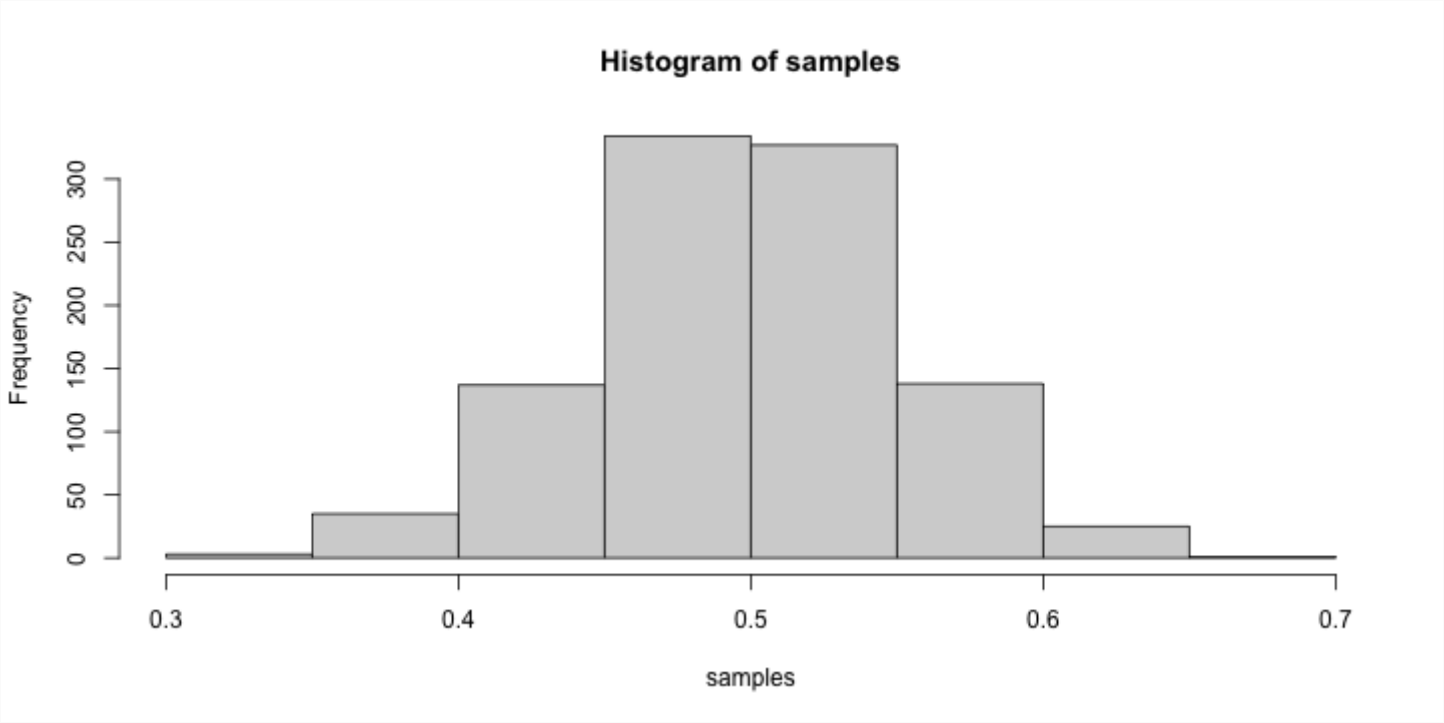
Lots of Random Samples

```
M <- 1000
samples <- numeric(length=M)
for(i in seq_len(M)) {
    samples[i] <- mean(sample(pop, size=30))
}
head(samples, n=8)
```

```
## [1] 0.5314329 0.4892158 0.5300205 0.5349329 0.5290058 0.4482886 0.3628467 0.4448132
```

Sampling Distribution

```
hist(samples)
```



Central Limit Theorem (CLT)

Let X_1, X_2, ..., X_n be independent, identically distributed random variables with mean μ and variance σ^2, both finite. Then for any constant z,

$$\lim_{n \to \infty} P\left(\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \leq z \right) = \Phi(z)$$

where Φ is the cumulative distribution function (cdf) of the standard normal distribution.

