Data Analysis with Python

Gregory M. Eirich

Columbia University

(Class #2)

1. Regression (and Correlation)

The Simplest Case:

Only 2 Variables,
Both Interval-Ratio/Ordinal

Importing packages

```
from __future__ import division
import numpy as np
import matplotlib
import matplotlib.pyplot as plt
import os
```

NumPy is a Python package for scientific computing.

Matplotlib is a (primarily 2D) desktop plotting package designed for creating publication-quality plots.

Pyplot (part of matplotlib) is the plotting package.

os is a module that allows us to do tasks like checking the working directory, changing the working directory, looking for files within a directory, and more.

The "import" command allows you to import the package and define a local name for it using the "as [name]" command. For example, when we "import numpy as np", "np" is the local name for the NumPy package going forward.

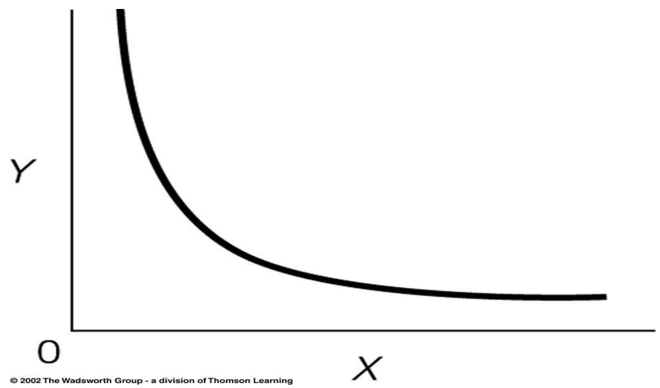
Correlating education and occupational prestige

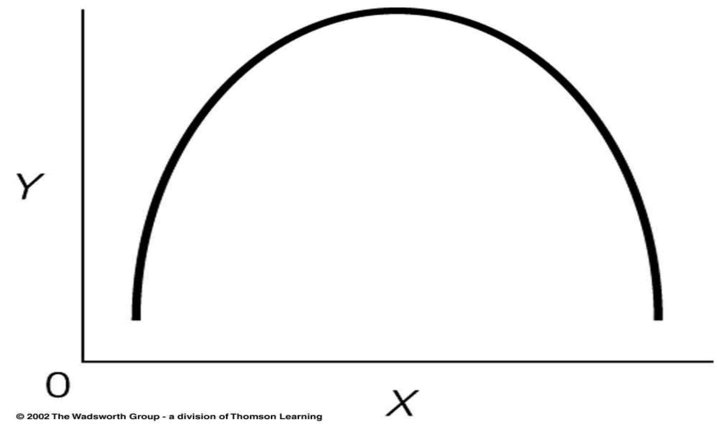
CASEID	EDUC	PRESTG80
20063076	10	45
2008 564	12	30
2006 605	12	34
20061650	12	50
20063014	12	51
20063352	14	55
2008 895	16	49
20063726	16	54
20081537	17	68

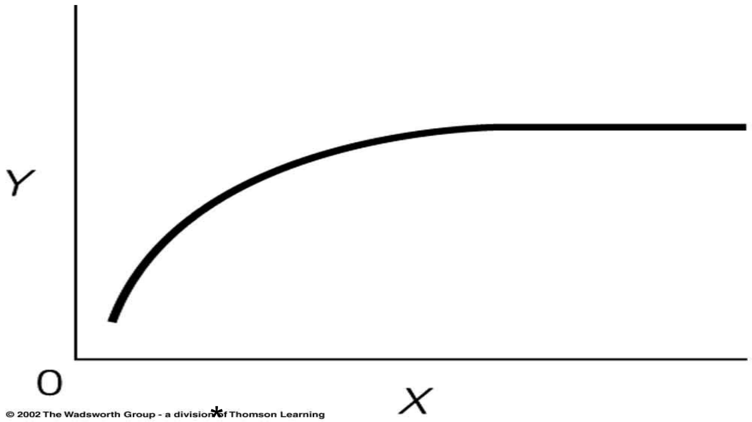
At first, we are looking to find linear relationships

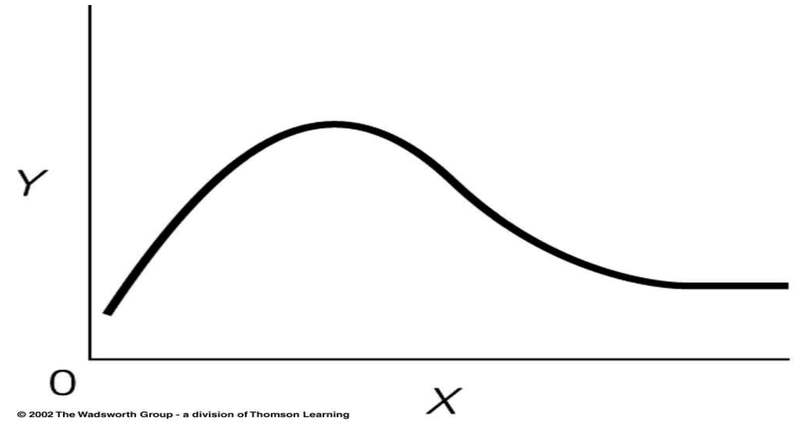
- That means that the relationship between EDUC and PRESTG80 remains constant at any value of EDUC or PRESTG80
- We can draw a straight line through the relationship
- Other relationships are possible …

Systematic BUT Nonlinear Associations





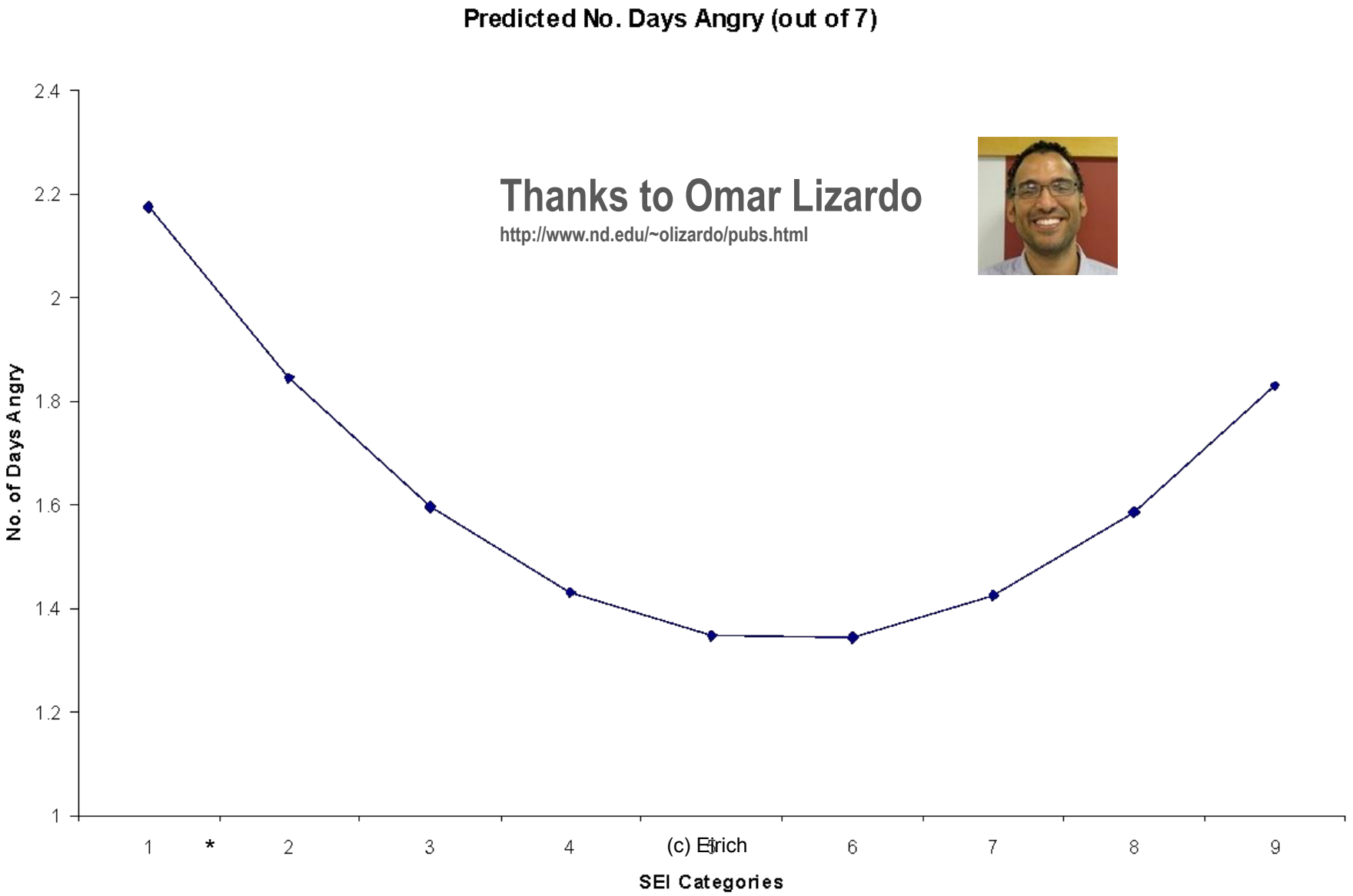




A non-linear example –

Who (rich, middle-class, or poor) is angry most often?

Curvilinear Relationship



Thanks to Omar Lizardo

http://www.nd.edu/~olizardo/pubs.html

2. Now, back to the linear model …

Questions

1. What is the **direction** of the relationship?
2. What is the **magnitude** of the relationship?
3. Is the relationship **statistically significant**?

Questions

1. What is the **direction** of the relationship?

   Positive or negative?

The model …

$$y = \alpha + \beta x$$

The population equation

$$\hat{y} = a + bx$$

The sample equation

The slope

$$b = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2}$$

b = Rise / Run

b = Cov(X,Y) / Var(X)

Think of cov(x,x) = var(x)

The slope

$$b = SP / SSx$$

SP=Sum of Products of Deviations from X and Y
SSx=Sum of Squared Deviations from X

What do we need to calculate the slope?

CASEID	EDUC (=X)	PRESTG80 (=Y)	X-Xbar	Y-Ybar	(X-Xbar)(Y-Ybar)	$(X\text{-}Xbar)^2$	$(Y\text{-}Ybar)^2$
20063076	10	45	-3.4	-3.4	11.9	11.9	11.9
2008 564	12	30	-1.4	-18.4	26.6	2.1	340.2
2006 605	12	34	-1.4	-14.4	20.9	2.1	208.6
20061650	12	50	-1.4	1.6	-2.2	2.1	2.4
20063014	12	51	-1.4	2.6	-3.7	2.1	6.5
20063352	14	55	0.6	6.6	3.6	0.3	43.0
2008 895	16	49	2.6	0.6	1.4	6.5	0.3
20063726	16	54	2.6	5.6	14.2	6.5	30.9
20081537	17	68	3.6	19.6	69.5	12.6	382.4

13.4 48.4 **SP=142.2** **SSx=46.2** **SSy=1026.2**

The slope

$$b = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2}$$

b = (142.2)/(46.2)

b = 3.08

Interpret the slope (b) of the regression line.

$b = 3.08$

An increase of 1 year of EDUC will result in a 3.08 point increase in occupational prestige.

Positive or Negative?

Positive relationship = both variables increase (or decrease) together

Negative relationship = one variable increases, while the other decreases, or vice versa

So the intuition is …

The slope measures how much the X and Y variables vary **together**, compared to how much the X variable varies on its own overall

The issue with the slope

- The interpretation of the slope is dependent on the units that the X and Y are measured in
- If you change the units, you can change the magnitude of the slope
- … we will use a correlation coefficient instead (later)

Computing *a, y-intercept, constant*

$$a = \bar{y} - b\bar{x}$$

The point at which the regression line crosses the vertical or Y axis when X is 0.

Computing *a, y-intercept, constant*

$$a = \bar{y} - b\bar{x}$$

= 48.45 – 3.08(13.45)

= 7.08

Computing *a, y-intercept, constant*

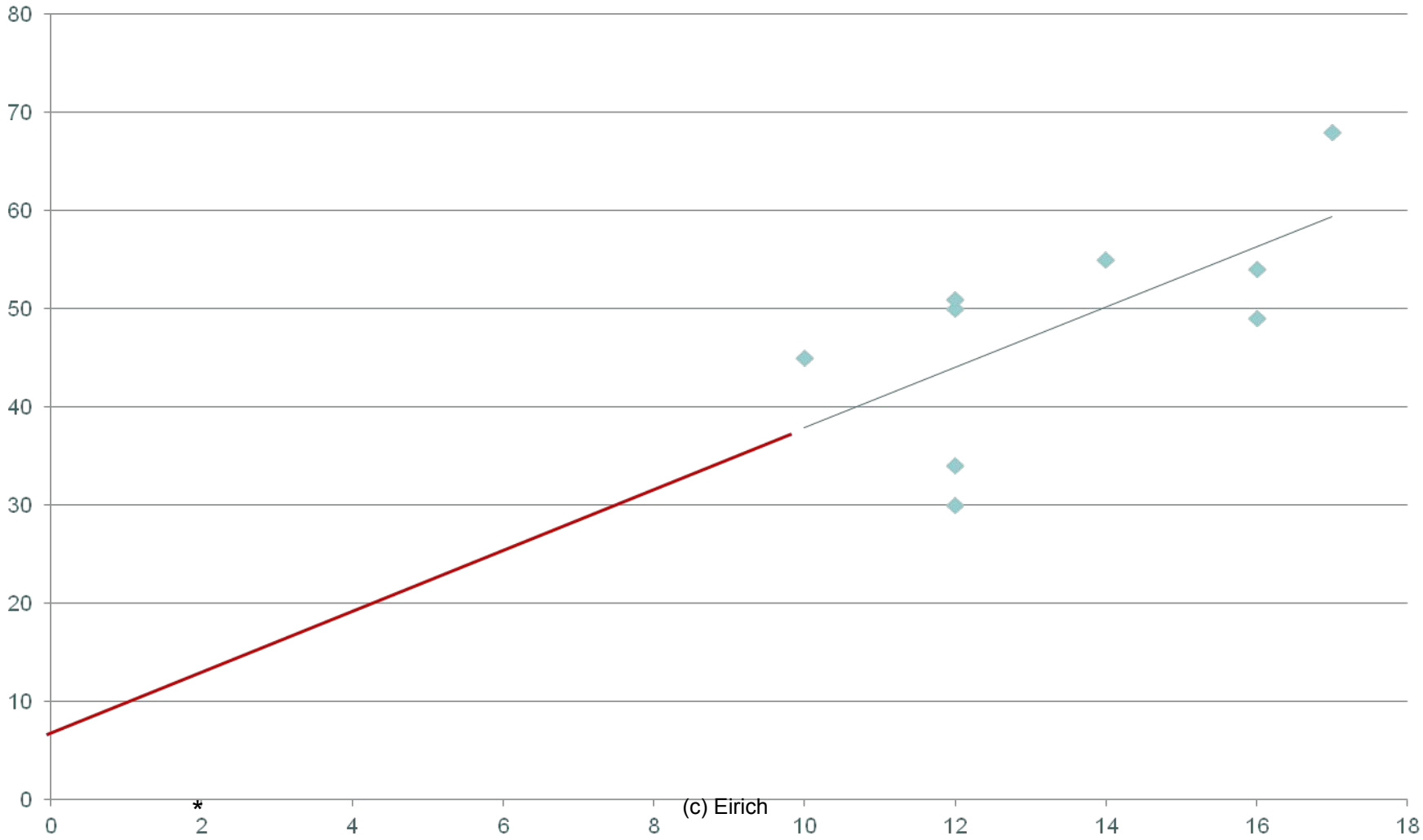
$$a = \bar{y} - b\bar{x}$$

= 48.45 – 3.08(13.45)

