

Final Review Section

Sam Frederick

Central Limit Theorem

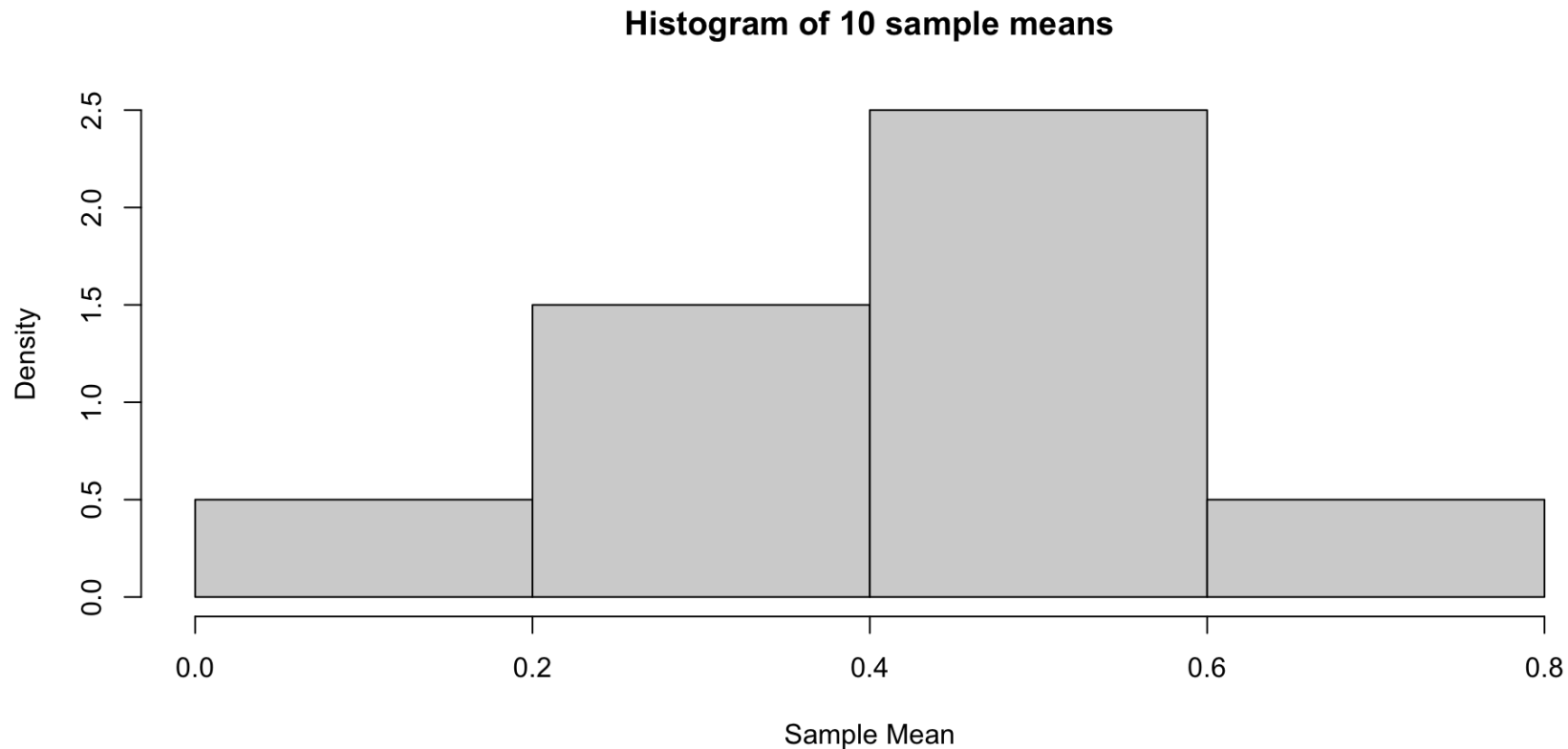
- What does the Central Limit Theorem tell us?
- Central Limit Theorem: the distribution of our sample means for n independent samples will approach a normal distribution as n goes to infinity with the mean equal to the population mean

Central Limit Theorem Example

- For each statement below, plot a histogram of the sample means
 - Simulate and take the mean of 10 random samples from a binomial distribution with $n = 10$
 - Simulate and take the mean of 100 random samples from a binomial distribution with $n = 10$
 - Simulate and take the mean of 1000 random samples from a binomial distribution with $n = 10$
 - Simulate and take the mean of 10000 random samples from a binomial distribution with $n = 10$

