2019 GSPIA Amazing Analytics Race

Wednesday Training Camp

Sera Linardi

Associate Professor of Economics

9:00am Getting ready: Your To-Do List

- IT: Tekky Bambang
- TA: Kristin Ronzi
- Get the name of 2 **new** people around you.

1. Register and find your ID number on your name tag
2. Get STATA if you haven't already. Open STATA.
3. Go to http://www.linardi.gspia.pitt.edu/?page_id=564 for the schedule. If you are unable to get online, talk to Tekky or Kristin.
4. Download all materials for Wednesday into a folder in your computer.
5. Click on Exercise: Baseline and try it. Use the ID # from your name tag.

We start lecture at 9:30am.

Welcome

Sera Linardi (linardi@pitt.edu)

- PhD in Social Science, California Institute of Technology
- Behavioral / experimental economist
- SP 20 classes: Behavioral Econ & Game Theory, R Data Visualization
- Starting a new center at GSPIA: Center for Analytics in Social Innovation (**CASI**) to bring analytical tools to public services.

Example of CASI projects:

- PittSmartLiving (NSF Transport project)
- Field experiment encouraging

  ex-inmates to use social services.

CPS: TTP Option: Medium:

Building a Smart City Economy and Information Ecosystem to Motivate Pro-Social Transportation Behavior

Team Members

University of Pittsburgh


Alexandros Labrinidis (PI)


Adam Lee


Yu-Ru Lin


Konstantinos Pelechrinis

School of Computing and Information


Kent Harries


Mark Magalotti

School of Engineering


Sera Linardi

GSPIA

TransitScreen


Matt Caywood

Partners

- Port Authority of Allegheny County (Bus Operator)
- Healthy Ride (Bike Share)
- City of Pittsburgh
- Oakland Business Improvement District
- Pittsburgh Downtown Partnership
- Envision Downtown
- Oakland Transportation Management Association
- Pittsburgh 2030 District
- Radius Networks
- UPMC

Project Goals

1. Design, develop, deploy, and evaluate a Cyber-Physical system that provides commuters with real-time information of arrival and utilization of all relevant options of public transit (e.g., bus, subway, shuttles, bikes, etc.)
2. Build a marketplace around **multimodal mobility**, where businesses can offer time-sensitive incentives connected to this transit information to nearby commuters (e.g., the next bus is too

First Results

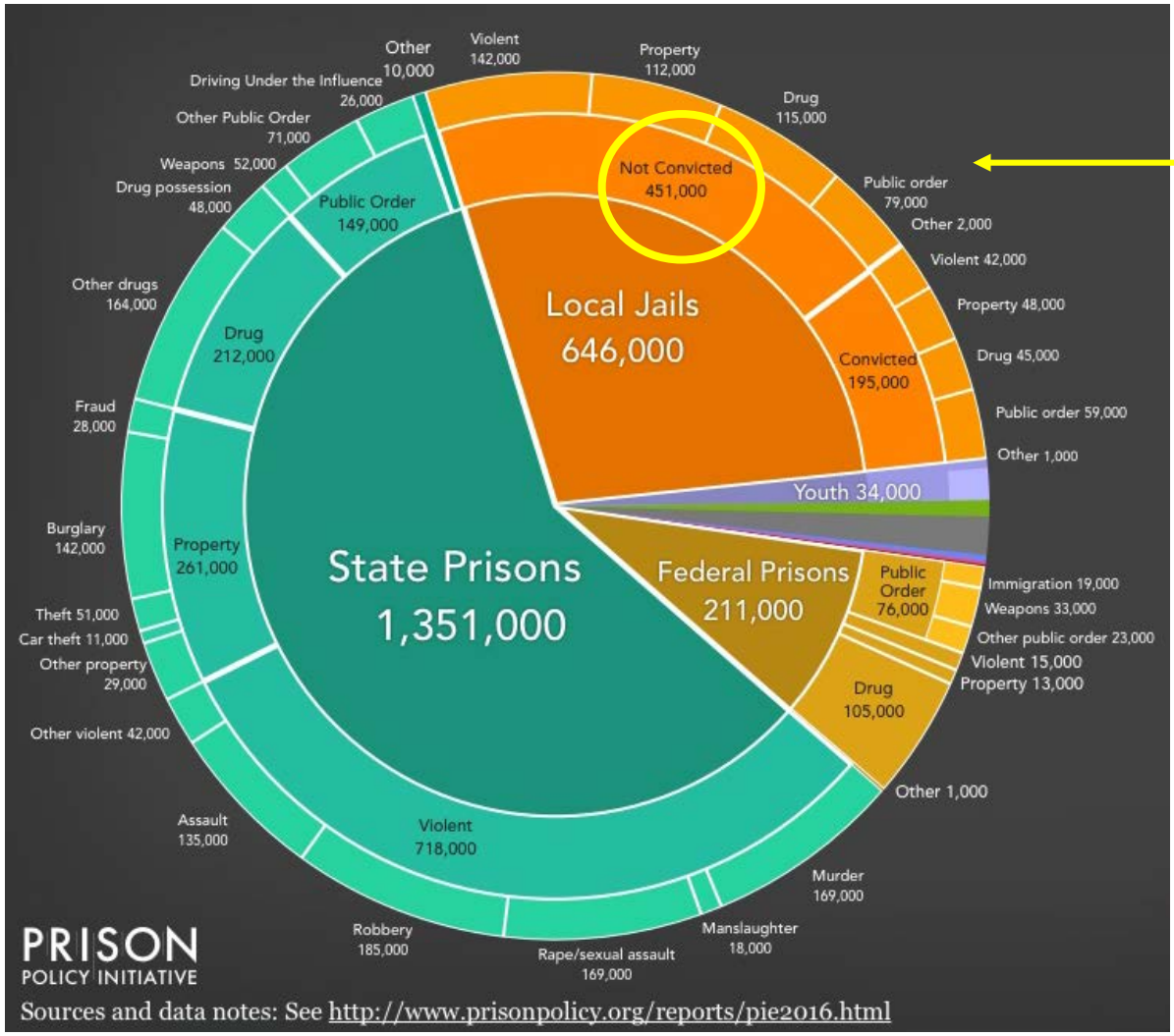

PittSmartLiving Display at City-County Building (Pittsburgh City Hall)


PittSmartLiving Display at Carnegie Library of Pittsburgh – Main (Oakland Branch)

https://www.youtube.com/watch?v=s90dp-8oA6o

Talk on Pitt Smart Living Human Behavior Lab

Allegheny County Jail, Pittsburgh, PA



Jail churn
Most people in jail has not been convicted.
The median stay was 10 days, and the mean stay was 58.3 days.

Jail | Rehabilitative Programs | All

The Changing Relationship Betw

Not secure | www.county.allegheny.pa.us/jail/rehabilitative-programs.aspx

REHABILITATIVE PROGRAMS

SEARCH

MY ALLEGHENY

CONTACT

CAREERS

SAVE PAGE

VIDEO

Type what you're looking for

» Visitor Information
» Visitation Schedules

» Inmate Phone System
» Contact

HOW DO I...
SEE MORE

Government » Legal and Public Safety » Jail » Rehabilitative Programs

Foundation of HOPE

The HOPE Pre-Release and Aftercare programs are inter-faith, faith-based, rehabilitative programs which work in collaboration with key community service providers and volunteers to empower incarcerated and released individuals to restore their relationship with their God, rebuild their lives, and reconcile to their community.

HOPE Pre-Release Program

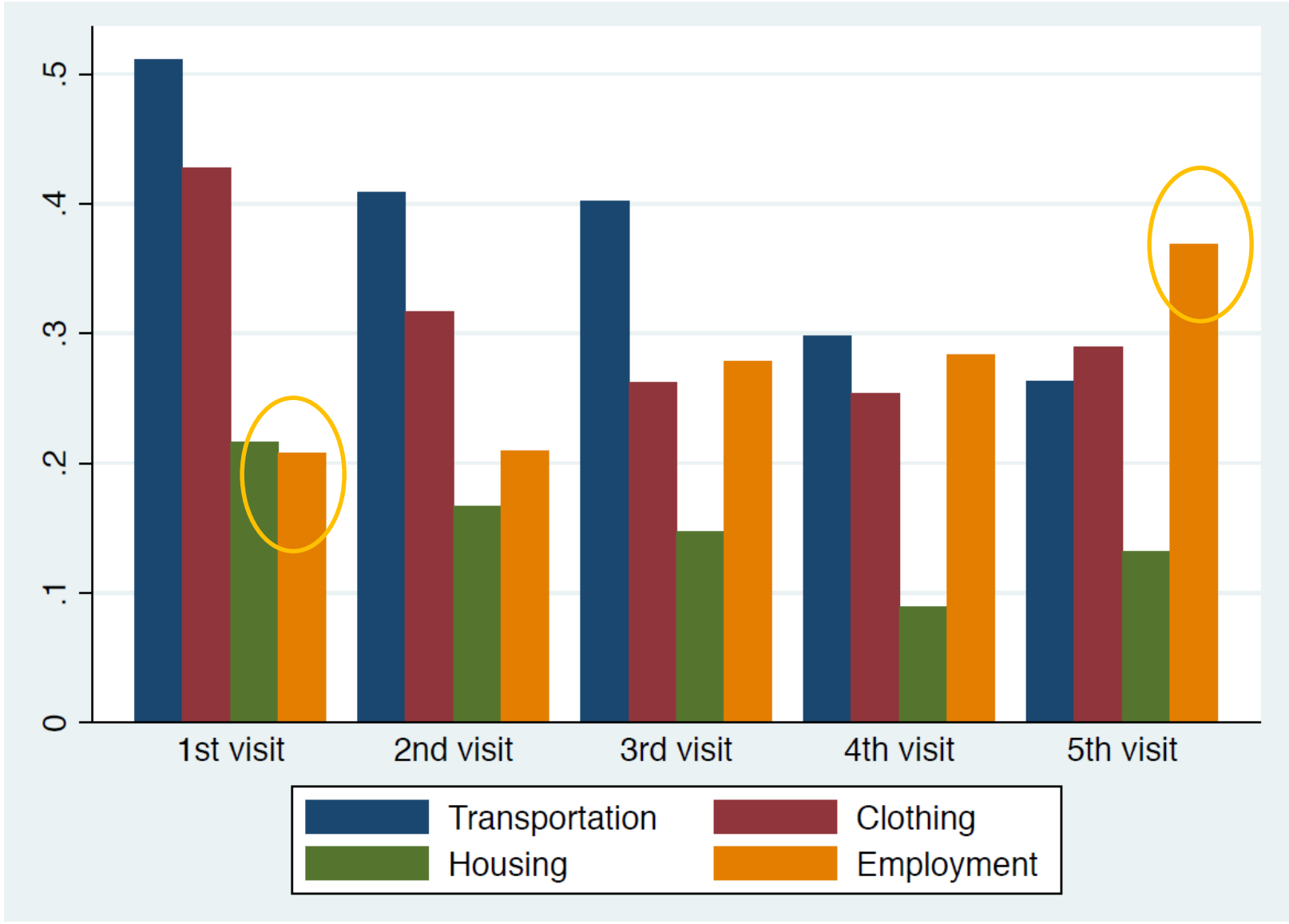
Over an eight week period, HOPE Pre-Release participants meet for over 120 hours of

JAIL

- Visitor Information
- Inmate Phone System
- Inmate Mail
- Inmate Funds
- Inmate Medical Supplies

Share


Foundation of
HOPE


0
.1
.2
.3
.4
.5
1st visit
2nd visit
3rd visit
4th visit
5th visit
Transportation
Clothing
Housing
Employment

Were you previously in jail?

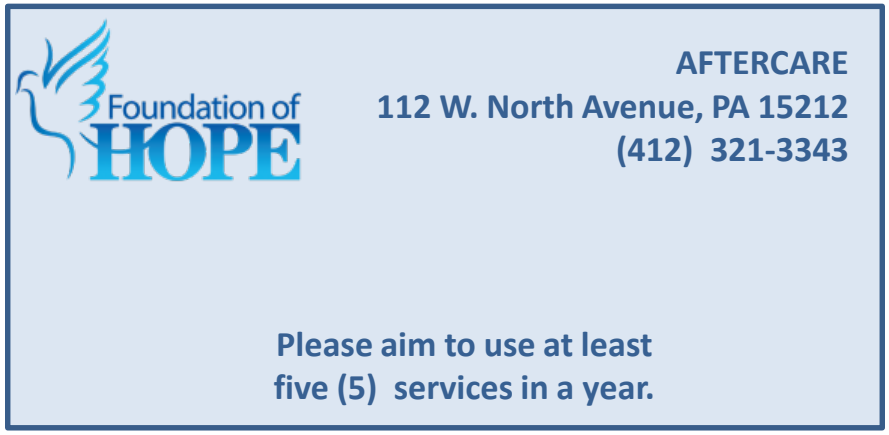
Get a $24 bus pass for taking a 10-minute survey

Call 412 321-3343 or stop by
Foundation of HOPE Aftercare
112 West North Avenue, Pittsburgh
during the Pitt Study Hours
to enroll in the study.