= 7.08

The regression line crosses the Y axis at 7.08 prestige points; when EDUC = 0, a person has 7.08 prestige points

Y-intercept or the constant

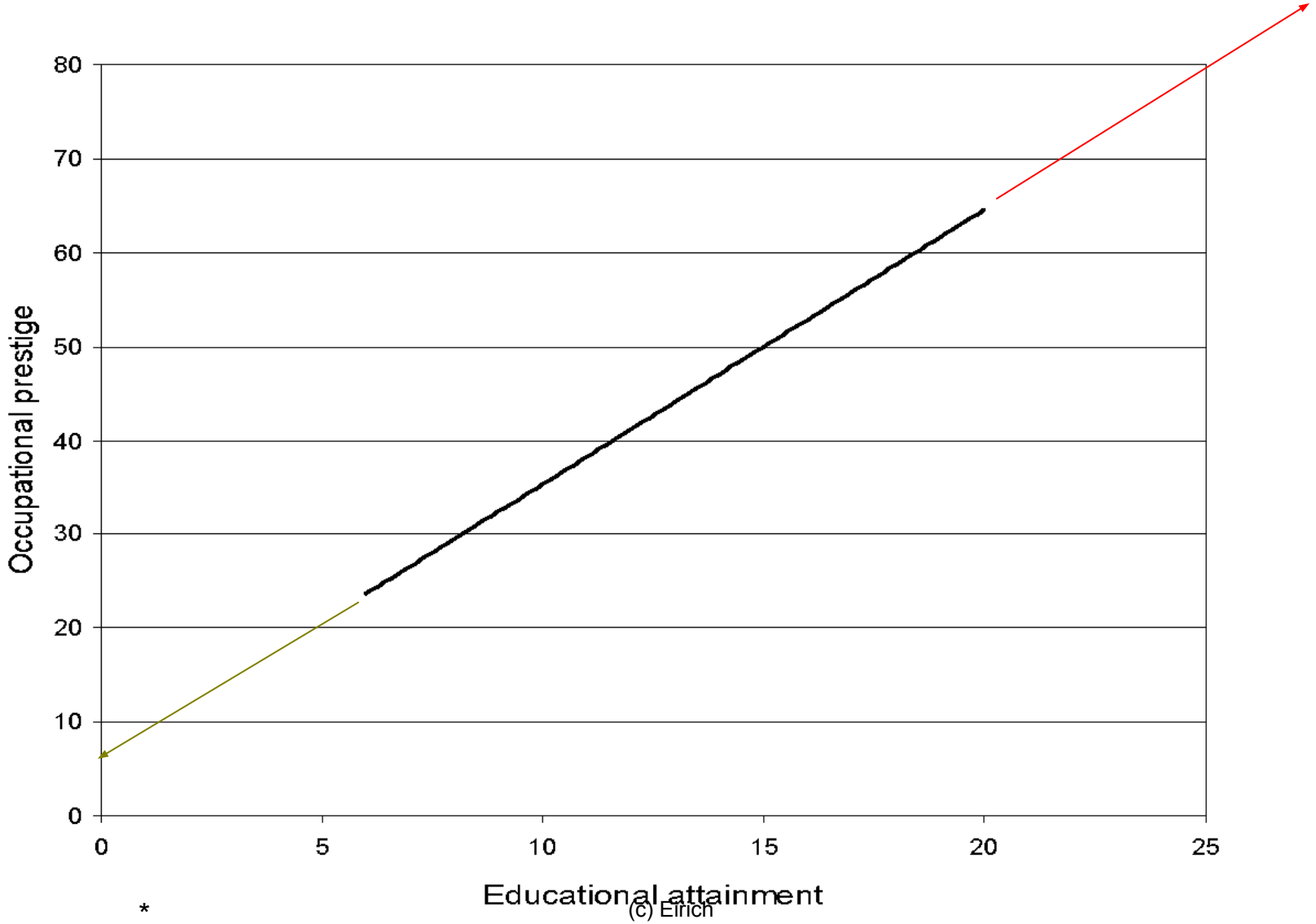


*

We can totally predict outside of the range of our data …

Regression allows us to predict forward or backward

Predict, using the slope (b) of the regression line.



*

Using regression line, predict Y, if X=?

If X=25:

$$Y = 7.08 + 3.08(x)$$

$$Y = 7.08 + 3.08(25)$$

$$Y = 7.08 + 77$$

$$Y = 84.08$$

People with 25 years of schooling, they would have an occupation whose prestige score is 84.08.

Pandas is better for handling data in Python

```
import pandas as pd

os.chdir('C:\Users\gme2101\Desktop\Data Analysis Data') # set working
directory

df = pd.read_csv('prestige.csv') # read in csv
df.head()   # check the first 5 lines
```

Out[15]:

	CASEID	EDUC	PRESTG80
0	20063076	10	45
1	2008 564	12	30
2	2006 605	12	34
3	20061650	12	50
4	20063014	12	51

Here is the regression set up in Python

Statsmodels is a great package for doing regression.

```
import statsmodels.api as sm
import statsmodels.formula.api as smf

df = pd.read_csv('prestige.csv')

lm = smf.ols(formula = 'PRESTG80~EDUC', data = df).fit()
```

Here is the regression output from Python

```
lm = smf.ols(formula = 'PRESTG80~EDUC', data = df).fit()
print (lm.summary())
```

OLS Regression Results

Dep. Variable:	PRESTG80	R-squared:	0.426
Model:	OLS	Adj. R-squared:	0.344
Method:	Least Squares	F-statistic:	5.204
Date:	Thu, 18 May 2017	Prob (F-statistic):	0.0565
Time:	19:23:58	Log-Likelihood:	-31.583
No. Observations:	9	AIC:	67.17
Df Residuals:	7	BIC:	67.56
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Intercept	7.0769	18.389	0.385	0.712	-36.407	50.561
EDUC	3.0769	1.349	2.281	0.057	-0.112	6.266

Omnibus:	2.156	Durbin-Watson:	1.728
Prob(Omnibus):	0.340	Jarque-Bera (JB):	1.083
Skew:	-0.518	Prob(JB):	0.582
Kurtosis:	1.654	Cond. No.	82.5

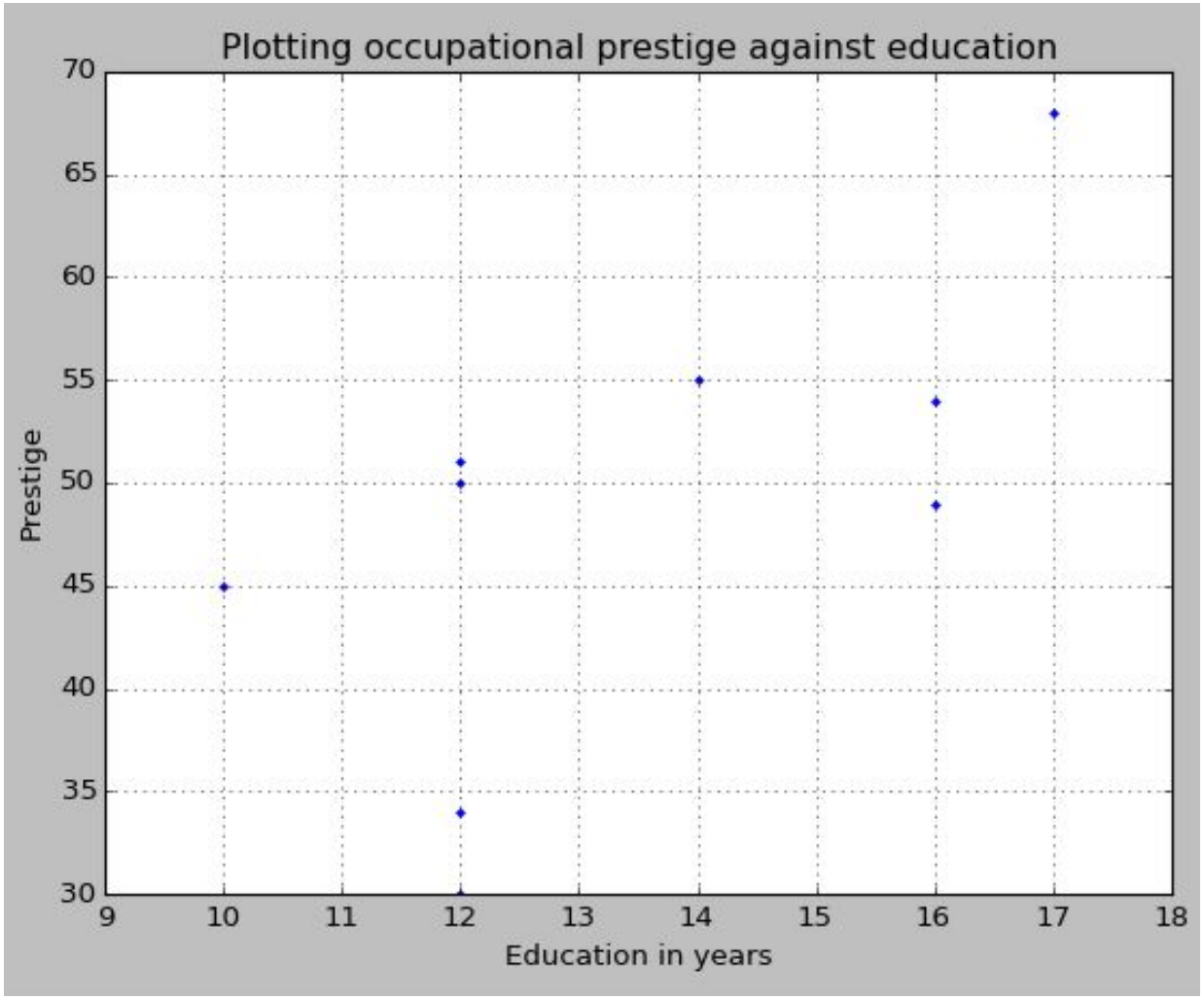
BTW, we can plot using pandas too

Here is the code:

```
df.plot(kind = 'scatter', x = 'EDUC', y = 'PRESTG80')
plt.title('Plotting occupational prestige against education')  # add title and
labels
plt.xlabel('Education in years')
plt.ylabel('Prestige')
plt.show()

# plt.savefig('prestige_plot_pandas.png',dpi=100) # saves figure to working
directory
```

Simple plot

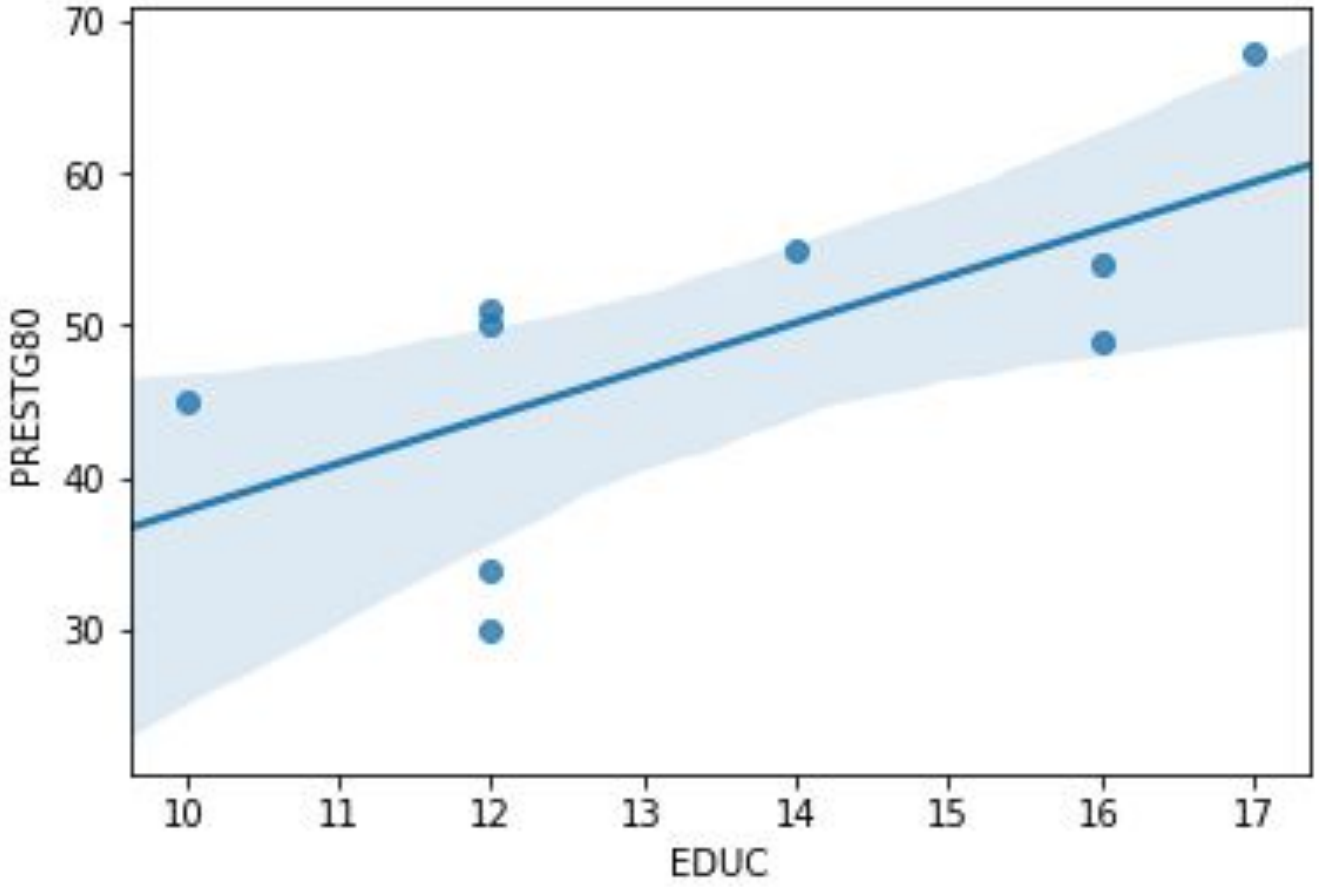

Plotting occupational prestige against education
Prestige
Education in years
70
65
60
55
50
45
40
35
30
9
10
11
12
13
14
15
16
17
18

BTW, we can plot using seaborn too

Here is the code:

```
import seaborn as sns
sns.regplot(df['EDUC'],df['PRESTG80'])
```

Simple plot



Questions

3. What is the **strength** of the relationship?

Correlation = a measure of how much one variable varies along with another variable

Correlation

- r can be thought of as a standardized b coefficient:

$$r = \left(\frac{s_x}{s_y} \right) b$$

where *s* means “standard deviation of”

Correlation

$$r = \left(\frac{s_x}{s_y}\right) b \quad = (2.4/11.3)*3.08 \quad = +\mathbf{0.65}$$

Where:
sx=2.4
sy=11.3

Here is the correlation output from Python

```
df.corr(method = 'pearson')
```

	EDUC	PRESTG80
EDUC	1.000000	0.653012
PRESTG80	0.653012	1.000000

Or - Here is the correlation output from Python

```
import scipy
scipy.stats.pearsonr(df["EDUC"], df["PRESTG80"])

# the first number is the correlation coefficient, and the second is p-value
```

```
(0.653012270802549 87, 0.05653095224376862 6)
```