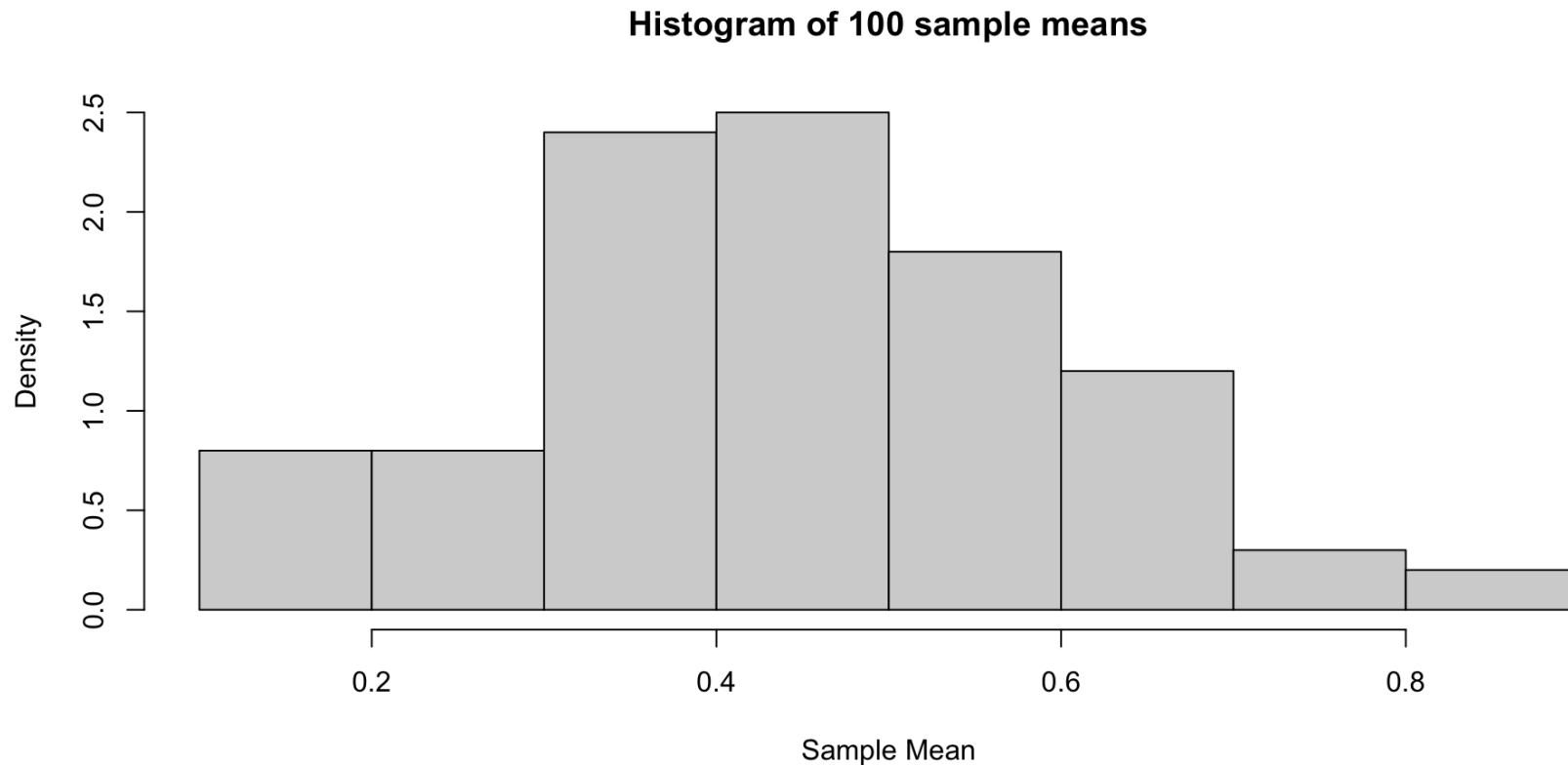
Central Limit Theorem

```
1 set.seed(123)
2 hist(replicate(10, mean(rbinom(10, size = 1, prob = 0.5))),
3       main = "Histogram of 10 sample means",
4       xlab = "Sample Mean", freq = F)
```