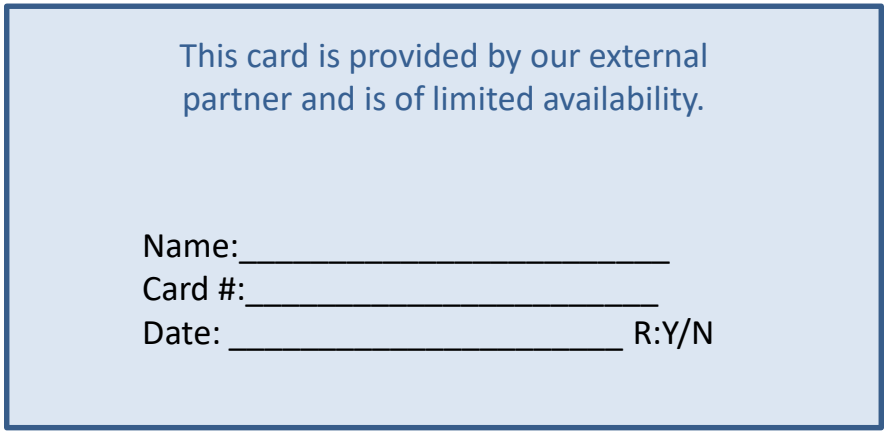
Pitt Study Hours

Tues 9-3:30 (staff: Vaib)
Wed 8:30-11 (staff: Vaib)
Thur 11:30-2 (staff: Bella)
Fri 9-3:30 (staff: Bella)

Punch card = No incentive

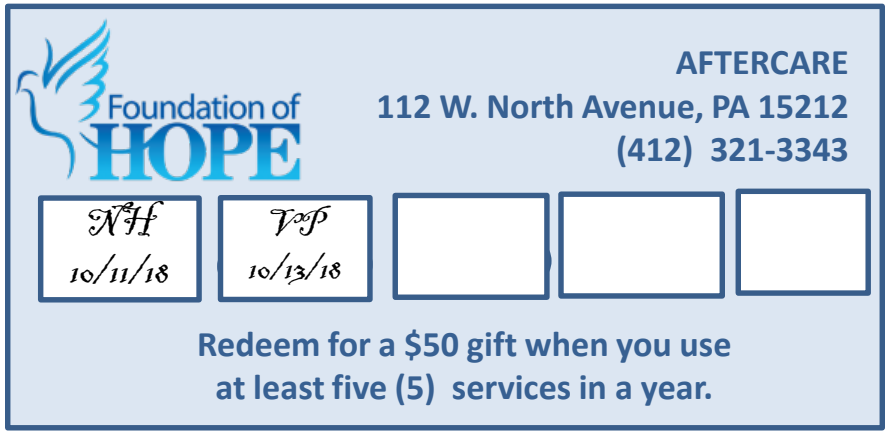


Front

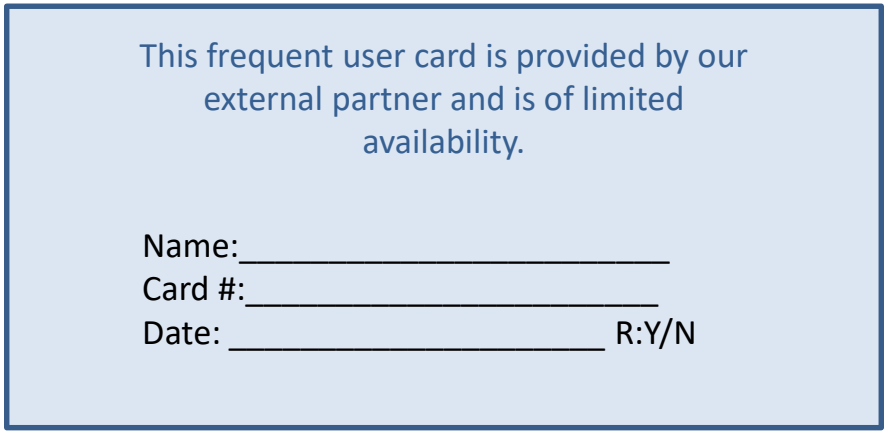


Back

Punch card: 3 visits to get incentive

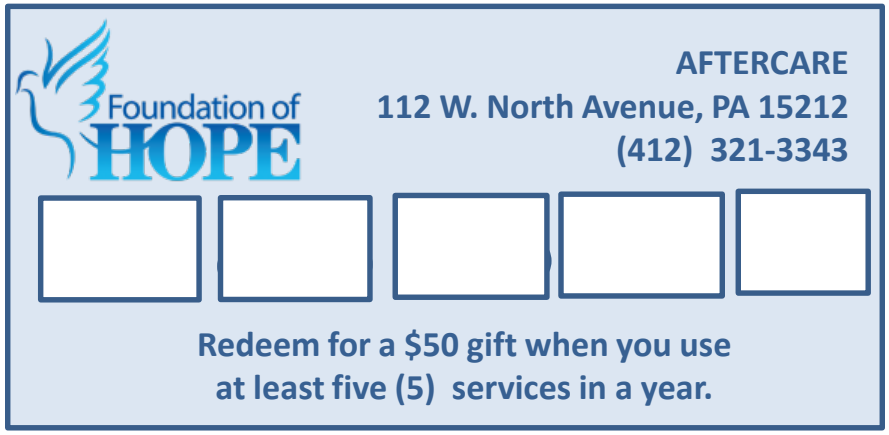


Front

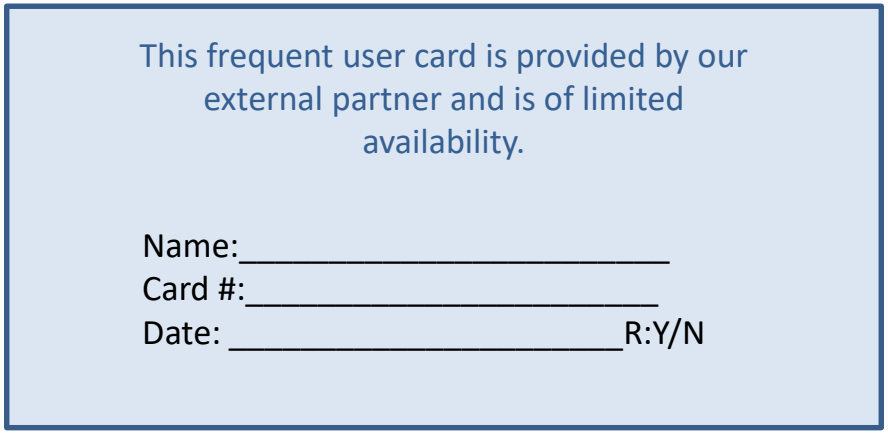


Back

Punch card: 5 visits to get incentive



Front



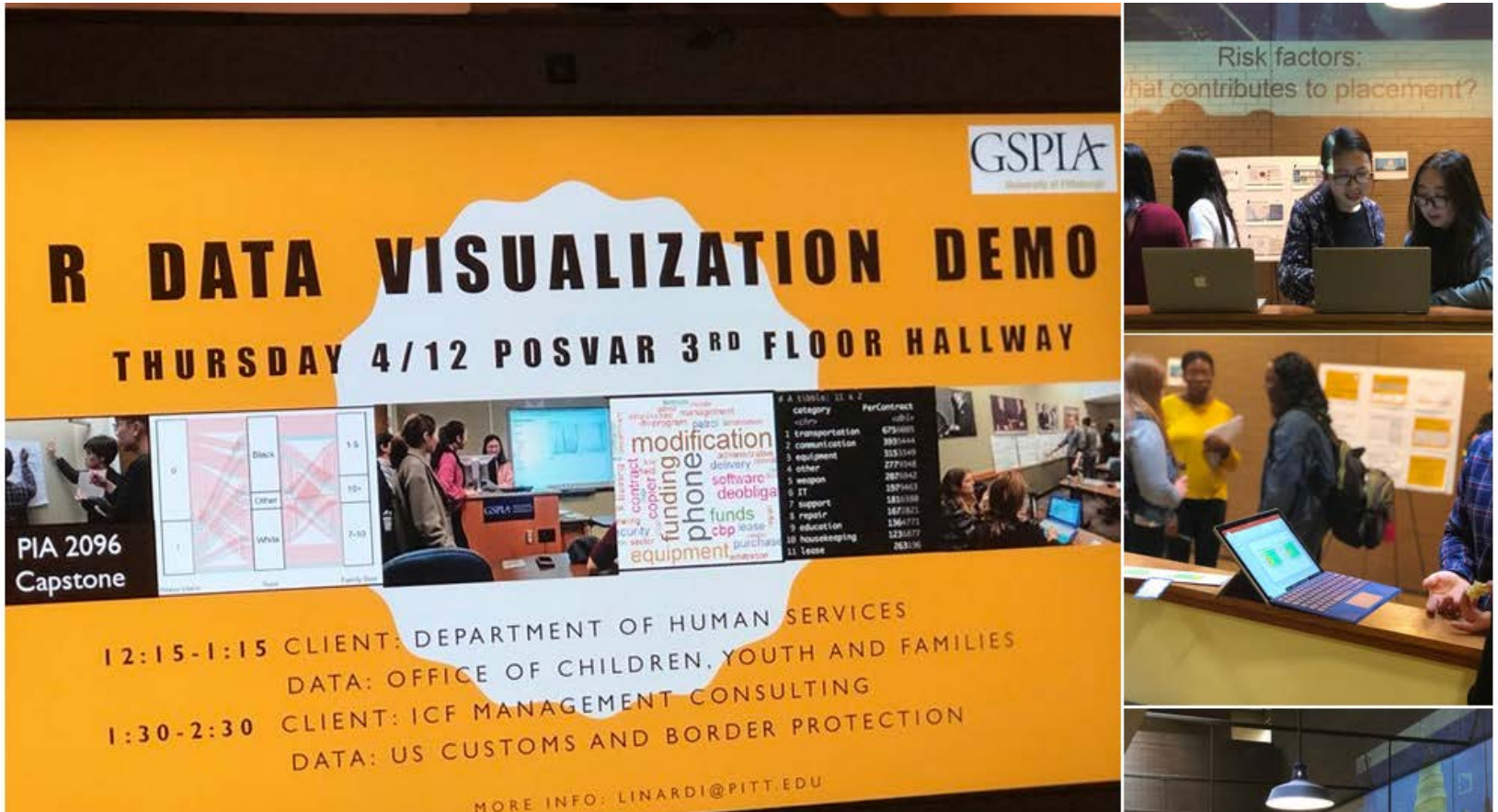
Back


A&M
UNIVERSITY OF
PITTSBURGH

Effect of incentive so far (ongoing)

Example of CASI activities:

- Amazing Analytics Race (tomorrow!)
- Reading group: Econ Meets CS (Tues 11-12), led by postdoc Jinyong Jeong. Will be offered as a class Spring 2020.
- R programming data visualization demo



What this workshop is and is NOT

What are we doing today? We are beginning your GSPIA journey with the end in mind: a career solving real world problems

First, let's define what this workshop will NOT do:

- Guarantee you an A in Quant I or Micro or any quant class
- Make you a math whiz
- Explain any mathematical concept in depth

<u>What this workshop aims to do:</u>

- Connect quant methods to the real world.
- Give you a preview of ALL the math you will see during your time here. You will most likely not encounter any math that you have not seen today.
- Provide a **quick-and-dirty**, hands-on experience of how quant methods give you an additional edge in tackling policy questions.

Schedule and people you will meet today

- 9:30-10:50 Intro, Lecture 1: Linear functions, Exercise 1
- 10:50-11:10 Meet your quant professors
- 11:10-11:20 Break
- 11:20-12:00 Lecture 2: Nonlinear functions and derivatives, Exercise 2
- 12:00-1:00 Lunch Break
- 1:00-1:15pm Amazing Analytics Race teams (TAs)
- 1:15-2:45 Lecture 3: Intro to Stats, Exercise 3
- 2:45-3:00pm Break
- 3:00pm-3:30pm Team exercise and alum Alex Heit

And.. what is GSPIA's Amazing Analytics Race ?

- At the end of today, you will be randomly split into pairs for tomorrow.
- Your mission will be explained tomorrow: you will have 3 hours to solve a puzzle by interlocking a series of 10 clues with your partner.
- You will use real world data, the quantitative methods you learn today, and lots of creativity.
- What's at stake: 1st place team = a $200 Bookstore gift certificate. 2nd place team = $100. 3rd place = $50.

- After teams are formed today, we will brief you on the rules of the race, and your team will get to practice working together.

Our Amazing Race Community Partner

WESTERN PENNSYLVANIA REGIONAL DATA CENTER

https://blogs.microsoft.com/newyork/2018/11/15/a

A data journey with Western Pennsylvania Regional Data Center

Nov 15, 2018 | Claire Suh, Microsoft Civic Tech Fellow in Pittsburgh

Claire Suh, Microsoft Civic Tech Fellow
and David Walker, WPRDC Data Scientist

How today's training camp works

- Data - Lecture (<1hr)– Exercise (10 mins) – Review the exercise (5-10 mins)

- You have the slides on your computer, so you can always go back / make notes, etc.

- Ask questions! There is no dumb question, this **is** a refresher workshop so forgetting basic stuff is totally okay. In completing exercise feel free to ask your neighbors/TAs/instructor for help.

- Please don't browse the internet/ phone for unrelated stuff. If you are waiting for others to finish, see if anyone near you needs help, or try new things with STATA.

Imagine you are an advisor to the mayor of Pittsburgh

- He is wondering whether or not to approve 10 new businesses on a strip of a crowded highway: businesses bring jobs but worsen congestion

- What you have to help you advise him:
  - Data on travel time on several highways given the number of cars on the highway (Cars.csv)
  - Data on number of cars given number of businesses along the highway (Business.csv)
  - Public opinion expert's estimated relationship between business development, traffic congestion and support for city government

Breaking down the question into mathematical concepts

1. how long does it take to travel the highway? (random variable)
2. how does the # (*number*) of cars affect travel time? (correlation, linear regression, slope)
3. can adoption of a different traffic system reduce congestion? (simultaneous equations)
4. how does the # of businesses affect # of cars?(nonlinear equations)
5. what is the optimal # of business to have? (optimization)

1. random variable

How long does it take to travel through the highway?