… or …

```
r = df['EDUC'].corr(df['PRESTG80'])
r
```

```
0.653012270802549 87
```

Correlation

- Ranges between -1 and +1
- 0 indicates no relationship between X and Y
- 1 (positive or negative) indicates perfect relationship between X and Y

Questions

4. Is the relationship **statistically significant**?

Statistical significance

To talk of statistical significance, we have to think of a probabilistic regression model

Statistical significance

What we had previously, was a **deterministic** model, where for any given x, we specified an exact y

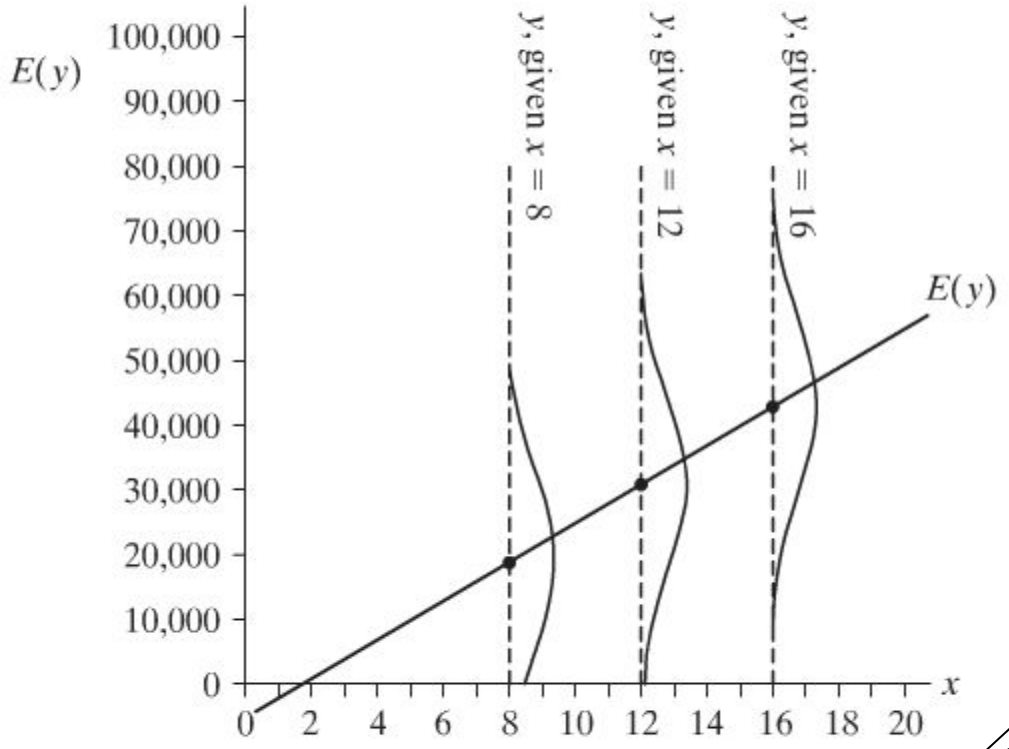
This is unrealistic …

Statistical significance

Instead, for any given X value, there is a typical range of Y values that correspond with that X

(Conditional probability.)

The expected values of y, conditioned on given values of x

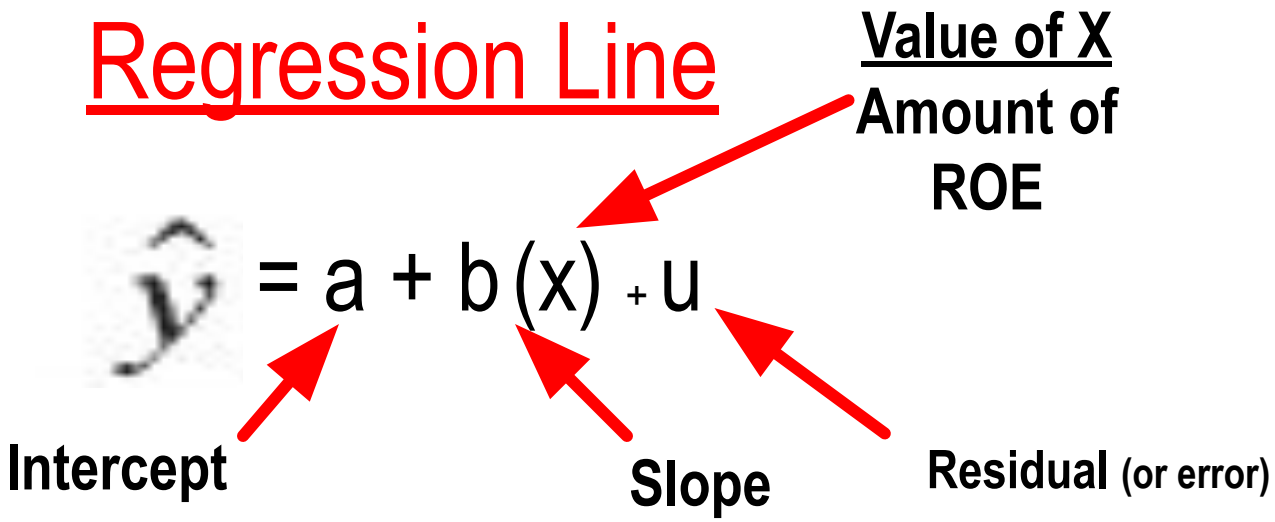


We will return to this sigma soon.

FIGURE 9.8: The Regression Model $E(y) = -5000 + 3000x$, with $\sigma = 13,000$, Relating y = Income (in Dollars) to x = Education (in Years)

5. A Regression Model that Includes The Residual (or Error) (or Disturbance) Term

Predicting CEO Pay (Y) from Company's ROE (X)



Mechanics of e

$$y_i = \hat{y}_i + \hat{u}$$

The actual value of y is made up a predicted value plus some error

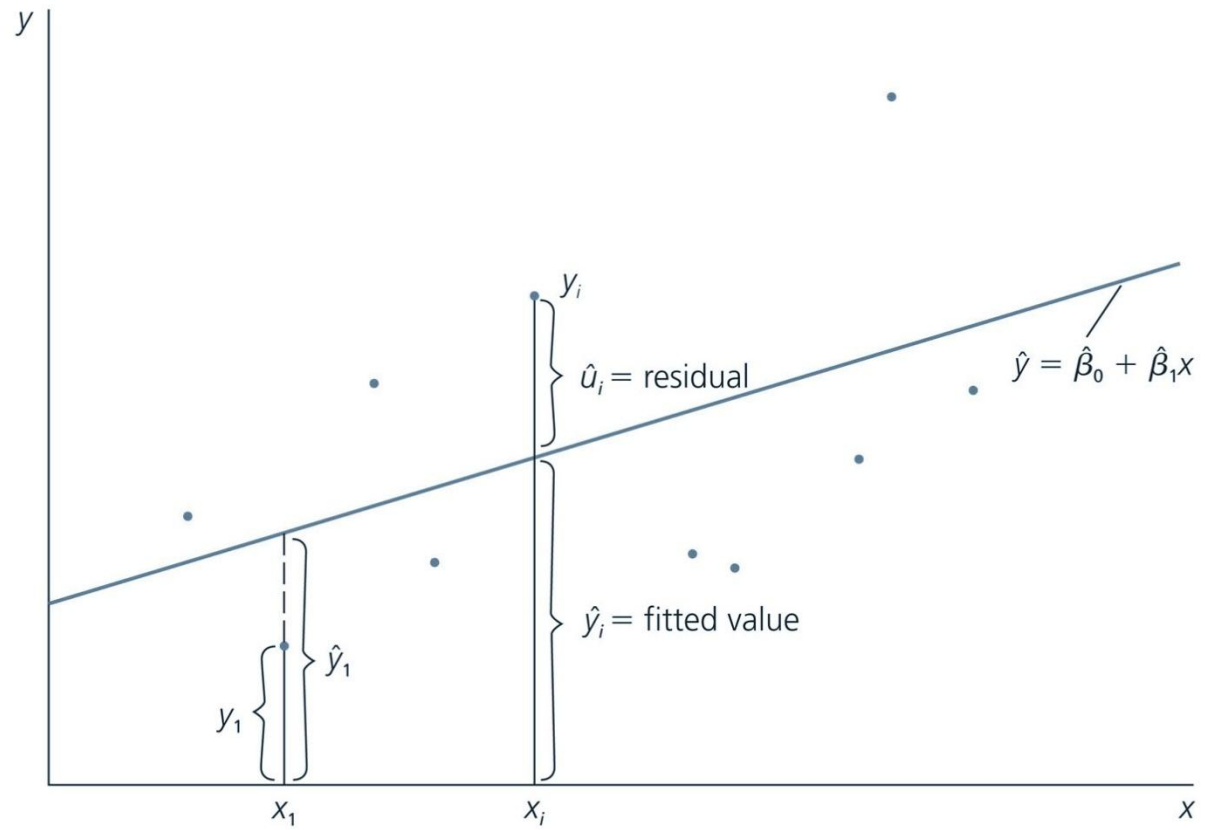
TABLE 2.2

Fitted Values and Residuals for the First 15 CEOs

obsno	*roe*	*salary*	*salaryhat*	*uhat*
1	14.1	1095	1224.058	−129.0581
2	10.9	1001	1164.854	−163.8542
3	23.5	1122	1397.969	−275.9692
4	5.9	578	1072.348	−494.3484
5	13.8	1368	1218.508	149.4923
6	20.0	1145	1333.215	−188.2151
7	16.4	1078	1266.611	−188.6108
8	16.3	1094	1264.761	−170.7606
9	10.5	1237	1157.454	79.54626
10	26.3	833	1449.773	−616.7726
11	25.9	567	1442.372	−875.3721
12	26.8	933	1459.023	−526.0231
13	14.8	1339	1237.009	101.9911

FIGURE 2.4

Fitted values and residuals.



Implications of e: Unpacking e

- Error term
- Disturbance
- Unobservables

Whatever affects our dependent variable but is not included in our equation is captured by *e*

(Much more to come on this next week)

Prediction equation has least squares property

Our goal is to draw a prediction line that has the smallest sum of squared errors (SSE)

$$\text{SSE} = \sum (y - \hat{y})^2$$

Where's the justification?

- Where's the justification that our *a* and *b* parameters from earlier actually occur when we minimize the sum of squared residuals?

- How do we know this is the best slope to choose?

6. SSE is part of R-square

A quick question on predicting height

The Calculation of R-sq

Sum of Squares Total – or TSS

Sum of Squares Error – or SSE

Sum of Squares Model – or Sum of Squares Regression

$$\sum (Y_i - \bar{Y})^2 = \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \bar{Y})^2$$

The Calculation of R-sq

Sum of Squares Total, or TSS

Sum of Squares Error, or SSE

Sum of Squares Model, or Sum of Squares Regression

$$\sum (Y_i - \bar{Y})^2 = \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \bar{Y})^2$$

How much total variance is there for Y?

How much does my prediction of Y_i differ from the actual value of Y_i?

How much better is my prediction of Y_i (because I have included the X_i) than just using the average of Y?

The Calculation of R-sq

$$R\text{-}sq = \frac{\text{Sum of Squares Total, SST} - \text{Sum of Squares Error, or SSE}}{\text{Sum of Squares Total, SST}}$$

How much of the total variance of Y, can I explain by using these Xs?

Or this …

$$R\text{-sq} = \frac{\text{Sum of Squares Model, or Sum of Squares Regression}}{\text{Sum of Squares Total, SST}}$$

How much of the total variance of Y, can I explain by using these Xs?

Calculate Coefficient of Determination

r^2 = Coefficient of Determination

$R^2 = r*r$

$R^2 = (.65)(.65)$

$R^2 = .43$

43% of the variation in occupational prestige is explained by the amount of education people get.

Here is the ANOVA output

```
lm = smf.ols(formula = 'PRESTG80~EDUC', data = df).fit()
print(sm.stats.anova_lm(lm, typ = 1))
```

	df	sum_sq	mean_sq	F	PR(>F)
EDUC	1	437.606838	437.606838	5.204159	0.056531
Residual	7	588.615385	84.087912	NaN	NaN

$$r^2 = \frac{\text{TSS} - \text{SSE}}{\text{TSS}} = \frac{1026.22 - 588.61}{1026.22} = 0.43$$

How much of the total variance of Y, can I reduce by considering these Xs?

What is R-sq if we put no Xs in the model at all?

A regression with no Xs (showing R, since Python won't do it)

```
> pelm2 = lm(prestige~prestige, data=d)
Warning messages:
1: In model.matrix.default(mt, mf, contrasts) :
  the response appeared on the right-hand side and was dropped
2: In model.matrix.default(mt, mf, contrasts) :
  problem with term 1 in model.matrix: no columns are assigned
> summary(pelm)

Call:
lm(formula = prestige ~ prestige, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-18.444  -3.444   1.556   5.556  19.556

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   48.444      3.775   12.83 1.28e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.33 on 8 degrees of freedom
```

What should be the slope from this model actually be?

A regression with no Xs

```
> pelm2 = lm(prestige~prestige, data=d)
> summary(pelm)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   48.444      3.775   12.83 1.28e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.33 on 8 degrees of freedom
```

If we specify no Xs, we still can get a constant. What does this constant mean? Does it look familiar? (How are we able to still get t-statistics?)

Compare to this ...

```
df["PRESTG80"].mean()
```

```
48.44444444444444
```

The mean = the intercept from an “empty” model

When You Can't Afford Sleep
By Olga Khazan

The Entrepreneur Who Wants to Save Paradise
By Diana Saverin

Rem
Char
By Ch

Why It's So Rare for a Wife to Be Taller Than Her Husband

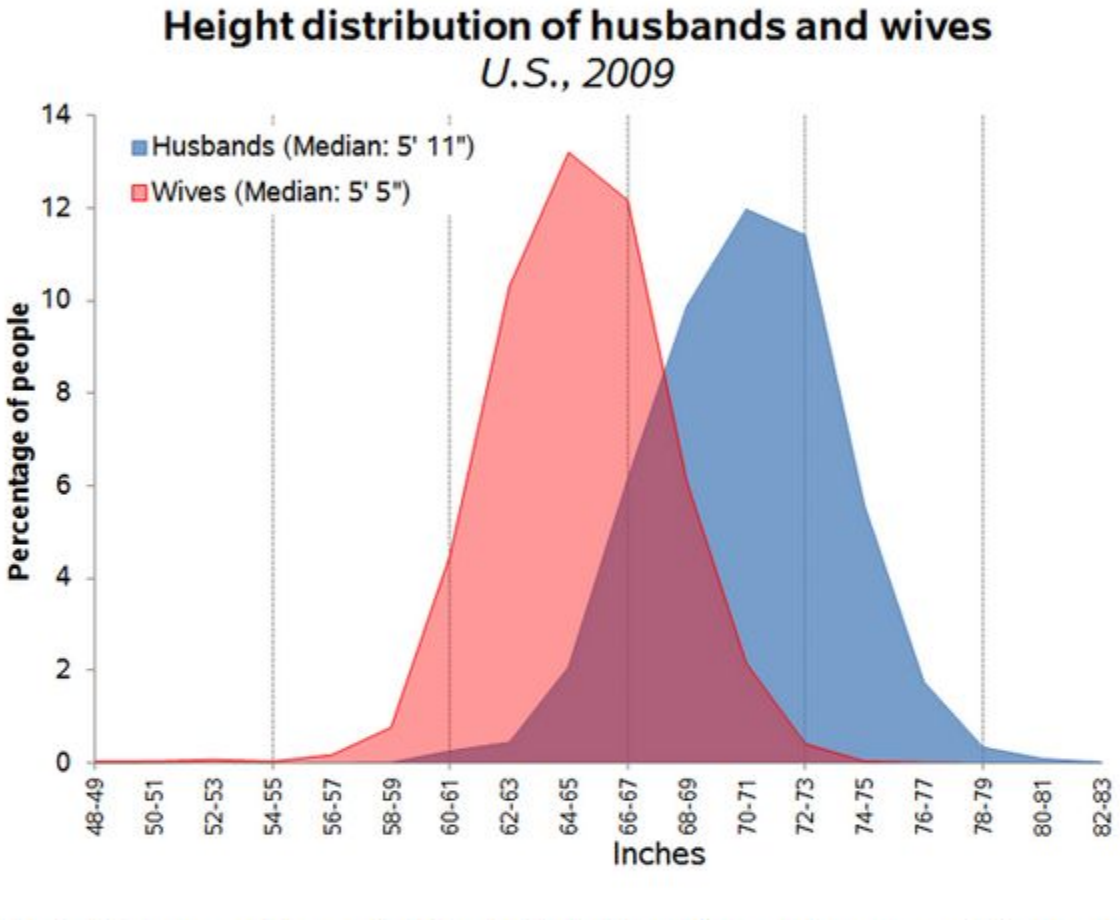
PHILIP COHEN | JAN 28 2013, 12:11 PM ET

It's not just because women are, on average, shorter than men.

