INTERPRETING PROGRAM IMPACT

Fundamentals of

PROGRAM EVALUATION

JESSE LECY

WHICH MODEL IS THE "RIGHT" ONE?

	Dependent Variable: Test Scores				
	Model 1	Model 2	Model 3	Model 4	Model 5
	(1)	(2)	(3)	(4)	(5)
tqual		57.771***			57.755***
		(0.263)			(0.259)
csize	−4.224***	−4.129***		−2.327	−2.781***
	(0.174)	(0.025)		(1.634)	(0.229)
ses			44.079***	19.920	14.160***
			(1.818)	(17.057)	(2.389)
Constant	702.385***	409.739***	504.439***	613.247***	346.459***
	(4.928)	(1.507)	(3.986)	(76.487)	(10.781)
Observations	1,000	1,000	1,000	1,000	1,000
Adjusted R^2	0.370	0.987	0.370	0.371	0.988
Note:					*p<0.1; **p<0.05; ***p<0.01

INTERPRETING PROGRAM IMPACT

WHICH BET WOULD YOU PREFER?

BET #1

The bet costs $1,000 to place
There is a 75% chance you win $1,500
There is a 25% chance you win $1,100

BET #2

The bet costs $1,000 to place
There is a 75% chance you win $4,000
There is a 25% chance you lose $2,000

WHICH BET WOULD YOU PREFER?

BET #1

The bet costs $1,000 to place
There is a 75% chance you win $1,500
There is a 25% chance you win $1,100

Expected value =
(0.75)(1500) + (0.25)(1100) =
$1,400

BET #2

The bet costs $1,000 to place
There is a 75% chance you win $4,000
There is a 25% chance you **lose $2,000**

Expected value =
(0.75)(4000) – (0.25)(2000) =
$2,500

WHICH BET WOULD YOU PREFER?

BET #1

The bet costs $1,000 to place
There is a 75% chance you win $1,500
There is a 25% chance you win $1,100

Expected value =
(0.75)(1500) + (0.25)(1100) =

$1,400

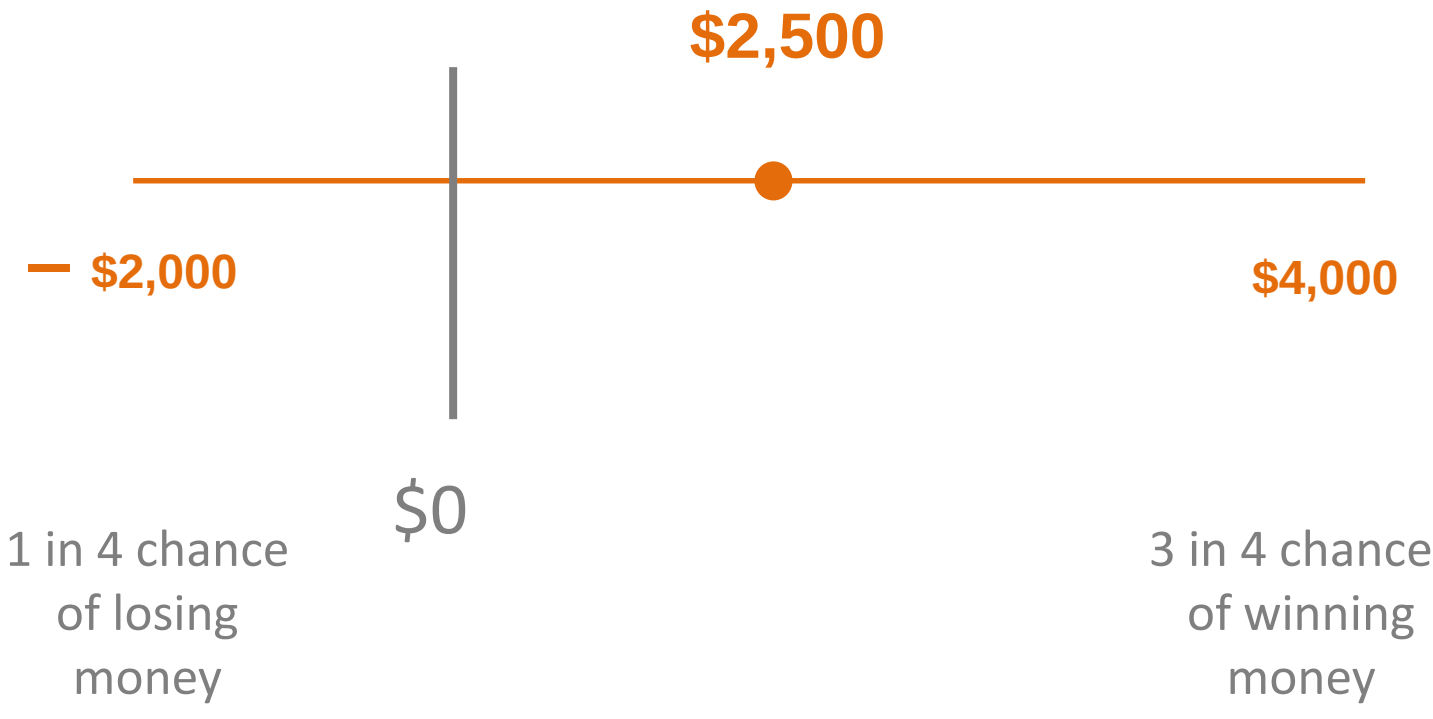
$1,100 **$1,500**

WHICH BET WOULD YOU PREFER?

BET #2

The bet costs $1,000 to place
There is a 75% chance you win $4,000
There is a 25% chance you **lose $2,000**

Expected value =
(0.75)(4000) – (0.25)(2000) =



WHICH BET WOULD YOU PREFER?

BET #1

100% chance of positive return

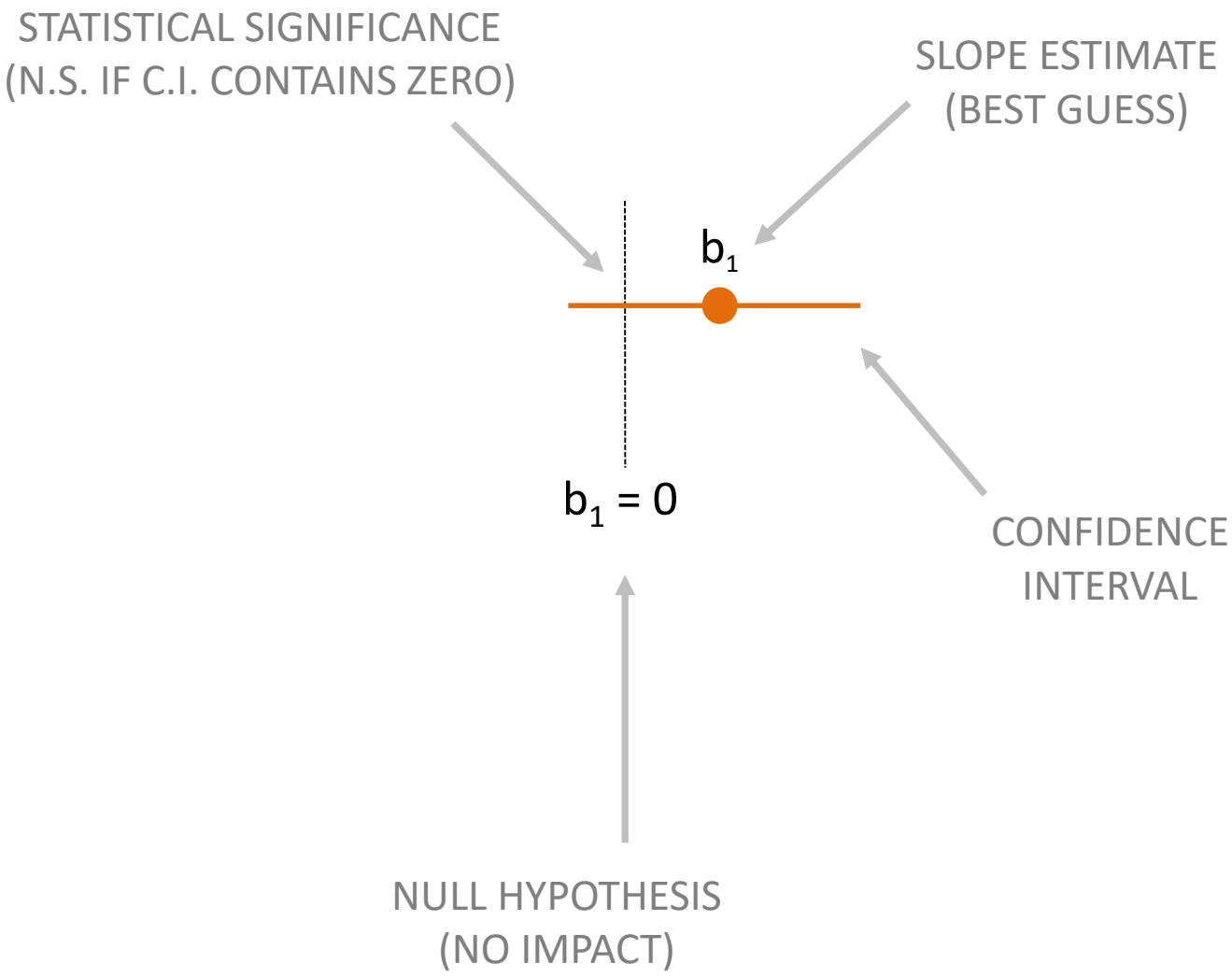
Expected value =
(0.75)(1500) + (0.25)(1100) =
$1,400

BET #2

75% chance of a positive return

Expected value =
(0.75)(4000) – (0.25)(2000) =
$2,500

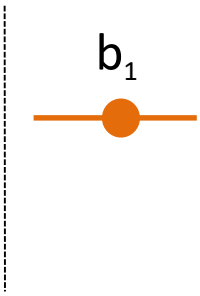
COEFFICIENT PLOTS



WHICH PROGRAM IS BETTER?

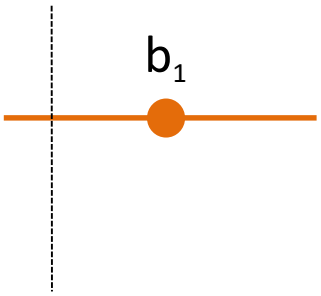
$$\text{Reading Speed} = b_0 + b_1 \text{Hours of Tutoring} + e$$

Program 1



$b_1 = 0$

Program 2



$b_1 = 0$

(assume these are all 95% confidence intervals)

The cost of the program is the bet we are making.

The expected value of the program is represented by the point estimate of the slope (b1).

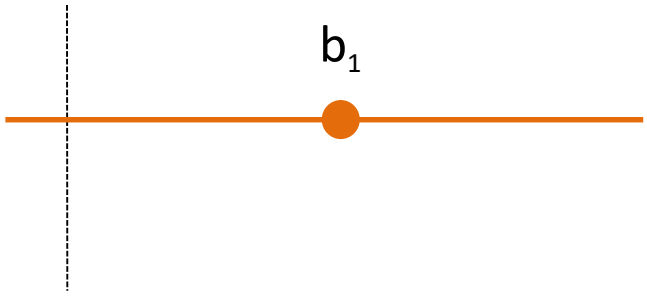
The risk (certainty) of the bet is symbolized by the confidence interval.

Preferences for bets is always a balance between expected pay-off and risk (uncertainty).

WHAT ABOUT NOW?