In other words...

The distribution of the sample mean is well approximated by a normal model:

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$$

where SE represents the **standard error**, which is defined as the standard deviation of the sampling distribution. In most cases σ is not known, so use s.

CLT Shiny App

```
library(DATA606)
shiny_demo('sampdist')
shiny_demo('CLT_mean')
```

Standard Error

```
samp2 <- sample(pop, size=30)
mean(samp2)
```

```
## [1] 0.5410922
```

```
(samp2.se <- sd(samp2) / sqrt(length(samp2)))
```

```
## [1] 0.05035122
```

Confidence Interval

The confidence interval is then $\mu \pm CV \times SE$ where CV is the critical value. For a 95% confidence interval, the critical value is ~1.96 since

$$\int_{-1.96}^{1.96} \frac{1}{\sigma\sqrt{2\pi}} d^{-\frac{(x-\mu)^2}{2\sigma^2}} \approx 0.95$$

```
qnorm(0.025) # Remember we need to consider the two tails, 2.5% to the left, 2.5% to the right
```

```
## [1] -1.959964
```

```
(samp2.ci <- c(mean(samp2) - 1.96 * samp2.se, mean(samp2) + 1.96 * samp2.se))
```

```
## [1] 0.4424038 0.6397806
```

Confidence Intervals (cont.)

We are 95% confident that the true population mean is between 0.4424038, 0.6397806.

That is, if we were to take 100 random samples, we would expect at least 95% of those samples to have a mean within 0.4424038, 0.6397806.

```
ci <- data.frame(mean=numeric(), min=numeric(), max=numeric())
for(i in seq_len(100)) {
    samp <- sample(pop, size=30)
    se <- sd(samp) / sqrt(length(samp))
    ci[i,] <- c(mean(samp),
                mean(samp) - 1.96 * se,
                mean(samp) + 1.96 * se)
}
ci$sample <- 1:nrow(ci)
ci$sig <- ci$min < 0.5 & ci$max > 0.5
```

Confidence Intervals

```
ggplot(ci, aes(x=min, xend=max, y=sample, yend=sample, color=sig)) +
    geom_vline(xintercept=0.5) +
    geom_segment() + xlab('CI') + ylab('') +
    scale_color_manual(values=c('TRUE'='grey', 'FALSE'='red'))
```

Hypothesis Testing

- We start with a null hypothesis (H_0) that represents the status quo.
- We also have an alternative hypothesis (H_A) that represents our research question, i.e. what we??? re testing for.
- We conduct a hypothesis test under the assumption that the null hypothesis is true, either via simulation or traditional methods based on the central limit theorem.
- If the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, we stick with the null hypothesis. If they do, then we reject the null hypothesis in favor of the alternative.

Hypothesis Testing (using CI)

H_0: The mean of `samp2` = 0.5

H_A: The mean of `samp2` $\neq$ 0.5

Using confidence intervals, if the *null* value is within the confidence interval, then we *fail* to reject the *null* hypothesis.

```
(samp2.ci <- c(mean(samp2) - 2 * sd(samp2) / sqrt(length(samp2)),
               mean(samp2) + 2 * sd(samp2) / sqrt(length(samp2))))
```

```
## [1] 0.4403897 0.6417946
```

Since 0.5 fall within 0.4403897, 0.6417946, we *fail* to reject the null hypothesis.

Hypothesis Testing (using *p*-values)