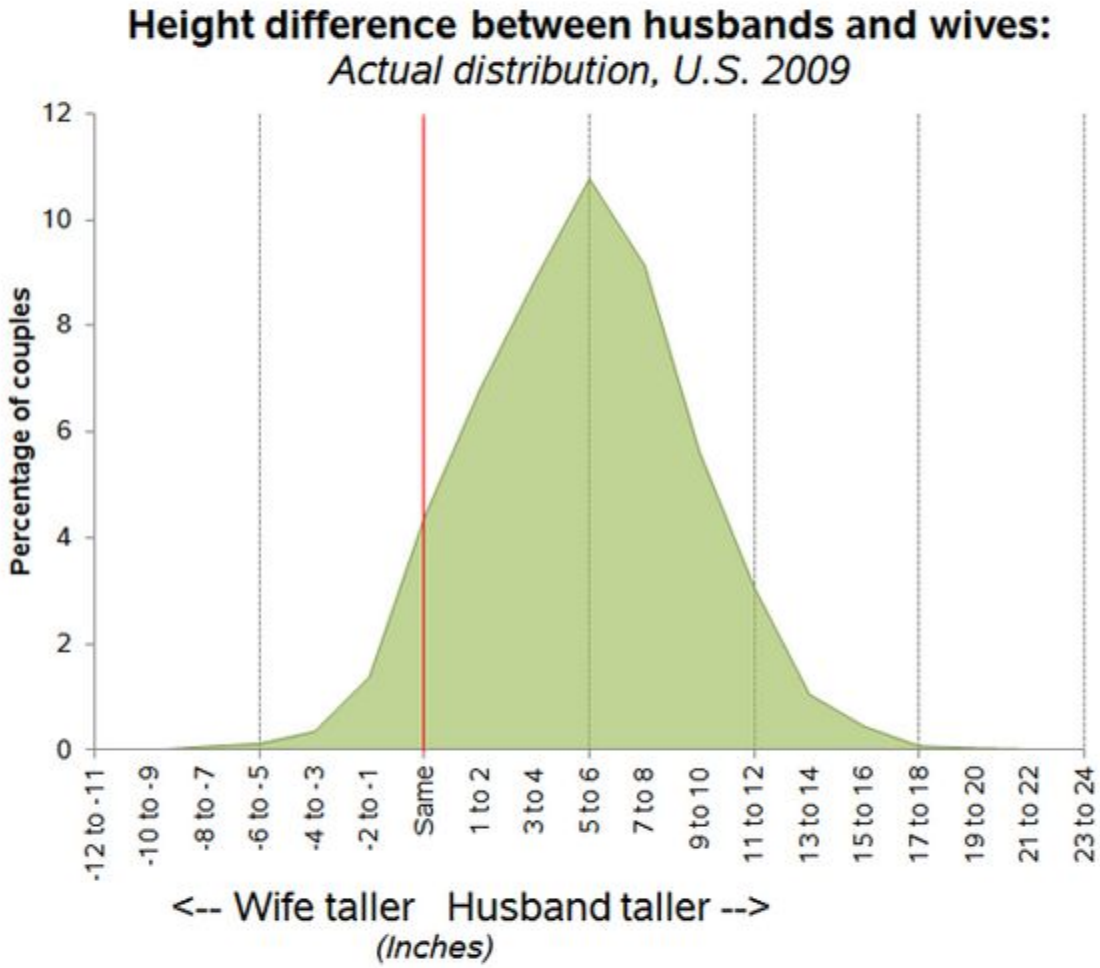
Keith Urban and his wife Nicole Kidman arrive at the 2009 American Music Awards. (Chris Pizzello/AP Images)

Men are bigger and stronger than women. That generalization, although true, doesn't adequately describe how sex affects our modern lives. In the first place, men's and women's size and strength are distributions. Strong women are

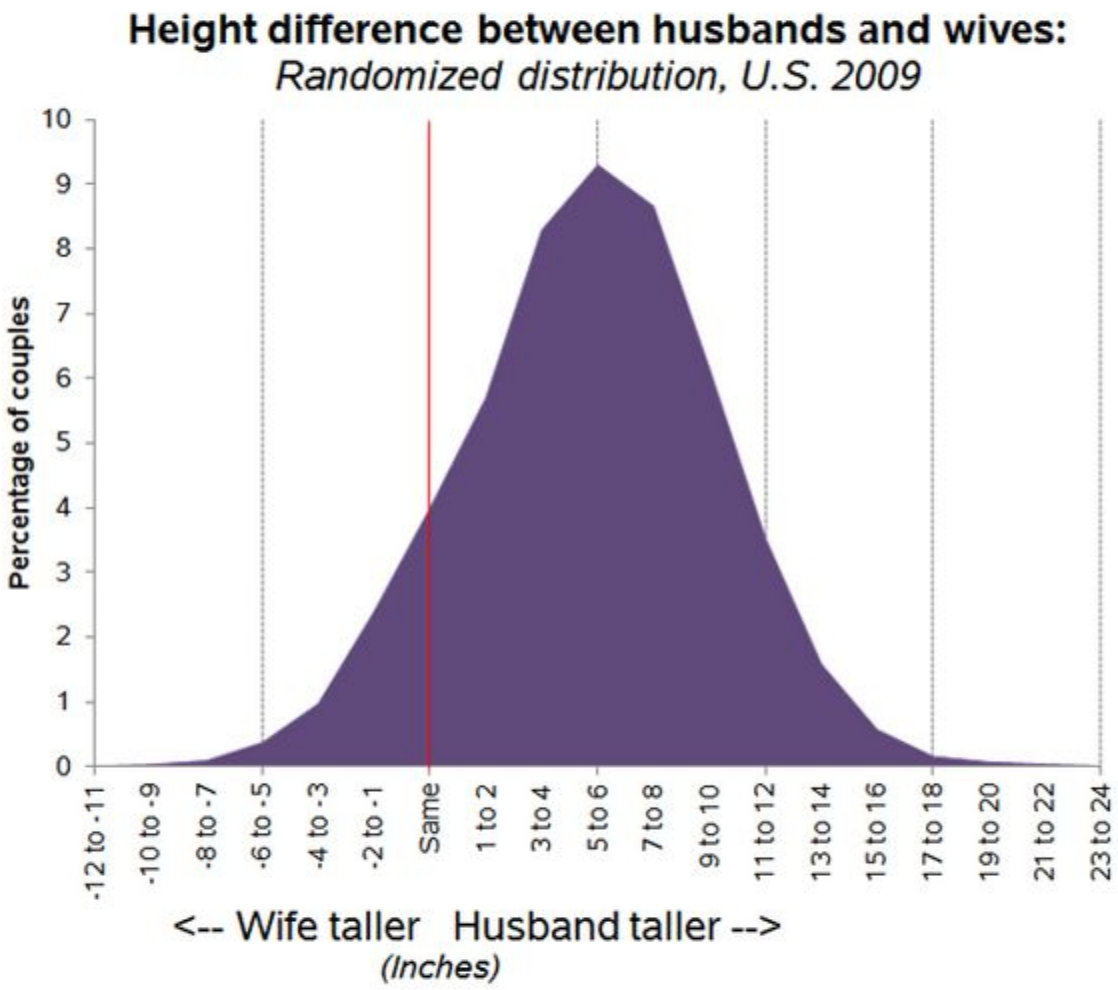
BTW ...


Height distribution of husbands and wives
U.S., 2009
Husbands (Median: 5' 11")
Wives (Median: 5' 5")
Percentage of people
0
2
4
6
8
10
12
14
48-49
50-51
52-53
54-55
56-57
58-59
60-61
62-63
64-65
66-67
68-69
70-71
72-73
74-75
76-77
78-79
80-81
82-83
Inches

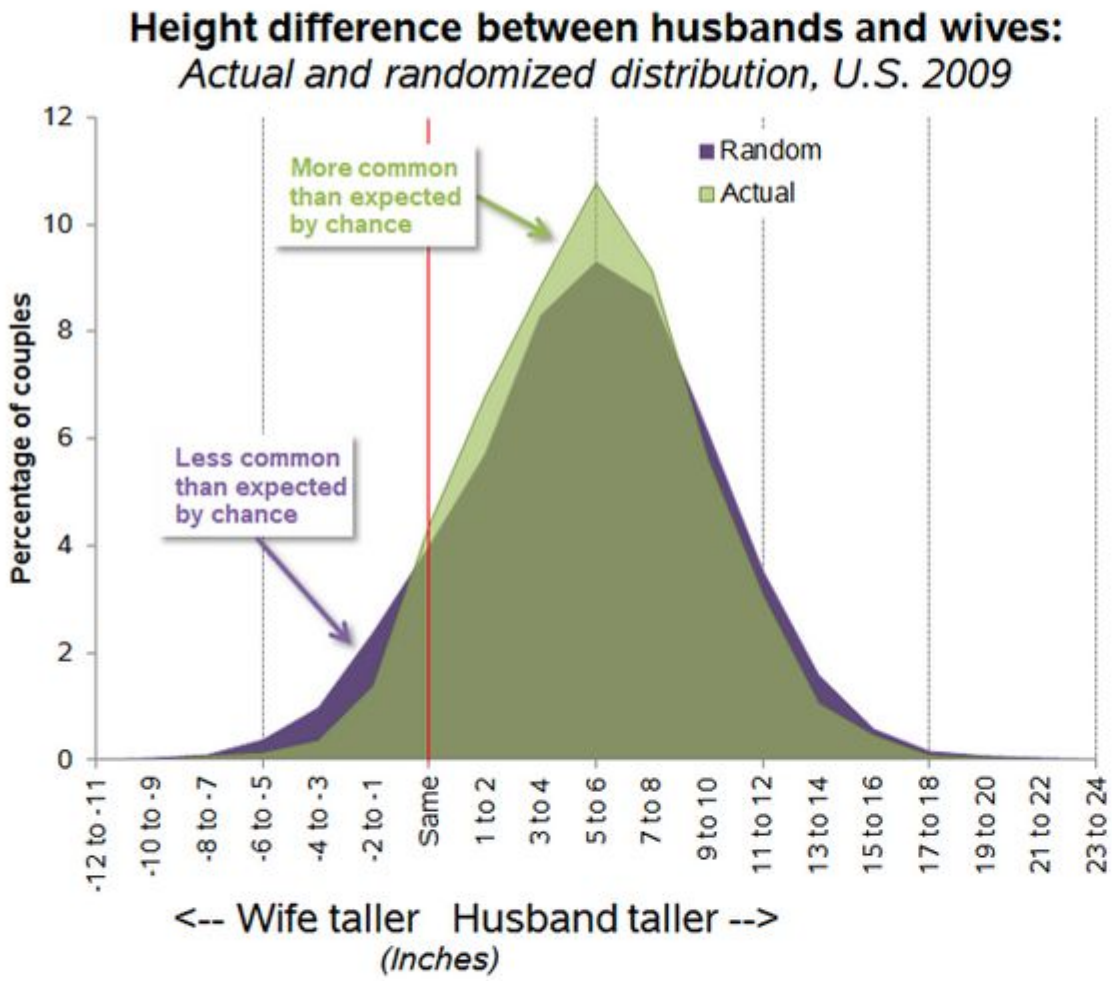
*


Height difference between husbands and wives:
Actual distribution, U.S. 2009
Percentage of couples
12
10
8
6
4
2
0
-12 to -11
-10 to -9
-8 to -7
-6 to -5
-4 to -3
-2 to -1
Same
1 to 2
3 to 4
5 to 6
7 to 8
9 to 10
11 to 12
13 to 14
15 to 16
17 to 18
19 to 20
21 to 22
23 to 24
<-- Wife taller Husband taller -->
(Inches)

*


Height difference between husbands and wives:
Randomized distribution, U.S. 2009
Percentage of couples
10
9
8
7
6
5
4
3
2
1
0
-12 to -11
-10 to -9
-8 to -7
-6 to -5
-4 to -3
-2 to -1
Same
1 to 2
3 to 4
5 to 6
7 to 8
9 to 10
11 to 12
13 to 14
15 to 16
17 to 18
19 to 20
21 to 22
23 to 24
<-- Wife taller Husband taller -->
(Inches)

*


Height difference between husbands and wives:
Actual and randomized distribution, U.S. 2009
Percentage of couples
12
10
8
6
4
2
0
Random
Actual
More common than expected by chance
Less common than expected by chance
-12 to -11
-10 to -9
-8 to -7
-6 to -5
-4 to -3
-2 to -1
Same
1 to 2
3 to 4
5 to 6
7 to 8
9 to 10
11 to 12
13 to 14
15 to 16
17 to 18
19 to 20
21 to 22
23 to 24
<-- Wife taller Husband taller -->
(Inches)

*

Calculate Coefficient of Determination – cont.

That also means that 57% of the variation in occupational prestige cannot be explained by the amount of education people get.

Other factors affects how prestigious people's jobs are.

Can you think of some factors?

7. Now that we know what SSE is, we can turn to t-scores for our coefficients

Not quite yet.