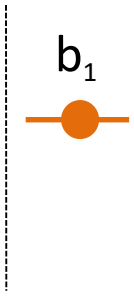
$$\text{Reading Speed} = b_0 + b_1 \text{Hours of Tutoring} + e$$

Program 1



$b_1 = 0$

Program 2

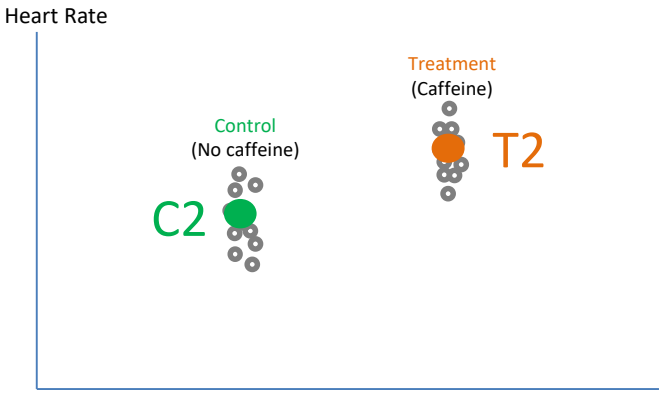


$b_1 = 0$

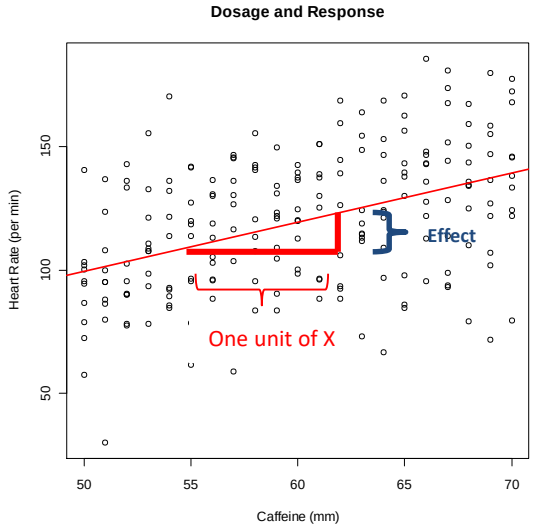
Which model is statistically significant?
Which program has more positive impact?

EXPERIMENT WITH GROUPS (TREATED: YES OR NO)

$$Heartrate = b_0 + b_1 \cdot Caffeine + \varepsilon$$



EFFECTS SIZES FOR LEVELS OF A TREATMENT



$$y = \beta_0 + \beta_1 x$$

β_1 =140

Slope=0

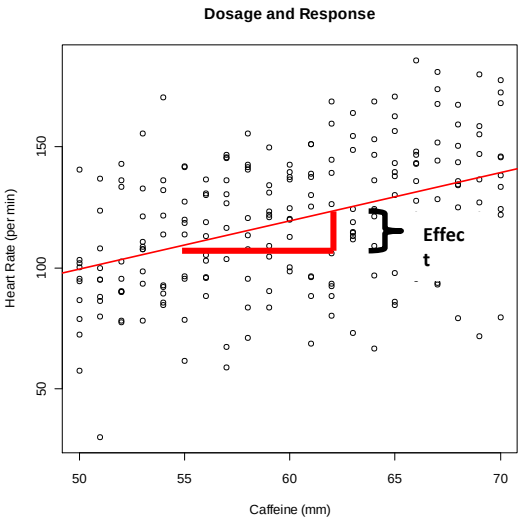
For a one-unit change in X, we expect a β_1 change in Y.

How big is the effect?

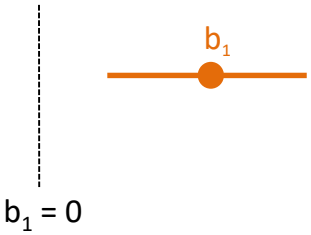
Is it significant?

EFFECT SIZE

$$Heartrate = b_0 + b_1 \cdot Caffeine + \varepsilon$$

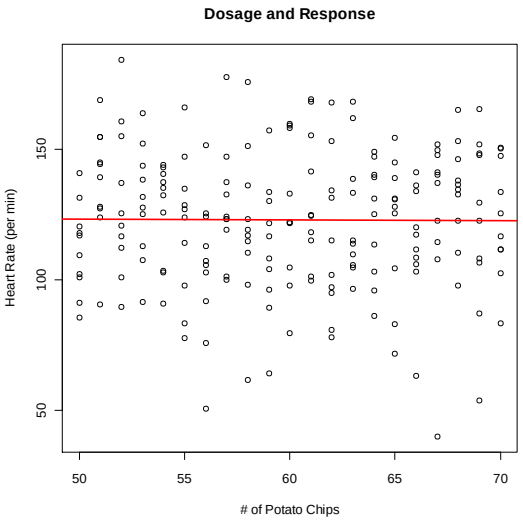


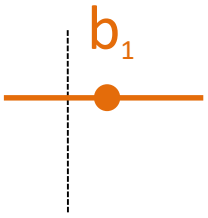
Positive & Significant Impact on Outcome



EFFECT SIZE

$$Heartrate = b_0 + b_1 \cdot Potato\ Chips + \varepsilon$$

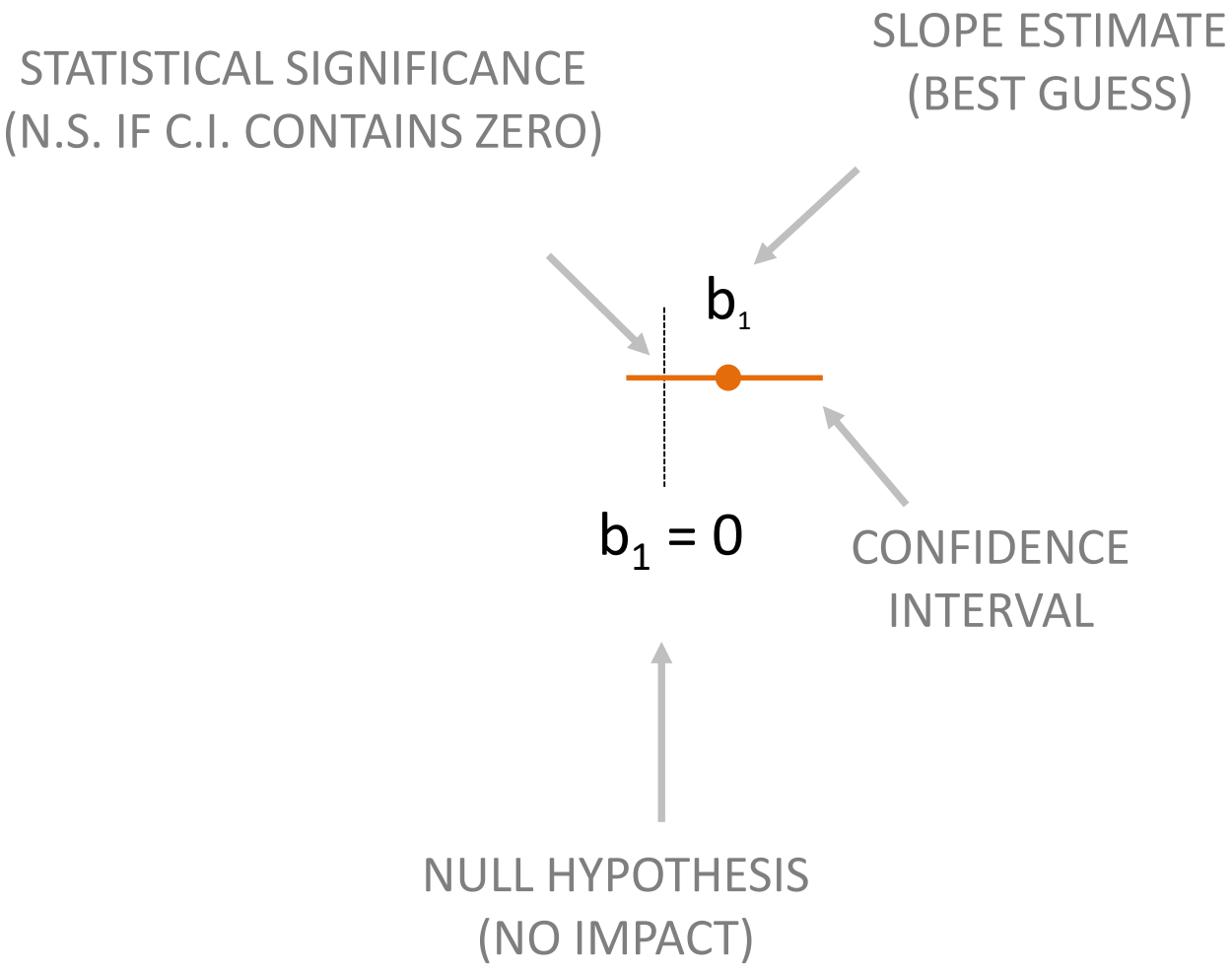




$b_1 = 0$

Not statistically significant – i.e. we can't tell whether the program has a positive or negative impact since the confidence interval is on both sides of zero.

HYPOTHESIS TESTING

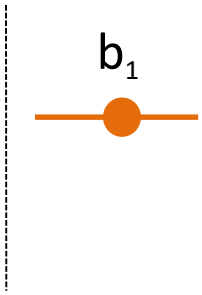


WHICH PROGRAM IS BETTER?

Consider two programs that are meant to improve reading comprehension. The dependent variable is a score on a reading comprehension exam (higher being better). Which program do you prefer and why?

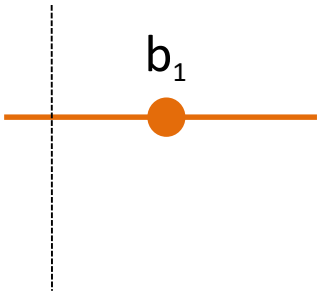
$$\text{Reading Speed} = b_0 + b_1 \text{Hours of Tutoring} + e$$

Program 1



$b_1 = 0$

Program 2



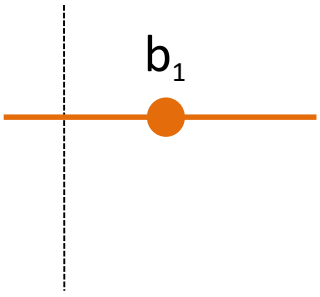
$b_1 = 0$

(assume these are all 95% confidence intervals)

WHAT ABOUT NOW?

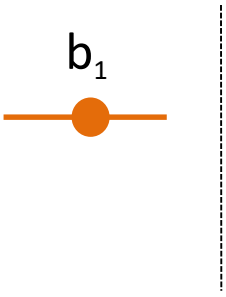
$$\text{Reading Speed} = b_0 + b_1 \text{Hours of Tutoring} + e$$

Program 1



$b_1 = 0$

Program 2



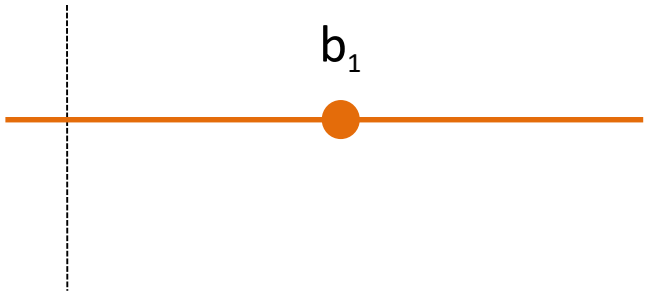
$b_1 = 0$

Which model is statistically significant?
Which program has more positive impact?

WHAT ABOUT NOW?