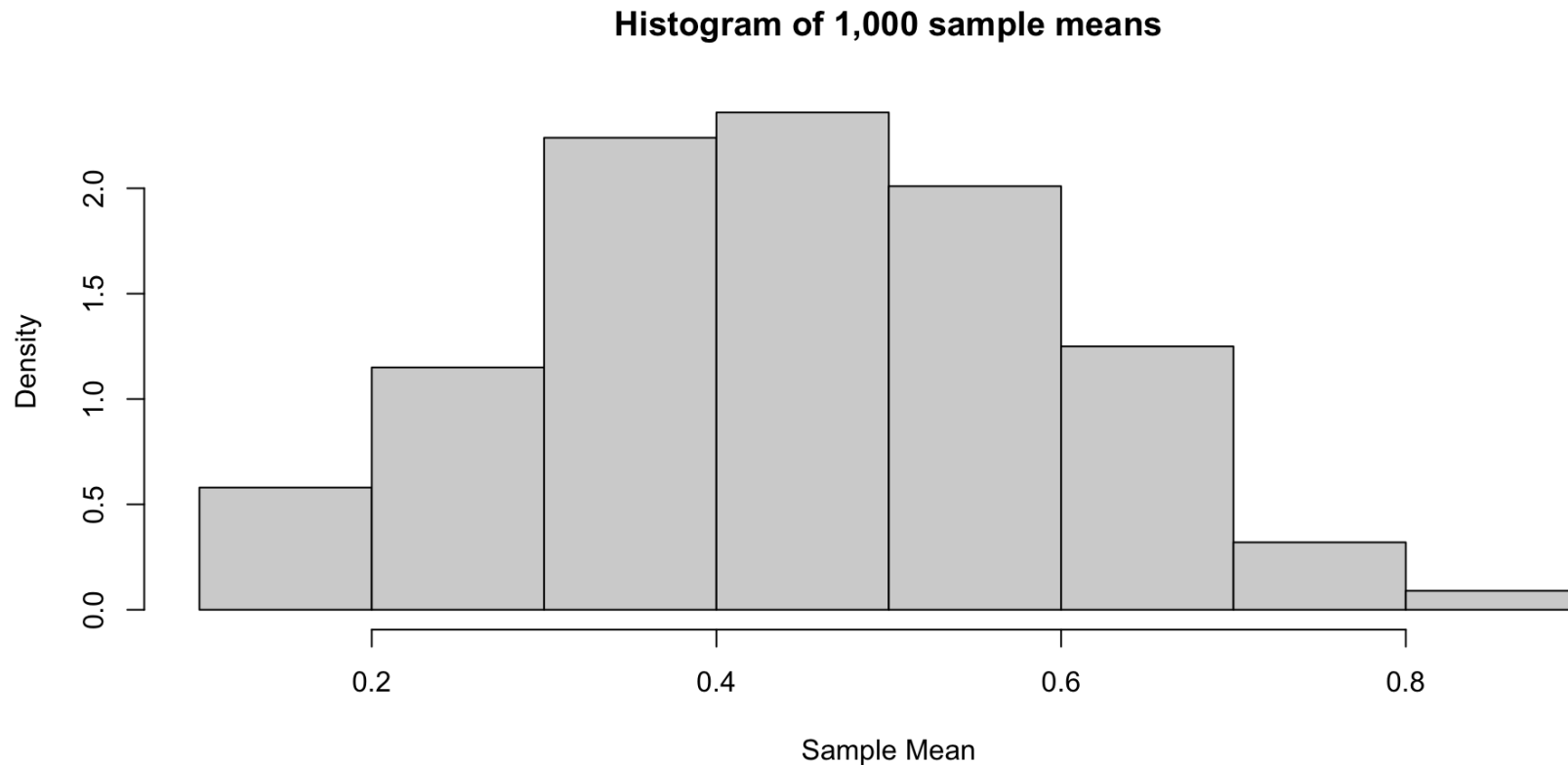
Central Limit Theorem

```
1 set.seed(123)
2 hist(replicate(100, mean(rbinom(10, size = 1, prob = 0.5))),
3      main = "Histogram of 100 sample means",
4      xlab = "Sample Mean", freq = F)
```