Interlude -- Crash course in hypothesis testing

Inferential Stats

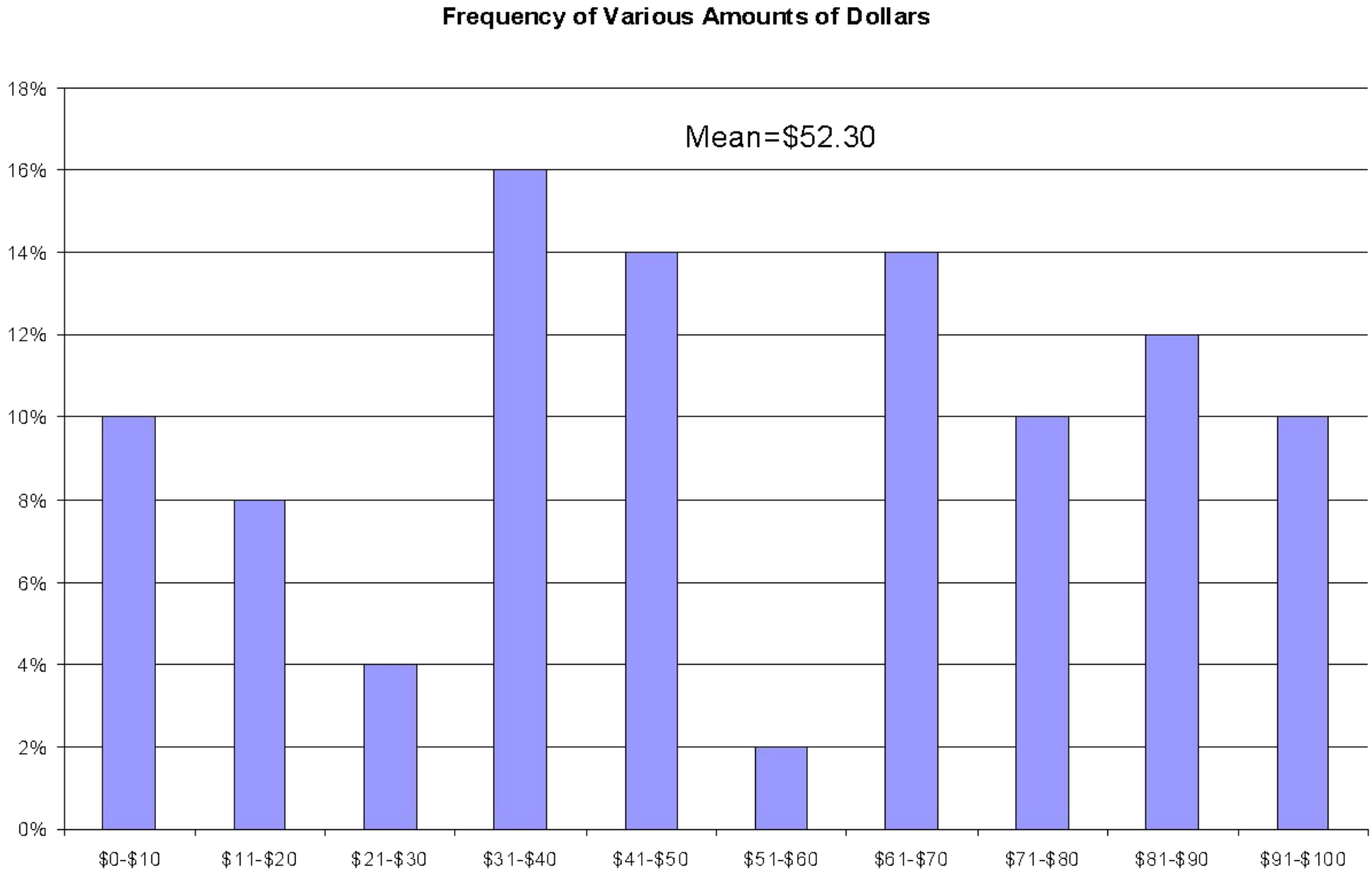
- Law of Large Numbers, the Normal Curve and Central Limit Theorem

The sampling distribution = just a probability distribution for a parameter

How much money do you have in your pocket right now? – My population

R's ID #	Amount of $ in Pocket
1	$ 38.69
2	$ 86.88
3	$ 67.95
4	$ 61.72
5	$ 36.86
6	$ 17.88
7	$ 39.83
8	$ 39.11
9	$ 88.91
10	$ 66.82
11	$ 90.93
12	$ 30.06
13	$ 45.23
…	…
50	$ 63.73

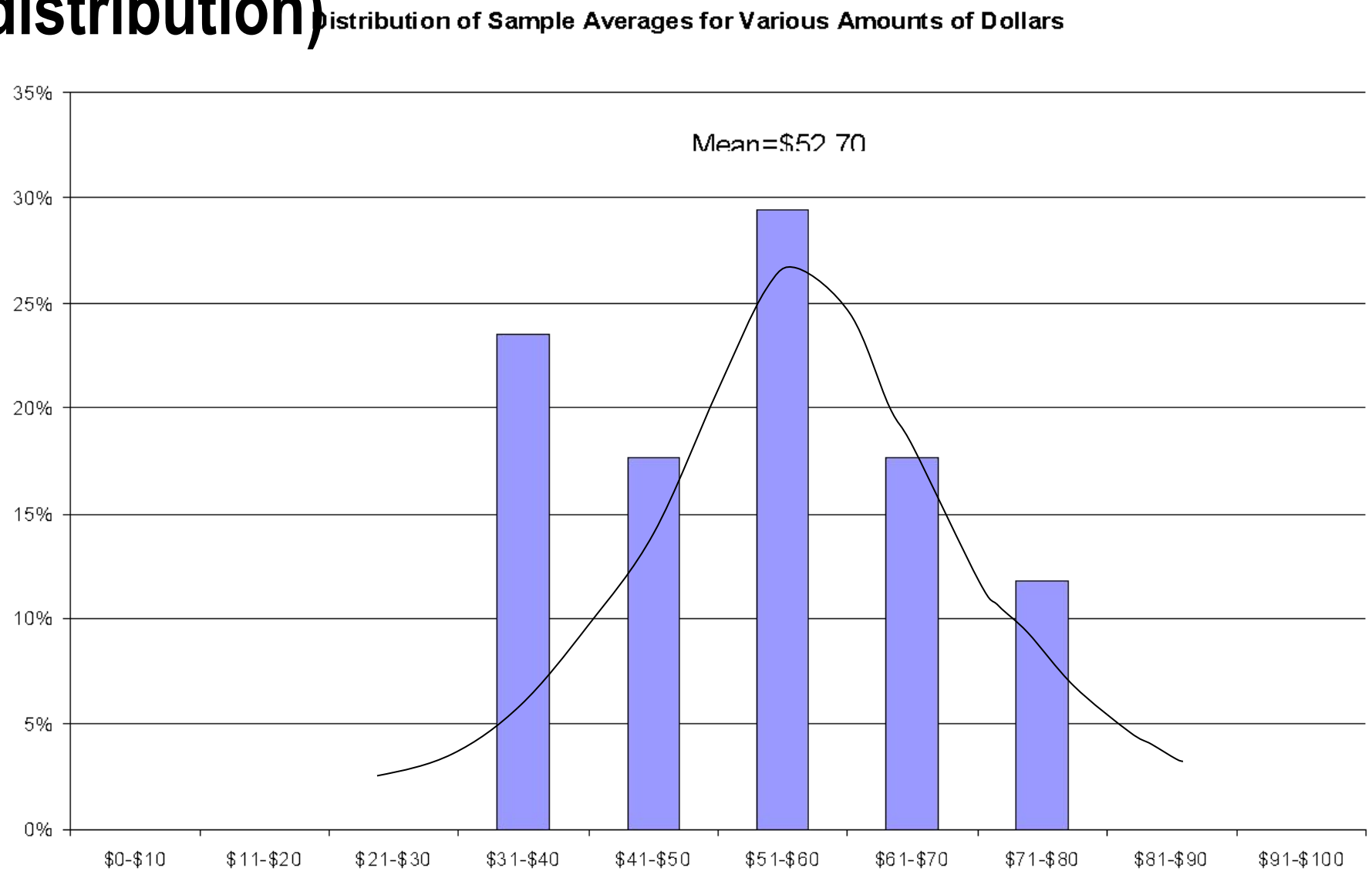
Histogram of the same data



Let's take samples of this data

	Mean of Each Sample	Sample		
		Value 1	Value 2	Value 3
N=3	$ 64.51	$ 38.69	$ 86.88	$ 67.95
N=3	$ 38.82	$ 61.72	$ 36.86	$ 17.88
N=3	$ 55.95	$ 39.83	$ 39.11	$ 88.91
N=3	$ 62.60	$ 66.82	$ 90.93	$ 30.06
N=3	$ 70.00	$ 45.23	$ 92.75	$ 72.04
N=3	$ 43.09	$ 2.52	$ 36.14	$ 90.60
N=3	$ 39.65	$ 32.52	$ 9.03	$ 77.39
N=3	$ 36.74	$ 42.95	$ 13.97	$ 53.31
N=3	$ 63.27	$ 45.46	$ 95.24	$ 49.10
N=3	$ 51.23	$ 31.80	$ 72.59	$ 49.31
N=3	$ 58.94	$ 26.90	$ 88.56	$ 61.36
N=3	$ 55.99	$ 5.43	$ 84.59	$ 77.96
N=3	$ 33.79	$ 2.46	$ 15.14	$ 83.77
N=3	$ 49.99	$ 45.27	$ 95.07	$ 9.63
N=3	$ 55.80	$ 70.55	$ 29.60	$ 67.25
N=3	$ 42.16	$ 18.47	$ 42.82	$ 65.21
N=2	$ 73.65	$ 83.57	$ 63.73	-

Histogram of sample averages (sampling distribution)



The Central Limit Theorem:
As sample size increases, the sampling distribution will approach normality, even if the population is not distributed normally, with mean, $\overline{y}$

So for instance …

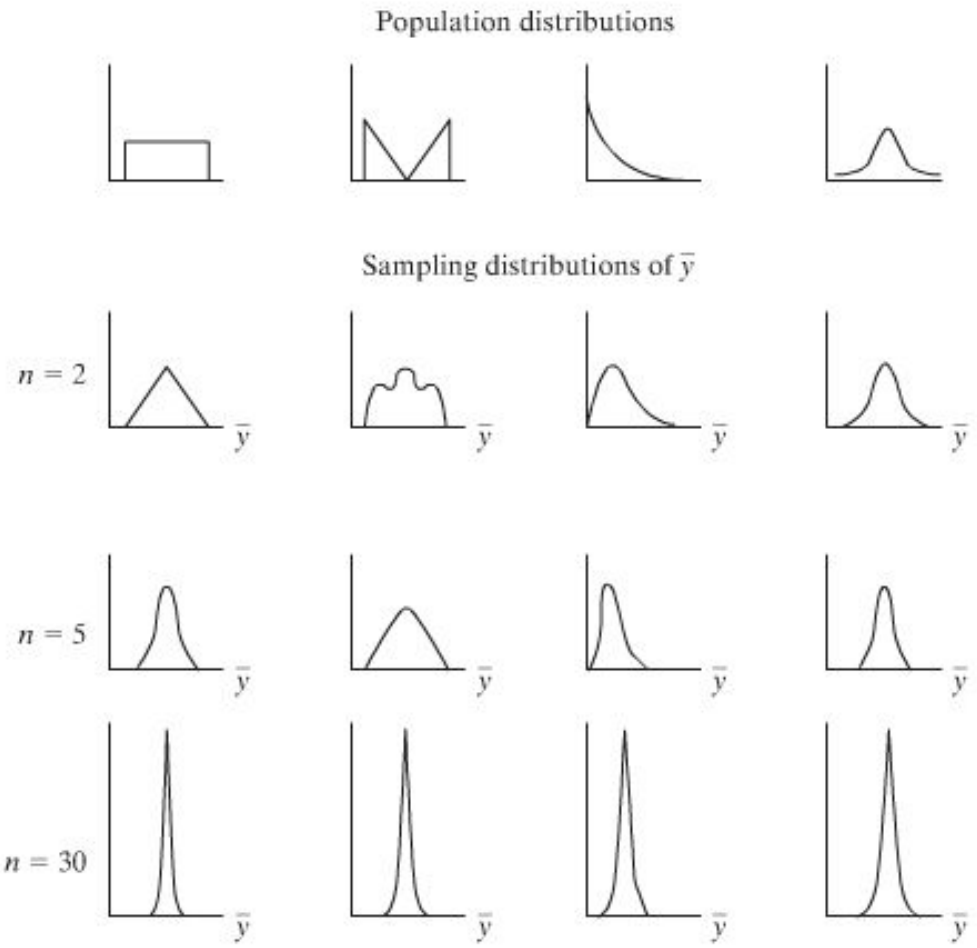


FIGURE 4.14: Four Different Population Distributions and the Corresponding Sampling Distributions of $\bar{y}$. As n increases, the sampling distributions get narrower and have more of a bell shape.

The sampling distribution's st. dev. (called the standard error) will approximate the population's st. dev., divided by the square root of the sample size, or:

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$$

As sample size increases, the sampling distribution's average (point estimate) will be a consistent estimator of the population's average

