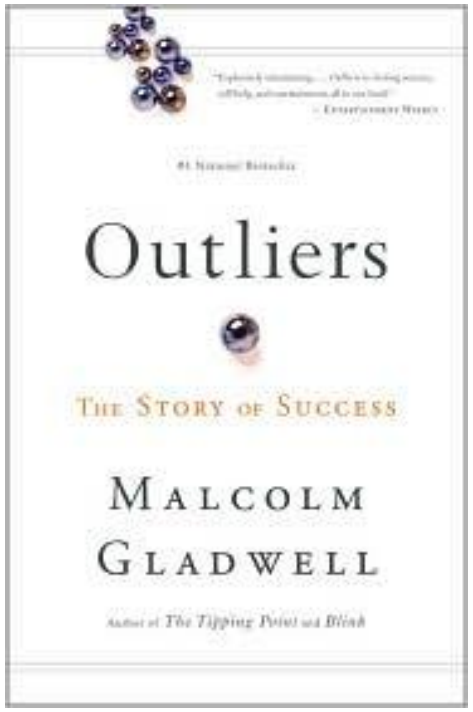
# Data Analysis for the Social Sciences

Gregory M. Eirich

*Lecturer in Discipline*

***Columbia University***




# What do kids' birthdays and NHL player success have in common?

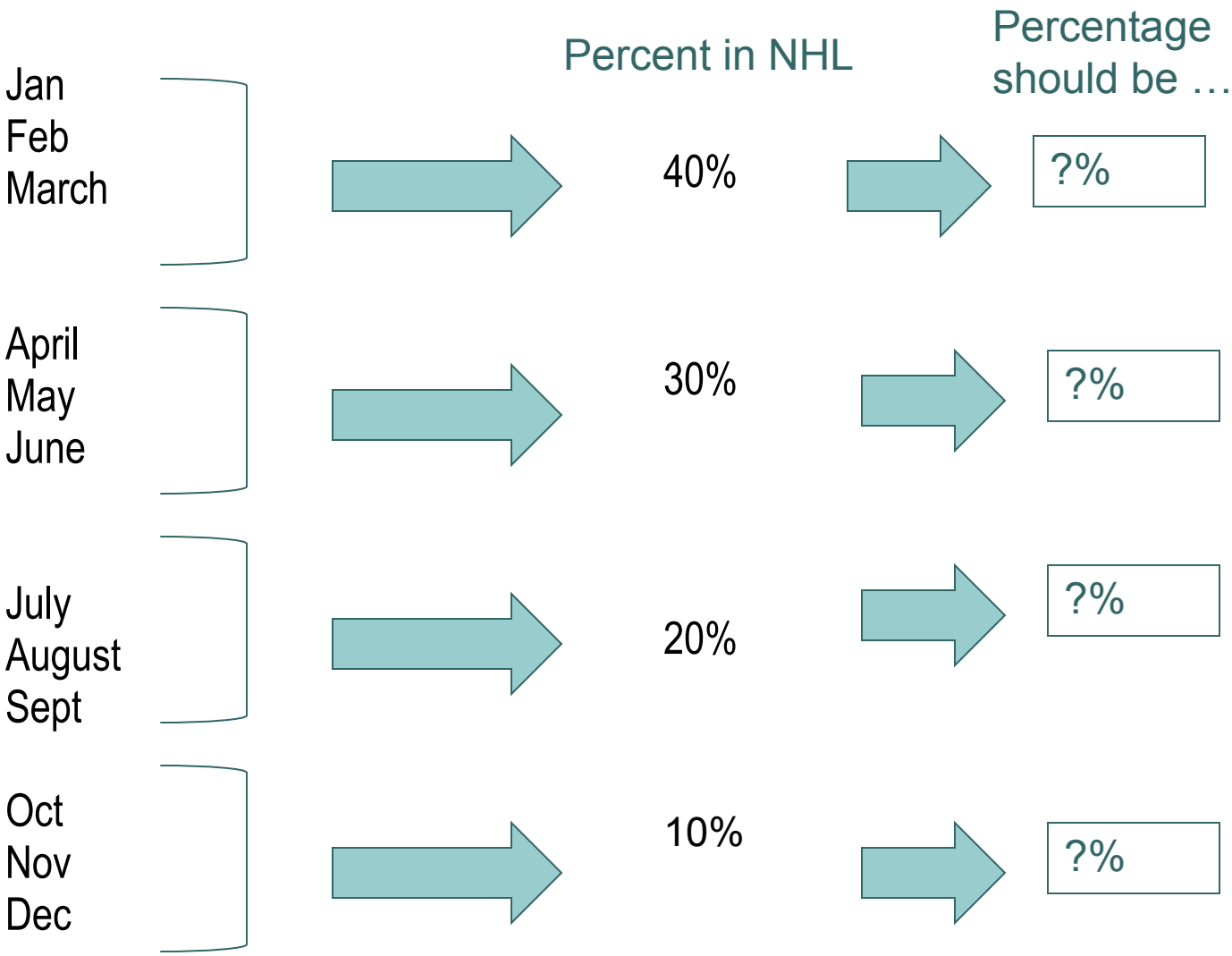
Here is the play-by-play for the first two goals in the Memorial Cup final, only this time I've substituted the players' birthdays for their names. It no longer sounds like the championship of Canadian junior hockey. It now sounds like a strange sporting ritual for teenage boys born under the astrological signs Capricorn, Aquarius, and Pisces.