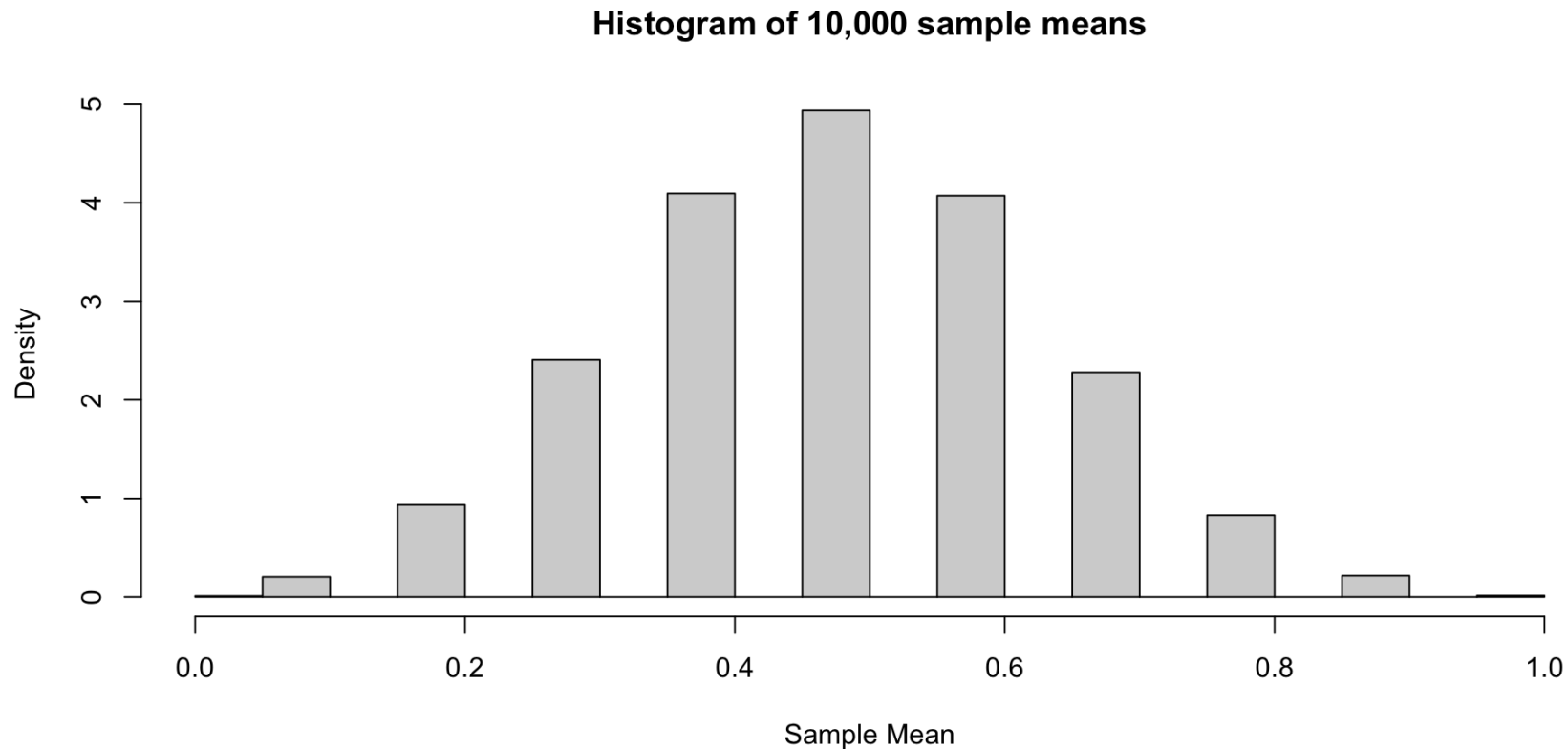
Central Limit Theorem

```
1 set.seed(123)
2 hist(replicate(1000, mean(rbinom(10, size = 1, prob = 0.5))),
3      main = "Histogram of 1,000 sample means",
4      xlab = "Sample Mean", freq = F)
```



Central Limit Theorem

```
1 set.seed(123)
2 hist(replicate(10000, mean(rbinom(10, size = 1, prob = 0.5))),
3      main = "Histogram of 10,000 sample means",
4      xlab = "Sample Mean", freq = F)
```