$$\bar{x} \sim N\left(mean = 0.49, SE = \frac{0.27}{\sqrt{30}} = 0.049\right)$$

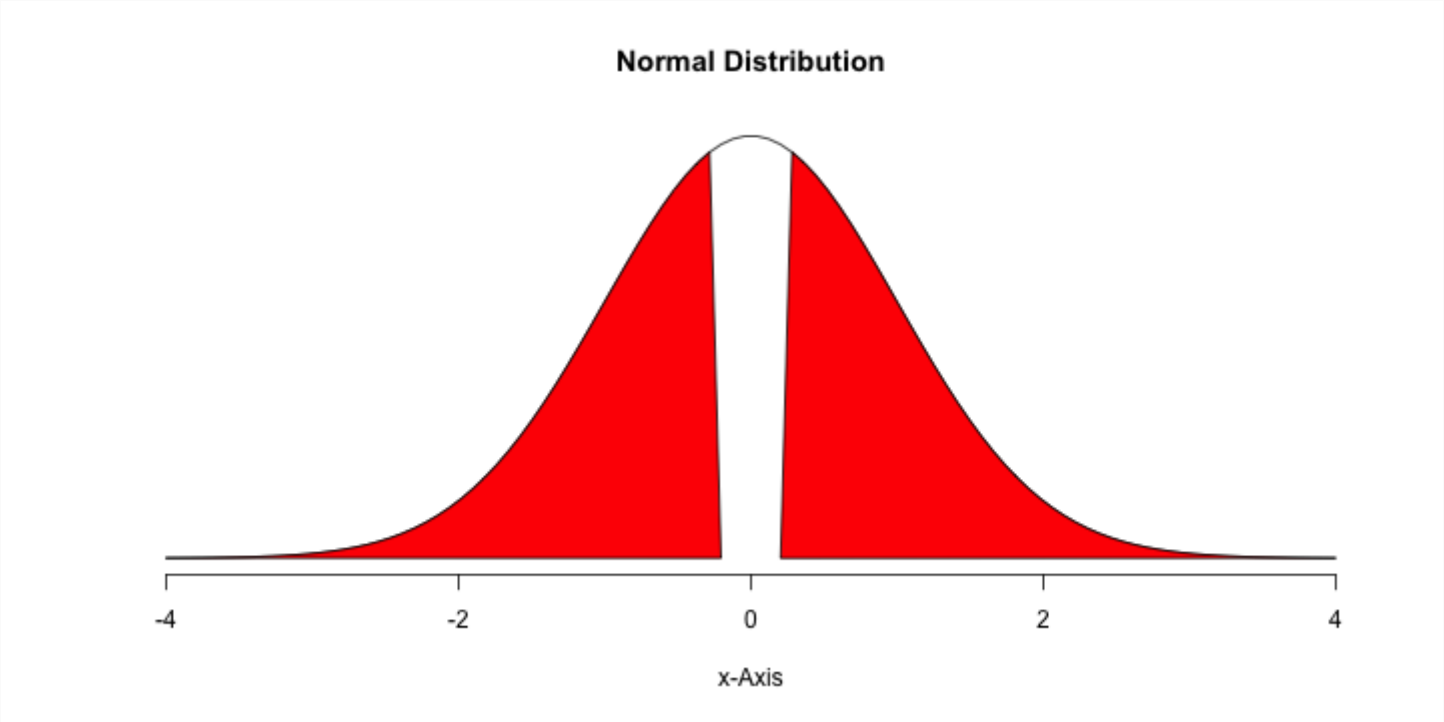
$$Z = \frac{\bar{x} - null}{SE} = \frac{0.49 - 0.50}{0.049} = -.204081633$$

```
pnorm(-.204) * 2
```

```
## [1] 0.8383535
```

Hypothesis Testing (using *p*-values)

```
normalPlot(bounds=c(-.204, .204), tails=TRUE)
```



Type I and II Errors

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

	fail to reject H_0	**reject H_0**
H_0 true	✔	Type I Error
H_A true	Type II Error	✔

- Type I Error: **Rejecting** the null hypothesis when it is **true**.
- Type II Error: **Failing to reject** the null hypothesis when it is **false**.