Random variable

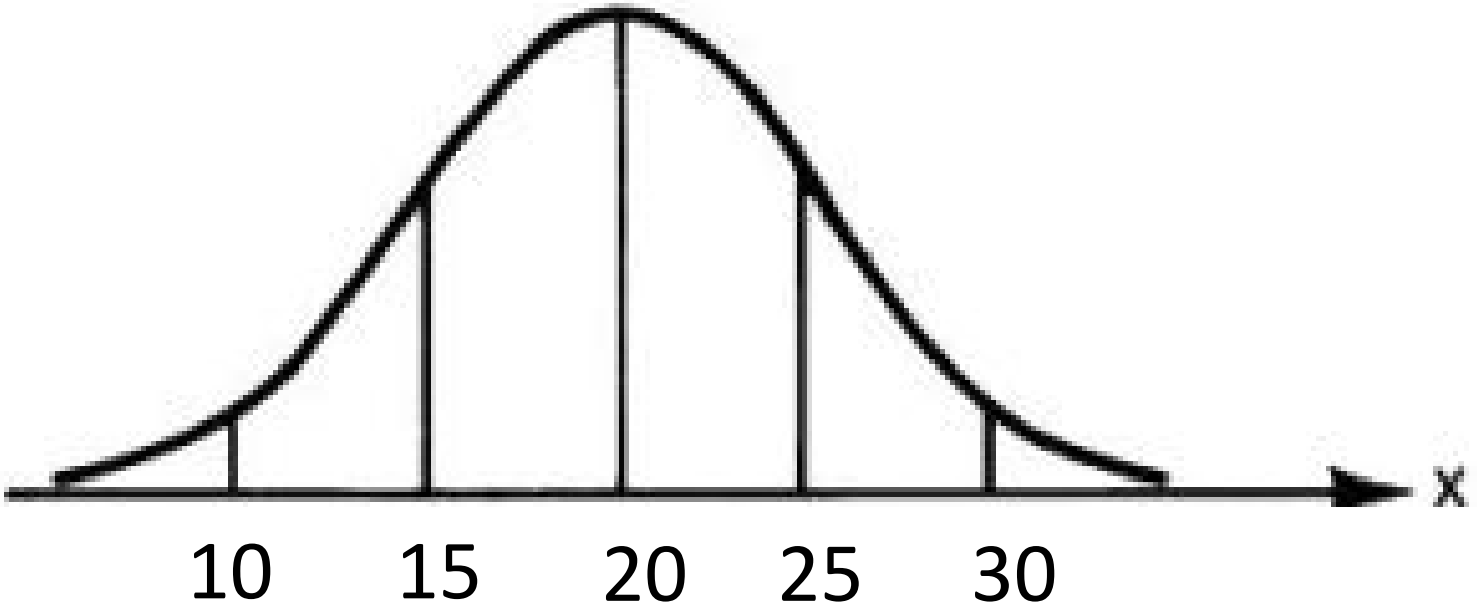
How long does it take to travel 20 miles on a city highway at 8am in the morning? Hands = 20 mins, 30 mins, 40 mins

- Different day, same highway, same hour in day = different travel time.
- Statistics is learning to get the information out of this uncertainty.
- 'Time needed to travel' is a random variable = the value is subject to variation due to chance.

- Is what is written on this board ALL the possible travel times for 20 miles? No. That would be the *population*. This is a *sample*. We usually only observe a sample of realizations of the random variable of interest.

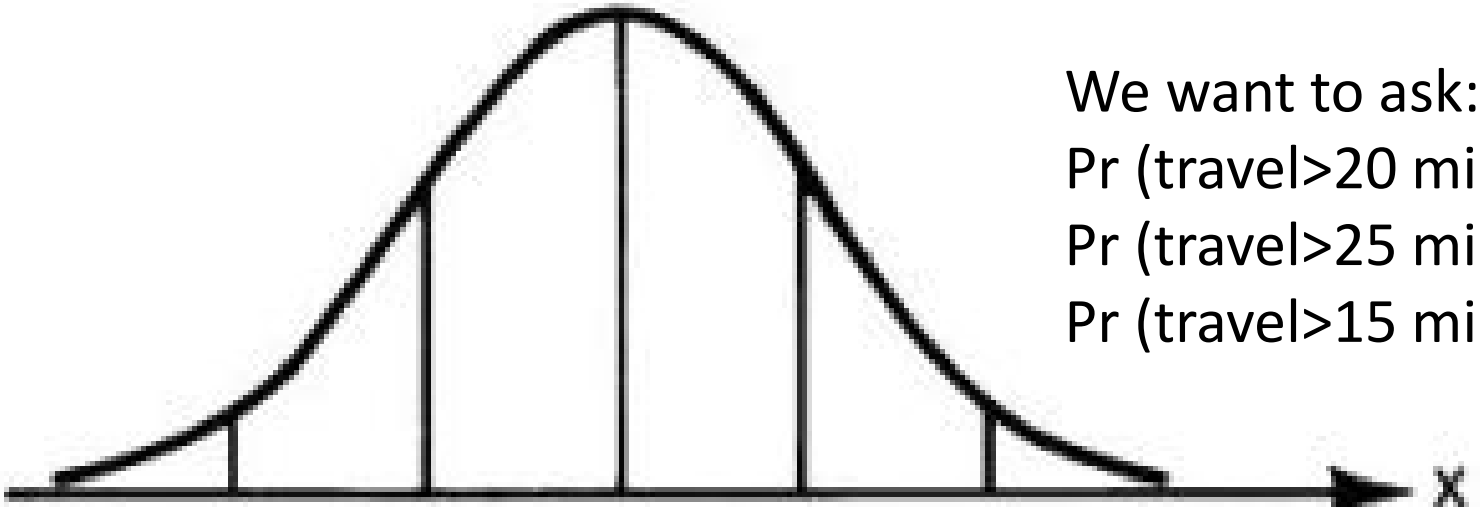
Distribution: what the population looks like

Suppose the distribution of travel time looks like this:



Distribution: what the population looks like

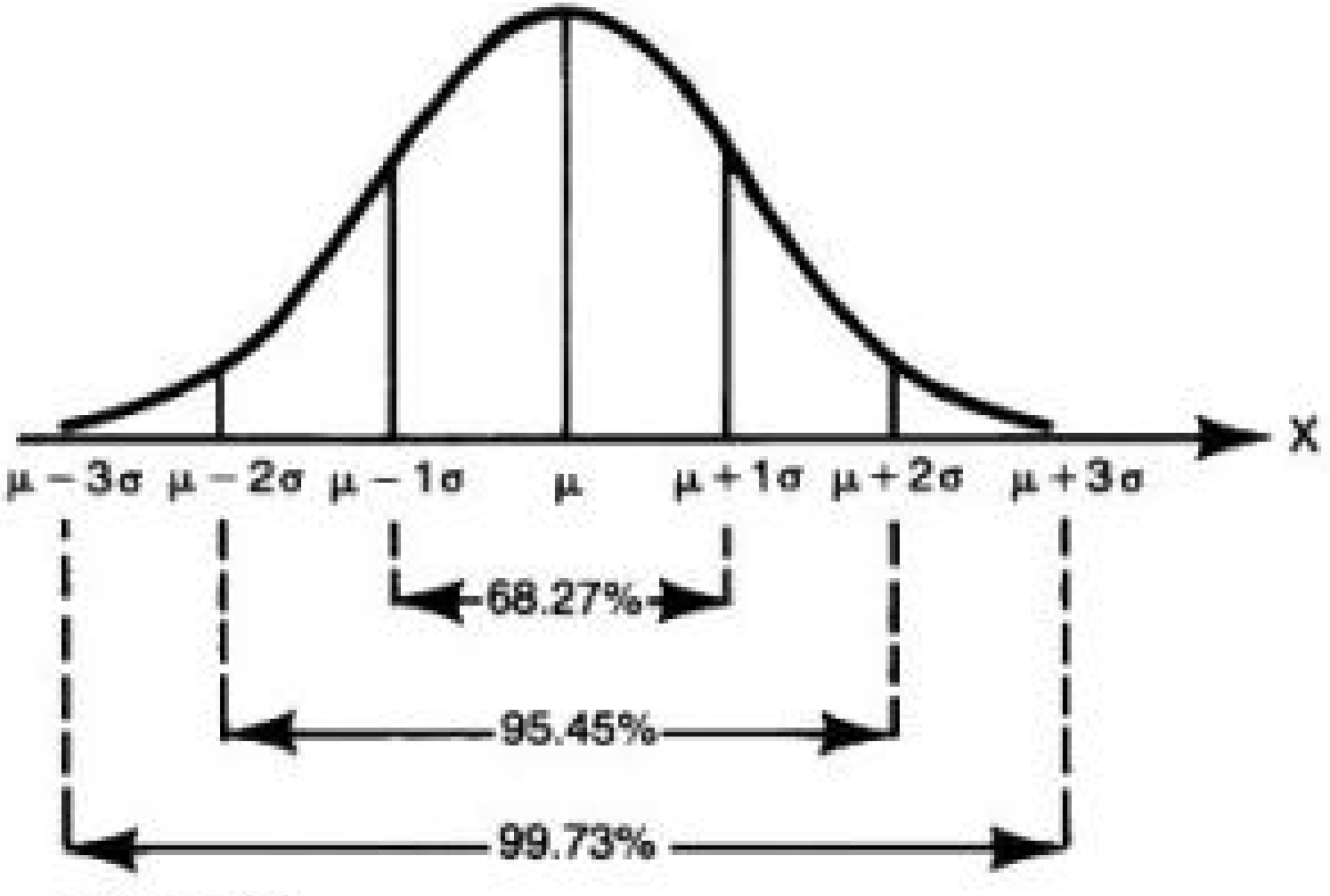
Then we can say that travel time is distributed normally with a mean 20 minutes and a std deviation of 5 minutes. (or a variance of $5^2=25$)



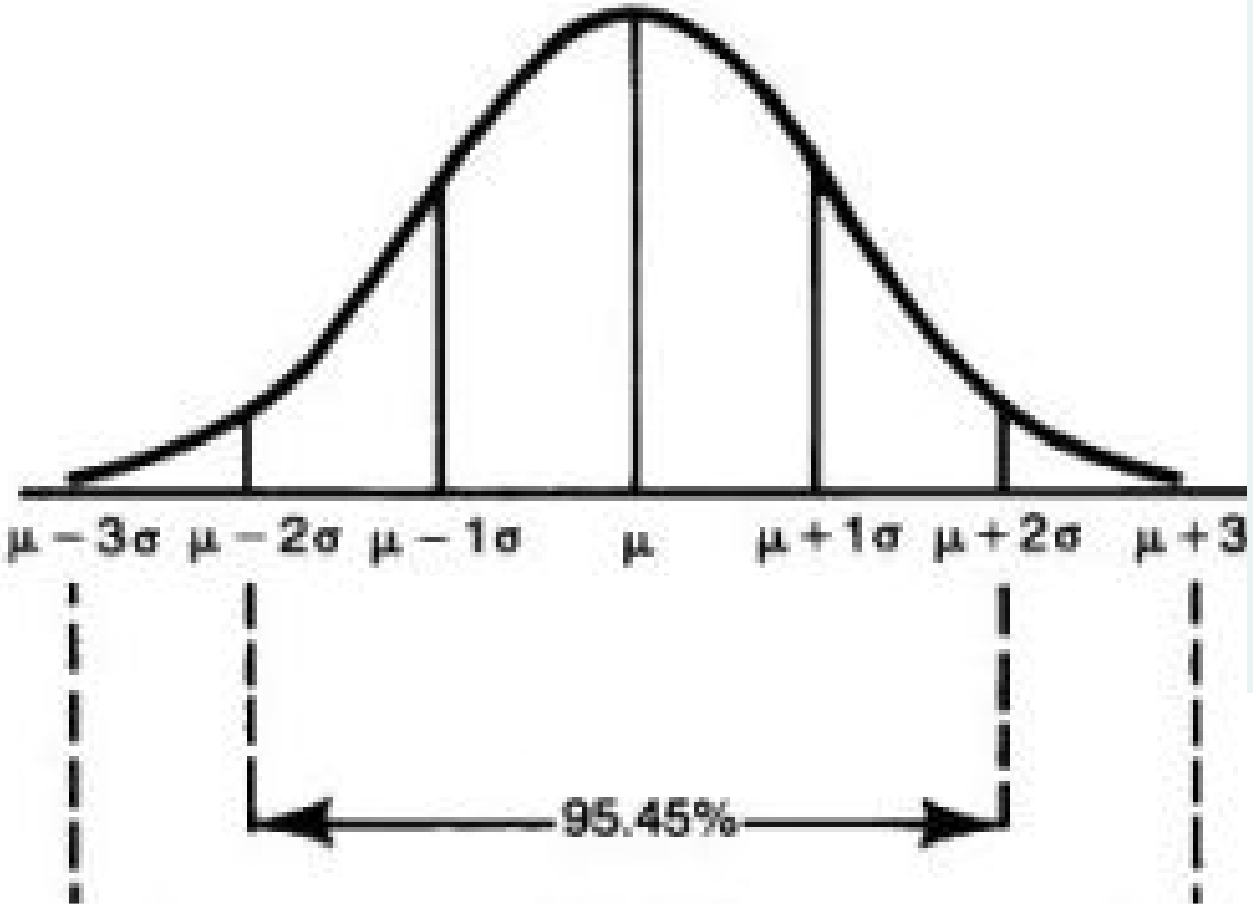
We want to ask:
Pr (travel>20 min) =?
Pr (travel>25 min) =?
Pr (travel>15 min) =?