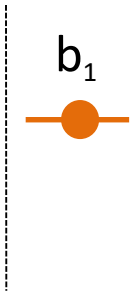
$$\text{Reading Speed} = b_0 + b_1 \text{Hours of Tutoring} + e$$

Program 1



$b_1 = 0$

Program 2

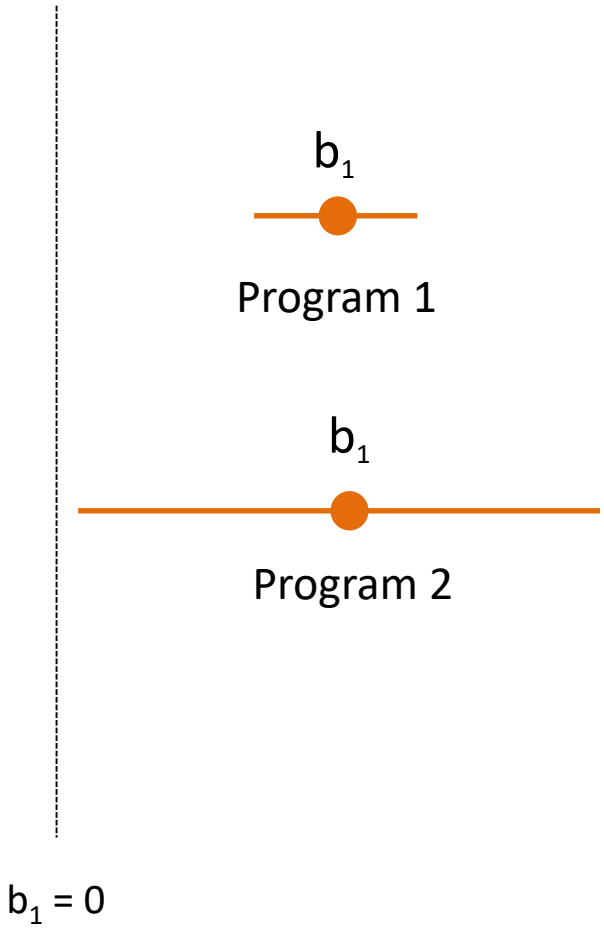


$b_1 = 0$

Which model is statistically significant?
Which program has more positive impact?

WHAT ABOUT NOW?

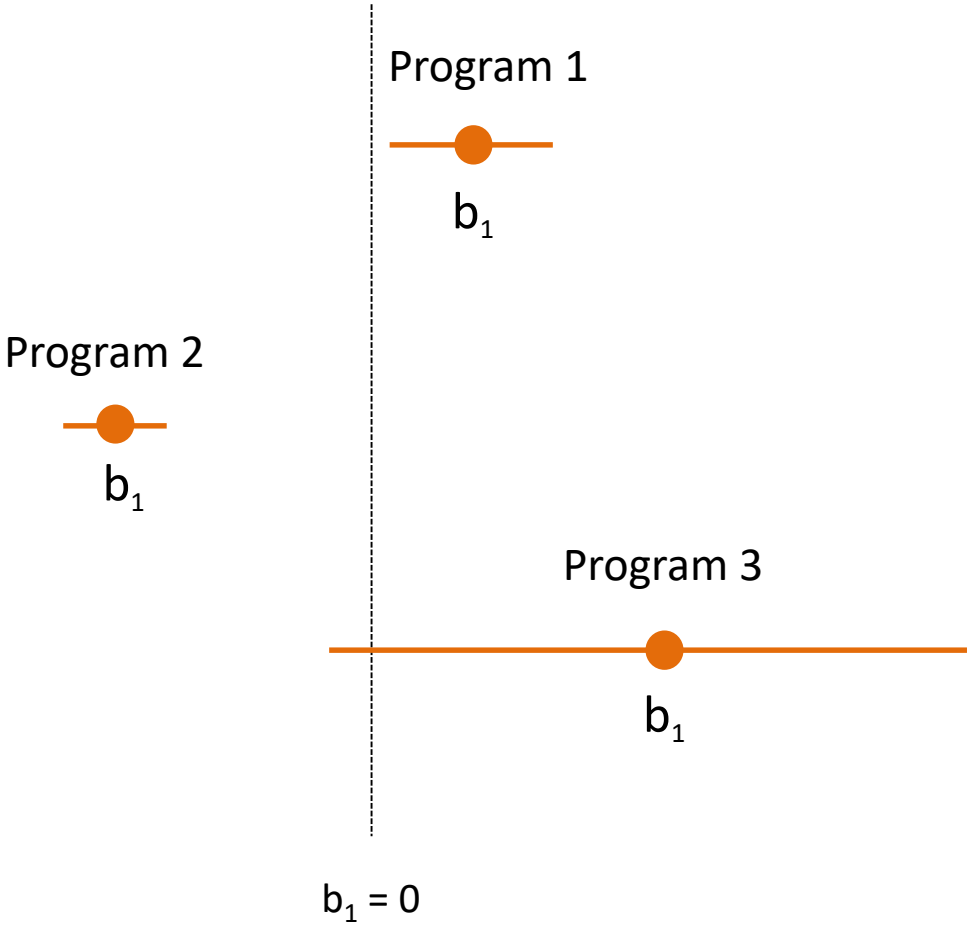
$$\text{Reading Speed} = b_0 + b_1 \text{Hours of Tutoring} + e$$



Which model is statistically significant?
Which program has more positive impact?

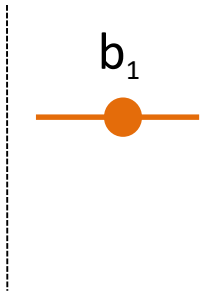
WHAT ABOUT NOW?

$$\text{Reading Speed} = b_0 + b_1 \text{Hours of Tutoring} + e$$



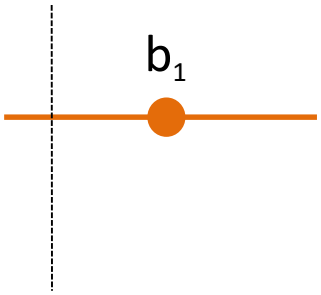
$$\text{Reading Speed} = b_0 + b_1 \text{Hours of Tutoring} + e$$

Program 1



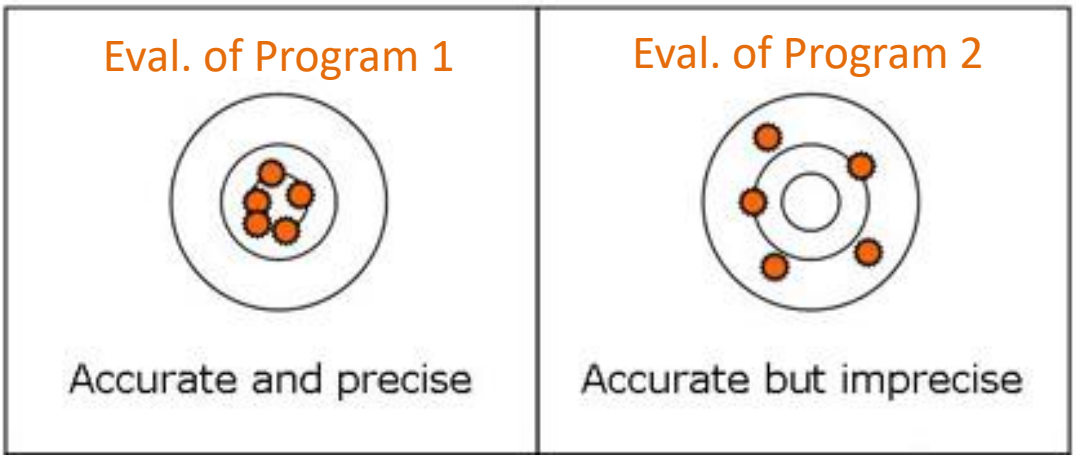
$b_1 = 0$

Program 2



$b_1 = 0$

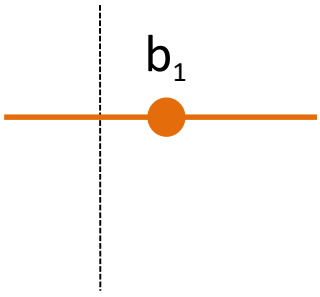
Model precision



LOOKING AHEAD

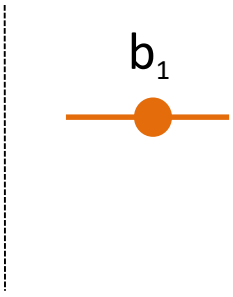
For now we are focusing on the interpretation of coefficient plots. But next week we will look at how adding control variables change models. They can shift coefficients, and change standard errors, changing the interpretations of program effectiveness.

Model 1



$b_1 = 0$

Model 2 w Controls

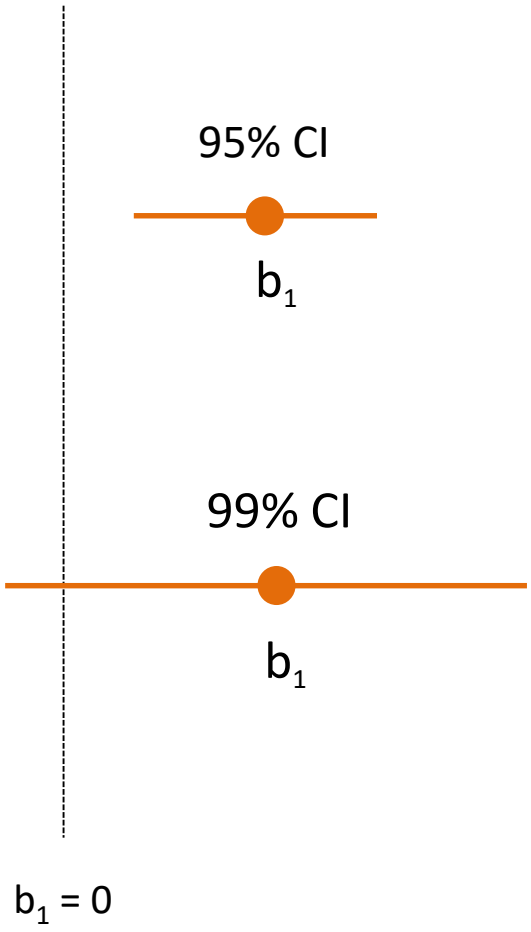


$b_1 = 0$

(assume these are all 95% confidence intervals)

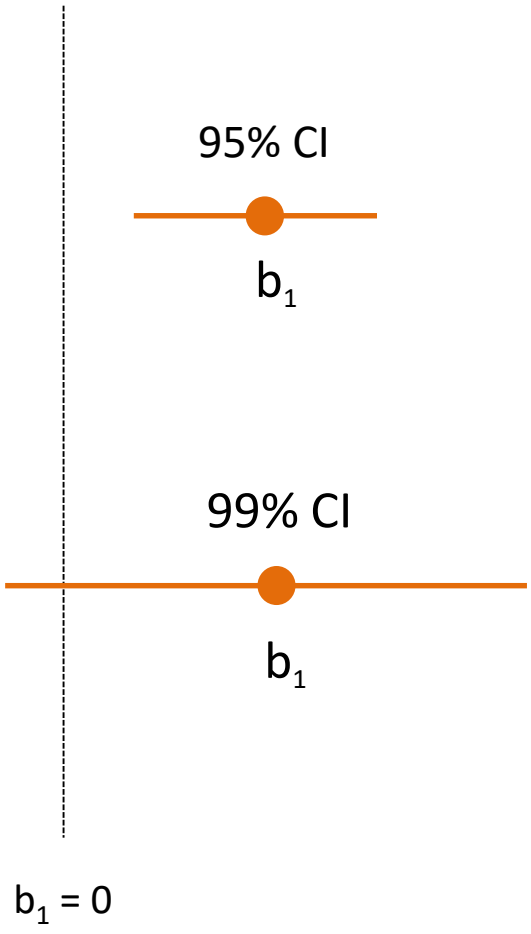
WHAT IS A P-VALUE?

WHICH OF THESE IS STATISTICALLY SIGNIFICANT?