Hypothesis Test

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

H_0 : Defendant is innocent

H_A : Defendant is guilty

Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty

Type 2 error

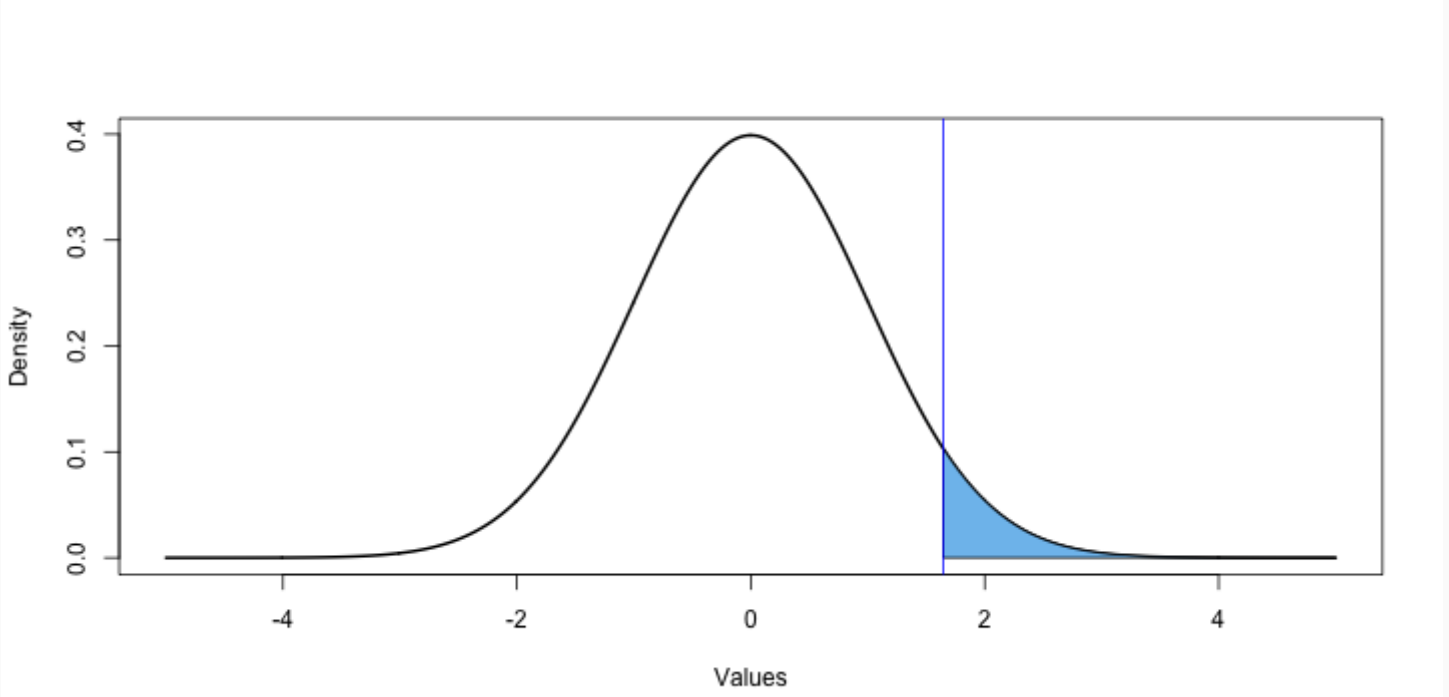
- Declaring the defendant guilty when they are actually innocent