Z-Scores

$$Z = \frac{\bar{X} - E[\bar{X}]}{\sqrt{Var(\bar{X})}}$$

Z-Scores

$$E[\bar{X}] = E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n E[X_i] = \frac{1}{n} \sum_{i=1}^n \mu = \frac{n}{n} \mu = \mu$$

$$\begin{aligned} Var(\bar{X}) &= Var\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n Var(X_i) \\ &= \frac{1}{n^2} n * \sigma^2 = \frac{1}{n} \sigma^2 \end{aligned}$$

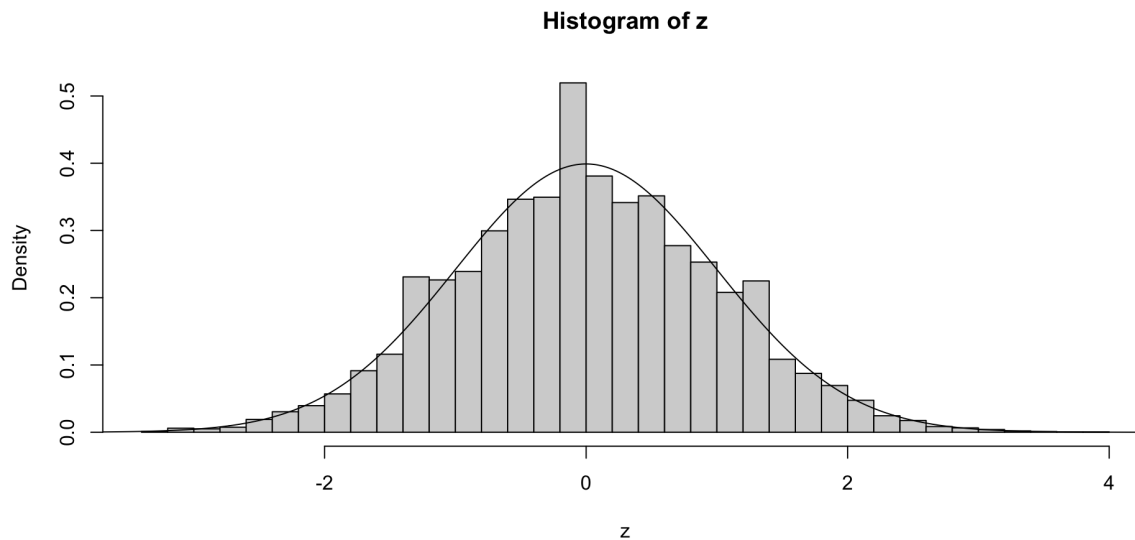
Central Limit Theorem

$$Z \xrightarrow{n \rightarrow \infty} N(0, 1)$$

As the number of samples goes to infinity, the distribution of Z is expected to be distributed approximately standard normal (i.e., $N(0,1)$).

Central Limit Theorem

```
1 p <- 0.5
2 n <- 1000
3 n.samp <- 10000
4 set.seed(123)
5 samp <- replicate(n.samp, mean(rbinom(n, 1, p)))
6 z <- (samp - 0.5)/sqrt(p*(1-p)/n)
7 stdnorm <- dnorm(seq(-4,40, 0.01))
8 hist(z, freq = F, breaks = 50)
9 lines(x = seq(-4,40, 0.01), y =stdnorm)
```



The Standard Normal Distribution

- Mean = 0, Standard Deviation = 1
- 68-95-99 approximation (*rough* approximation)
 - 68% of density within about 1 standard deviation on either side of mean
 - 95% of density within about 2 standard deviations on either side of mean
 - 99% of density within about 3 standard deviations on either side of mean

Standard Errors

$$SE = \sqrt{Var(\bar{X})} = \sqrt{\frac{Var(X)}{n}} = \frac{\sigma}{\sqrt{n}}$$

- What's the Standard Error of our Sample Mean in the simulation above?

```
1 sqrt(p*(1-p)/n)
```

```
[1] 0.01581139
```

Confidence Intervals

What is the formula for a 95% Confidence Interval for a sample mean?

$$[\bar{X} - |z_{0.05/2}| * SE, \bar{X} + |z_{0.05/2}| * SE]$$

How do we calculate $|z_{0.05/2}|$?

```
1 abs(qnorm(0.025, mean = 0, sd = 1))
```

```
[1] 1.959964
```