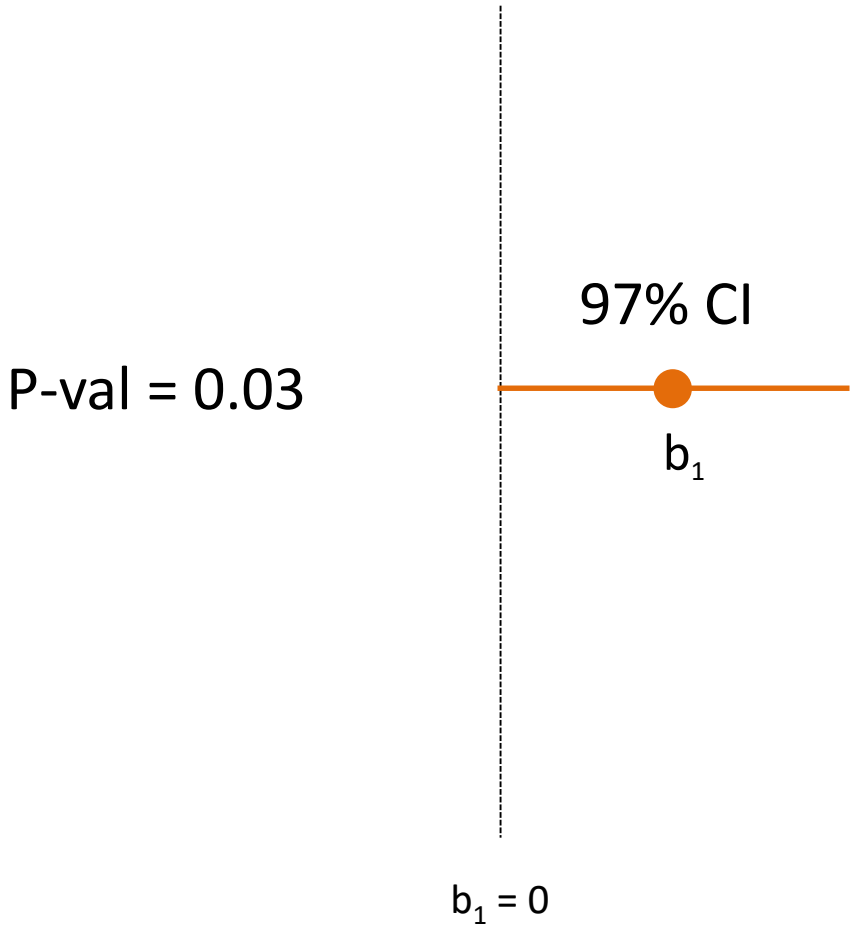
These are both estimates from the same model.

WHAT IS THE P-VALUE IN THIS CASE?



These are both estimates from the same model.

WHAT IS THE P-VALUE IN THIS CASE?

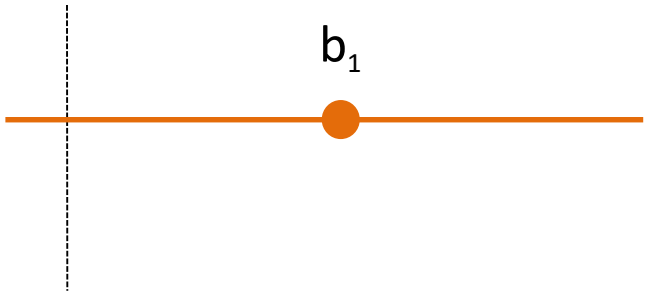


The p-value tells you how large you can draw your confidence interval before it contains the null.

WHICH PROGRAM IS BETTER?

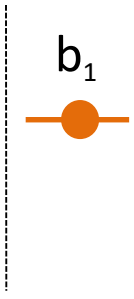
$$\text{Reading Speed} = b_0 + b_1 \text{Hours of Tutoring} + e$$

Program 1



$b_1 = 0$

Program 2



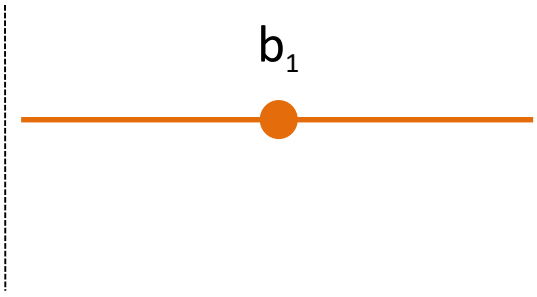
$b_1 = 0$

95% CONFIDENCE INTERVALS

WHAT ABOUT NOW?

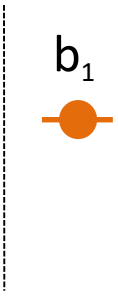
$$\text{Reading Speed} = b_0 + b_1 \text{Hours of Tutoring} + e$$

Program 1



$b_1 = 0$

Program 2



$b_1 = 0$

90% CONFIDENCE INTERVALS

WHICH BET WOULD YOU PREFER?

BET #1

The bet costs $1,000 to place
There is a 75% chance you win $1,500
There is a 25% chance you win $1,100

BET #2

The bet costs $1,000 to place
There is a 75% chance you win $4,000
There is a 25% chance you win $0

MECHANICS OF CONFIDENCE INTERVALS

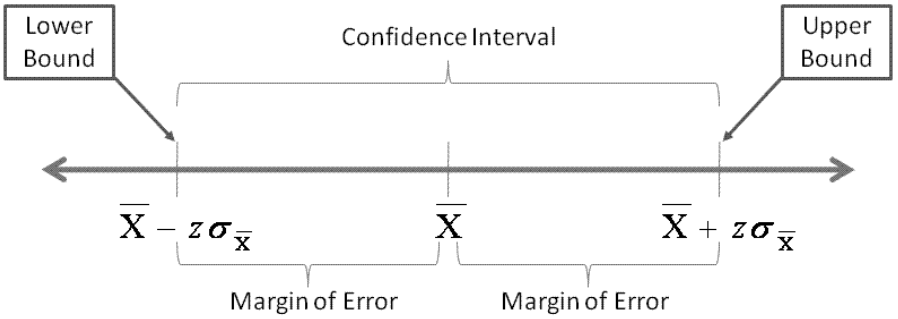
THE ROAD MAP (AGAIN)

	Of the Mean:	Of the Slope:
Sampling Variance:	$\sigma_x^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$ (for x)	$\sigma_\varepsilon^2 = \frac{SSE}{n-2} = \frac{\sum e_i^2}{n-2}$ (using the residual)
↓		
Standard Deviation:	$\sigma_x = \sqrt{\sigma_x^2}$	$\sigma_\varepsilon = \sqrt{\sigma_\varepsilon^2}$
↓		
Standard Error:	$SE_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$	$SE_{b_1} = \sqrt{\frac{\sigma_\varepsilon^2}{\sum (x_i - \bar{x})^2}}$
↓		
Confidence Interval	$\mu = \bar{x} \pm t \cdot SE_{\bar{x}}$ (of the mean)	$\beta_1 = b_1 \pm t \cdot SE_{b_1}$ (of the slope)

THE FORMULA

If we were sure of ourselves we wouldn't need a margin of error!

We only have a sample, though, so we can't be certain.

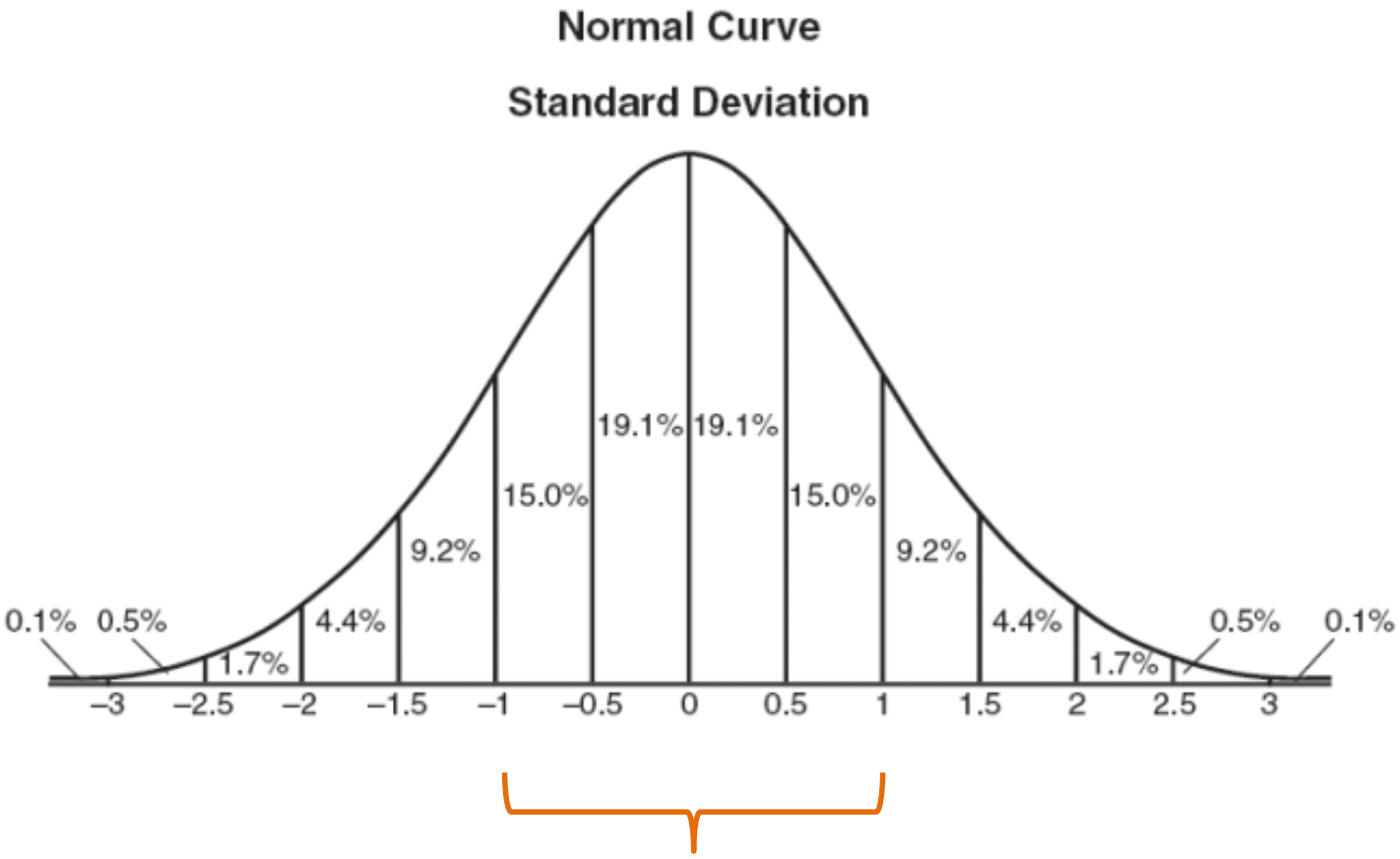


$$CI \ \ for \ \mu = \bar{x} \pm t\frac{s}{\sqrt{n}}$$

(CI of the mean)