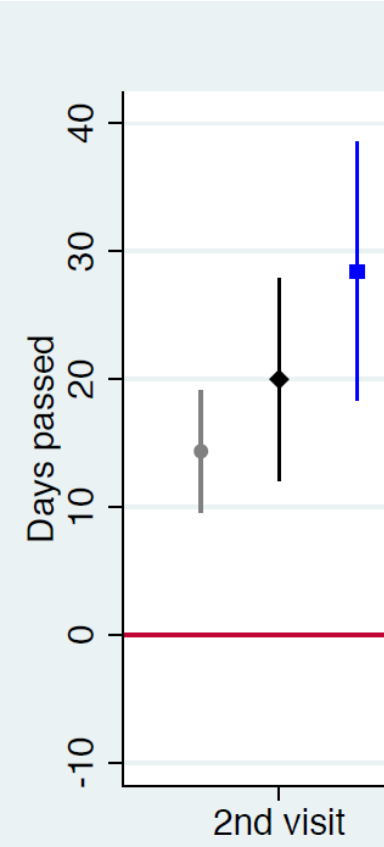
Let's look at the normal distribution more generally:

Earlier we drew this
for $\mu=20$, $\sigma=5$



Recall this? 95% confidence interval

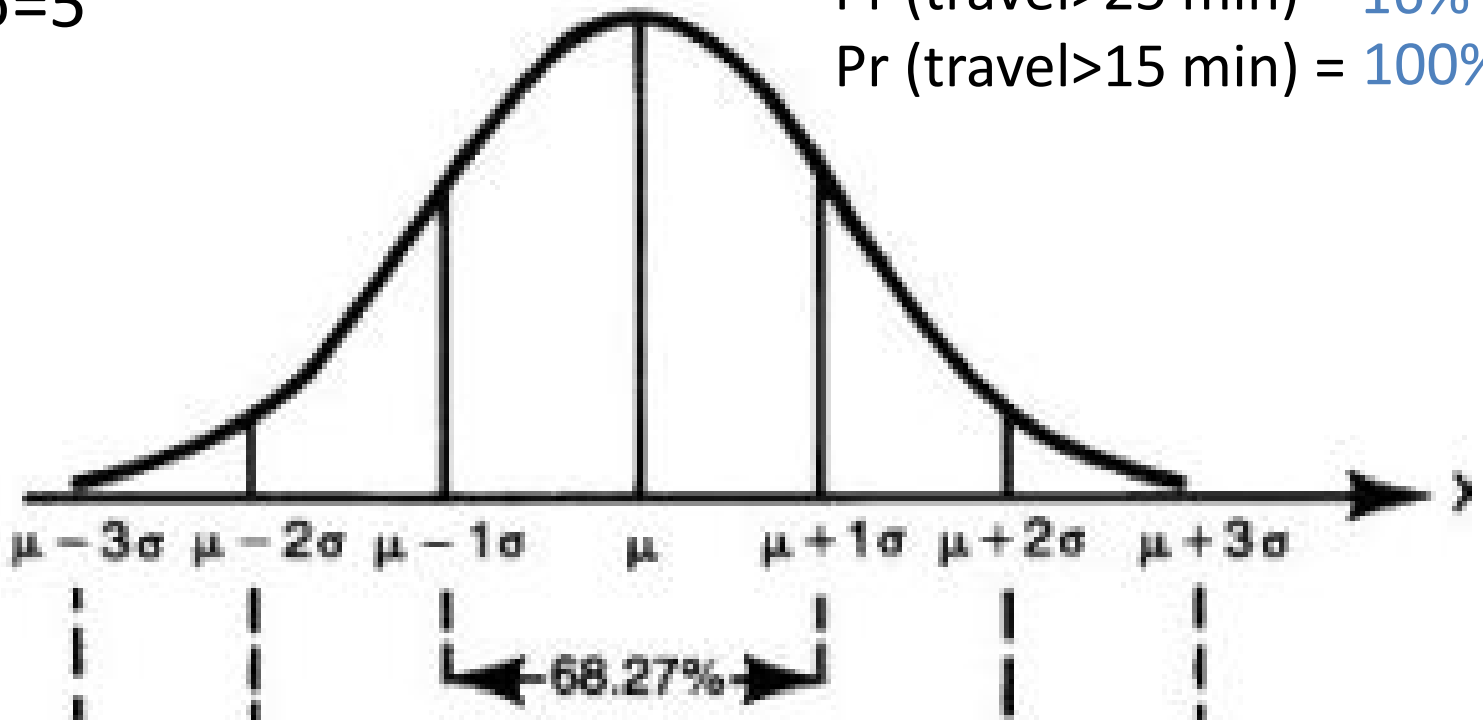




Let's look at the normal distribution more generally:

Travel time:
μ=20, σ=5

Pr (travel>20 min) = 50%
Pr (travel>25 min) = 16%
Pr (travel>15 min) = 100%-16%



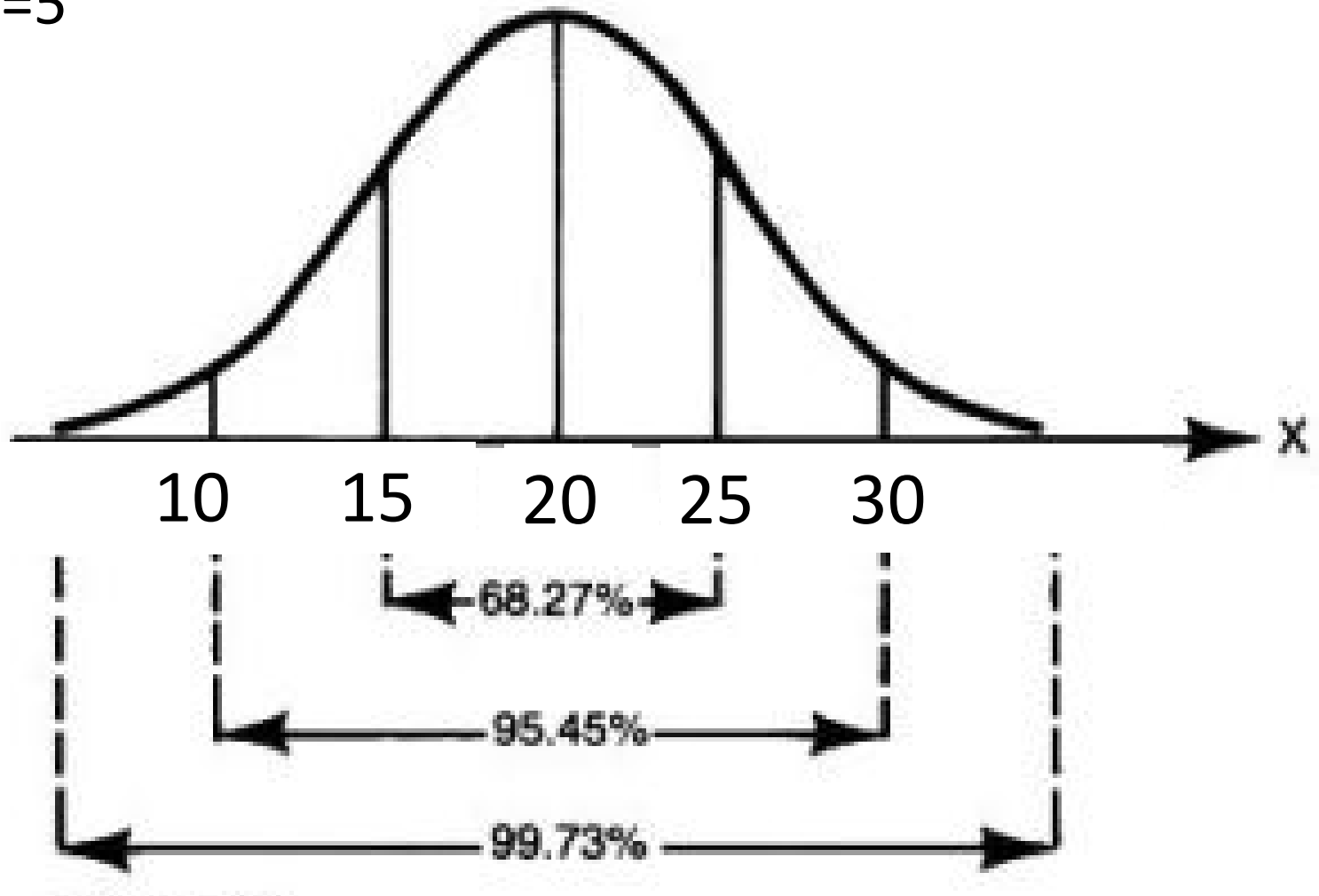
16%

(100%-68%) /2 = 16%

Your turn

Travel time:
μ=20, σ=5

Pr (travel>30 min) = ?



Joint probabilities

Probability that John & Beth (who travelled separately) are both late when

- they each gave themselves 20 minutes?

- John left 20 minutes ago while Beth left 25 minutes ago?

SIDEBAR: joint vs conditional probability

JOINT: Tossing a coin and a dice what is the probability you get a H and an even number?

CONDITIONAL : You toss a dice and got an even number. What is the probability that the number is < 4?

Looking at data in STATA

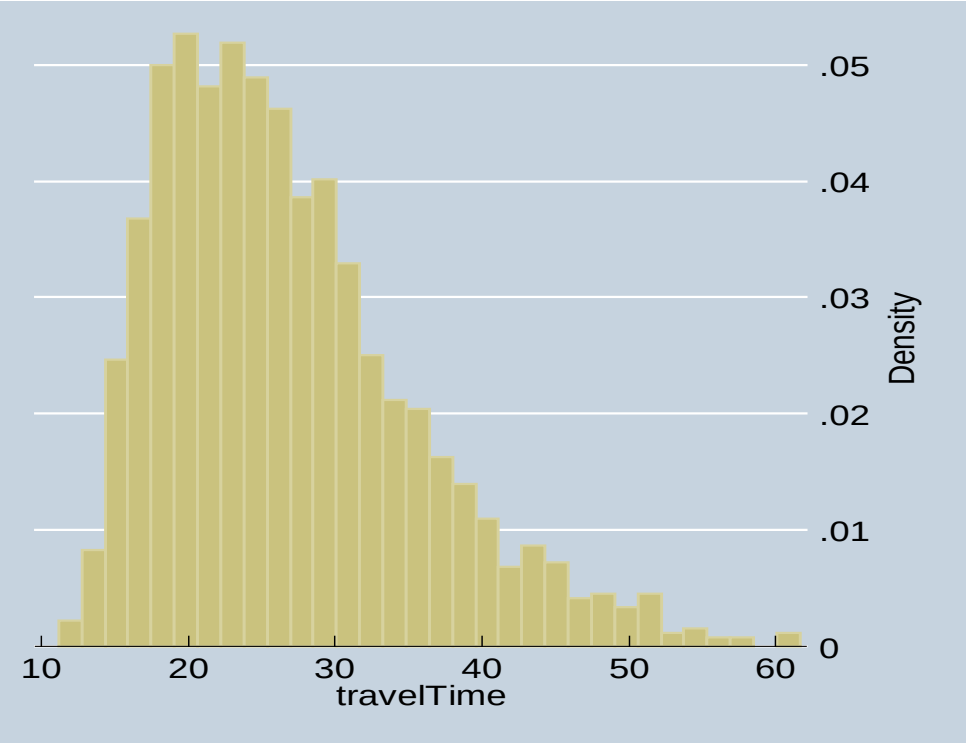
- Suppose cars.csv contains a random sample of travel time and # of cars on Pittsburgh highways.

- Let's load it with Data Editor. Open cars.csv in Excel. Highlight, copy. Open data editor. Click on first cell and paste. Treat first row as variable name.

- Again: all just quick and dirty today!

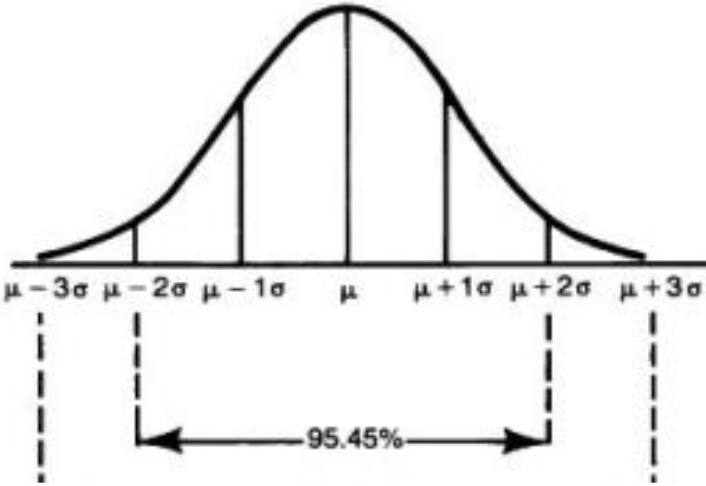
Travel time

Histogram



hist traveltime (not normal, but we'll treat it as such today)

Graphics → Histogram ->Variable: traveltime




```
. mean traveltime

Mean estimation                   Number of obs    =      1674
```

	Mean	Std. Err.	[95% Conf.	Interval]
traveltime	26.71808	.2103392	26.30553	27.13064

According to this sample, the average time needed to travel 20 miles on a Pittsburgh highway is 26.7 minutes.

```
. mean traveltime

Mean estimation                   Number of obs    =      1674
```

	Mean	Std. Err.	[95% Conf.	Interval]
traveltime	26.71808	.2103392	26.30553	27.13064

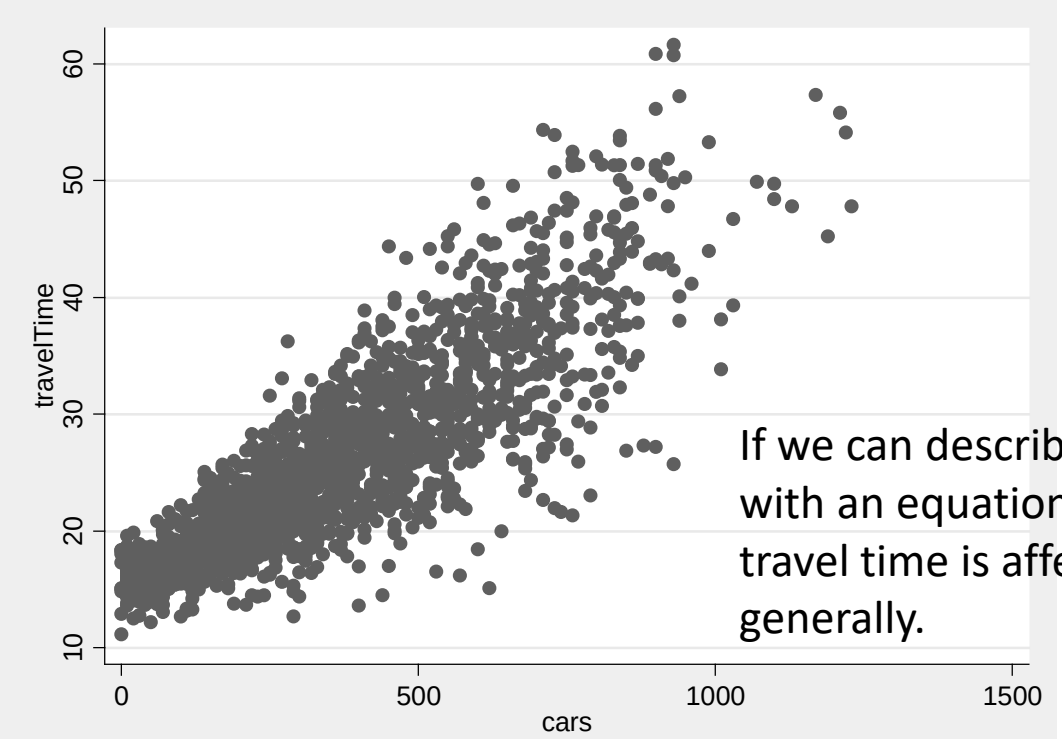
The standard error of the mean travel time is 0.21 minutes. This means when we take random sample of 1674 car trips down this highway, the average travel time will fluctuate by 0.21.