> *March 11 starts around one side of the Tigers' net, leaving the puck for his teammate January 4, who passes it to January 22, who flips it back to March 12, who shoots point-blank at the Tigers' goalie, April 27. April 27 blocks the shot, but it's rebounded by Vancouver's March 6. He shoots! Medicine Hat defensemen February 9 and February 14 dive to block the puck while January 10 looks on helplessly. March 6 scores!*

Let's go to the second period now.

> *Medicine Hat's turn. The Tigers' scoring leader, January 21, charges down the right side of the ice. He stops and circles, eluding the Vancouver defenseman February 15. January 21 then deftly passes the puck to his teammate December 20—wow! what's* he *doing out there?!—who shrugs off the onrushing defender May 17 and slides a cross-crease pass back to January 21. He shoots! Vancouver defenseman March 12 dives, trying to block the shot. Vancouver's goalie, March 19, lunges helplessly. January 21 scores! He raises his hands in triumph. His teammate May 2 jumps on his back with joy.*

# Birthdays determine NHL access



# Course Expectations

# You will be able to …

1. Interpret/critique statistical results
2. Improve “number sense,” esp. w/ descriptive data
3. Understand quantitative data structures and look for data-sets
4. Perform statistical operations on the computer
5. Be confident in ability to write a thesis

# Goal 1 - Interpret/critique statistical results

# Can you interpret this table?

Table 1: Effect of Daughters on Party Identification – Linear Probability Models

| VARIABLES | (1) Democrat | (2) Republican | (3) Republican Scale | (4) Democrat[@] | (5) Republican[@] | (6) Republican Scale[@] |
|---|---|---|---|---|---|---|
| Proportion Female | -0.146** | 0.110** | 0.532** | -0.140** | 0.099* | 0.493** |
| | (0.042) | (0.041) | (0.172) | (0.042) | (0.040) | (0.171) |
| Constant | 0.548** | 0.330** | -0.444** | 0.943** | -0.275** | -2.548** |
| | (0.025) | (0.024) | (0.103) | (0.099) | (0.095) | (0.404) |
| Observations | 1076 | 1076 | 1062 | 1072 | 1072 | 1058 |
| R-squared | 0.011 | 0.007 | 0.009 | 0.038 | 0.054 | 0.048 |

** p<0.01, * p<0.05, + p<0.1
Standard errors in parentheses

[@]Controls for: Female; Protestant; Age; Education; Married
All models are un-weighted, limited to those with biological children, and the proportion of female children excludes non-biological children. There is no significant interaction between gender and proportion girls.
Results are the same when using Washington's methods – including number of girls and controlling for total number of children as opposed to using proportion girls. However, the magnitude is smaller; the coefficient on number of girls is about half that of proportion girls in each model. Number of biological children is significantly associated with Democratic identification and may be endogenous.

Source: THE EFFECT OF DAUGHTERS ON PARTISANSHIP. Dalton Conley et al.

# What about with the abstract

ABSTRACT

Washington (2008) finds that, controlling for total number of children, each additional daughter makes a member of Congress more likely to vote liberally and attributes this finding to socialization. However, daughters' influence could manifest differently for elite politicians and the general citizenry, thanks to the selection gradient particular to the political process. This study asks whether the proportion of female biological offspring affects political party identification. Using nationally-representative data from the General Social Survey, we find that female offspring induce more conservative political identification. We hypothesize that this results from the change in reproductive fitness strategy that daughters may evince.

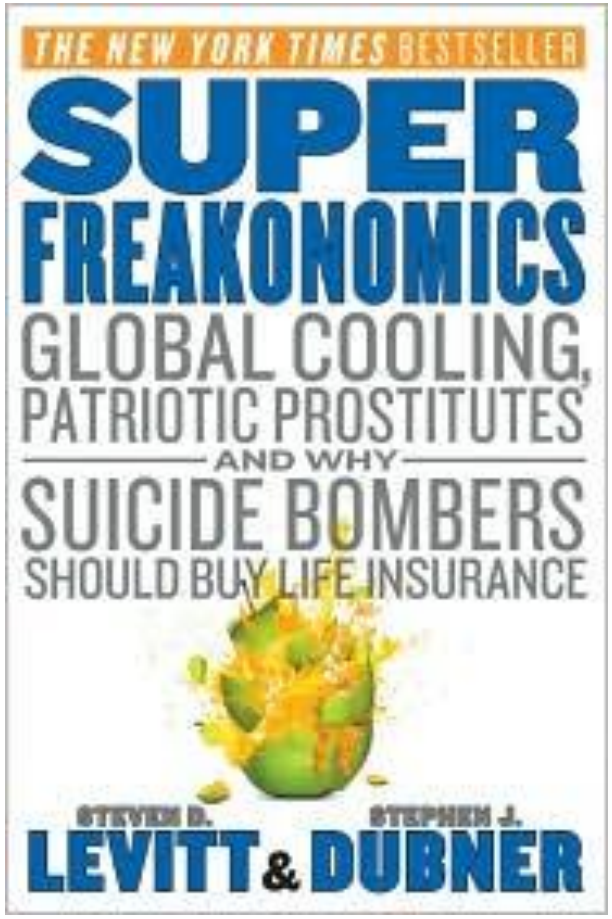
Source: THE EFFECT OF DAUGHTERS ON PARTISANSHIP. Dalton Conley et al.