Confidence Interval Example

What is the 95% confidence interval for the mean of a sample of 1000 from the binomial distribution?

```
1 set.seed(123)
2 samp <- rbinom(1000, size = 1, prob = 0.5)
3 samp.mean <- mean(samp)
4 samp.mean
```

```
[1] 0.493
```

```
1 samp.mean - abs(qnorm(0.025))*sqrt(p*(1-p)/n)
```

```
[1] 0.4620102
```

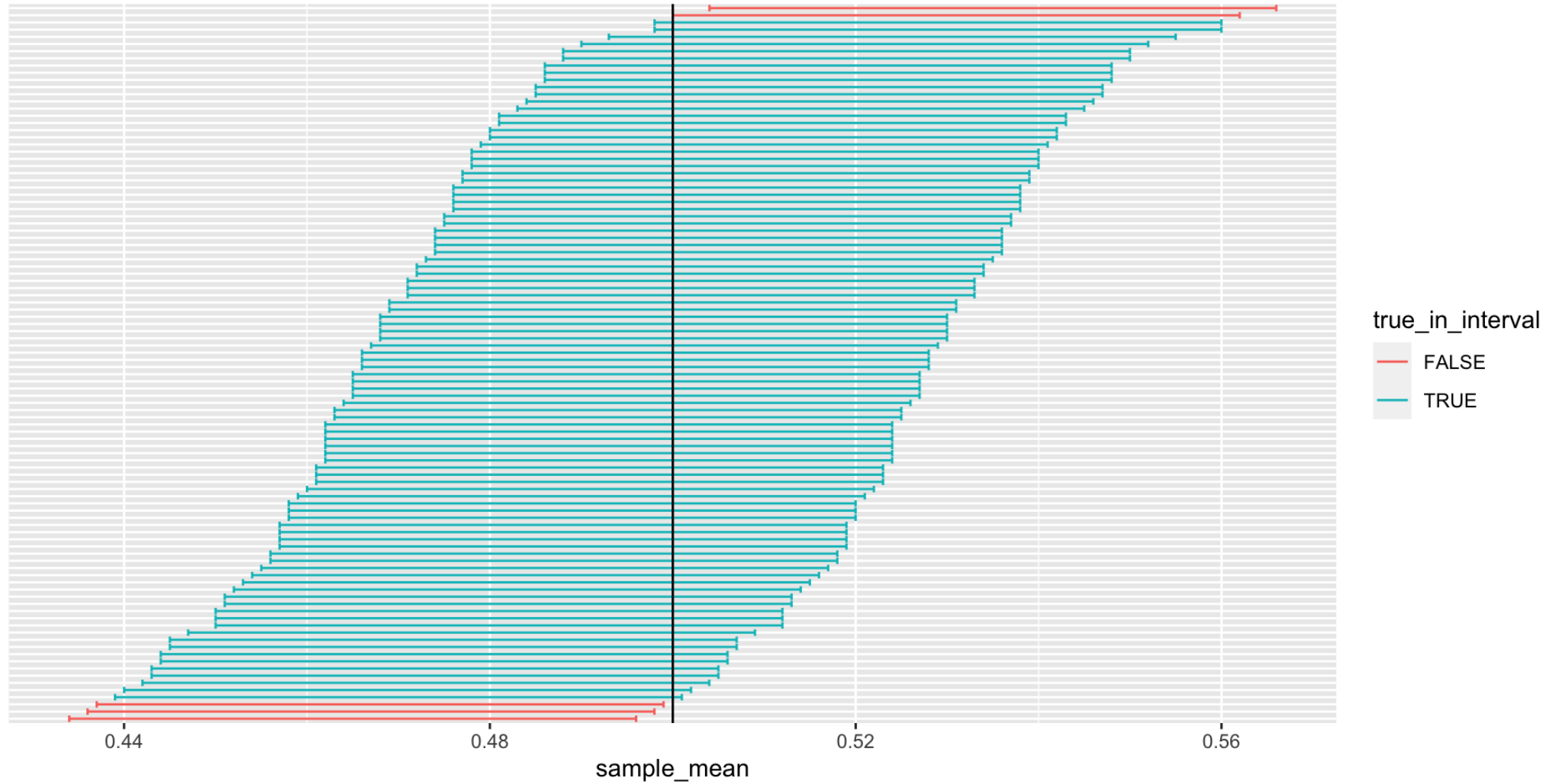
```
1 samp.mean + abs(qnorm(0.025))*sqrt(p*(1-p)/n)
```

```
[1] 0.5239898
```

95% Confidence Interval

- Does *not* mean we are 95% confident the true population mean is within the interval
- It means if we repeated the sampling process many times, about 95% of the confidence intervals would contain the true population mean

95% Confidence Interval



Hypothesis Testing

Two-Sided Hypothesis Test:

- Null Hypothesis: $H_0 : \mu = 0$
- Alternative Hypothesis: $H_1 : \mu \neq 0$
- How do we test the null hypothesis?