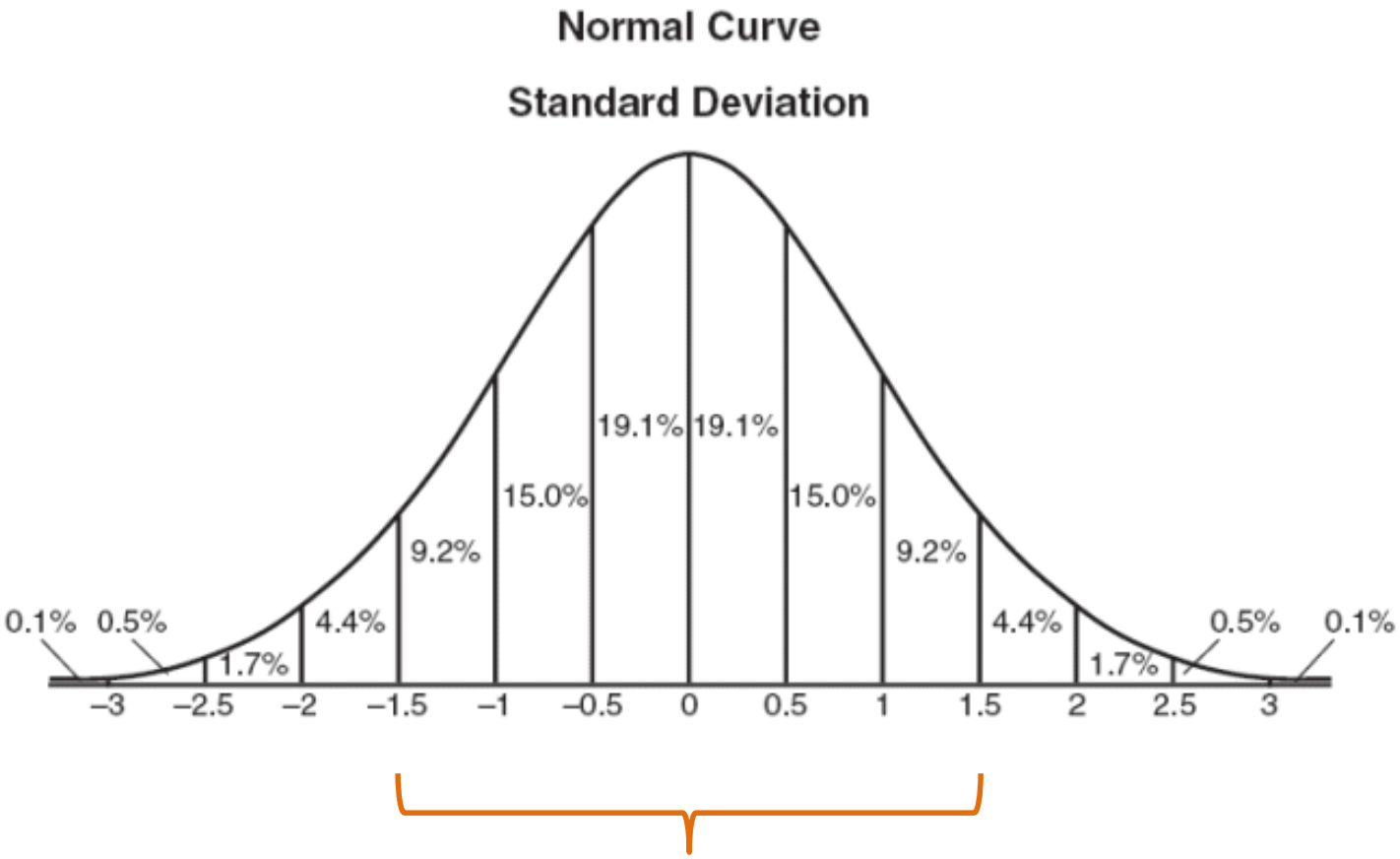
The population parameters are never known, so we use t-stats and the formula for the sample standard error.

WHAT IS A T-VALUE?



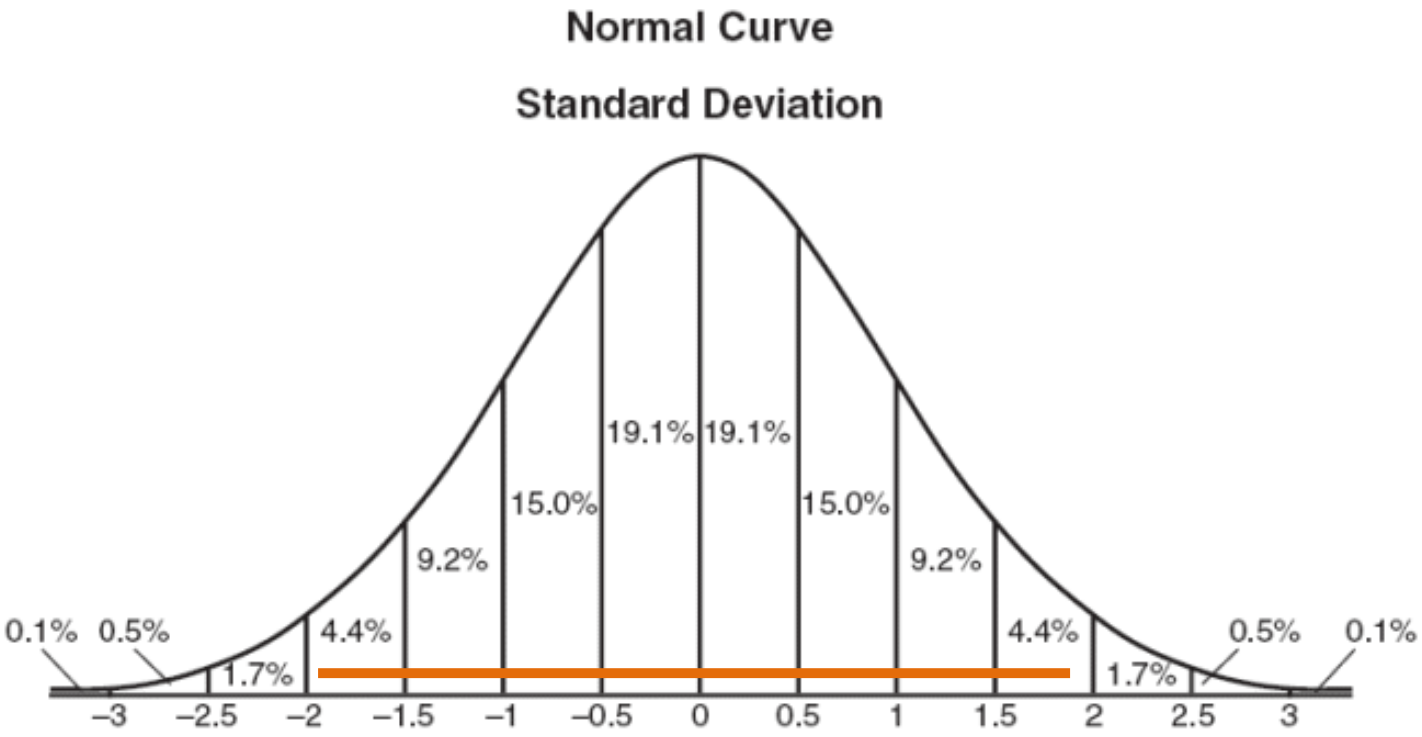
If we examine an interval that is 1 standard deviation from the mean in both directions, we know that this will include 68.2% of the cases.

WHAT IS A T-VALUE?



If we examine an interval that is 1.5 standard deviations from the mean in both directions, we know that this will include 86.6% of the cases.

WHAT IS A T-VALUE?



t=1.96

I want a 95% confidence interval, so I find the t-value where 95% of the data falls within the interval (in a 2-sided test).

DETERMINE T-VALUE:

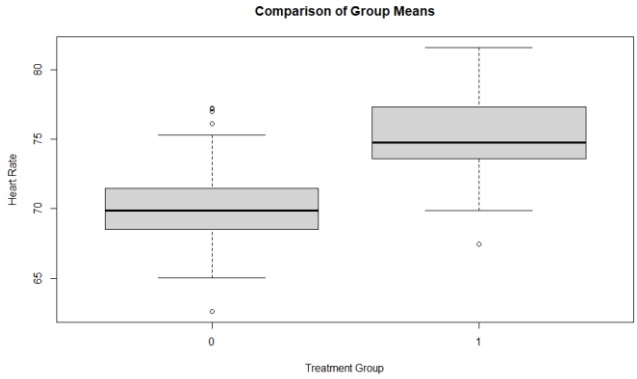
$$CI \ for \ \beta_1 = b_1 \pm t \cdot SE_{b_1}$$

$$CI \ for \ \mu = \bar{x} \pm t \cdot SE_{\bar{x}}$$

1) Select a level of confidence
2) Figure out your sample size
3) Find a t-table
4) Match level of confidence to sample size

Or just use software like a normal person

```
# create some fake data
# y = heartrate
# t = treatment (e.g. caffeine)

t <- rep( c(0,1), 50 )
y <- 70 + rnorm(100,0,3) + 5*t

plot( factor(t), y,
+       main="Comparison of Group Means",
+       xlab="Treatment Group",
+       ylab="Heart Rate" )
```



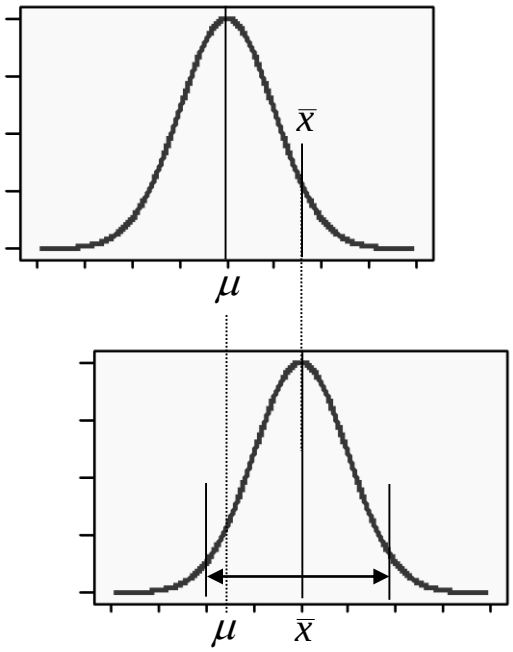

```
head( data.frame( y, t ) )
         y t
1 67.67494 0
2 71.19501 1
3 69.81960 0
4 73.58000 1
5 65.47680 0
6 74.04290 1
```

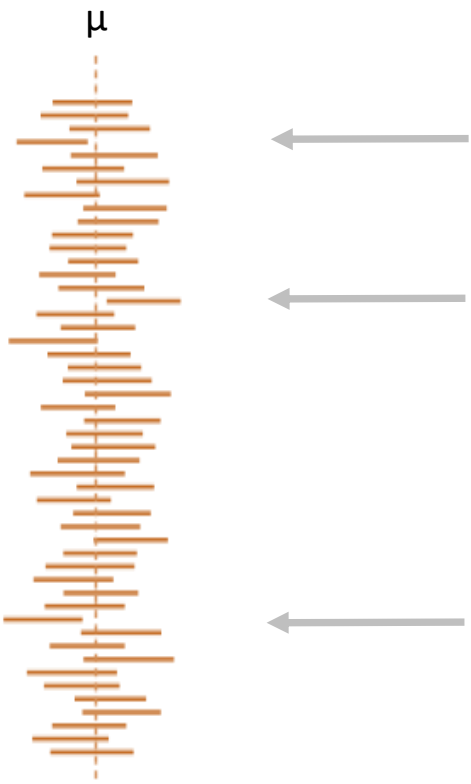
```
mean.t <- mean( y[ t == 1 ] )
mean.c <- mean( y[ t == 0 ] )

mean.c
[1] 70.29385
mean.t
[1] 75.25103

# effect size
mean.t - mean.c
[1] 4.957187

t.test( y ~ t )

        Welch Two Sample t-test

data:  y by t
t = -8.1231, df = 97.927, p-value = 1.391e-12
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.168235 -3.746139
sample estimates:
mean in group 0 mean in group 1
       70.29385        75.25103
```

HOW OFTEN ARE WE WRONG?

If **alpha=0.05**, what is our level of confidence?

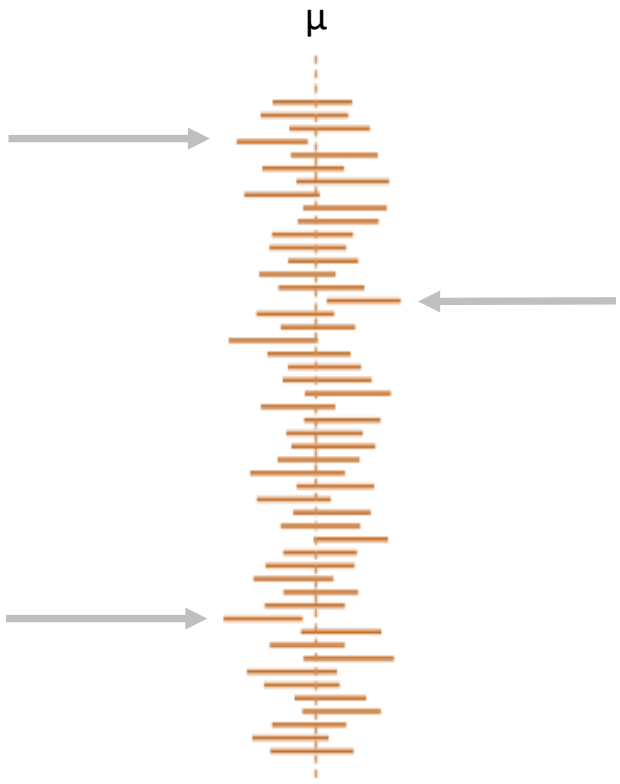
HOW OFTEN ARE WE WRONG?