# How could this study be critiqued or improved on?

- Quantitative studies can be critiqued in many ways: study design, variable creation, model selection, robustness, etc. – you should be comfortable with these criticisms

# Goal 2 - Improve “number sense,” esp. w/ descriptive data

# Guesstimating




- Which is safer for you … to drive drunk or walk drunk?

# Everybody says: driving drunk is more unsafe for the driver

- A drunk driver is 13x more likely to cause an accident than a sober one
- 30%+ of all fatal crashes involve at least one drunk driver. At late night, that proportion rises to 60%.
- 1 of every 140 miles is driven drunk, or 21 billion miles each year.

# But …

- Each year, 1,000+ drunk pedestrians die in traffic accidents (out of 13k total traffic deaths)

- But on a per-mile basis, is it more dangerous to drive drunk or walk drunk?

# Estimating miles walked drunk

- Americans walk ~1/2-mile per day outside the home or workplace. 237 million Americans are aged 16+. That's 43 billion miles walked each year by people of driving age.
- Let's *assume* that 1 of every 140 of those miles are walked drunk (the same proportion of miles that are driven drunk)
- That's 307 million miles are walked drunk each year.

# The big punch-line …

| | Miles Drunk | Deaths | Death Rate Per Mile Drunk |
|---|---|---|---|
| Drive | 21,000,000,000 | 8,550 | 0.0000004 |
| Walk | 307,000,000 | 1,000 | 0.0000033 |

- On a per-mile basis, a drunk walker is *8*x *more likely* to get killed than a drunk driver

Source:
http://www.huffingtonpost.com/2009/10/26/superfreakonomics-on-drun_n_333490.html

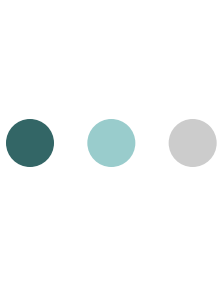
# Caveat

- This math applies to the driver's chances of death, not other people on the road
- Probably more likely that drunk driver will kill others than a drunk walker
- Other comments?

Source:
http://www.huffingtonpost.com/2009/10/26/superfreakonomics-on-drun_n_333490.html

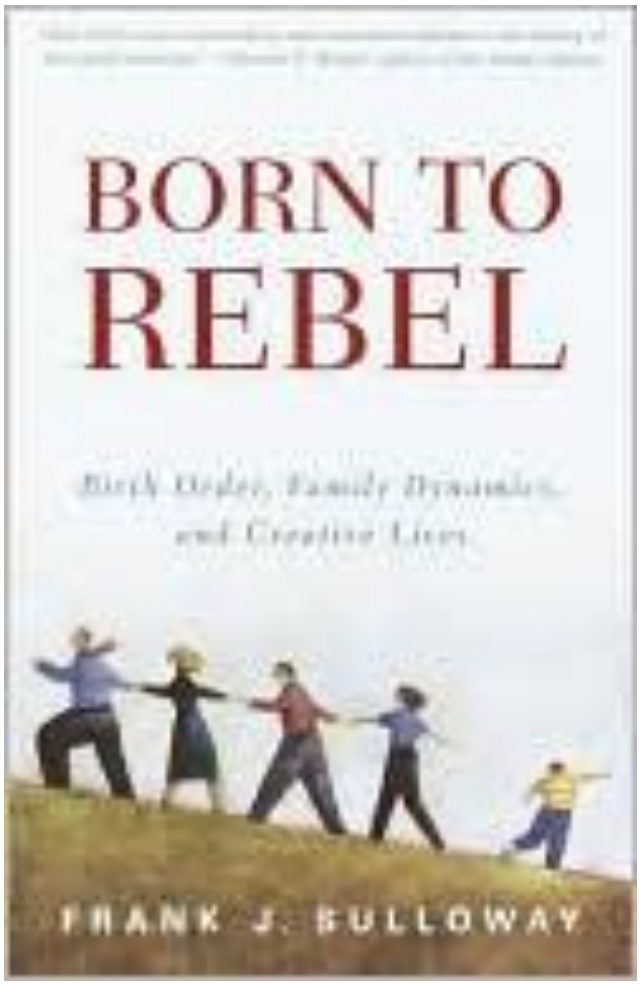
# Goal 2 - Improve "number sense," esp. w/ descriptive data

- You should gain an intuition about numbers – ratios, percentages, decimals, odds, probabilities, coefficients, p-values, etc.
- You should be able to "guess-timate" the answer based on a sense of quantities – this is critical to capture basic errors and typos



# Goal 3 - Understand quantitative data structures and look for data-sets

# A tantalizing idea




- Historical evidence suggests that oldest born tend to be conservative, severe and leaders
- Younger born tend to be innovative, creative and liberal-minded

# Looking to test Sullaway

- Jeremy Freese finds “true love” in the basement of the UW-Madison library
- Freese wanted to see if Sullaway’s theory held up for “regular” people, not just famous people throughout history
- He needed a dataset to do this
- His object of affection: Special Module of the General Social Survey (1994)

# A story of dataset "romance"

- He had criteria for a "ideal companion" – Nationally representative, current, information on birth order, interviewed multiple siblings, and psychological/personality varaibles
- GSS had everything except psychological/personality varaibles … but it solicited opinions that could be categorized as "liberal" *vs*. "conservative"