Type 1 error

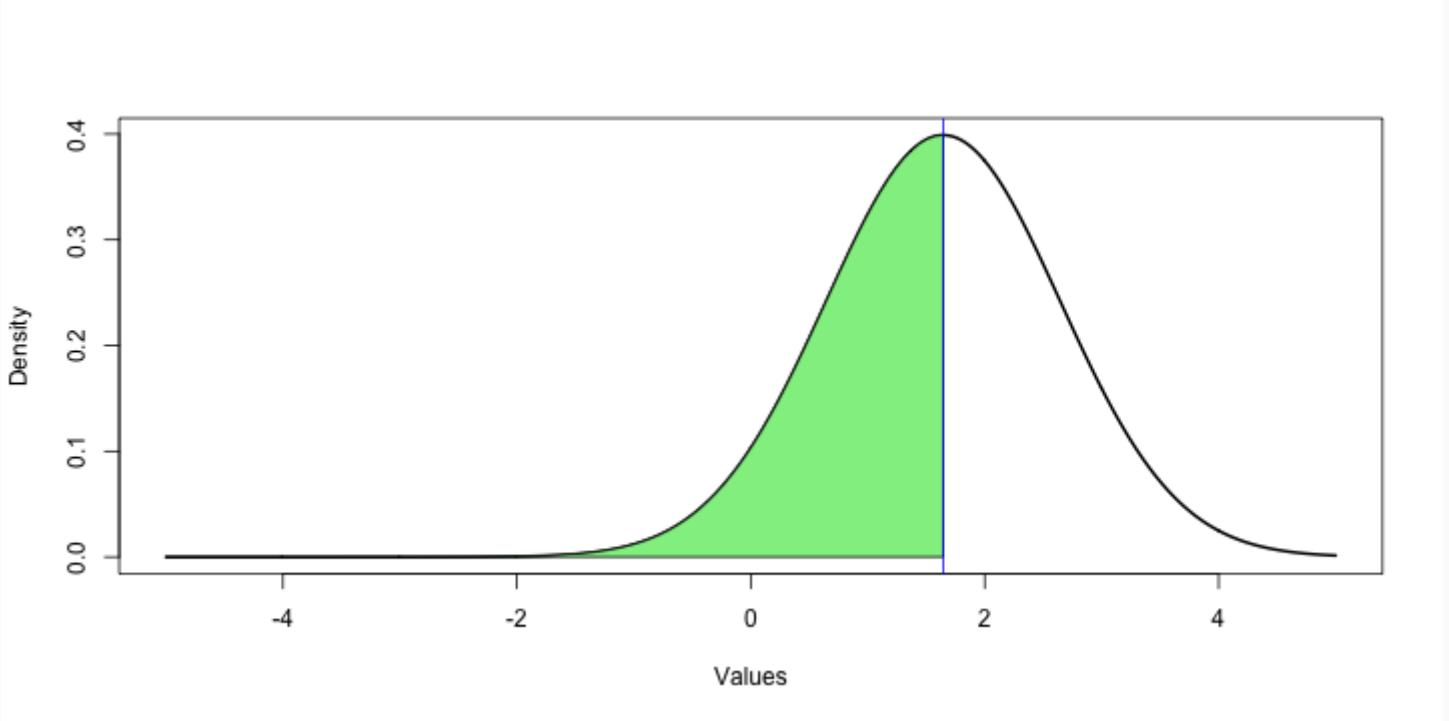
Null Distribution

```
(cv <- qnorm(0.05, mean=0, sd=1, lower.tail=FALSE))
```

```
## [1] 1.644854
```



Alternative Distribution



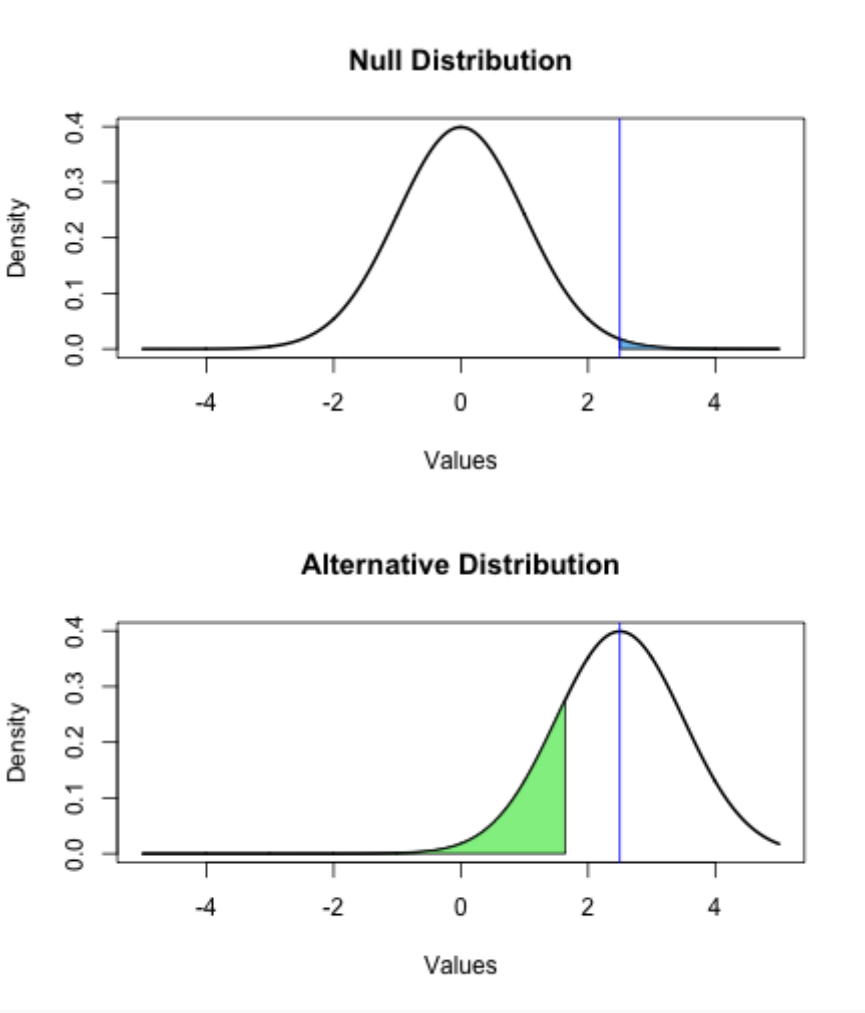

```
pnorm(cv, mean=cv, lower.tail = FALSE)
```

```
## [1] 0.5
```