Hypothesis Testing

- How do we test the null hypothesis?
- If the null hypothesis is true, the Z-Score is $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \frac{\bar{X}-0}{\sigma/\sqrt{n}}$

Using the Central Limit Theorem...

- Z-Scores should be distributed approximately standard normal
- Calculate the probability of observing a Z-Score at least as large in magnitude as observed *if the null hypothesis is true*
 - This is the “p-value”

Hypothesis Testing: Binomial

$$H_0 : p = 0.5$$

$$H_1 : p \neq 0.5$$

```
1 set.seed(123)
2 samp <- rbinom(1000, size = 1, prob = 0.5)
3 samp.mean <- mean(samp)
4 z.score <- (samp.mean - 0.5)/sqrt(0.5^2/1000)
5 z.score
```

```
[1] -0.4427189
```

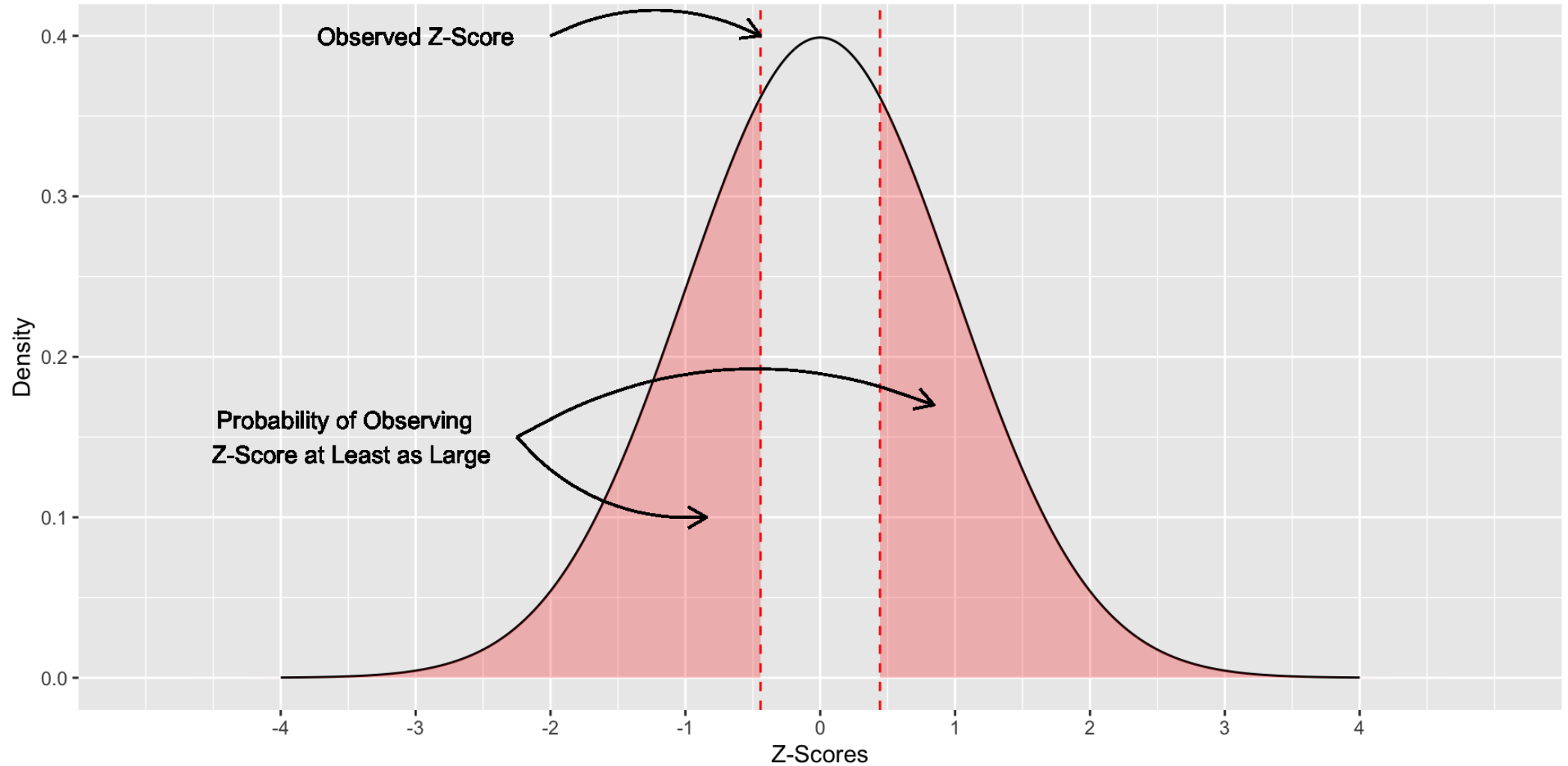
```
1 2*abs(pnorm(z.score))
```

```
[1] 0.6579691
```