# A story of dataset "romance"

## REBEL WITHOUT A CAUSE OR EFFECT: BIRTH ORDER AND SOCIAL ATTITUDES

**Jeremy Freese**
*Indiana University*

**Brian Powell**
*Indiana University*

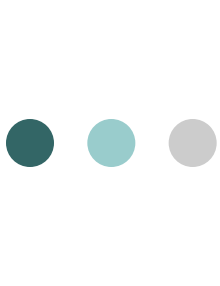
**Lala Carr Steelman**
*University of South Carolina*

*The enduring effects of an individual's birth order have been subject to a long and lively debate in sociology and other disciplines. Recently, in response to Sulloway's (1996)* Born to Rebel: Birth Order, Family Dynamics, and Creative Lives, *interest has increased in the possible effects of birth order on social attitudes. Using quantitative, historical data, Sulloway found that birth order is a better predictor of social attitudes than is gender, class, or race. His novel, evolutionary theory asserts the universal influence of birth order across eras and cultures. We use contemporary data to test Sulloway's contention that firstborn adults are more conservative, supportive of authority, and "tough-minded" than laterborns. Examining 24 measures of social attitudes from the General Social Survey (GSS), we find no support for these claims, either in terms of significant effects or even the direction of nonsignificant coefficients. An expanded inquiry using all (202) relevant attitudinal items on the GSS yields similar results. In our analysis, variables discounted by Sulloway—gender, race, social class, and family size—are all linked to social attitudes more strongly than is birth order. Our findings suggest that birth-order theories may be better conceptualized in terms of modest effects in limited domains and in specific societies.*

Sociologists of the family have long tried and of being raised in single- versus two-par-

# Goal 3 - Understand quantitative data structures and look for data-sets

- There is a logic to quantitative data structures – you should learn to "see" it, both when you design your own and in understanding others
- You would not believe the amount of data that is out there … you should be able to search for it and understand what variables are in it



# Goal 4 - Perform statistical operations on the computer

# Goal 4 - Perform statistical operations on the computer

- You will learn Python to view, recode and manipulate data on the computer
- You will learn how to evoke commands
- You will learn how to interpret read-outs

# Goal 5 - Confident in ability to write a thesis

# Theses

- One of the keys to a good thesis is to tell a story with numbers
- You want to try to convince your reader as you would, if you were a lawyer in a trial

# Actually, a lots of theses started out here … for instance …

- Political intolerance of marginalized groups in the US
- The effect of the financial crisis on American happiness
- Role of educational attainment on generalized trust in the US
- The role of organizational culture in job satisfaction and employee turnover intention
- A comparison of the economic performance of Jewish World War II veterans to non-Jewish World War II veterans

# You will be able to …

1. Interpret/critique statistical results
2. Improve “number sense,” esp. w/ descriptive data
3. Understand quantitative data structures and look for data-sets
4. Perform statistical operations on the computer
5. Confident in ability to write a thesis

# Data Analysis for the Social Sciences

(Part 2)

Gregory M. Eirich

*Lecturer in Discipline*

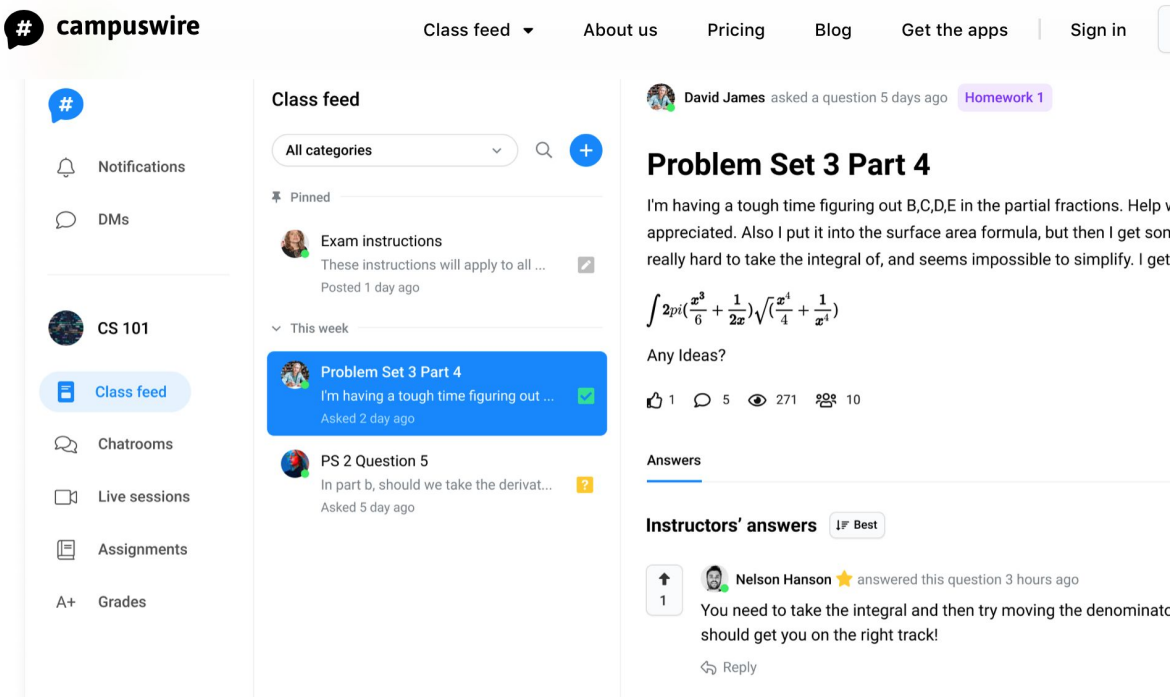
***Columbia University***

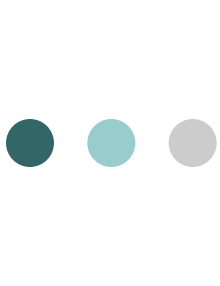
# Expect a survey from me

This will help me start to get to know you.