Chose an **alpha-level**, which determines the size of the confidence interval. This example uses alpha=0.05. We would expect five samples in one-hundred to result in confidence intervals that do not contain the true mean. We see 3 in 50 draws here, which is consistent with expectations.

HOW OFTEN ARE WE WRONG?

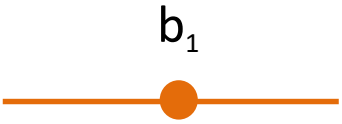


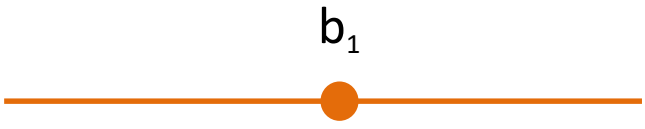
What if we change our alpha from 0.05 to 0.10, how many of these confidence intervals would not contain the true mean?

HOW OFTEN ARE WE WRONG?

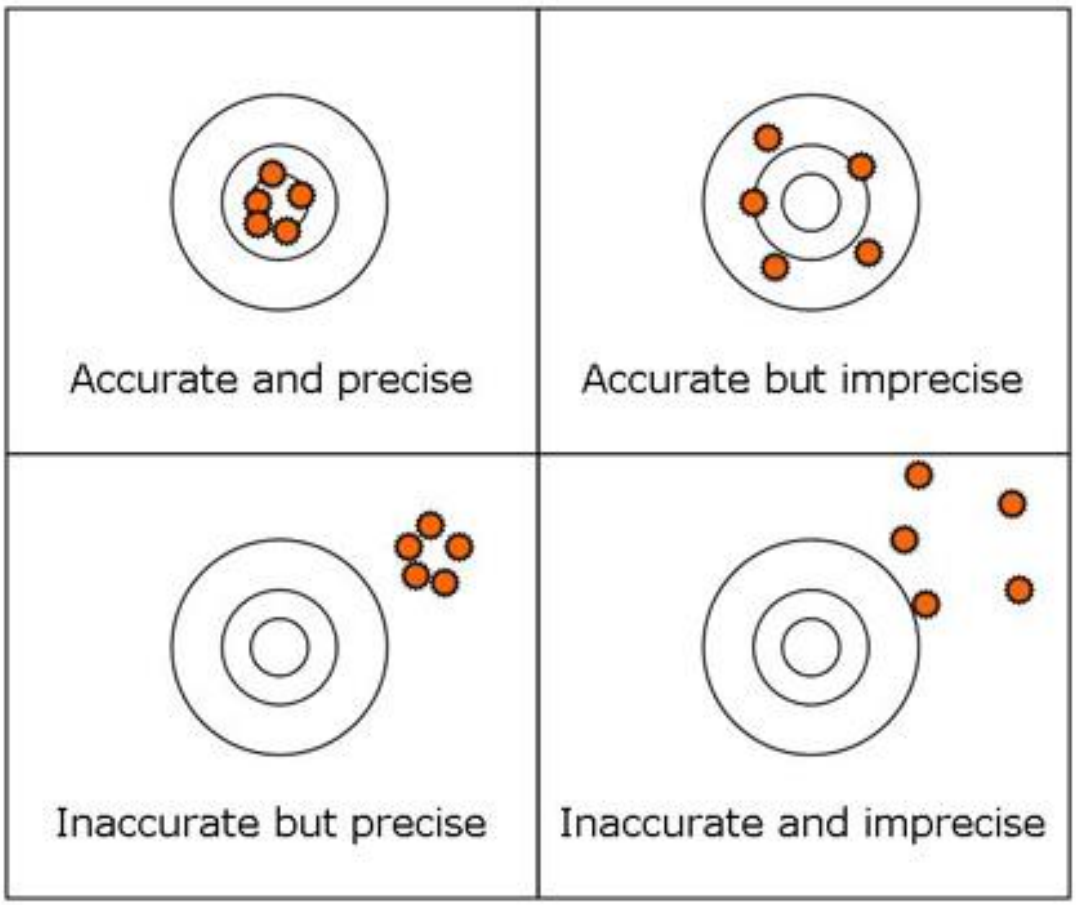
Is a **90% confidence interval** bigger

or small than a **95% confidence interval**?





WHERE WE ARE GOING:



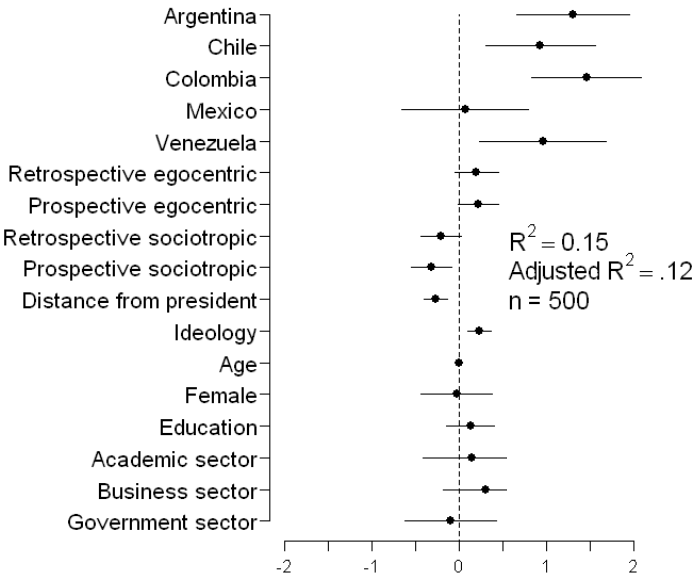
Regression estimates should be:

1. **UNBIASED** (accurate)
2. **EFFICIENT** (precise)

COEFFICIENT PLOTS AS AN ALTERNATIVE TO DENSE REGRESSION TABLES

Variable	Coefficient (Standard Error)
Constant	.41 (.93)
Countries	
Argentina	1.31 (.33)$^{\#\#\#\ B,M}$
Chile	.93 (.32)$^{\#\#\#\ B,M}$
Colombia	1.46 (.32) $^{\#\#\#\ B,M}$
Mexico	.07 (.32)^{A,CH,CO,V}
Venezuela	.96 (.37)$^{\#\#\ B,M}$
Threat	
Retrospective egocentric economic perceptions	.20 (.13)
Prospective egocentric economic perceptions	.22 (.12)$^{\#}$
Retrospective sociotropic economic perceptions	-.21 (.12)$^{\#}$
Prospective sociotropic economic perceptions	-.32 (.12)$^{\#\#}$
Ideological Distance from president	
Ideology	
Ideology	.23 (.07) $^{\#\#\#}$
Individual Differences	
Age	.00 (.01)
Female	-.03 (.21)
Education	.13 (.14)
Academic Sector	.15 (.29)
Business Sector	.31 (.25)
Government Sector	-.10 (.27)
R^2	.15
Adjusted R^2	.12
n	500
$^{\#\#\#}p < .01$, $^{\#\#}p < .05$, $^{\#}p < .10$ (two-tailed)	

Table 2 from Stevens (2006): Determinants of Authoritarian Aggression



This figure has everything we need to interpret the regression. It shows the magnitude of the relationship between each variable and Y, the standard error of each coefficient (encoded in the confidence interval), and statistical significance (does it cross zero?).

Which do you prefer?

INTERPRETING PROGRAM IMPACT

What should be clear in my mind?

1. **Our interpretation of program impact involves an understanding of the "effect size" (regression slope) and the precision with which it is estimated (the confidence interval).**

2. **The level of confidence we select determines the t-value, which determines the size of the confidence interval.**

3. **For a program to be statistically significant, the confidence interval around the slope should not contain the null hypothesis (slope=0).**

4. **We can choose an arbitrary level of confidence such that our confidence interval will not contain the null.**

5. **The p-value tells us the largest confidence interval that we can draw that does not contain the null.**

6. **Program investments are bets that balance effect size plus confidence.**

ASIDE:
STATISTICAL POWER

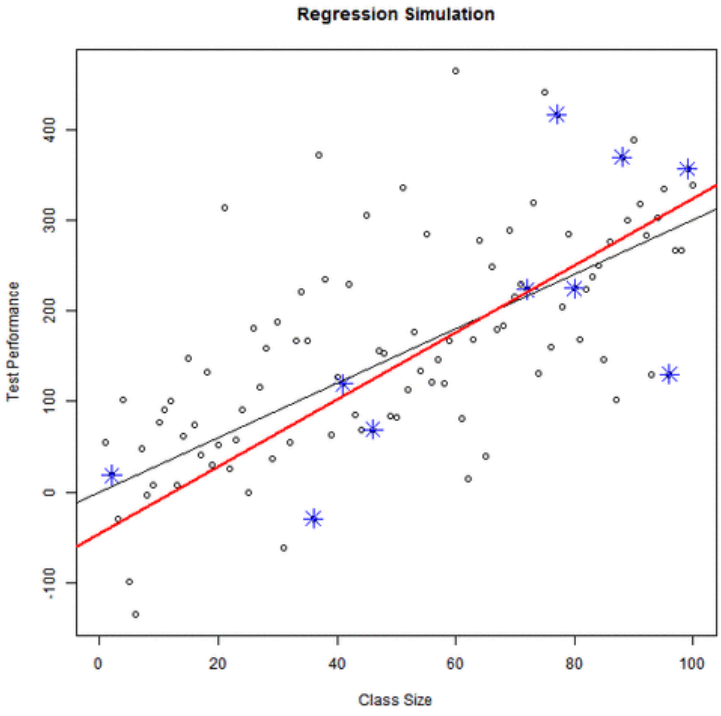
STATISTICAL POWER

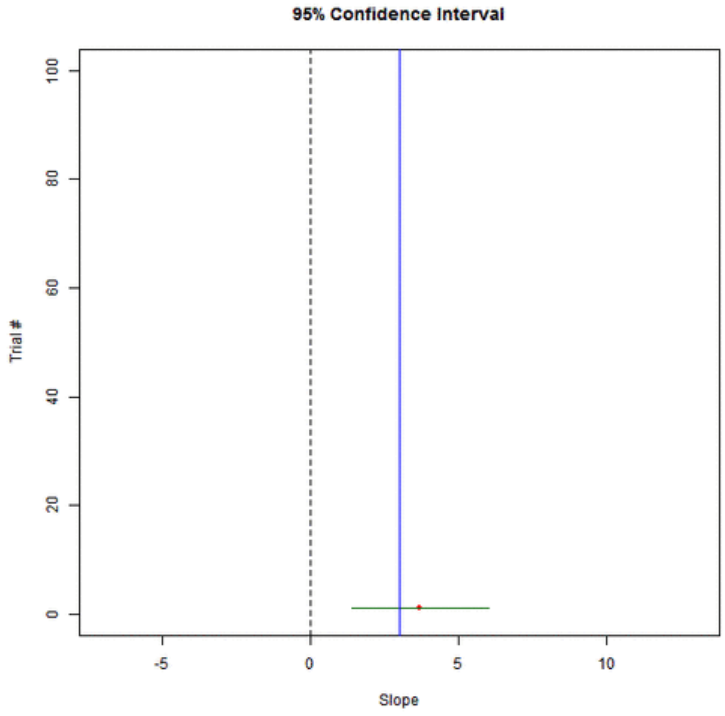
Power: the ability to detect a program impact when it exists.

Type I Error: Claiming a program has impact when it doesn't (false positive). This type of error is usually caused by bias in our estimate of impact.