2. correlation, linear regression, slope / rate / derivative

how does the # of cars affect travel time?

Relationship between two random variables

correlation between travel time and # of cars

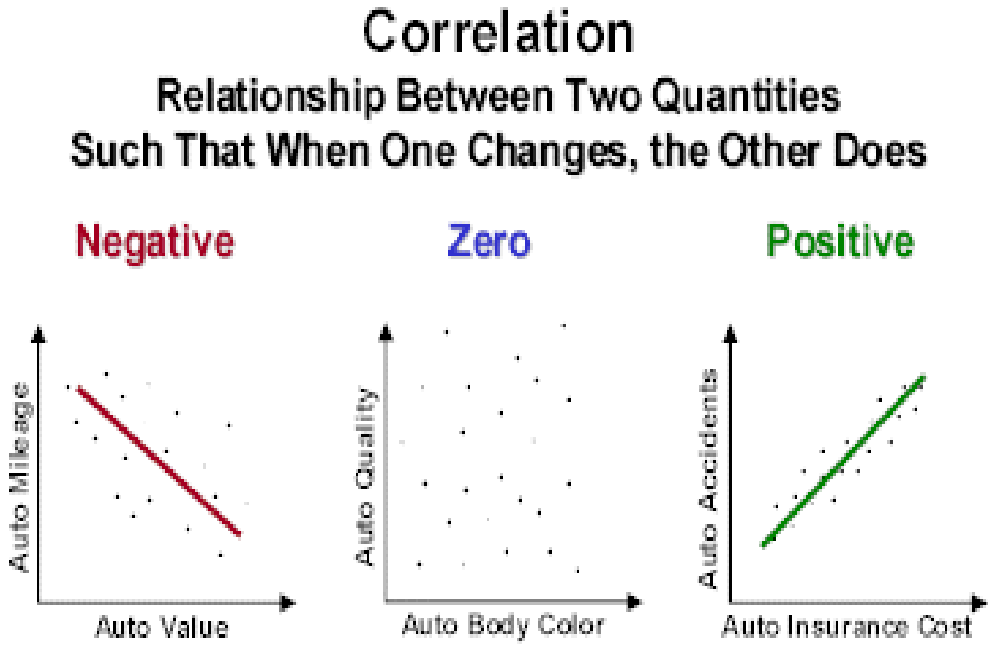


If we can describe this relationship with an equation, we can tell how travel time is affected by cars more generally.

scatter traveltime cars

Graphics → Twoway ->Create->Y variable: traveltime, X variable: cars

Scatterplot shows correlation between two variables.



To find the relationship, we can try to fit a line across this scatterplot that is the closest possible to ALL the points. This is a regression line.

Regression

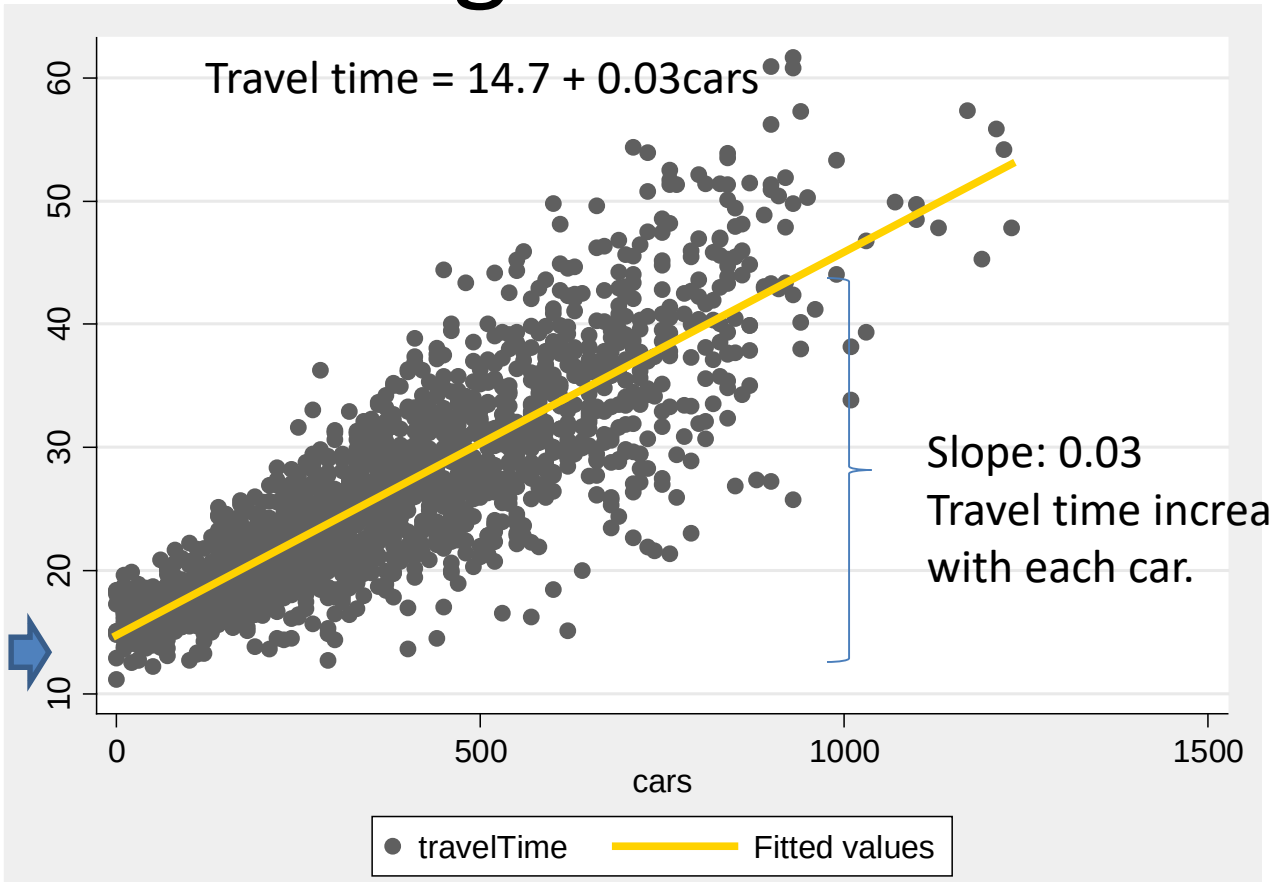
reg traveltime cars

traveltime	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Int]
cars	.031	.000489	63.67	0.000	.03019	.0321
_cons	14.714	.220157	66.83	0.000	14.28	15.146

traveltime = 14.7 + 0.03cars

What does it mean?

Drawing a linear function



With an increase of 1000 cars, travel time increases by 1000*0.03 = 30 minutes. So with a thousand cars on the highway, total travel time is 14.7+30 = 44.7 minutes

Linear Functions

With linear functions, an increase in X always increases Y by the same amount. For example, one additional car increase travel time by 0.03 minutes, regardless of whether there's 100 or 1000 cars on the freeway.

Hint: marginal analysis useful in Cost-Benefit Analysis and in Micro. The regression finds that the marginal impact of 1 car on congestion is 0.03 minutes in travel time.

Inverting a linear function

- traveltime = 14.7 + 0.03*cars
- If it takes you 20 minutes to travel, how many cars are on a freeway?

Hint: useful in microeconomics in inverting demand curves

Inverting a linear function

You know travel time as a function of cars
traveltime = 14.7 + 0.03*cars

You want cars as a function of travel time:
Traveltime- 14.7 = 0.03*cars
Cars = (Traveltime- 14.7) / 0.03
Cars = Traveltime/0.03- 14.7/0.03
Cars = 33.3*Traveltime - 490

Now, it's easier to answer this question:
If it takes you 20 minutes to travel, how many cars are on a freeway?
Cars = 33.3*20 - 490 =176

(BTW: what is the intercept and slope of this inverted function?
Intercept = -490Slope 33.3)

Confidence interval in regressions

```
reg traveltime cars
------------------------------------------------------------------------------
  traveltime |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Int]
-------------+----------------------------------------------------------------
       cars |        .031   .000489    63.67   0.000      .03019    .0321
      _cons |      14.714   .220157    66.83   0.000       14.28    15.146
------------------------------------------------------------------------------
```

The confidence interval tells us that we can be 95% confident that every car increases travel time by between 0.03 or 0.032 minutes.

P values in regressions

reg traveltime cars

traveltime	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Int]
cars	.031	.000489	63.67	0.000	.03019	.0321
_cons	14.714	.220157	66.83	0.000	14.28	15.146

The p-value tells us that the probability of finding a coefficient of 0.03 in this data when there is actually no relationship between travel time and number of cars is 0.000

Contrast: p value when there is no correlation

Relationship between the ID # of public works official recording the data and travel time.

traveltime	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ID	-.0524	.043	-1.21	0.227	-.1375	.0327
_cons	27.16	.420	64.66	0.000	26.3	27.9

The probability of finding a coefficient of -0.05 in this data when there is actually no relationship between travel time and the ID of the person recording the data is 22.7%

SIDEBAR:

By convention, cutoffs p-value is noted with *:

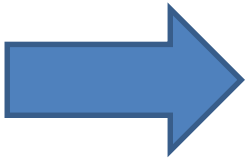
Table 1: Regression table

	(1) Price	(2) Price
Weight (lbs.)	1.747**	3.465***
	(2.72)	(5.49)
Mileage (mpg)	−49.51	21.85
		(0.29)
		3673.1***
		(5.37)
Constant	1946.1	−5853.7
	(0.54)	(−1.73)
Observations	74	74

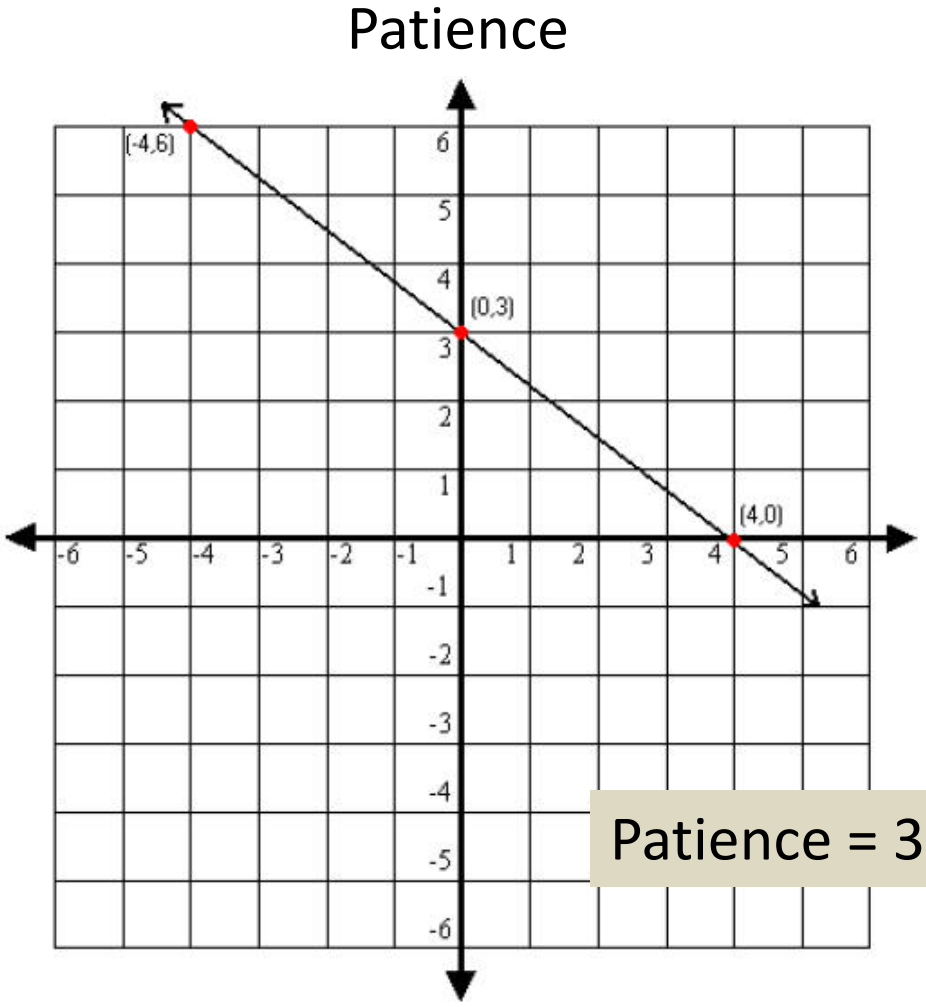
t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Hint: useful in any class where you need to read papers with empirical results.

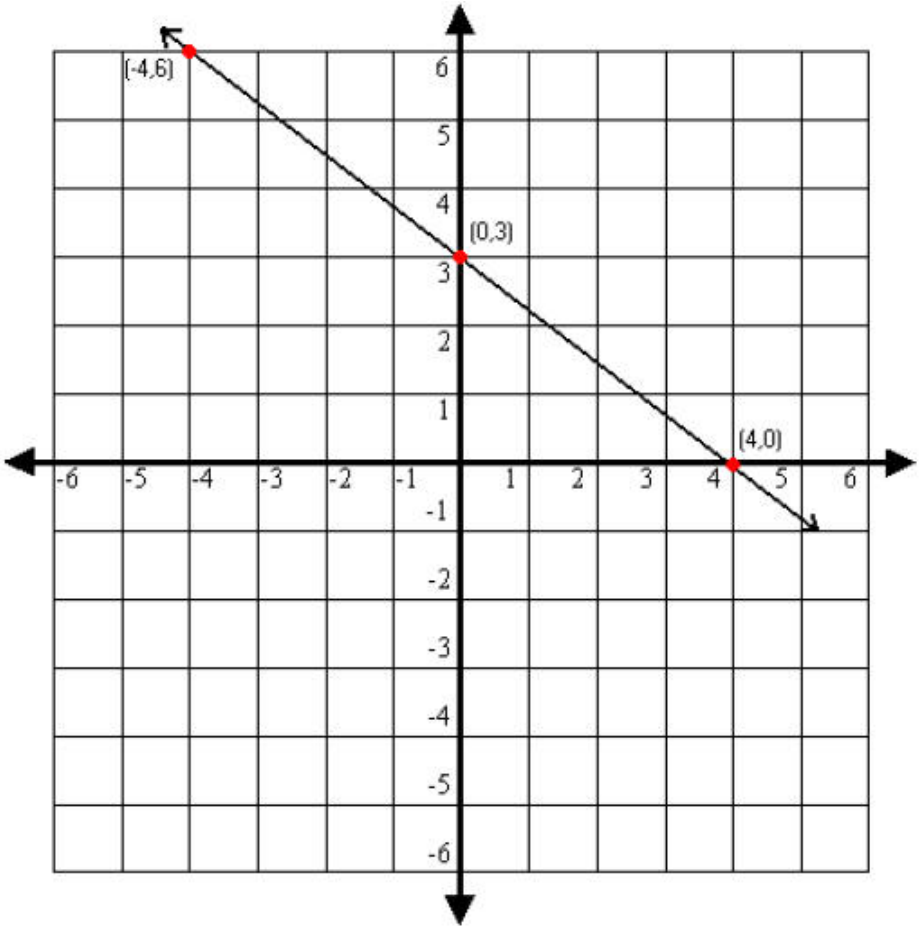


Looking at a graph and identifying the linear equation: y=a+bx



Patience = 3 – 3Traffic jam/4

Looking at a graph and identifying the linear equation



- Steps:

- Linear equations take the form of y=a+bx. So:
- Step 1: Identify the vertical intercept (0,3) a=3
- Step 2: Identify the horizontal intercept (4,0)
- Step 3: calculate the slope
- increase in y/increase in x

  b = -3/4

  (or rise over run)
- So, function is y=3-3x/4

How many additional businesses should be allowed along a busy highway to maximize citizens satisfaction?