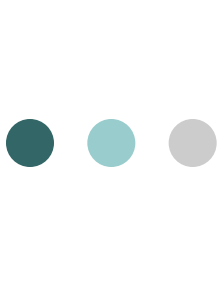
# Expect a Campuswire invite

We will use Campuswire for almost all of our communications, moving forward.





# Let’s talk about statistics …



# 1. How statisticians view the world

# How statisticians view the world

1. VARIABLES
   - We take things that are “whole” and rip them apart into variables
   - E.g., We don’t care about Kate as a whole person … We care about her components – What color hair does she have? How tall is she? Etc.
   - We compare and correlate variables together

# How statisticians view the world

2. APPROXIMATIONS

- We have concepts in our theories, but we cannot directly "see" them – How do we see *power* or *status* or *class*?
- We instead must measure indicators that approximate our concepts
- As such, statistics is always a means to test our concepts and theories – not an end in itself … and always approximate

# How statisticians view the world

3. UNCERTAINTY AND PROBABILITY

- Everything has a certain probability of happening
- We care about things work “on average” – not that we can predict the outcome of any given case
- All outcomes are compared against the idea of “could have happened by chance”

# Gelman talks about it too



## The fallacy of the excluded middle — statistical philosophy edition

Posted by Andrew on 18 August 2018, 9:12 am

I happened to come across this post from 2012 and noticed a point I'd like to share again. I was discussing an article by David Cox and Deborah Mayo, in which Cox wrote:



> The three challenges of statistical inference are:
>
> 1. *Generalizing from sample to population* and from past to future, problems which are associated with survey sampling and forecasting but actually arise in nearly every application of statistical inference;
> 2. *Generalizing from control to treatment group*, a problem which is associated with causal inference, which is implicitly or explicitly part of the interpretation of most regressions we have seen; and
> 3. *Generalizing from observed measurements to the underlying constructs of interest*, as most of the time our data do not quite record exactly what we would ideally like to study.
>
> All three of these challenges can be framed as problems of prediction (for new people or new items that are not in the sample, for future outcomes under different potentially assigned treatments, and for underlying constructs of interest, if they could be measured exactly).

# "Identifying testable implications of their own assumptions"- Judea Pearl

Not secure | causality.cs.ucla.edu/blog/index.php/2014/10/27/are-economists-smarter-than-epidemiologists-comments-on-imbenss-recent-paper/

Causal Analysis in Theory and Practice

October 27, 2014

## Are economists smarter than epidemiologists? (Comments on Imbens's recent paper)

Filed under: Discussion,Economics,Epidemiology,General — eb @ 4:45 pm

In a recent survey on Instrumental Variables (link), Guido Imbens fleshes out the reasons why some economists "have not felt that graphical models have much to offer them."

His main point is: "In observational studies in social science, both these assumptions [exogeneity and exclusion] tend to be controversial. In this relatively simple setting [3-variable IV setting] I do not see the causal graphs as adding much to either the understanding of the problem, or to the analyses." [page 377]

What Imbens leaves unclear is whether graph-avoiding economists limit themselves to "relatively simple settings" because, lacking graphs, they cannot handle more than 3 variables, or do they refrain from using graphs to prevent those "controversial assumptions" from becoming transparent, hence amenable to scientific discussion and resolution.

When students and readers ask me how I respond to people of Imbens's persuasion who see no use in tools they vow to avoid, I direct them to the post "The deconstruction of paradoxes in epidemiology", in which Miquel Porta describes the "revolution" that causal graphs have spawned in epidemiology. Porta observes: "I think the "revolution — or should we just call it a renewal"? — is deeply changing how epidemiological and clinical research is conceived, how causal inferences are made, and how we assess the validity and relevance of epidemiological findings."

So, what is it about epidemiologists that drives them to seek the light of new tools, while economists (at least those in Imbens's camp) seek comfort in partial blindness, while missing out on the causal revolution? Can economists do in their heads what epidemiologists observe in their graphs? Can they, for instance, identify the testable implications of their own assumptions? Can they decide whether the IV assumptions (i.e., exogeneity and exclusion) are satisfied in their own models of reality? Of course the can't; such decisions are intractable to the graph-less mind. (I have challenged them repeatedly to these tasks, to the sound of a pin-drop silence)