Type II Error: Failure to detect program impact when it exists (false negative). This type of error is usually caused by a lack of adequate statistical power (large standard errors).

LOW POWER

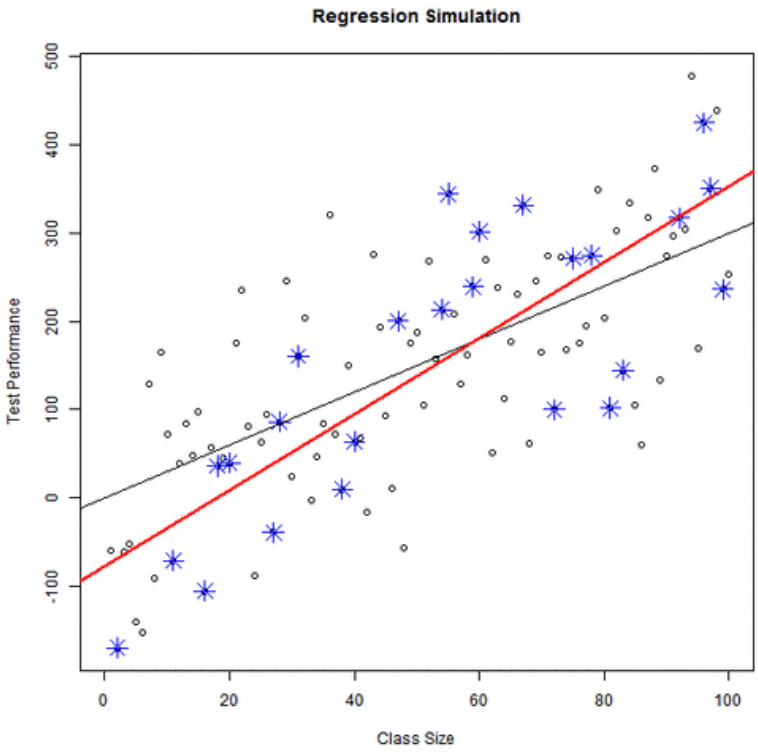


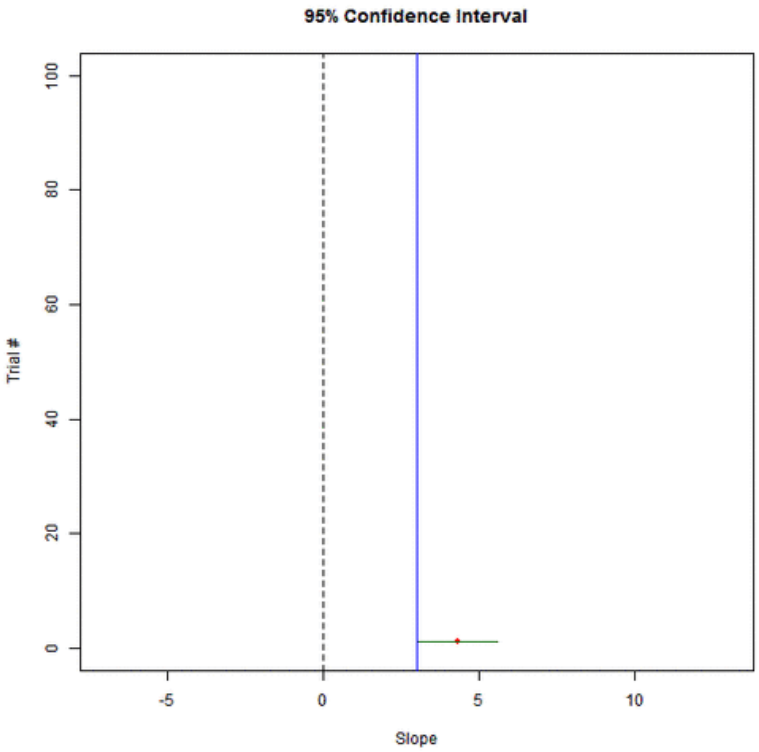


In many cases we fail to reject the null, even though our true program impact is a slope of 3.

Note that our model is unbiased – our estimates all cluster around the true slope.

HIGH POWER

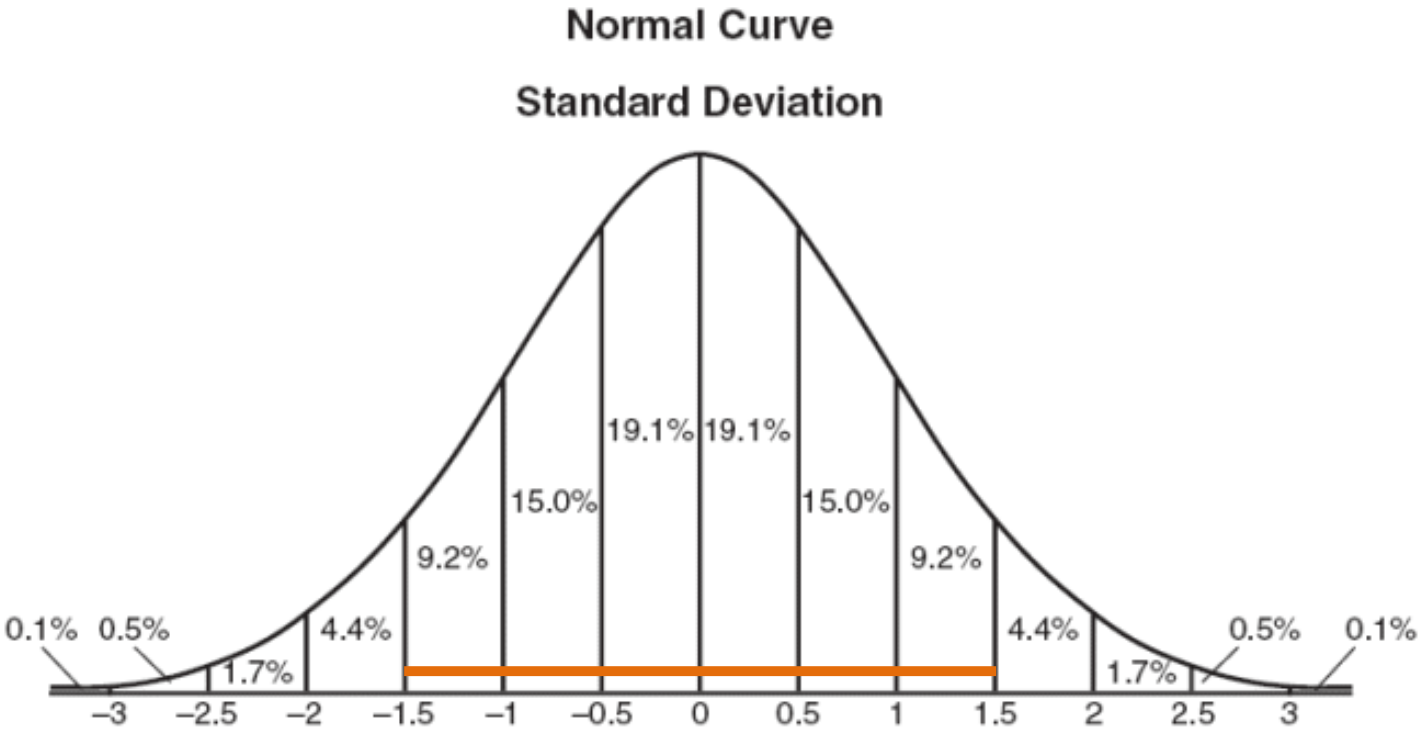




We are sure here that we have enough power to say something concrete about the program impact. We do not worry about Type II Errors in this evaluation.

What has changed?

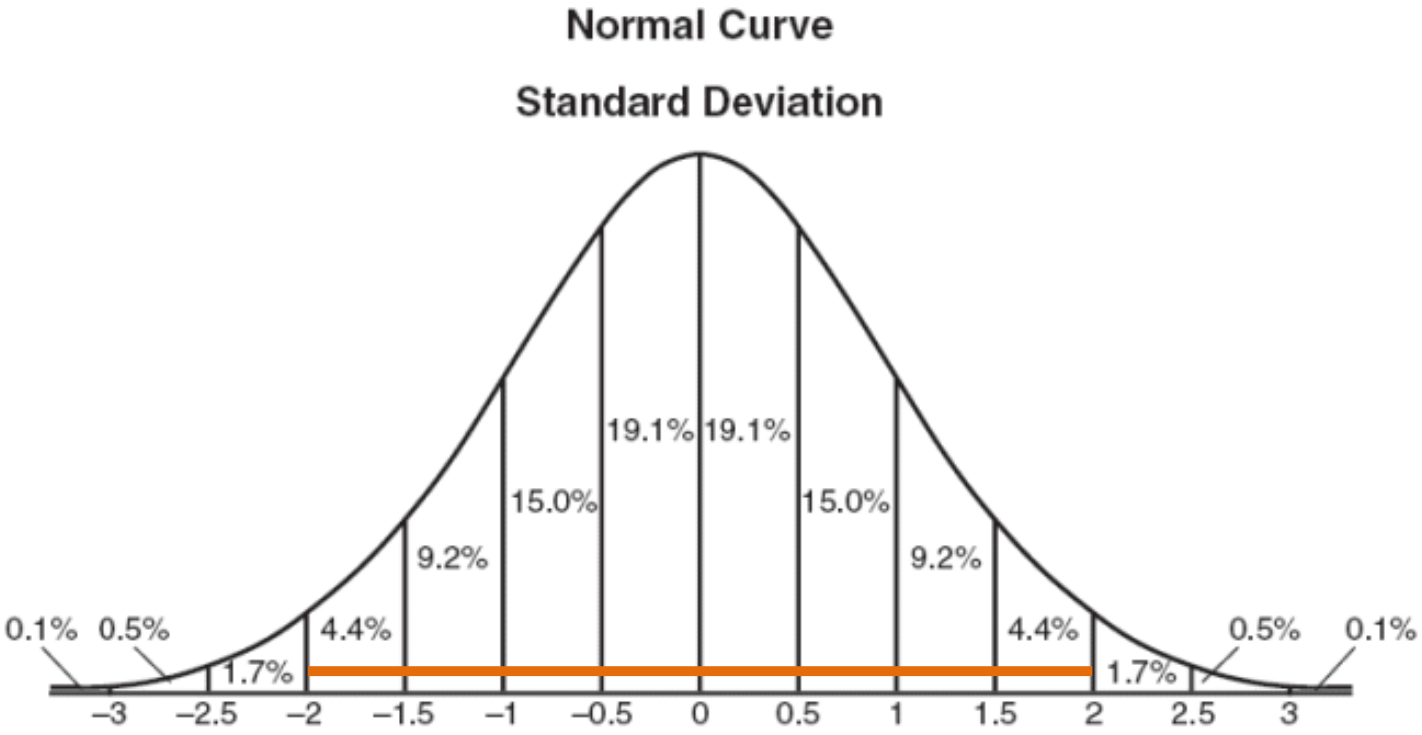
WHAT IS THE "COST" OF GAINING MORE CONFIDENCE?



An interval with a width of of 1.5 stan. dev.'s ensures that we capture 86.6% of the data

There is an increasing marginal cost of gaining confidence.