Breaking down the question into mathematical concepts

1. How long does it take to travel the highway? (random variable) On average 26.7 minutes.
2. How does the # of cars affect travel time? (correlation, linear regression, slope) Travel time = 14.7+0.03 cars
3. Can adoption of a different traffic system reduce congestion? (simultaneous equations)
4. How does the # of businesses affect # of cars?(nonlinear equations)
5. What is the optimal # of business to have? (optimization, derivatives, chain rule)

3. comparing two highways: should you adopt another traffic system?

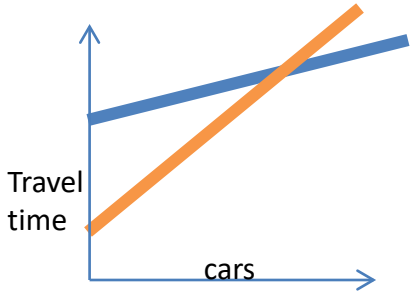
(simultaneous equations, or, systems of equations)

- Previously you learned that for Pittsburgh highways, traveltime = 14.7 + 0.03*cars.

- A colleague suggested that in anticipation of congestion from the new businesses, you should consider a traffic system that has been adopted by Cleveland to reduce travelling time. There, traveltime = 8.7 + 0.05 cars.

- Should you do that? What is the maximum # of cars such that travelling with the Cleveland system is faster than the Pittsburgh system?

- **Pittsburgh:** Traveltime = 14.7 + 0.03 cars
- **Cleveland :** Traveltime = 8.7 + 0.05 cars
- The question asks for what is cars such that traveltime is equal to each other.



Traveltime = 8.7 + 0.05 cars

14.7 + 0.03 cars = 8.7 + 0.05 cars

6 = 0.02 cars.

Cars = 300

- What's the average # of cars in the Pittsburgh highway of interest?
- Which system is better?

Review Exercise 1

- Questions?

Schedule and people you will meet today

- 9:30-10:50 Intro, Lecture 1: Linear functions, Exercise 1
- 10:50-11:10 Meet your quant professors
- 11:10-11:20 Break
- 11:20-12:00 Lecture 2: Nonlinear functions and derivatives, Exercise 2
- 12:00-1:00 Lunch Break
- 1:00-1:15pm Amazing Analytics Race teams (TAs)
- 1:15-2:45 Lecture 3: Intro to Stats, Exercise 3
- 2:45-3:00pm Break
- 3:00pm-3:30pm Team exercise and alum Alex Heit

How many additional businesses should be allowed along a busy highway to maximize citizens satisfaction?

Breaking down the question into mathematical concepts

1. How long does it take to travel the highway? (random variable) On average 26.7 minutes.
2. How does the # of cars affect travel time? (correlation, linear regression, slope) Travel time = 14.7+0.03 cars
3. Can adoption of a different traffic system reduce congestion? (simultaneous equations) No.
4. How does the # of businesses affect # of cars?(nonlinear equations)
5. What is the optimal # of business to have? (optimization, derivatives)

4. Nonlinear function

We will now use our other data set, “business.csv”

This data set has the # of businesses on a highway and the # of commuter cars associated with these businesses.

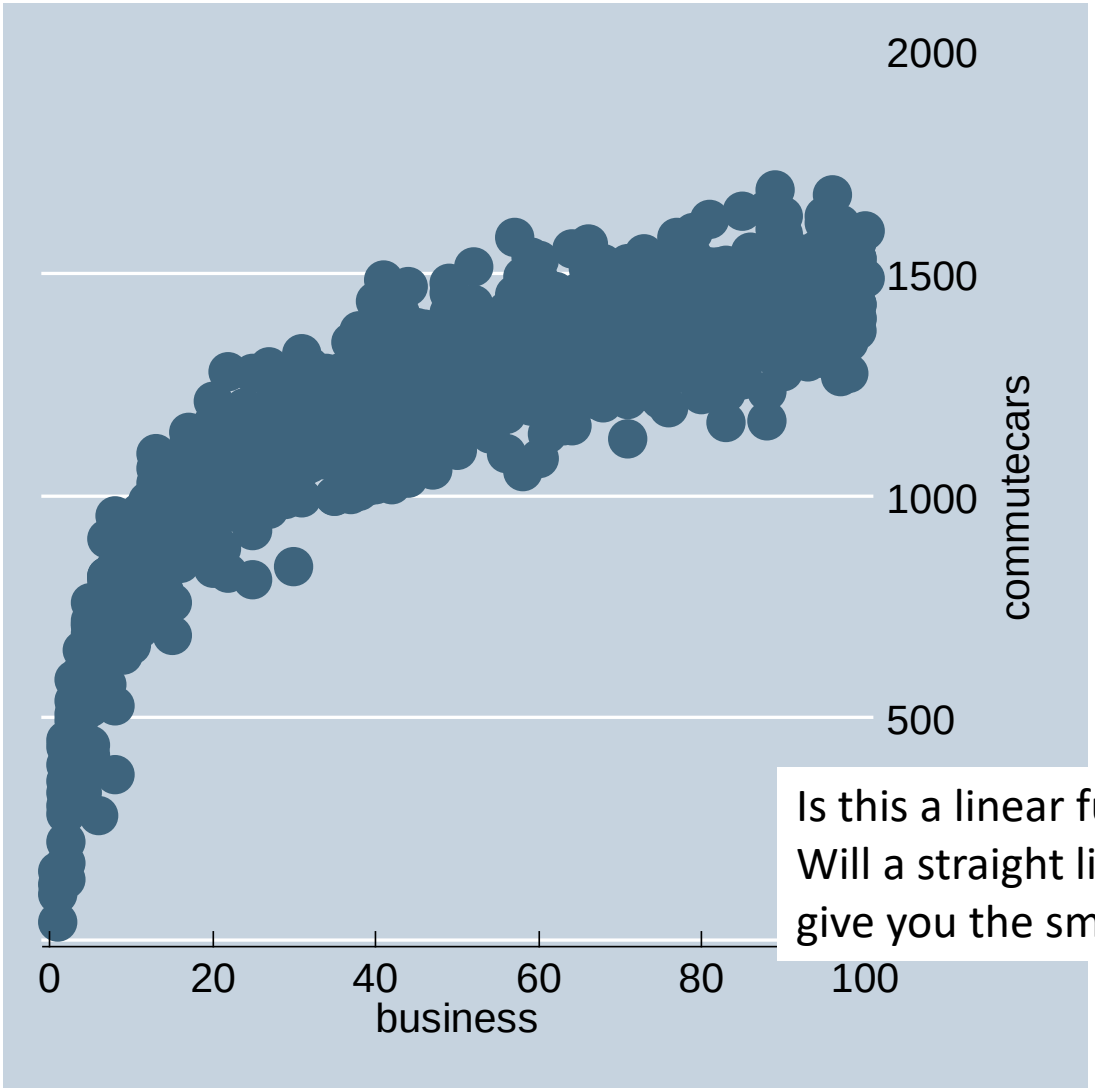
clear (you must clear out the old data)

Load new Business.csv

Look in data editor

What relationship are we trying to figure out?

scatter commutecars business



Is this a linear function?
Will a straight line
give you the smallest error?

Nonlinear functions

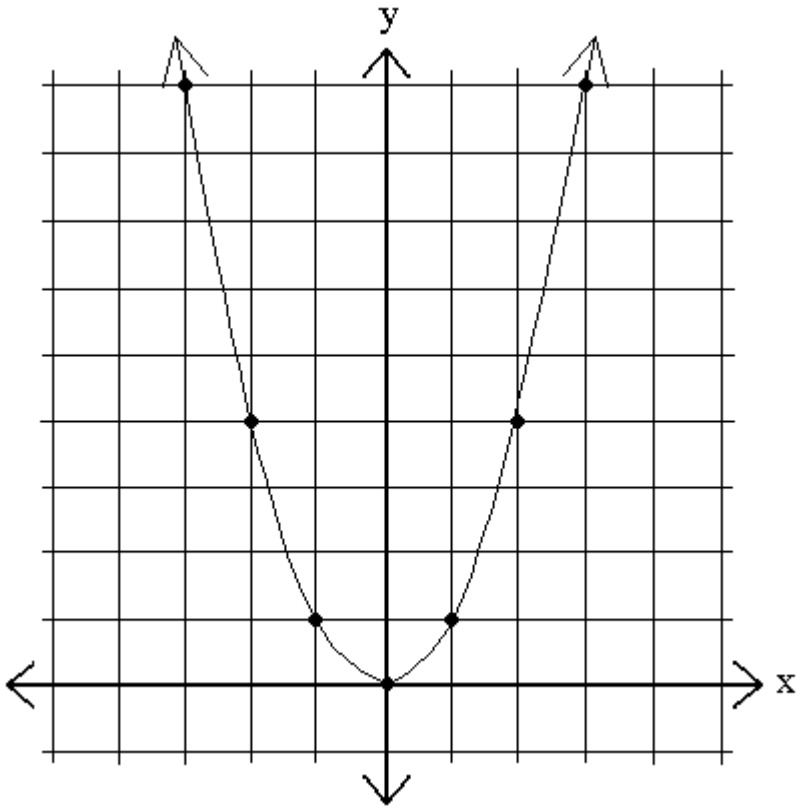
Let’s find what our function resembles:

- Quadratic function
- Logarithmic function
- Exponential function

Quadratic function

$y=x^2$

x	*Y*
-3	9
-2	4
-1	1
0	0
1	1
2	4
3	9

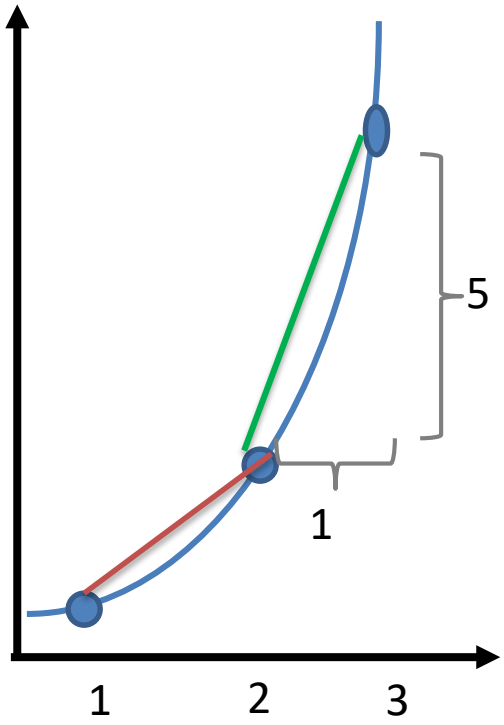


Notice how y changes as x changes.
The slope is no longer the same (“not a constant”) as we travel through the x axis: increasing x by 1 changes y by -5 at x=-3, by 1 at x=0, and by 3 if x=1

Slopes and derivatives

$y=x^2$

x	Y
-3	9
-2	4
-1	1
0	0
1	1
2	4
3	9

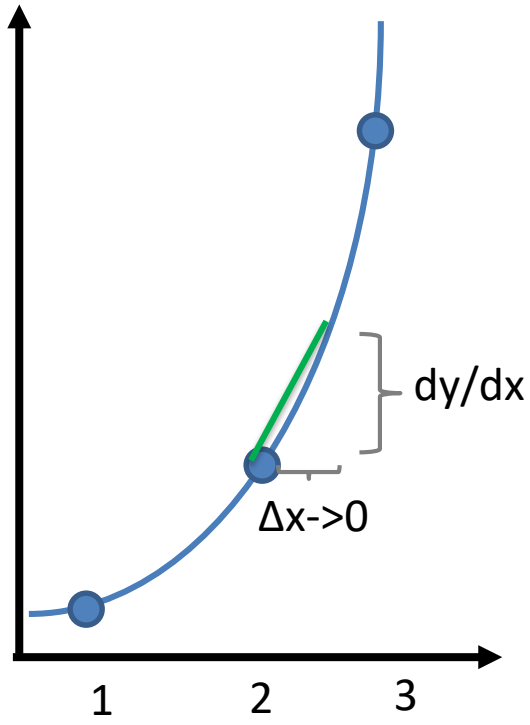


As we discussed earlier, slope is the increase in y / increase in x.
So we can find an average slope between two points.
Average slope as x goes from 2 to 3 is (9-4)/(3-2) = 5
Average slope as x goes from 1 to 2 is (4-1)/(2-1) = 3
But how do we find the slope at the single point (x=2)? There's nothing to measure!

Slopes and derivatives

$y=x^2$

x	Y
-3	9
-2	4
-1	1
0	0
1	1
2	4
3	9



But how do we find the slope at a single point (x) ? We can make up another point x+Δx where Δx is very small, so we have two points (x and x+Δx)and calculate the slope there.