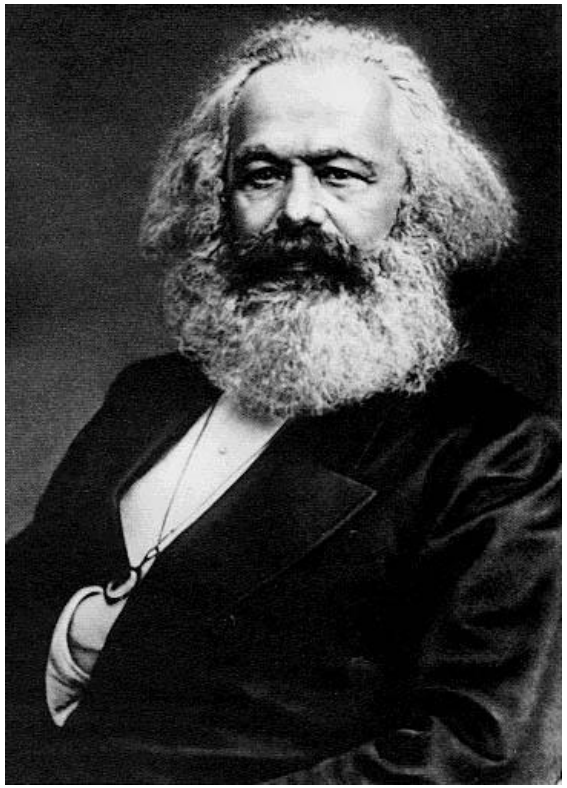
Or, are problems in economics different from those in epidemiology? I have examined the structure of typical problems in the two fields, the number of variables involved, the types of data available, and the nature of the research questions. The problems are

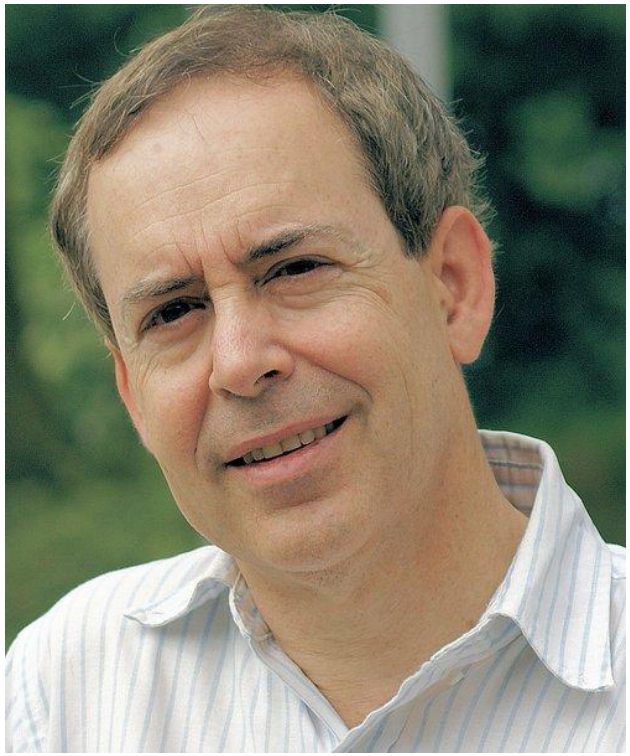
# Types of variables, part 1

There are many types of variables:

- *Nominal/Categorical* – Names of categories, with no ranking. E.g., race, religion, sex
- *Ordinal* – A ranked (or ordered) set of responses that go from more to less. E.g., very dissatisfied (1) -> very satisfied (5)
- *Interval/Ratio* – A numerical value that is counted/measured, with a "true zero." E.g., no. of siblings, annual income, hrs. worked last week

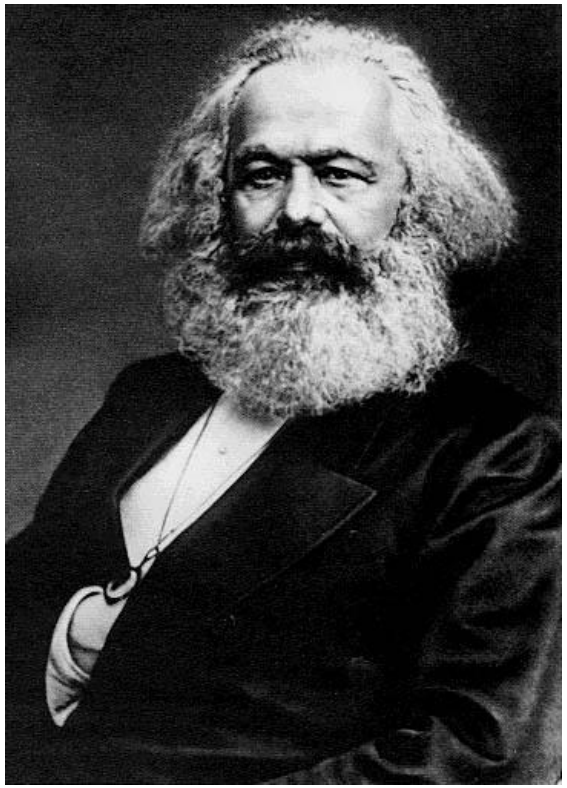
# Q: In stratification, how'd we get from Marx to DiPrete?