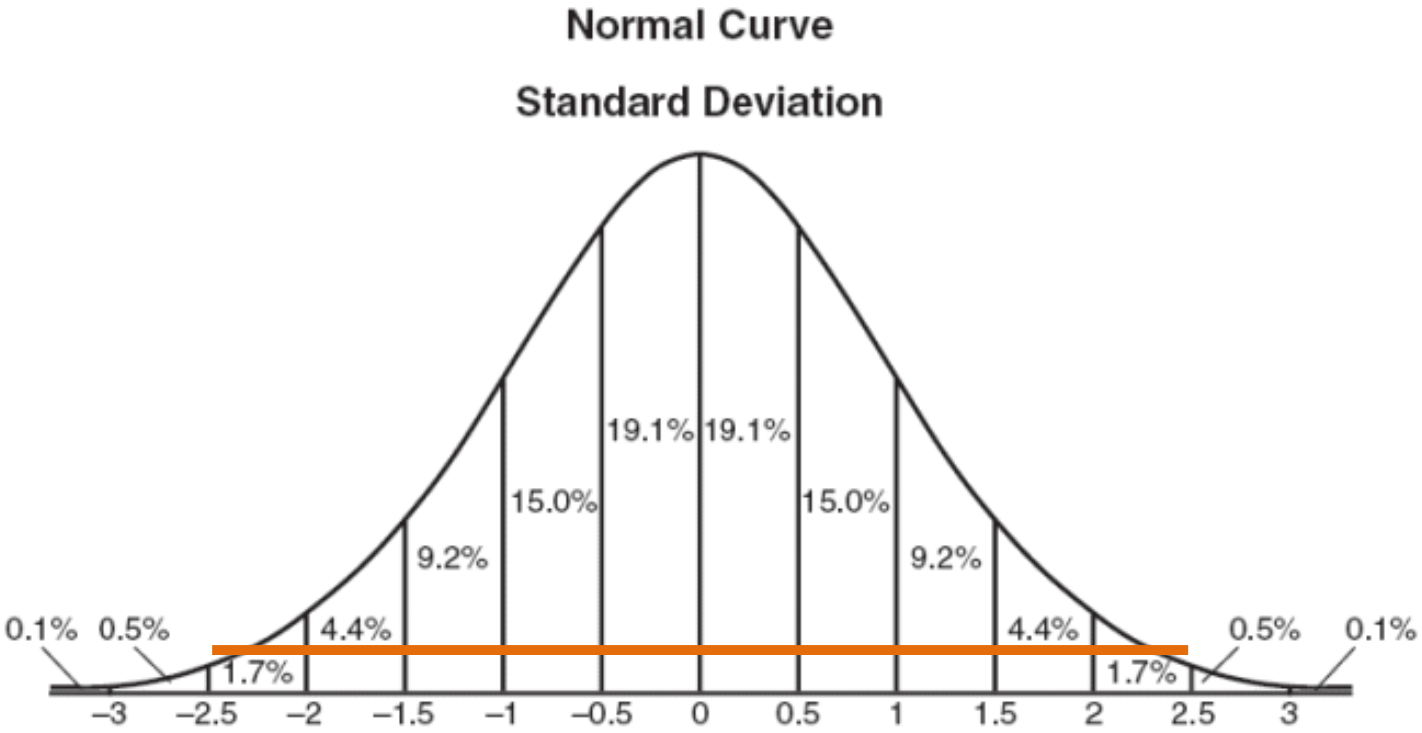
WHAT IS THE "COST" OF GAINING MORE CONFIDENCE?



If we increase the interval to 2 standard deviations we now capture 95.4% of the data for a gain of 8.8 points of confidence.

There is an increasing marginal cost of gaining confidence.

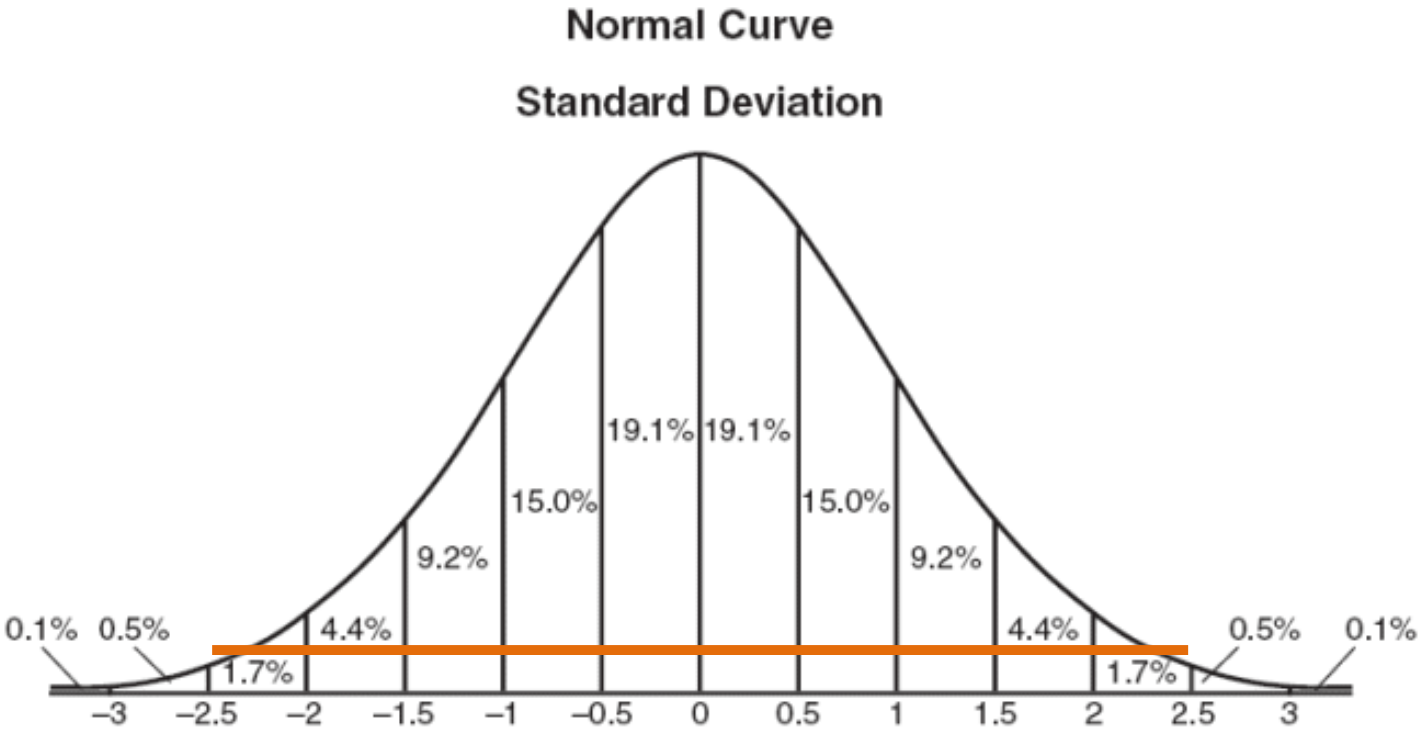
WHAT IS THE "COST" OF GAINING MORE CONFIDENCE?



If we take another half-unit step to 2.5 standard deviations from the mean we now capture 98.8% of the data, but we gain only 3.4 points from the same increase in interval size, less than half the confidence gain as before.

Increasing the interval from 2.5 to 3 standard deviations results in only 1 more point of confidence gained.

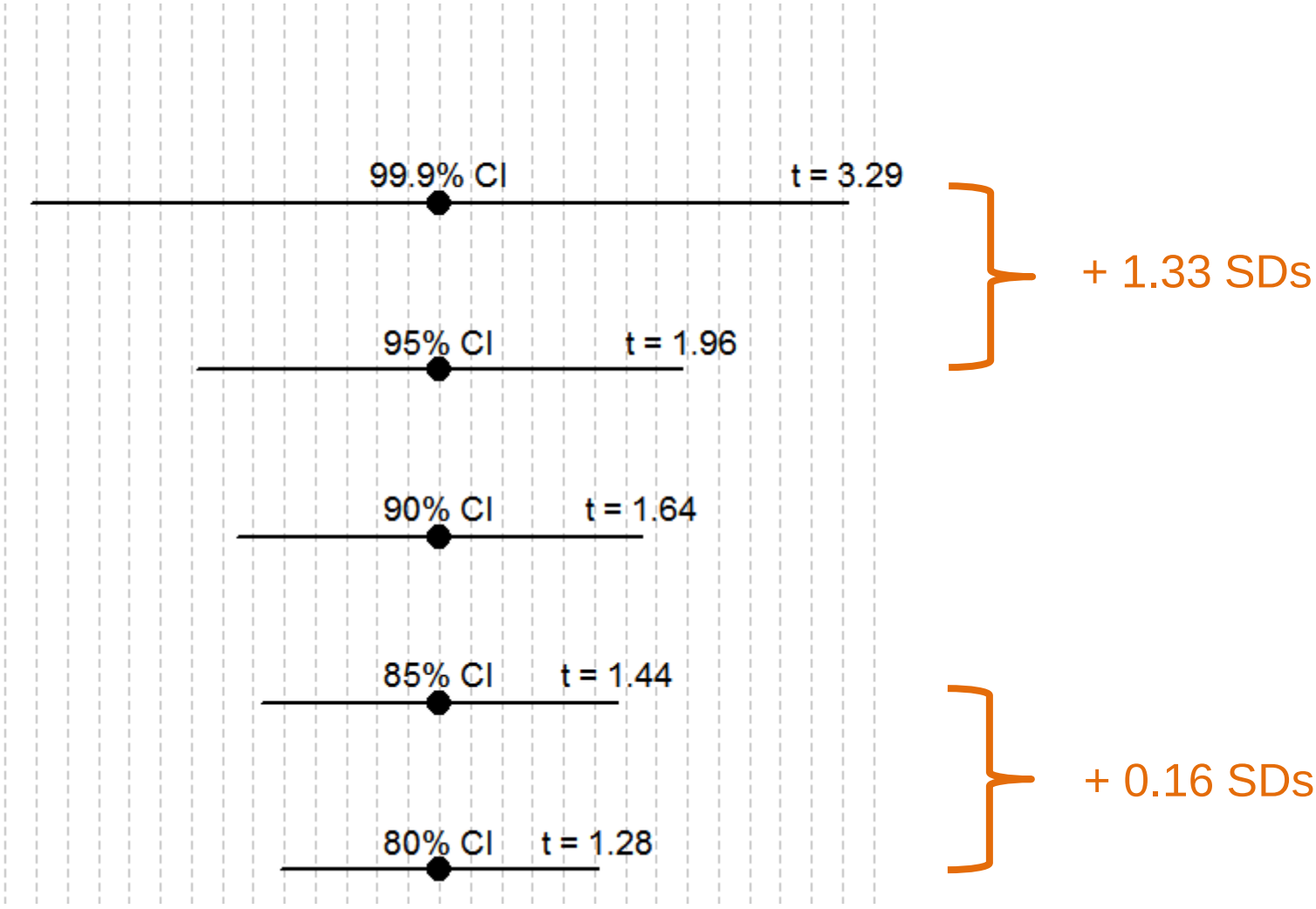
WHAT IS THE "COST" OF GAINING MORE CONFIDENCE?



Each additional unit of confidence become more and more expensive as you approach 100%.

What is the relationship between a "unit of confidence" and a confidence interval?

WHAT IS THE "COST" OF GAINING MORE CONFIDENCE?



There is an increasing marginal cost of gaining confidence. The

```
x.85 <- round( qnorm( 0.075, mean = 0, sd = 1 ), 2 )

x.90 <- round( qnorm( 0.05, mean = 0, sd = 1 ), 2 )

x.95 <- round( qnorm( 0.025, mean = 0, sd = 1 ), 2 )

x.999 <- round( qnorm( 0.0005, mean = 0, sd = 1 ), 2 )

ci.lower <- c(x.80,x.85,x.90,x.95,x.999)

par( mar=c(0,0,0,0) )

plot.new()
plot.window( xlim=c(-3.5,3.5), ylim=c(1,6) )

abline( v=seq(-3.5,3.5,by=0.25), lty=2, col="gray" )
points( rep(0,5), 1:5, pch=19, cex=2 )
segments( x0=ci.lower, x1=abs( ci.lower ), y0=1:5,
          lwd=2 )

text( rep(0,5), 1:5, c("80% CI","85% CI","90% CI","95% CI","99.9% CI"),
      cex=1.2, pos=3 )

text( abs( ci.lower ), 1:5,
      paste("t = ",abs( ci.lower ),sep=""), cex=1.2, pos=3 )
```