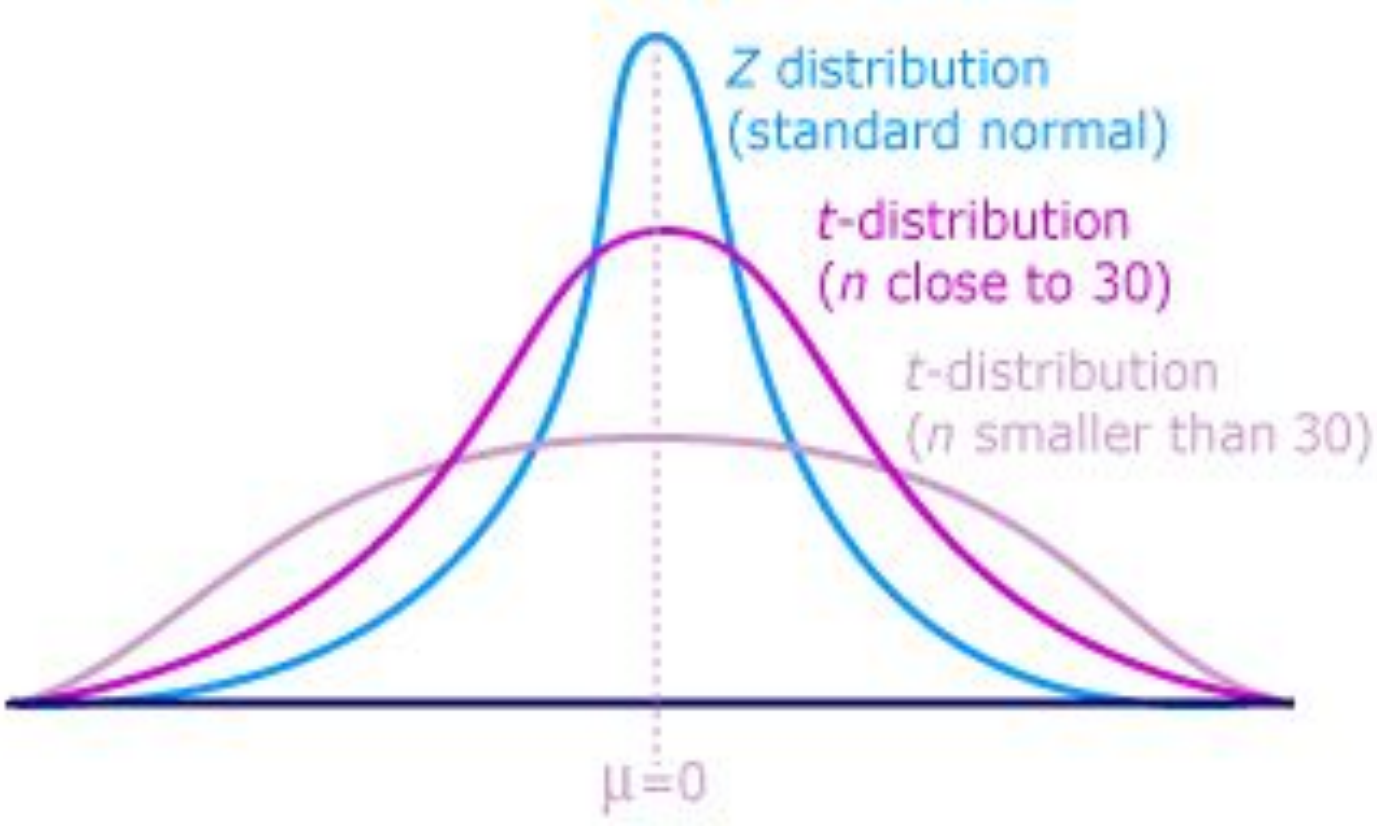
- The Law of Large Numbers

Example: simple t-test

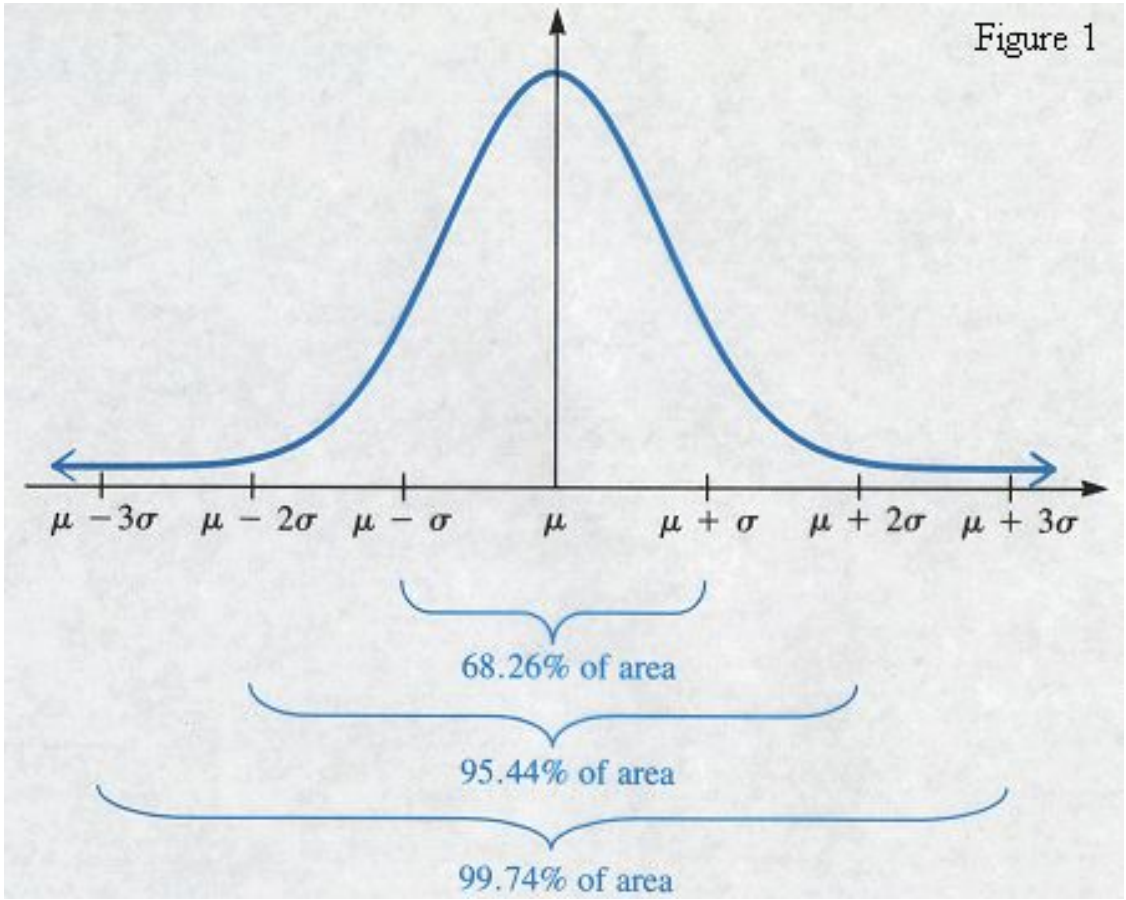
Difference between the means of 2 independent samples (with equal variances assumed)

What is a t-distribution?

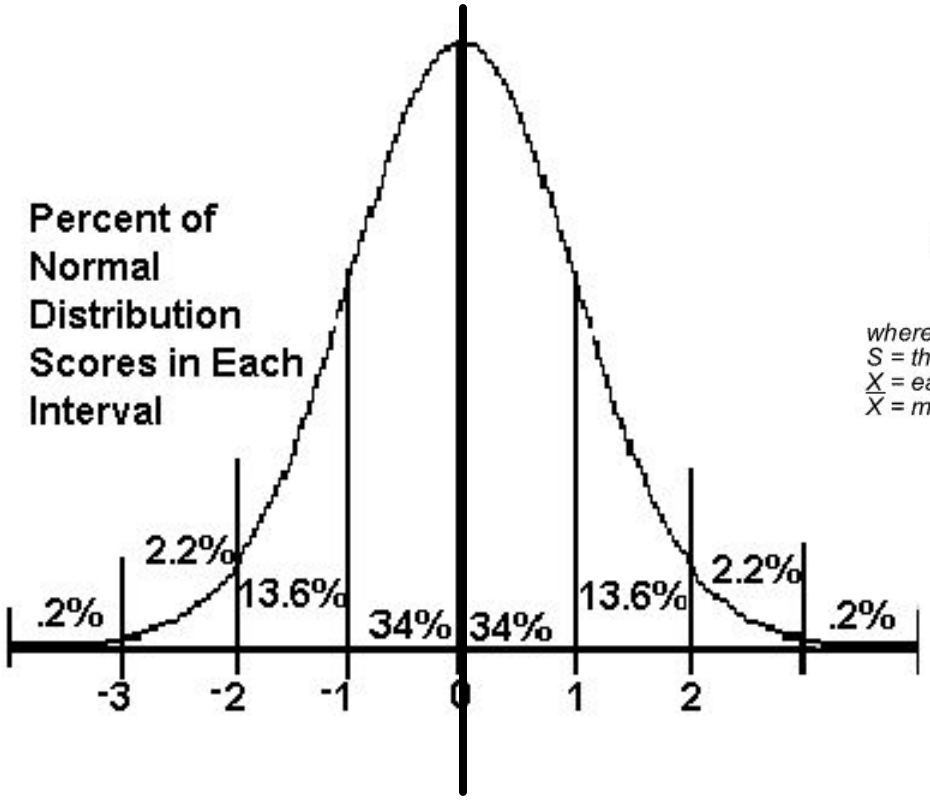
A normal curve that becomes the Z-distribution, as *n* approaches infinity



The normal curve



The standard normal



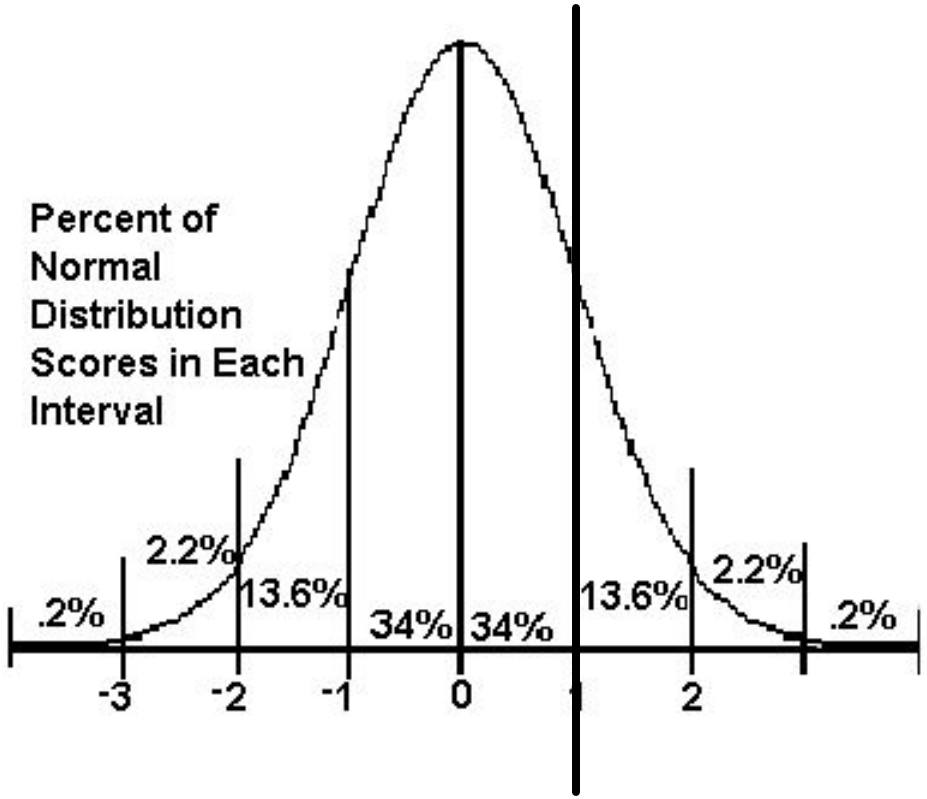
$$z = \frac{X - \overline{X}}{S}$$

where z is the standard score,
S = the standard deviation of a sample,
X = each value in the data set,
$\overline{X}$ = mean of all values in the data set.

For instance, if you have a Z-score of 0

Then 50% of the cases are below you and 50% of the cases are above you

The standard normal



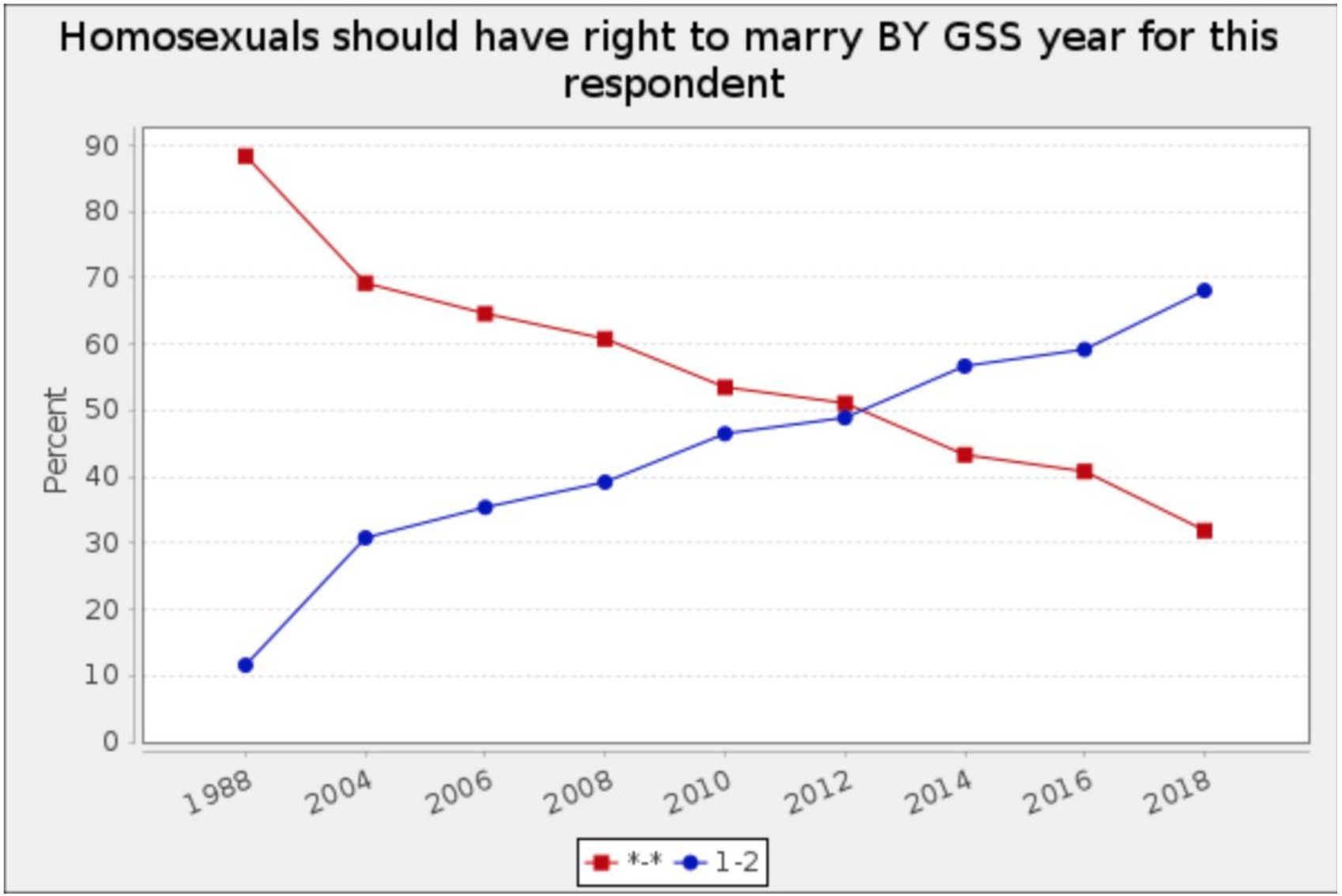
For instance, if you have a Z-score of +1

Then 84% (50+34) of the cases are below you and 16% (1-84) of the cases are above you

An example-- Do Americans with gay family members feel the same way about gay marriage as Americans without gay family members?

Is the difference between the means of these 2 groups really different from 0?

Overall context



Some recodes

Next, we are going to ask questions about people in your family, including relatives and in-laws. How many are gay men or women? 0? 1? 2-5? 6-10? More than 10?

```
## NOT:  GSS['anygayfam'] = np.where(GSS['acqfmgay']!= 1, 1, 0) b/c it does not handle missings well ##

col         = 'acqfmgay'
conditions  = [ GSS[col] >1, GSS[col] == 1 ]
choices     = [ 1, 0 ]

GSS["anys"] = np.select(conditions, choices, default=np.nan)
```

Do you disagree or agree? Homosexual couples should have the right to marry one another. 1. Strongly disagree … through 5. Strongly agree

```
GSS['agreegaymarr'] = 6- GSS['marhomo']
```

Summary stats first

It looks like the difference of opinion over gay marriage between Americans with gay relatives and Americans without gay relatives is -0.72 of a point (i.e., 2.42 - 3.14). Is 0.72 really ≈ 0, as the null suggests?

```
GSS[GSS['anys'] == 1]['agreegaymarr'].describe()

count    85.000000
mean      3.141176
std       1.582290
min       1.000000
25%       2.000000
50%       4.000000
75%       5.000000
max       5.000000
Name: agreegaymarr, dtype: float64
```

```
GSS[GSS['anys'] == 0]['agreegaymarr'].describe()

count    229.000000
mean       2.419214
std        1.530043
min        1.000000
25%        1.000000
50%        2.000000
75%        4.000000
max        5.000000
Name: agreegaymarr, dtype: float64
```

The t-test

Is -0.72 really ≈ 0, as the null suggests?

```
group1 = GSS[GSS['anys'] == 1]['agreegaymarr'].astype(float)
group2 = GSS[GSS['anys'] == 0]['agreegaymarr'].astype(float)


scipy.stats.ttest_ind(group1.dropna(), group2.dropna())

(3.6808654700551435, 0.00027379305308427215)
```

Is -0.72 really ≈ 0, as the null suggests?

It is highly unlikely we got such a large difference between these two groups just by chance. The t-value associated with this difference is -3.68. **If the null were true**, I would only get a t-value that large by chance 3.68 times in 10,000 tries due to sampling variability

```
scipy.stats.ttest_ind(group1.dropna(), group2.dropna())
```

```
(3.6808654700551435, 0.00027379305308427215)
```

where $t = \frac{M_1 - M_2}{\sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}}}$

Now for slopes ...

For slopes, the same logic as before ...