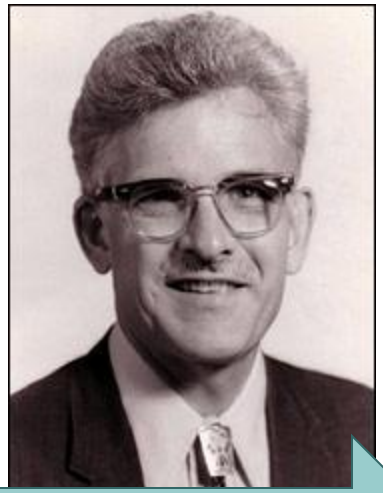




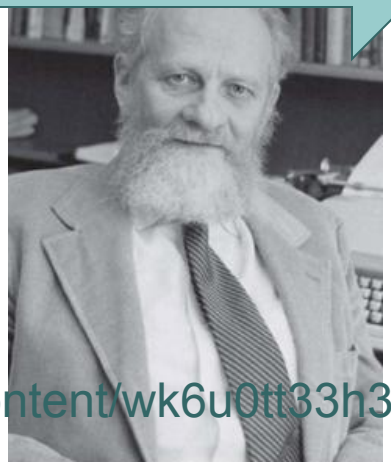
# A: Changing *measures* of class

Nominal

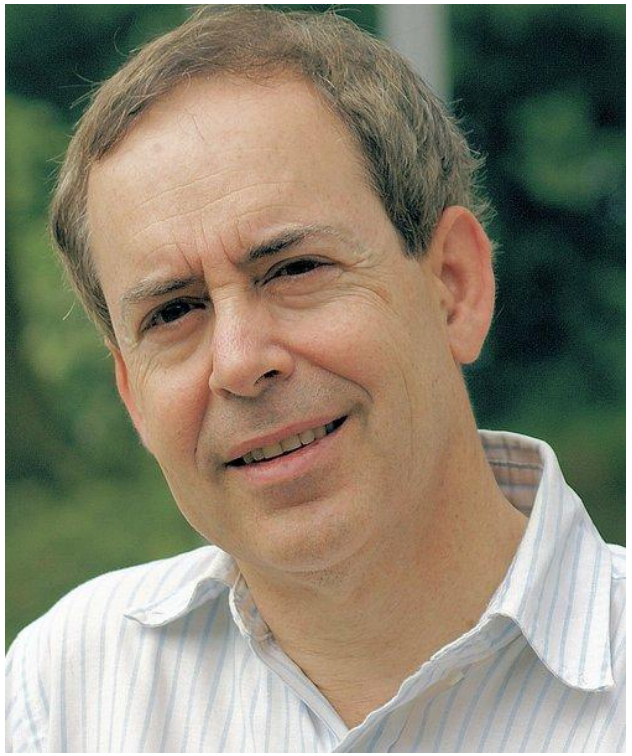




Ordinal



Interval-Ratio



http://www.springerlink.com/content/wk6u0tt33h375002/

# Why do we care about types of variables, part 1?

1. Different types of variables have different levels of precision (1 category vs. 5 categories)
2. Certain statistical operations can only be done on certain types of variables. E.g., You cannot use regression with nominal-only variables
3. We are going to learn to recode variables from one type to another

# Types of variables, part 2

Another type of variables:

- *Independent* – This is thought to be the "cause" variable; E.g., education
- *Dependent* – This is thought to be the "effect" variable; E.g., income
- *Control* – This is thought to be another possible causal factor in addition to the independent variable; E.g., parental SES

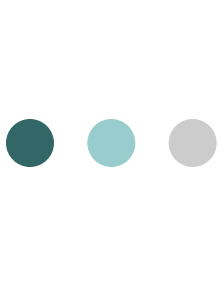
# Why do we care about types of variables, part 2?

1. The essence of social statistics, as it is practiced, is to pick an independent variable and examine its impact on some outcome (dependent variable)
2. We must be careful because it is often hard to figure out which is really the independent and which is the dependent variable

# CAUTION: Ordinal variables can be "double agents" between #s and categories

Ordinal variables rely on a key assumption:

- *Equal distance between responses*



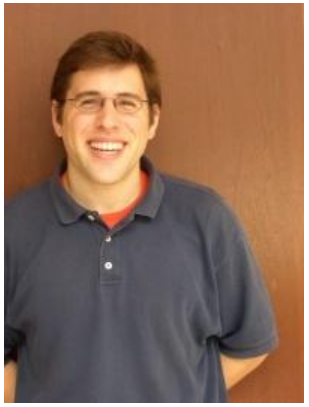
# 2. It’s all about randomization

# It's all about randomization

- All statistics works because we believe that the outcomes should be randomly distributed
- … if not, then we argue that a social structure or a social process is at work

# Simple randomization

- As in a experiment …
- An example: Matt Salganik …



- … & Duncan Watts

# Simple randomization, cont.

- More than 14,000 participants at Music Lab (www.musiclab.columbia.edu)

- Listened to, rated and, if they chose, downloaded songs by bands they had never heard of

- Some of the participants saw only the names of the songs and bands

- But others also saw how many times the songs had been downloaded by previous participants

# Simple randomization, cont.

• This second group — in the “social influence” condition — was further split into eight parallel “worlds” such that participants could see the prior downloads of people only in their own world.

• S & W didn’t manipulate any of these rankings — all the artists in all the worlds started out identically, with zero downloads — but because the different worlds were kept separate, they subsequently evolved independently of one another.

# Simple randomization, cont.

- In all the social-influence worlds, the most popular songs were much more popular (and the least popular songs were less popular) than in the independent condition.