Note: if z.score is positive, the p-value can be calculated:

```
1 2*(1 - pnorm(z.score))
```

Hypothesis Testing: Binomial



Regression

Example from Assignment 5:

```
1 progtax <- read.csv("progtax.csv")
```

```
1 fit <- lm(toprate~wwi + gdppc+left_seatshare + factor(country) +  
2         factor(year), data = progtax)  
3 round(summary(fit)$coefficients[c("wwi", "gdppc", "left_seatshare"),],3)
```

	Estimate	Std. Error	t value	Pr(> t)
wwi	32.918	2.550	12.910	0.000
gdppc	-5.854	2.129	-2.750	0.007
left_seatshare	0.041	0.093	0.444	0.657

$$Z. Score_{wwi} = \frac{32.918-0}{2.55} = 12.9$$

```
1 1-pnorm(12.9)
```

```
[1] 0
```

Regression

```
1 fit <- lm(toprate~wwi + gdppc+left_seatshare + factor(country) +  
2           factor(year), data = progtax)  
3 round(summary(fit)$coefficients[c("wwi", "gdppc", "left_seatshare"),],3)
```

	Estimate	Std. Error	t value	Pr(> t)
wwi	32.918	2.550	12.910	0.000
gdppc	-5.854	2.129	-2.750	0.007
left_seatshare	0.041	0.093	0.444	0.657

95% Confidence Interval for $\hat{\beta}_1$:

```
1 32.918 - qnorm(0.975)*2.55
```

```
[1] 27.92009
```

```
1 32.918 + qnorm(0.975)*2.55
```

```
[1] 37.91591
```

```
1 confint(fit)[rownames(confint(fit))%in%  
2           c("wwi", "gdppc", "left_seatshare"),]
```

	2.5 %	97.5 %
wwi	27.8880533	37.9478276
gdppc	-10.0532222	-1.6540168
left_seatshare	-0.1420066	0.2245598

Regression Tables: Stargazer

```
1 install.packages("stargazer")  
2 library(stargazer)  
3 fit1 <- lm(toprate~wwi, data = progtax)  
4 fit2 <- lm(toprate~wwi + gdp pc+left_seatshare+factor(country)+  
5           factor(year), data = progtax)  
6 stargazer(fit1, fit2, type = "text",  
7          omit = "factor",  
8          add.lines = list(c("Country Fixed Effects", "No", "Yes"),  
9                          c("Year Fixed Effects", "No", "Yes")))
```


Stargazer

Dependent variable:		
	toprate	
	(1)	(2)
wwi	34.616*** (2.024)	32.918*** (2.550)
gdppc		-5.854*** (2.129)
left_seatshare		0.041 (0.093)
Constant	8.231*** (1.036)	29.573*** (8.241)
Country Fixed Effects	No	Yes
Year Fixed Effects	No	Yes
Observations	248	228
R2	0.543	0.839
Adjusted R2	0.541	0.804
Residual Std. Error	14.019 (df = 246)	9.262 (df = 187)
F Statistic	292.421*** (df = 1; 246)	24.282*** (df = 40; 187)
Note: *p<0.1; **p<0.05; ***p<0.01		

Fixed Effects

Why might we include country fixed effects?

- Account for omitted country-specific confounders that do not vary over time

Why might we include year fixed effects?

- Account for omitted year-specific confounders that do not vary across countries

Questions