Another Example (mu = 2.5)

```
mu <- 2.5
(cv <- qnorm(0.05,
             mean=0,
             sd=1,
             lower.tail=FALSE))
```

```
## [1] 1.644854
```



Numeric Values

Type I Error

```
pnorm(mu, mean=0, sd=1, lower.tail=FALSE)
```

```
## [1] 0.006209665
```

Type II Error

```
pnorm(cv, mean=mu, lower.tail = TRUE)
```

```
## [1] 0.1962351
```

Shiny Application

Visualizing Type I and Type II errors: https://bcdudek.net/betaprob/

Why p < 0.05?

Check out this page: https://r.bryer.org/shiny/Why05/

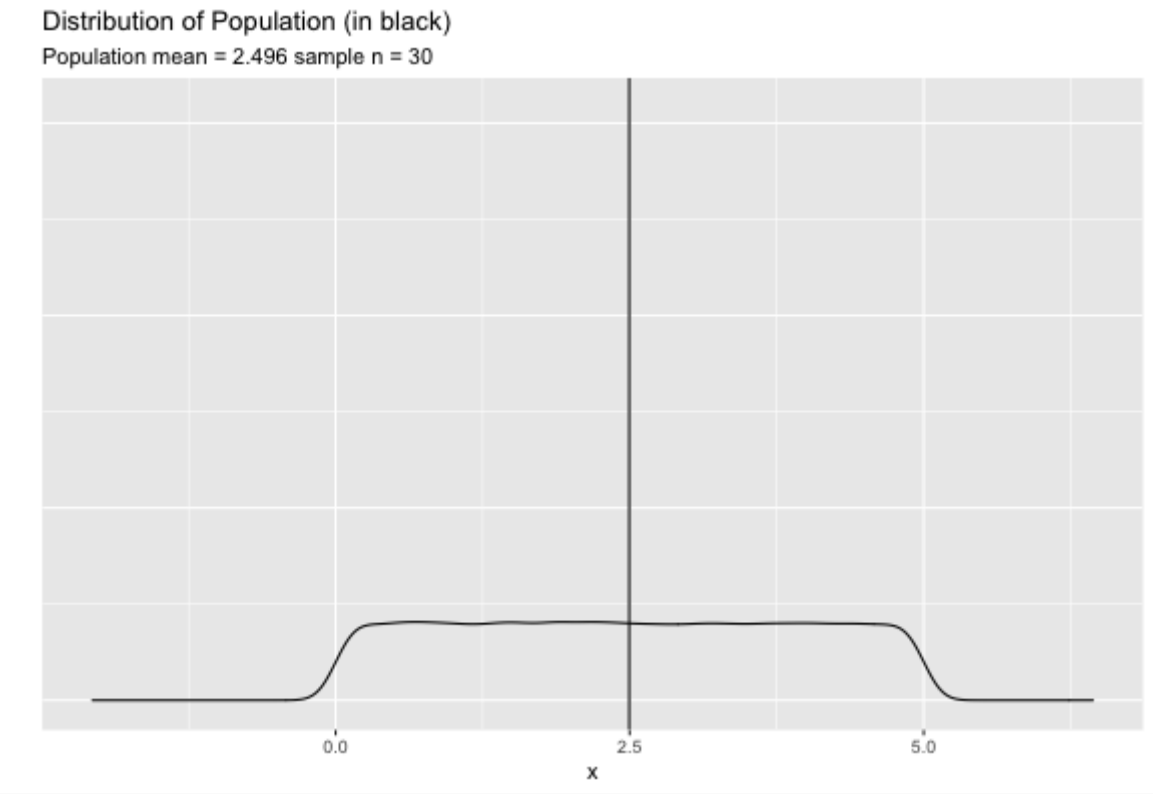
See also:

Kelly M. *Emily Dickinson and monkeys on the stair Or: What is the significance of the 5% significance level?* Significance 10:5. 2013.

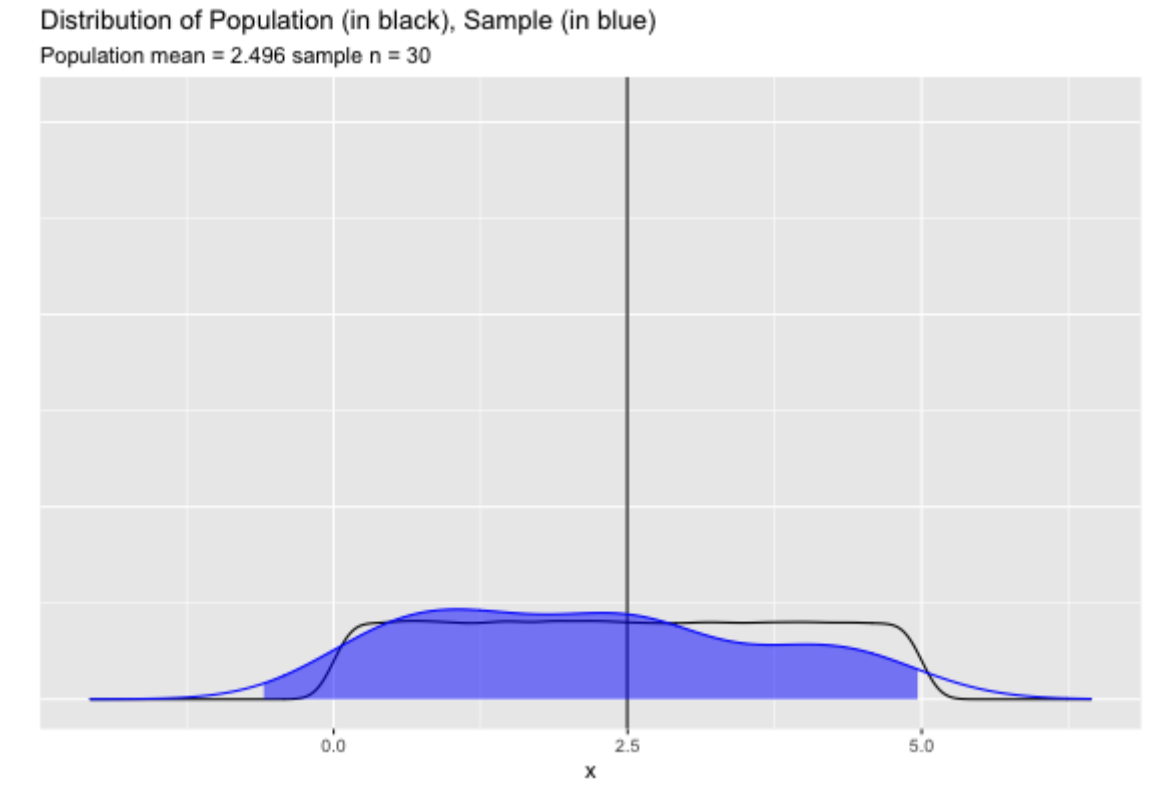
Statistical vs. Practical Significance

- Real differences between the point estimate and null value are easier to detect with larger samples.
- However, very large samples will result in statistical significance even for tiny differences between the sample mean and the null value (effect size), even when the difference is not practically significant.
- This is especially important to research: if we conduct a study, we want to focus on finding meaningful results (we want observed differences to be real, but also large enough to matter).
- The role of a statistician is not just in the analysis of data, but also in planning and design of a study.