The slope at x **as we make Δx shrink to 0 is the derivative of y at x.** We write **dx** instead of "as Δx shrink to 0" so the derivative of y over x is usually written as dy/dx.

The Recipe for Derivatives

<u>the power rule</u>:

Identify: m (constant), x (variable), c (exponent)

$$\text{if } y=mx^c \text{ , } dy/dx= mcx^{c-1}$$

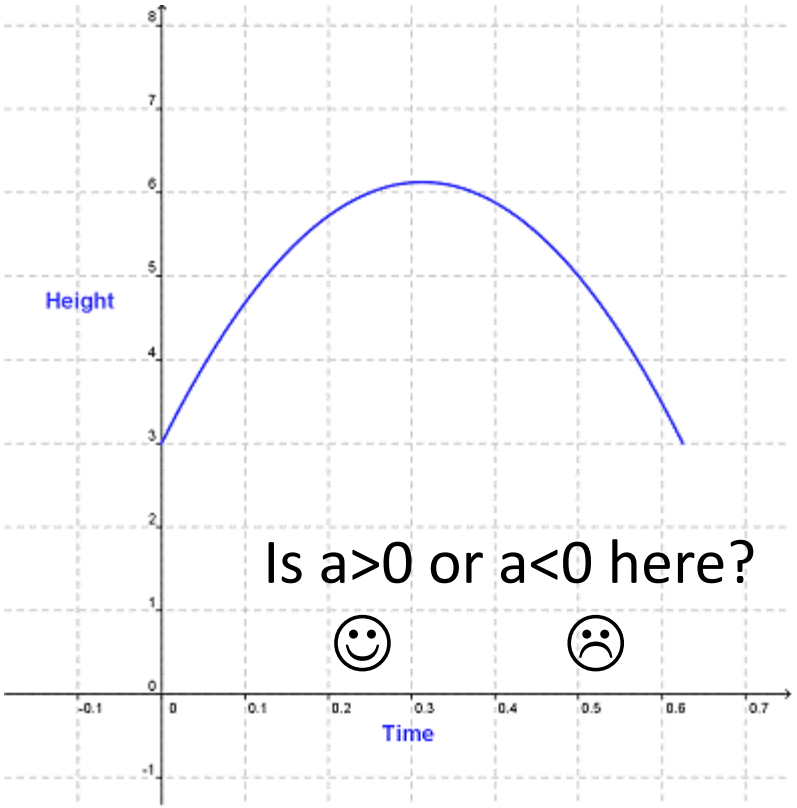
- $y=x^2 = 1x^2$ constant=1, var =x, exponent=2.
  $dy/dx=1*2x^{(2-1)} =2x$

So the derivative of x^2at x=2 is 2*2 = 4

(note this is between 3 and 5 from slide 41)

Other quadratic functions

$y = ax^2 + bx + c$



Quadratic functions are one type of polynomial functions:

$$a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0,$$

Rules for simplifying polynomials

Product rules	$x^n \cdot x^m = x^{n+m}$	$2^3 \cdot 2^4 = 2^{3+4} = 128$
	$x^n \cdot b^n = (x \cdot b)^n$	$3^2 \cdot 4^2 = (3 \cdot 4)^2 = 144$
Quotient rules	$x^n / x^m = x^{n-m}$	$2^5/2^3 = 2^{5-3} = 4$
	$x^n / b^n = (x / b)^n$	$4^3/2^3 = (4/2)^3 = 8$
Power rules	$(x^n)^m = x^{n \cdot m}$	$(2^3)^2 = 2^{3 \cdot 2} = 64$

Hint: Maybe useful when working with utility functions in microeconomics.

mx^c in general polynomials

Example:

- $Y=3x^8+ 4\sqrt{x} - 5x + \frac{2}{x}+ 9 + 2x^8$

For each term, identify: m constant, x variable, c exponent (mx^c)

- $Y=3x^8+ 4x^{1/2} - 5x^1 + 2x^{-1} + 9x^0+2x^8$

the power rule:

if $y=mx^c$, $dy/dx= mcx^{c-1}$

- Suppose y is the spread of a disease, x is % of population below poverty line, and z is temperature. How does a 1 unit increase in poverty affect the disease?

- $y=3x^3 + 4x^3$. Simplify first: $7x^3$. Then identify constant=7, var =x, exponent=3.

  $dy/dx=7*3x^{(3-1)} =21x^2$

- $y=3x^2+ 8$ constant=8, var =x, exponent=0.

  $dy/dx=6x + 8*0x^{(0-1)} =6x + 0 = 6x$

- $y=3x^2+ 8z$ constant=8z, var =x, exponent=0.

  $dy/dx=6x + 8z*0x^{(0-1)} =6x + 0 = 6x$

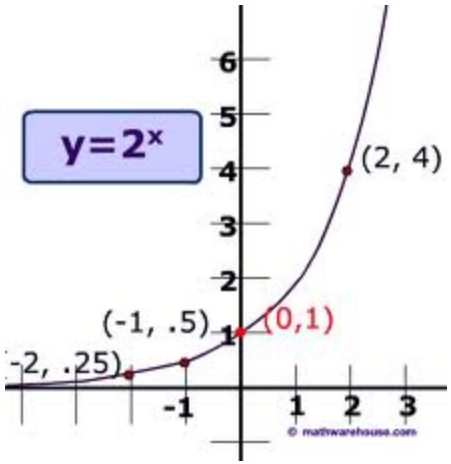
When will you use this in class? When you're trying to figure out the rate of change in an outcome due to the implementation of a policy.

Exponential function

- The growth of a terrorist cell:
- At month 0 there's 1 person 1
- At month 1 this person recruited 2 people 2
- At month 2 each persons recruited 2 people 4
- What is the function that describe the growth?
- $f=2^x$ where x is time (month)

This is an exponential functions

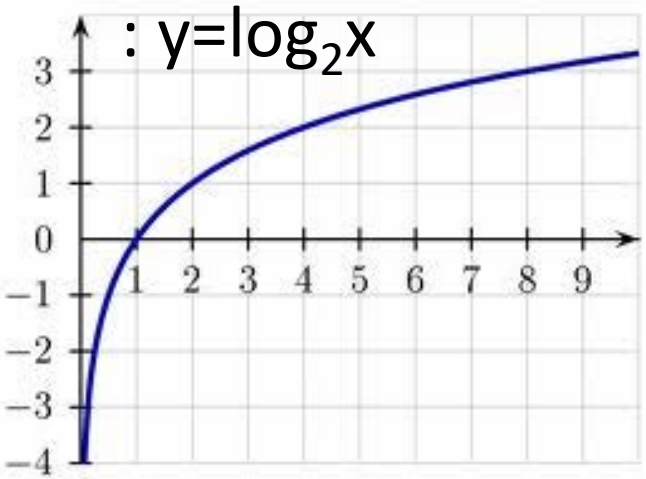
Notice it "asymptotes" at the y axis.



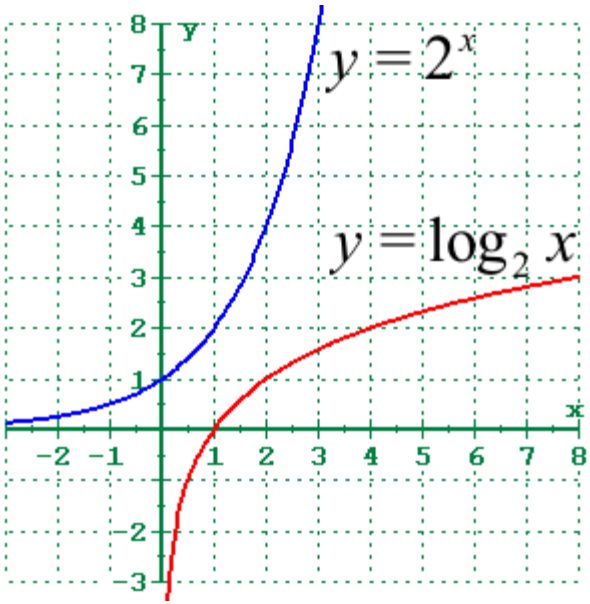
Logarithmic function

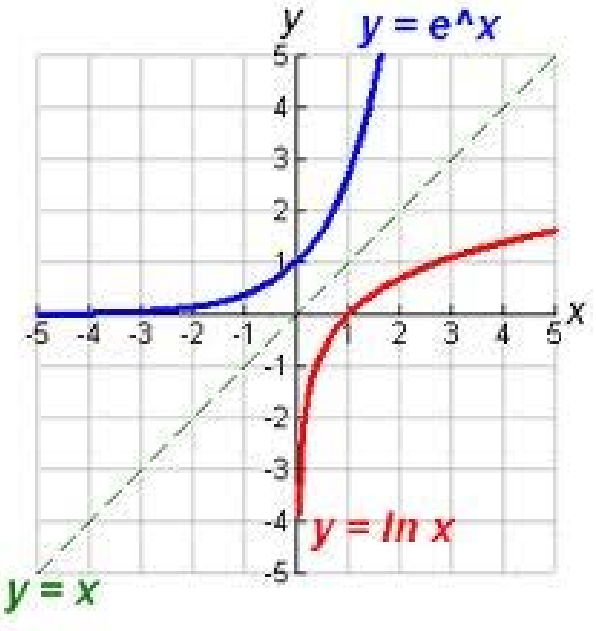
- Time since the inception of the terrorist cell
- If there is 1 member (x=1) , y=0 months
- If there are 2 members (x=2) , y=1
- If there are 8 members (x=3)

- inverse of the exponential funct
- Notice it “asymptotes” at the x a



2^x and exp(x) = 2.72^x
Logs and natural logs

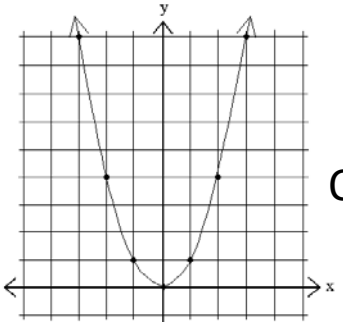




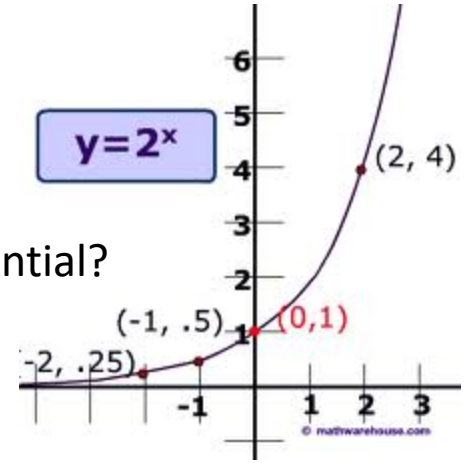
In practice when you see y=log(x) it's usually y=ln(x) than $y=\log_k(x)$. This is because **if y=ln(x), dy/dx is just 1/x**

When will you use this? When you're learning about logistic regressions.

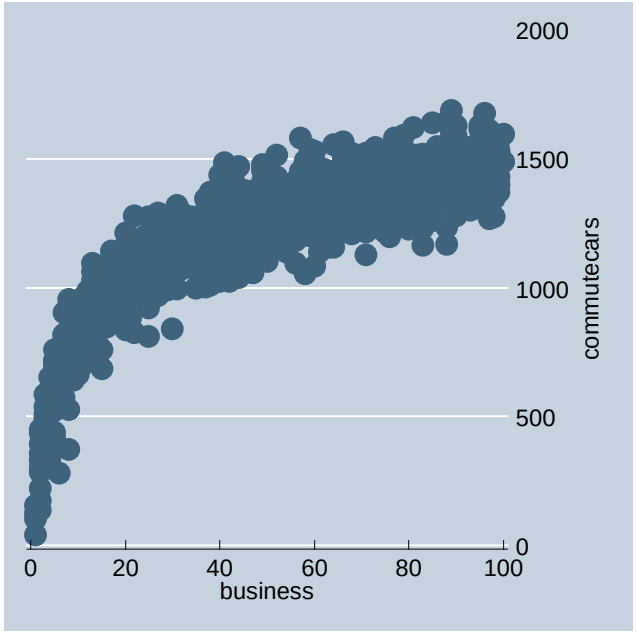
So back to your data:



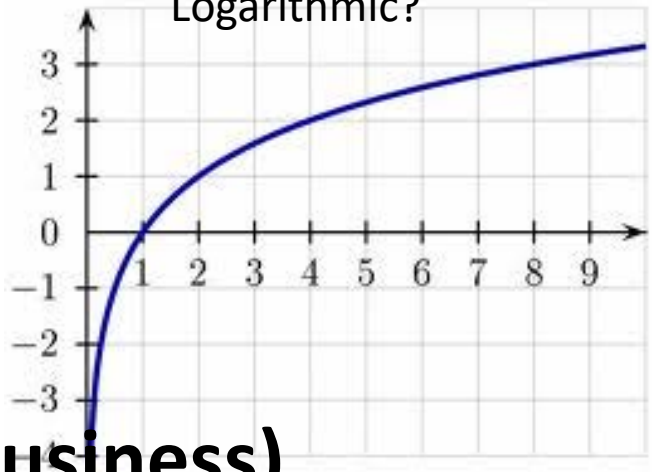
Quadratic?



Exponential?



Logarithmic?



Indeed, cars = 130+290*ln(business)

5. what is the optimal # of business to have? (optimization, compound functions)

How many businesses should be on the highway?