Based on the previous facts, we can argue that our sampling distribution of slopes for education is approximated by the t-distribution

For the B_{educ} coefficient from our regression, if there is a theoretical sampling distribution of B_{educs} if we kept running the same regression over new data over and over again. We want to test that sampling distribution of B_{educ}.

Remember our regression

```
lm = smf.ols(formula = 'PRESTG80~EDUC', data = df).fit()
print (lm.summary())
```

OLS Regression Results

Dep. Variable:	PRESTG80	R-squared:	0.426
Model:	OLS	Adj. R-squared:	0.344
Method:	Least Squares	F-statistic:	5.204
Date:	Thu, 18 May 2017	Prob (F-statistic):	0.0565
Time:	19:23:58	Log-Likelihood:	-31.583
No. Observations:	9	AIC:	67.17
Df Residuals:	7	BIC:	67.56
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Intercept	7.0769	18.389	0.385	0.712	-36.407	50.561
EDUC	3.0769	1.349	2.281	0.057	-0.112	6.266

Omnibus:	2.156	Durbin-Watson:	1.728
Prob(Omnibus):	0.340	Jarque-Bera (JB):	1.083
Skew:	-0.518	Prob(JB):	0.582
Kurtosis:	1.654	Cond. No.	82.5

What we are trying to do:

Testing the statistical significance of our coefficient

The t-test is to test the null hypothesis:

$B_{educ} = 0$ *[or, 3.08 ≈ 0]*

Against our alternative hypothesis:

$B_{educ} \neq 0$ *[or, 3.08 ≉ 0]*

How does this relate to hypothesis testing for my b_i?

- We can compare the t-statistic we actually get for our B_{educ} to a theoretical distribution of all the B_{educs} we could be expected to get if we kept rerunning the same regression model over new samples over and over again

- The null hypothesis says that the particular B_{educ} we got this time (B=3.08) is not very unusual; that we would have gotten a B_{educ} that large (in absolute value) by chance many, many, many times

Remember our regression results

```
lm = smf.ols(formula = 'PRESTG80~EDUC', data = df).fit()
print (lm.summary())
```

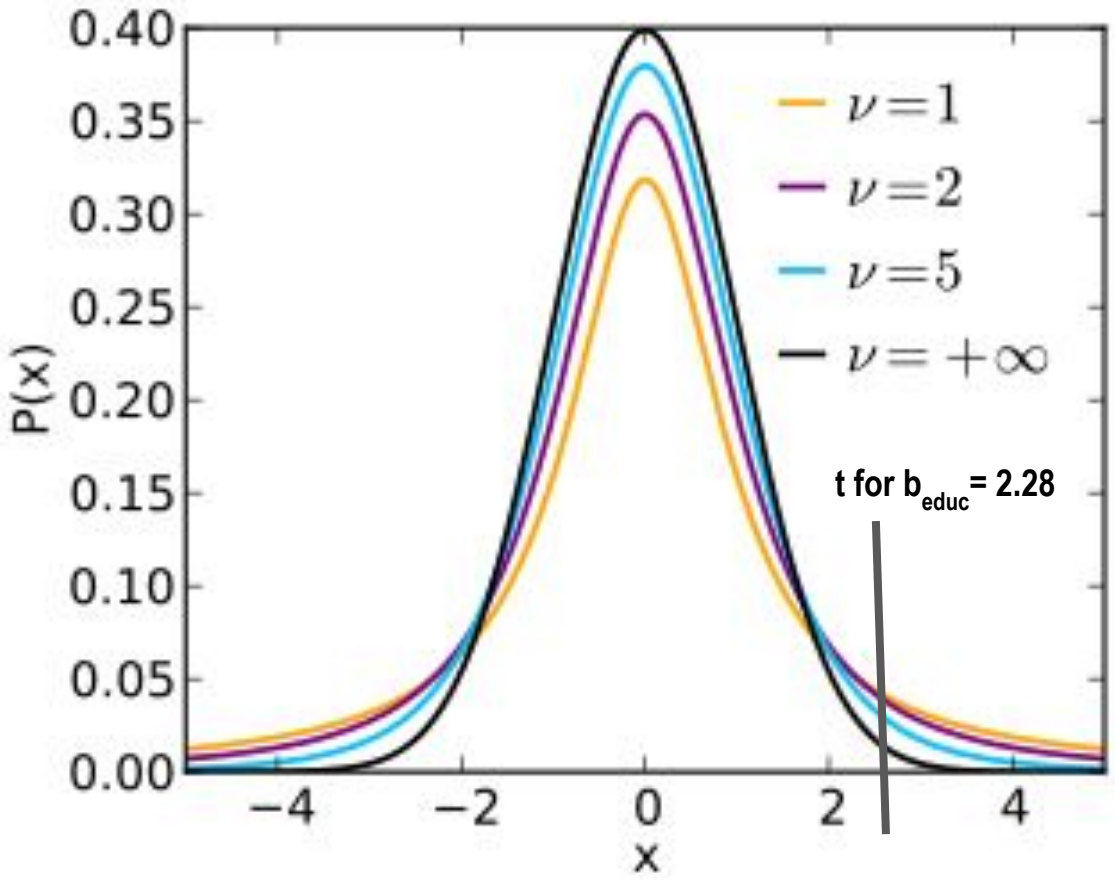
OLS Regression Results

Dep. Variable:	PRESTG80	R-squared:	0.426
Model:	OLS	Adj. R-squared:	0.344
Method:	Least Squares	F-statistic:	5.204
Date:	Thu, 18 May 2017	Prob (F-statistic):	0.0565
Time:	19:23:58	Log-Likelihood:	-31.583
No. Observations:	9	AIC:	67.17
Df Residuals:	7	BIC:	67.56
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Intercept	7.0769	18.389	0.385	0.712	-36.407	50.561
EDUC	3.0769	1.349	2.281	0.057	-0.112	6.266

Omnibus:	2.156	Durbin-Watson:	1.728
Prob(Omnibus):	0.340	Jarque-Bera (JB):	1.083
Skew:	-0.518	Prob(JB):	0.582
Kurtosis:	1.654	Cond. No.	82.5

How does our t-stat compare?



To get a t-stat as large as 2.28 (for n=9) for B_{educ} is pretty unlikely to have happened by chance

How does our t-stat compare?

Exactly how unlikely is t=2.28 (for n=9) to have occurred, just by chance, for B_{educ}?

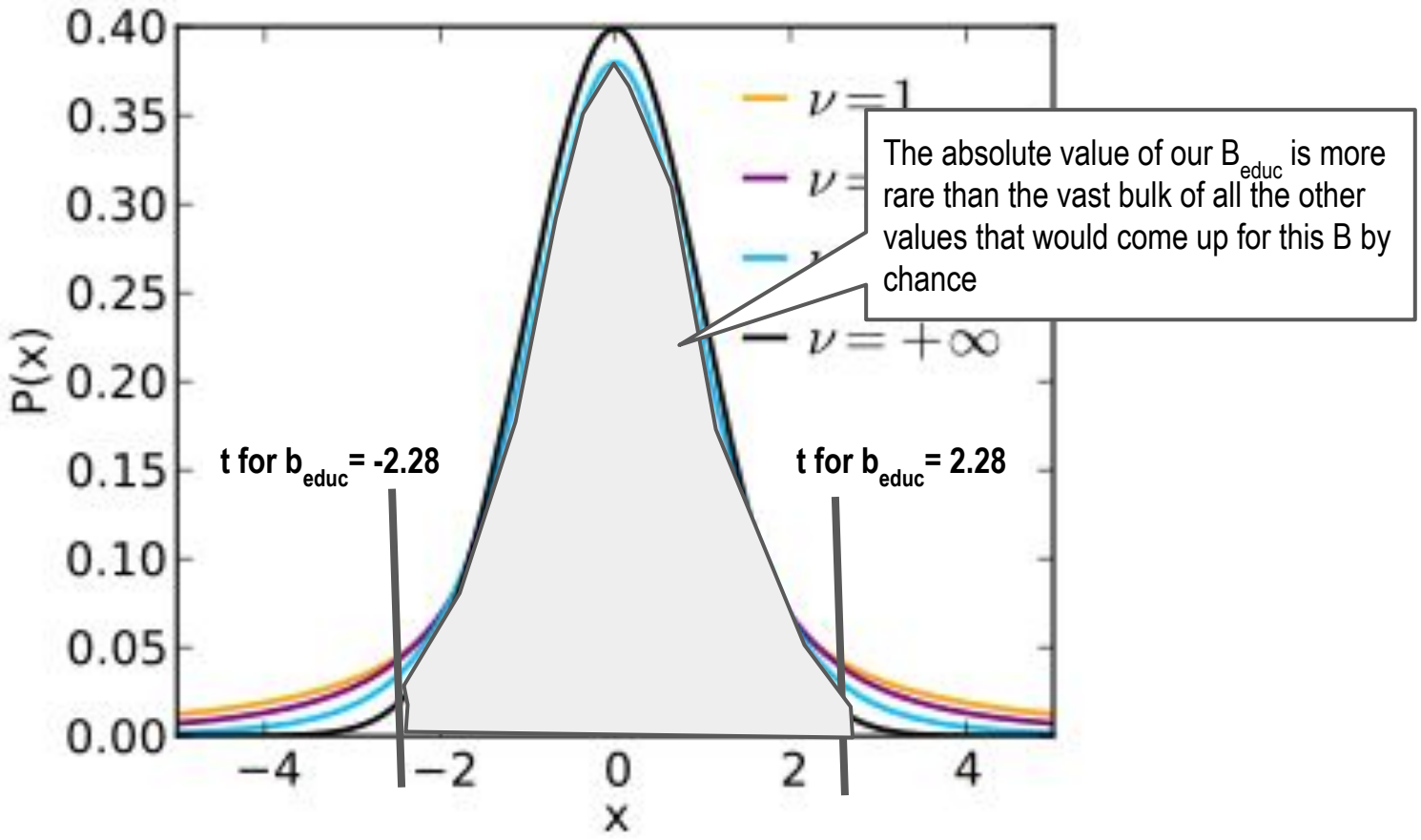
The p-value for that is found in the regression output and is $p<0.0565$. That means that there is less than a 6% chance that I would get a number as positive as 2.28 or as negative as -2.28 just by chance, assuming the null is correct.

How does our t-stat compare?

Given that a t=2.28 (for n=9) is only likely to happen by chance less than 6% of the time, we have strong evidence against the null.

Remember, the null argued that our B=3.08 is essentially really just 0. We can say with a fair bit of confidence that our B=3.08 does not equal 0.

Two-tailed test



We assume a two-tailed possibility for our t-stat of 2.28 (for n=9). That means that we consider extremity of outcomes in both directions possible.

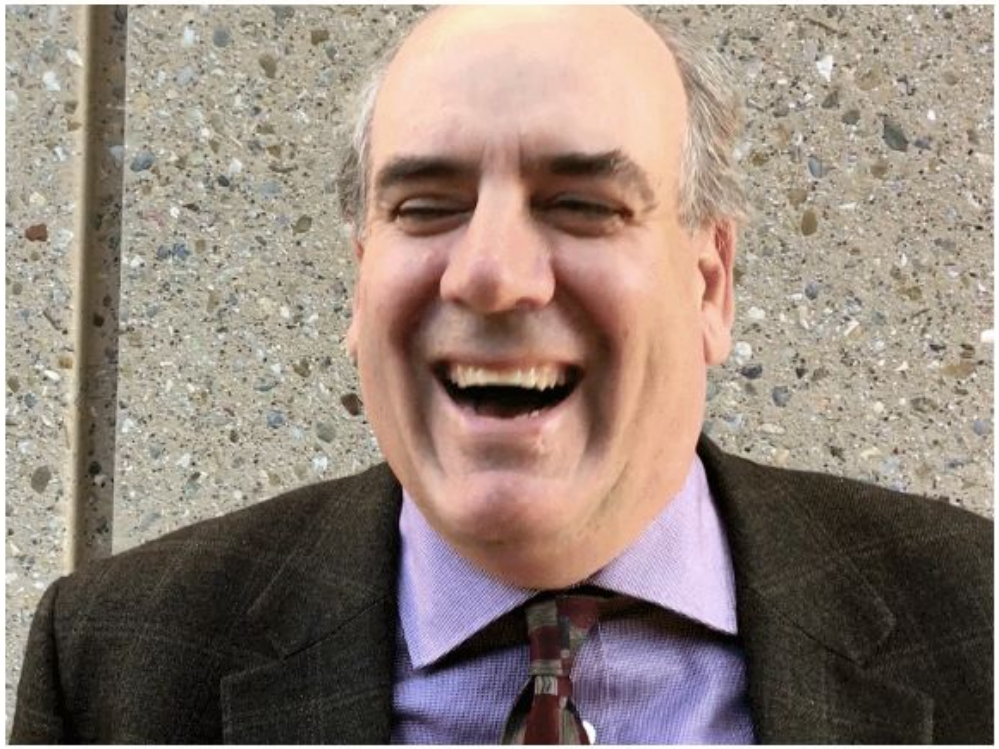
Remember … (of course)

NOV. 24, 2015 AT 12:12 PM

Not Even Scientists Can Easily Explain P-values

By Christie Aschwanden

Filed under Scientific Method



P-values have taken quite a beating lately. These widely used and commonly misapplied statistics have been blamed for giving a veneer of legitimacy to

t-value

ient?

8. How do I calculate a t-value for a regression coefficient?

Wait!

But I still don't know how we got the t-statistic of 2.28 to begin with. **Where did that come from?**

We need the Sum of Squared Errors (SSE) to figure out our t-statistic for B_{educ}.

Here is the regression output from R again

```
lm = smf.ols(formula = 'PRESTG80~EDUC', data = df).fit()
print (lm.summary())
```

OLS Regression Results

Dep. Variable:	PRESTG80	R-squared:	0.426
Model:	OLS	Adj. R-squared:	0.344
Method:	Least Squares	F-statistic:	5.204
Date:	Thu, 18 May 2017	Prob (F-statistic):	0.0565
Time:	19:23:58	Log-Likelihood:	-31.583
No. Observations:	9	AIC:	67.17
Df Residuals:	7	BIC:	67.56
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Intercept	7.0769	18.389	0.385	0.712	-36.407	50.561
EDUC	3.0769	1.349	2.281	0.057	-0.112	6.266

Omnibus:	2.156	Durbin-Watson:	1.728
Prob(Omnibus):	0.340	Jarque-Bera (JB):	1.083
Skew:	-0.518	Prob(JB):	0.582
Kurtosis:	1.654	Cond. No.	82.5

T-Test to Determine Statistical Significance of B_{educ}

$$t = \frac{b}{se} \quad = 3.08/1.35 = \mathbf{2.28}$$

where

$$se = \frac{s}{\sqrt{\sum (x - \bar{x})^2}}, \quad \text{where} \quad s = \sqrt{\frac{\text{SSE}}{n - 2}}$$

What is *s*?

$$s = \sqrt{\frac{\text{SSE}}{n - 2}}$$

This is otherwise known as the **root mean squared error**, or also **residual standard error** (in our case, it equals 9.17)

That means that our t-statistic is measured in quasi-standard deviation units

Here is the ANOVA output

```
lm = smf.ols(formula = 'PRESTG80~EDUC', data = df).fit()
print(sm.stats.anova_lm(lm, typ = 1))
```

	df	sum_sq	mean_sq	F	PR(>F)
EDUC	1	437.606838	437.606838	5.204159	0.056531
Residual	7	588.615385	84.087912	NaN	NaN

sqrt(84.09) = 9.17

The root mean squared error, or also residual standard error, is simply the square root of SSE

Is b statistically significant?

```
lm = smf.ols(formula = 'PRESTG80~EDUC', data = df).fit()
print (lm.summary())
```

OLS Regression Results

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Intercept	7.0769	18.389	0.385	0.712	-36.407	50.561
EDUC	3.0769	1.349	2.281	0.057	-0.112	6.266

Omnibus:	2.156	Durbin-Watson:	1.728
Prob(Omnibus):	0.340	Jarque-Bera (JB):	1.083
Skew:	-0.518	Prob(JB):	0.582
Kurtosis:	1.654	Cond. No.	82.5

The null claims that this coefficient (for educ) is the same as zero (or, t=0), but we find that this coefficient has a t-score of 2.28.