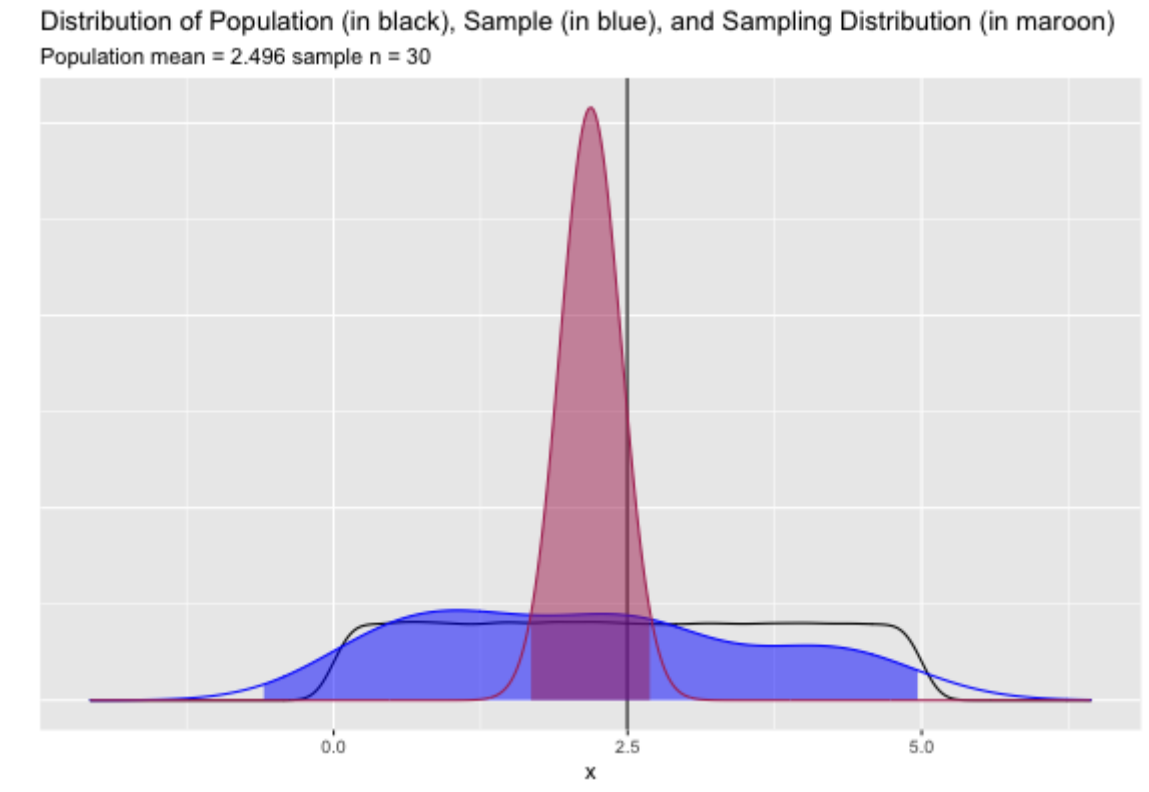
Review: Sampling Distribution



Review: Sampling Distribution



Review: Sampling Distribution



Assignments

Lab 5 is in two parts: A) Sampling Distributions and B) Confidence Levels. To get started, run the following commands:

```
DATA606::startLab('Lab5a')
DATA606::startLab('Lab5b')
```

Chapter 5 homework: https://epsy630.bryer.org/assignments/homework/

One Minute Paper

Complete the one minute paper:

https://forms.gle/yB3ds6MYE89Z1pURA

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