Let's break this down:

1. How does business affect travel time?

   We know: **cars = 130+290*ln(business)**

   And: **traveltime = 14.7 + 0.03*cars**

2. Suppose his public opinion expert says:

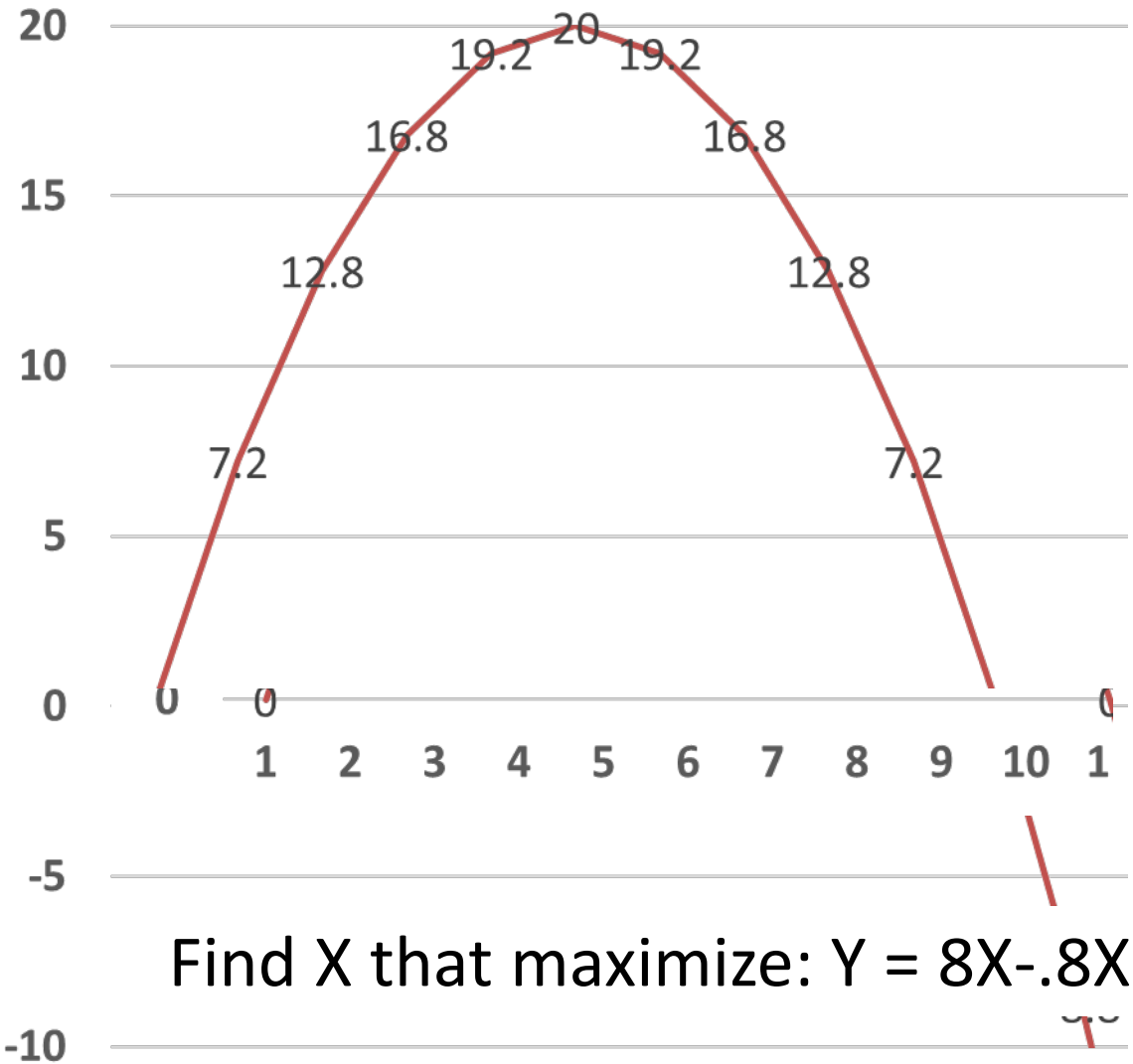
   complaints = travel time,

   praise = # of business^2/2

Then how can he maximize:

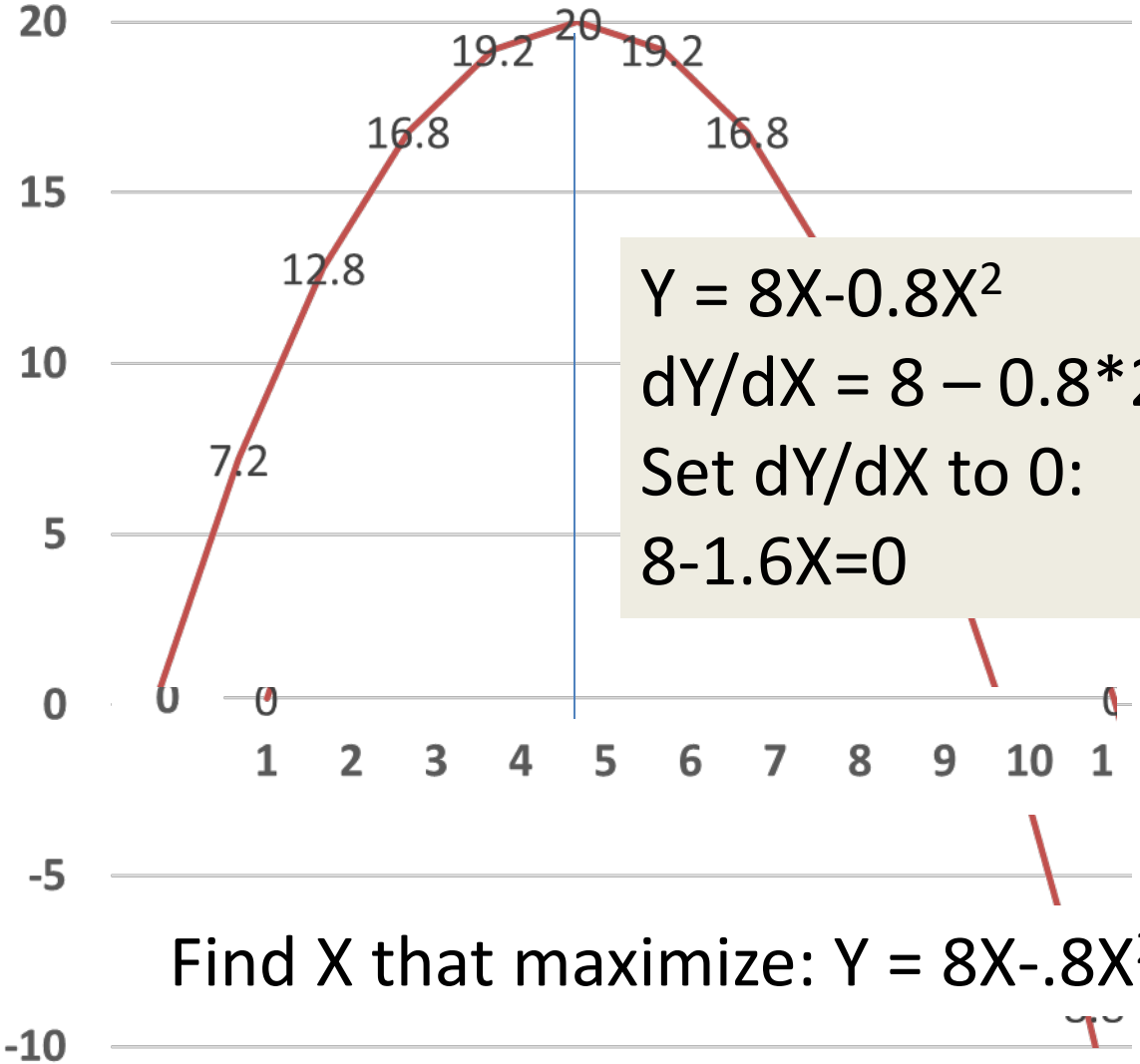
praise – complaints ?

First, how do we optimize a function?



Find X that maximize: $Y = 8X-.8X^2$

Y is maximized when dY/dX=0



$Y = 8X-0.8X^2$

$dY/dX = 8 – 0.8*2*X = 8-1.6X$

Set dY/dX to 0:

$8-1.6X=0$ $X=5$

Find X that maximize: $Y = 8X-.8X^2$

How many businesses should be on the highway?

Let's break this down:

1. How does business affect travel time?

   cars = 130+290*ln(business)

   traveltime = 14.7 + 0.03*cars

   So: traveltime = 14.7 + 0.03*(130+290*ln(business))

   Simplifying: t= 18.6+8.7ln(b)

How many businesses should be on the highway?

Let's break this down:

1. How does business affect travel time?

   $t = 18.6+8.7\ln(b)$

2. Suppose his public opinion expert says:

   complaints = travel time,

   praise = # of business^2/2

So: praise –complaints = $b^2/2 - t$

$b^2/2 - (18.6+8.7\ln(b))$

Take the derivative and set it to 0: $b-8.7/b = 0$

$b=8.7/b \quad b^2=8.7 \quad b = 3\text{-ish}$

Imagine you are an advisor to the mayor of Pittsburgh

Should I approve 10 new businesses on a strip of a crowded highway?

Mayor Peduto, GSPIA'11

The data suggests 10 is too many, Mr. Mayor. Let me walk you through my reasoning.

YOU

Exercise 2

- Break.

- BEFORE LEAVING FOR LUNCH, PACK UP. WE WILL FORM TEAMS AND YOU WILL SIT WITH YOUR NEW TEAMMATE AFTER LUNCH.

On to the Race!

Analytics

- Show Race Packet Materials.
- Tomorrow: you will absolutely need your computer.
- You will be coding and thinking and racing from room to room, so make sure you are comfortable.
- There will be 10 **clues**. Solving each clue in three tries or less will earn your team 1 point. The team with the highest number of points wins the race. **Ties** are broken by how quickly you complete the race.
- There will be Roadblocks. In Roadblocks each person in the team must solve a puzzle individually. The point will only be given if both team members successfully solve their puzzle.

Ok that’s enough excitement…

- Any questions about Exercise 2?
- Let’s dive into STATA now!

Schedule and people you will meet today

- 9:30-10:50 Intro, Lecture 1: Linear functions, Exercise 1
- 10:50-11:10 Meet your quant professors
- 11:10-11:20 Break
- 11:20-12:00 Lecture 2: Nonlinear functions and derivatives, Exercise 2
- 12:00-1:00 Lunch Break
- 1:00-1:15pm Amazing Analytics Race teams (TAs)
- 1:15-2:45 Lecture 3: Intro to Stats, Exercise 3
- 2:45-3:00pm Break
- 3:00pm-3:30pm Team exercise and alum Alex Heit

Writing and saving commands in STATA

- In your classes (and in your job in the future) you will want more control over what you did to the data and replicability.
- This is so you can remember what you did and that others can replicate your results.
- This is harder to do with the menu bar.

  - Go to Window, Do File Editor, and choose New Do-file Editor.
  - This will open a new .do file.
  - Write your commands in it.
  - Highlight one of the commands and click the “Execute (do)” icon. It should run the command. You can also copy and paste directly to the command window.
  - Save this file as MathCamp.do
  - Continue adding commands into this file.

Loading and exploring

- Clearing memory: clear
- Loading .csv file: cars.csv
- See all variables: sum
- String variables (highway) vs numeric variables
- Tab highway

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        cars |     1,674    385.3883     232.479         0       1230
  traveltime |     1,674    26.71808    8.605933   11.16805   61.63294
     highway |         0
          ID |     1,674    8.378136    4.848058         0         17

.
```

Relationship between variables

- Pwcorr (correlation)

```
. pwcorr cars traveltime

             |     cars travel~e
-------------+------------------
        cars |   1.0000
  traveltime |   0.8414   1.0000
```

- Reg traveltime cars
- Scatter traveltime cars

Sorting and Viewing Data

- **Tab (tabulate types)**
- **Tab then use that as new data to sort**

- **Moving things in and out of STATA to Excel**

- gsort
- Sort in both order
- Ascending: gsort cars
- Descending: gsort -cars
- Sort
- Only sort in ascending order

- List

```
gsort cars

. list in 1/5

+---------------------------------+
v1   cars   travel~e   highway
---------------------------------
1.  1153     0   18.31223    Roscoe
2.  1532     0   15.03788    Roscoe
3.  1321     0   15.07491    Roscoe
4.   170     0   18.04261     Robb
5.   822     0    12.8783   Jemison
+---------------------------------+

. list in -5/L

+---------------------------------+
v1   cars   travel~e   highway
---------------------------------
1670.   220   1170   57.35289     Robb
1671.   253   1190   45.25637     Robb
1672.   554   1210   55.85953   Clarion
1673.  1390   1220   54.14872    Roscoe
1674.    59   1230   47.82588    Roscoe
+---------------------------------+
```

Conditional statements and working with strings (if, and (&), or (|), ==, !=)

sum traveltime if cars < 100

mean traveltime if cars > 150 & cars < 200

reg traveltime cars if highway !="SqHill"

sum traveltime if highway =="SqHill" | highway == "Clarion"

list traveltime if highway =="SqHill" & cars > 400

Tab + multiple logical expression

Generating new variables

- gen: simple transformations of other variables

  gen travelsq = traveltime^2

- What if you mess up making a variable and want to recreate it? Eg. You want travelsq to be ½*traveltime^2

  drop travelsq

  gen travelsq = (1/2)* traveltime^2

Can combine gen with logical statements :

gen toocrowded = (cars>400)

Using your new variable:

reg traveltime cars if toocrowded

reg traveltime cars if !toocrowded

reg traveltime toocrowded

Graphing

Comparing two subgroups:

twoway (scatter traveltime cars if toocrowded) (scatter traveltime cars if !toocrowded)

twoway (scatter traveltime cars if highway==“Roscoe”) (scatter traveltime cars if highway==“Robb”)

Comparing two version of traveltime:

twoway (scatter traveltime cars) (scatter travelsq cars)

How to save your graphs?

File– Save As – (I usually do .pdf)

Or: Win users: right click and click Copy and then paste into your word doc.

Review Exercise 3

- Any questions?

Ok but what I really want to know is .. How can I win the Amazing Analytics Race?

How to increase your winning probability

- Review all the material tonight with your teammate and decide on how you want to handle roadblocks and other scenarios. The math will be simple but will require creative applications.
- Stata commands: You MUST get familiar with all the commands we did today.
- When getting your answers checked you can send just one person so one of you can continue working.
- Tomorrow: you can setup starting from 12:30pm. We will distribute materials for the race at 1pm
- On to work with your teammate! (Group exercise)

Alumni analytics career talk

Alexander S. Heit

Principal, Partnerships & Strategy at VISIMO

A 615 Fifth Avenue, Suite 202 - Coraopolis PA 15108

P 412-528-1958 **M** 412-576-6129

E alex@visimoconsulting.com

W www.visimoconsulting.com

Training done!
But you can stay to work on the group exercise and talk to us until 4pm.