Is b statistically significant?

```
lm = smf.ols(formula = 'PRESTG80~EDUC', data = df).fit()
print (lm.summary())
```

OLS Regression Results

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Intercept	7.0769	18.389	0.385	0.712	-36.407	50.561
EDUC	3.0769	1.349	2.281	0.057	-0.112	6.266

Omnibus:	2.156	Durbin-Watson:	1.728
Prob(Omnibus):	0.340	Jarque-Bera (JB):	1.083
Skew:	-0.518	Prob(JB):	0.582
Kurtosis:	1.654	Cond. No.	82.5

I would get a coefficient with a t-score of 2.28 by chance only 5.7 times out of 100 (when my sample size is 9). This is just shy of providing at least 95% confidence in my estimate (p-value = 0.057).

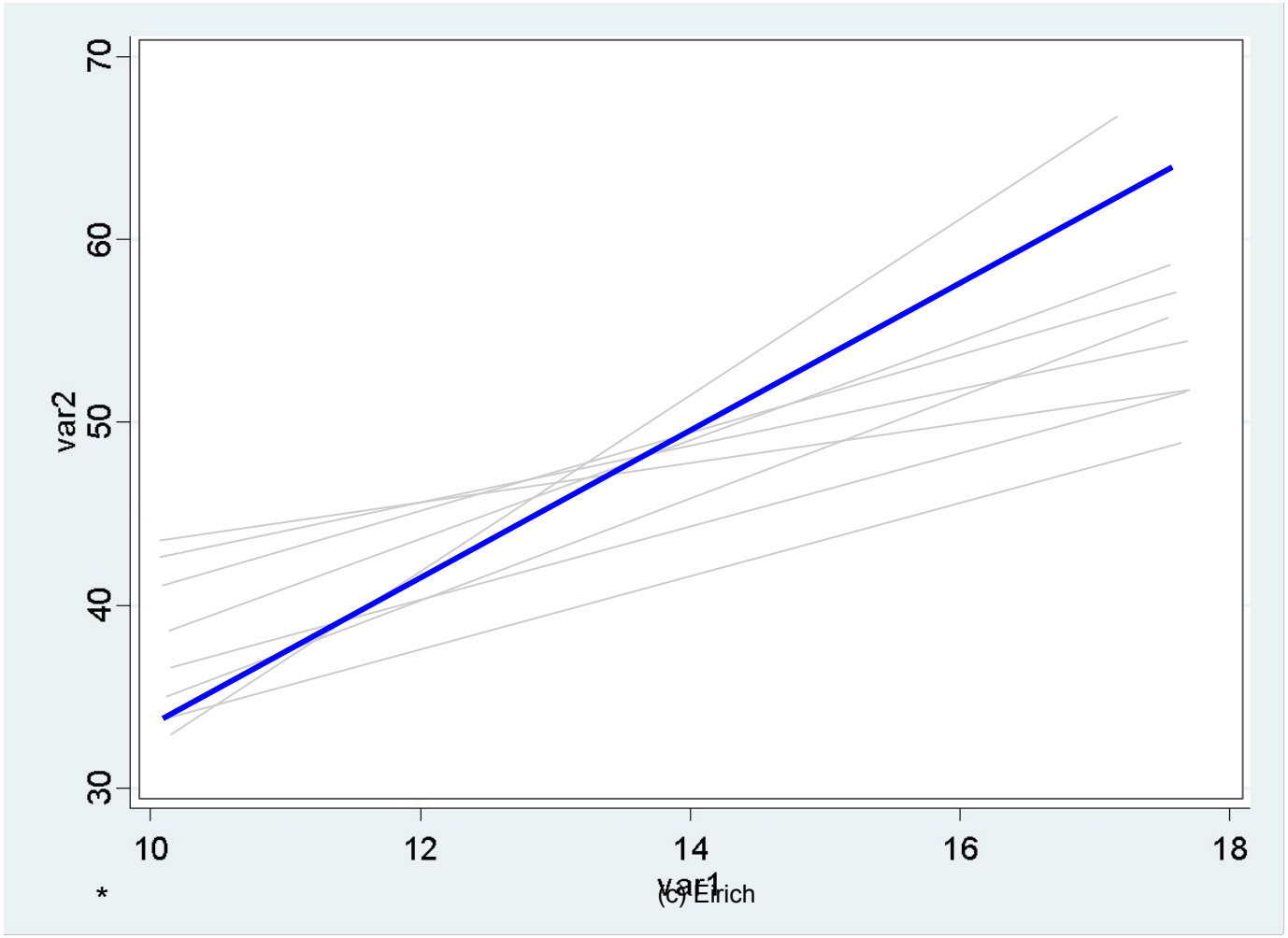
Is b statistically significant?

```
conf_95 = lm.conf_int(alpha = 0.05)
conf_95.columns = ["2.5%", "97.5%"]
conf_95
```

	2.5%	97.5%
Intercept	-36.407151	50.560997
EDUC	-0.112437	6.266283

Likewise, the 95% confidence intervals around this educ coefficient are (-0.11, 6.27). Because the estimate straddles zero, we cannot be sufficiently (i.e., 95%) certain that the coefficient is not really zero -- but it is barely ever less than zero, too. Hence, it is marginally significant.

We have just the blue line, but all the grey lines are quite possible too



Remember … (of course)

« How tall is Jon Lee Anderson?
"I have no idea who Catalina Garcia is, but she makes a decent ruler": I don't know if John Lee "little twerp" Anderson actually suffers from tall-person syndrome, but he is indeed tall »

Misunderstanding the p-value

Posted by Andrew on 12 March 2013, 11:55 am

he New York Times has a feature in its Tuesday science section, Take a Number, to which I occasionally contribute (see here and here).

Today's column, by Nicholas Balakar, is in error. The column begins:

> When medical researchers report their findings, they need to know whether their result is a real effect of what they are testing, or just a random occurrence. To figure this out, they most commonly use the p-value.

This is wrong on two counts. First, whatever researchers might feel, this is something they'll never know. Second, results are a combination of real effects and chance, it's not either/or.

Perhaps the above is a forgivable simplification, but I don't think so; I think it's a simplification that destroys the reason for writing the article in the first place. But in any case I think there's no excuse for this, later on:

> By convention, a p-value higher than 0.05 usually indicates that the results of the study, however good or bad, were probably due only to chance.

This is the old, old error of confusing p(A|B) with p(B|A). I'm too rushed right now to explain this one, but it's in just about every introductory statistics textbook ever written. For more on the topic, I recommend my recent paper, P Values and Statistical Practice, which begins:

> The casual view of the P value as posterior probability of the truth of the null hypothesis is false and not even close to valid under any reasonable model, yet this misunderstanding persists even in high-stakes settings (as discussed, for example, by Greenland in 2011). The formal view of the P value as a probability conditional on the null is mathematically correct but typically irrelevant to research goals (hence, the popularity of alternative—if wrong—interpretations). . . .

Our interpretation of p-values does not mean what we think it means

9. t-test for correlation

Is r statistically significant?

```
scipy.stats.pearsonr(df["EDUC"], df["PRESTG80"])

# the first number is the correlation coefficient, and the second is
p-value
```

```
(0.653012270802549 87, 0.05653095224376862 6)
```

These are the exact same t-statistics (and p-values) from our earlier regression for B_{educ} because the square root of R2 = |ρ|, in bivariate regressions (i.e., with only 1 *X* and 1 *Y*) … sqrt(0.426) = 0.653

Where's the justification?

- Where's the justification for (a) B being unbiased, (b) efficient, (c) that the residuals should sum to 0, (d) that the residuals have constant variance, and so on?

10. An example with ordinal variables

One ordinal variable …

DEGREE

Percent	N	Value	Label
22.8	12,074	0	LT HIGH SCHOOL
51.6	27,310	1	HIGH SCHOOL
5.2	2,774	2	JUNIOR COLLEGE
13.8	7,273	3	BACHELOR
6.5	3,447	4	GRADUATE
	30	8	DK
	135	9	NA
100	**53,043**		**Total**

… with another ordinal …

SPANKING: FAVOR SPANKING TO DISCIPLINE CHILD.

Do you strongly agree, agree, disagree, or strongly disagree that it is sometimes necessary to discipline a child with a good, hard, spanking?

Percent	N	Value	Label
27.6	5,264	1	STRONGLY AGREE
47.4	9,016	2	AGREE
18.3	3,485	3	DISAGREE
6.7	1,274	4	STRONGLY DISAGREE

The regression

```
## Here we use our previously imported Pandas and Statsmodels packages.##

GSS = pd.read_csv("GSS_Cum.csv")
d = GSS[GSS.year == 2006] # subset on the year 2006
spank_lm = smf.ols(formula = 'spanking ~ degree', data = d).fit()
print (spank_lm.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                spanking   R-squared:                       0.021
Model:                            OLS   Adj. R-squared:                  0.020
Method:                 Least Squares   F-statistic:                     41.85
Date:                Thu, 18 May 2017   Prob (F-statistic):           1.24e-10
Time:                        19:25:50   Log-Likelihood:                -2501.9
No. Observations:                1970   AIC:                             5008.
Df Residuals:                    1968   BIC:                             5019.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept      1.9137      0.032     59.568      0.000          1.851     1.977
degree         0.1043      0.016      6.469      0.000          0.073     0.136
==============================================================================
Omnibus:                       87.112   Durbin-Watson:                   1.874
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               91.313
Skew:                           0.500   Prob(JB):                     1.48e-20
Kurtosis:                       2.664   Cond. No.                         3.87
==============================================================================

Warnings:
```

Interpretation

By moving up one educational category, a person (on average) will score 0.104*** higher points on the spanking scale, expressing increasing disagreement.

Last key point: You have lots of discretion in regression

Article Menu Close

Download PDF

Full Article

Content List

Abstract

Crowdsourcing Data Analysis: Skin Tone and Red Cards in Soccer

Disclosures

Stages of the Crowdsourcing

Supplemental Material

Figures & Tables

Article Metrics

Cite | Share | Request Permissions

Related Articles

Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results

R. Silberzahn, E. L. Uhlmann, D. P. Martin, more...

Show all authors

First Published August 23, 2018 | Research Article | Check for updates

https://doi.org/10.1177/2515245917747646

Article information

Abstract

Twenty-nine teams involving 61 analysts used the same data set to address the same research question: whether soccer referees are more likely to give red cards to dark-skin-toned players than to light-skin-toned players. Analytic approaches varied widely across the teams, and the estimated effect sizes ranged from 0.89 to 2.93 (*Mdn* = 1.31) in odds-ratio units. Twenty teams (69%) found a statistically significant positive effect, and 9 teams (31%) did not observe a significant relationship. Overall, the 29 different analyses used 21 unique combinations of covariates. Neither analysts' prior beliefs about the effect of interest nor their level of expertise readily explained the variation in the outcomes of the analyses. Peer ratings of the quality of the analyses also did not account for the variability. These findings suggest that significant variation in the results of analyses of complex data may be difficult to avoid, even by experts with honest intentions. Crowdsourcing data analysis, a strategy in which numerous research teams are recruited to simultaneously investigate the same research question, makes transparent how defensible, yet subjective, analytic choices influence research results.

Keywords

crowdsourcing science, data analysis, scientific transparency, open data, open materials

*

Is the stats glass half empty or half full?

Team	Analytic Approach	Odds Ratio
12	Zero-Inflated Poisson Regression	0.89
17	Bayesian Logistic Regression	0.96
15	Hierarchical Log-Linear Modeling	1.02
10	Multilevel Regression and Logistic Regression	1.03
18	Hierarchical Bayes Model	1.10
31	Logistic Regression	1.12
1	OLS Regression With Robust Standard Errors, Logistic Regression	1.18
4	Spearman Correlation	1.21
14	WLS Regression With Clustered Standard Errors	1.21
11	Multiple Linear Regression	1.25
30	Clustered Robust Binomial Logistic Regression	1.28
6	Linear Probability Model	1.28
26	Hierarchical Generalized Linear Modeling With Poisson Sampling	1.30
3	Multilevel Logistic Regression Using Bayesian Inference	1.31
23	Mixed-Model Logistic Regression	1.31
16	Hierarchical Poisson Regression	1.32
2	Linear Probability Model, Logistic Regression	1.34
5	Generalized Linear Mixed Models	1.38
24	Multilevel Logistic Regression	1.38
28	Mixed-Effects Logistic Regression	1.38
32	Generalized Linear Models for Binary Data	1.39
8	Negative Binomial Regression With a Log Link	1.39
20	Cross-Classified Multilevel Negative Binomial Model	1.40
13	Poisson Multilevel Modeling	1.41
25	Multilevel Logistic Binomial Regression	1.42
9	Generalized Linear Mixed-Effects Models With a Logit Link	1.48
7	Dirichlet-Process Bayesian Clustering	1.71
21	Tobit Regression	2.88
27	Poisson Regression	2.93

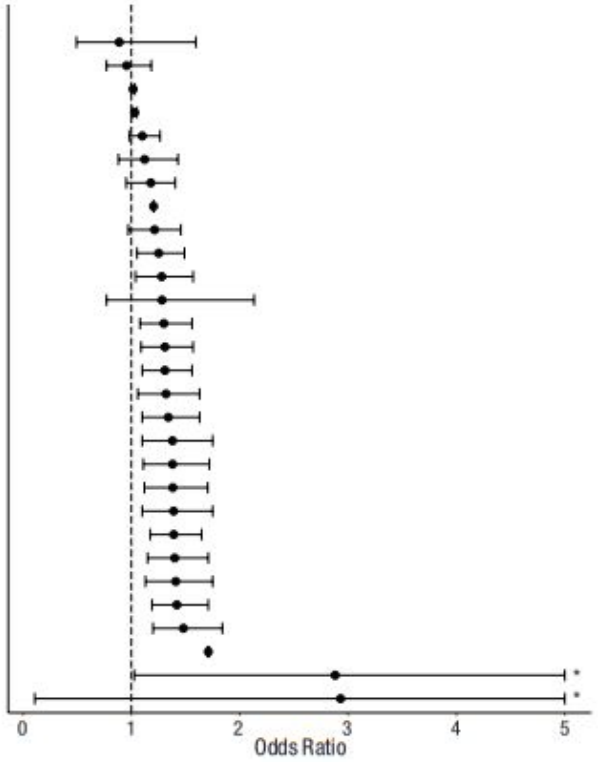


Fig. 2. Point estimates (in order of magnitude) and 95% confidence intervals for the effect of soccer players' skin tone on the number of red cards awarded by referees. Reported results, along with the analytic approach taken, are shown for each of the 29 analytic teams. The teams are ordered so that the smallest reported effect size is at the top and the largest is at the bottom. The asterisks indicate upper bounds that have been truncated to increase the interpretability of the plot; the actual upper bounds of the confidence intervals were 11.47 for Team 21 and 78.66 for Team 27. OLS = ordinary least squares; WLS = weighted least squares.