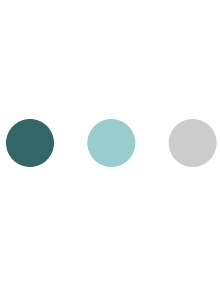
- The particular songs that became hits were different in different worlds

# Simple randomization, cont.

- The impact of a listener's own reactions is easily overwhelmed by his or her reactions to others. The song "Lockdown," by 52metro, for example, ranked 26th out of 48 in quality; yet it was the No. 1 song in one social-influence world, and 40th in another.

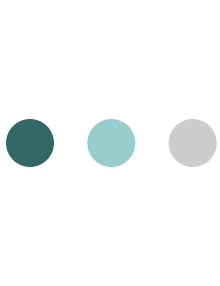
- Overall, a song in the Top 5 in terms of quality had only a 50 percent chance of finishing in the Top 5 of success.


BIT
SOCIAL RESEARCH
BY
in the DIGITAL AGE
BIT
MATTHEW J. SALGANIK



# So that's the benefit of experiments

# We usually work with observation data, though.



# 3. Descriptive Stats

# Central tendency of the data

- Mean
- Median
- Mode

# Mean

- Uses as much information as each observation can provide

HOW MUCH MONEY CONTRIBUTED TO RELIGIOUS ORG

**$1,383.60**

# Median

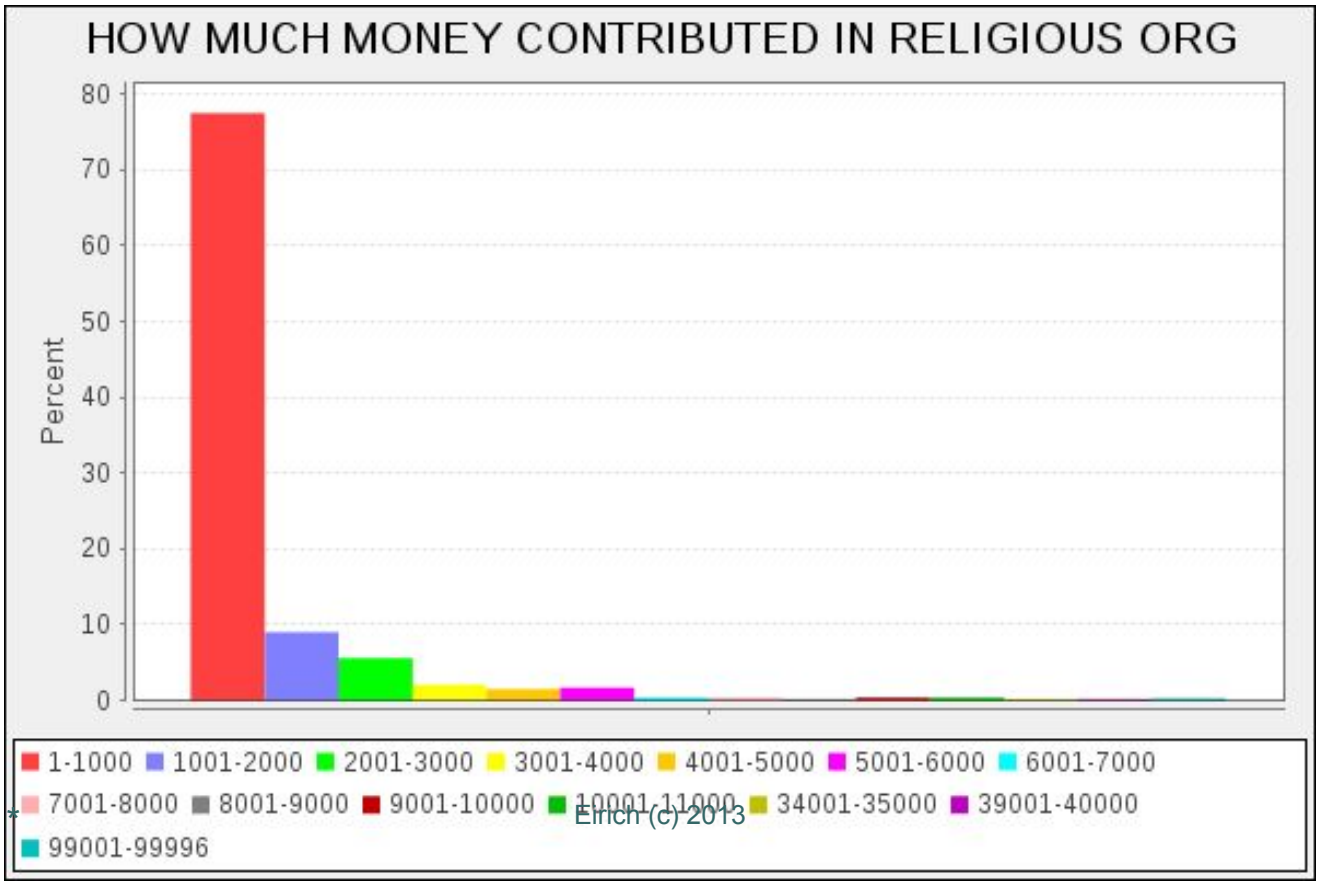
- Most appropriate with very skewed distributions

HOW MUCH MONEY CONTRIBUTED TO RELIGIOUS ORG

**$300**

# Median

- Why? **Without the most extreme outlier, mean=$1,084.49**



# Skewed distributions

- Positive skew vs. negative skew

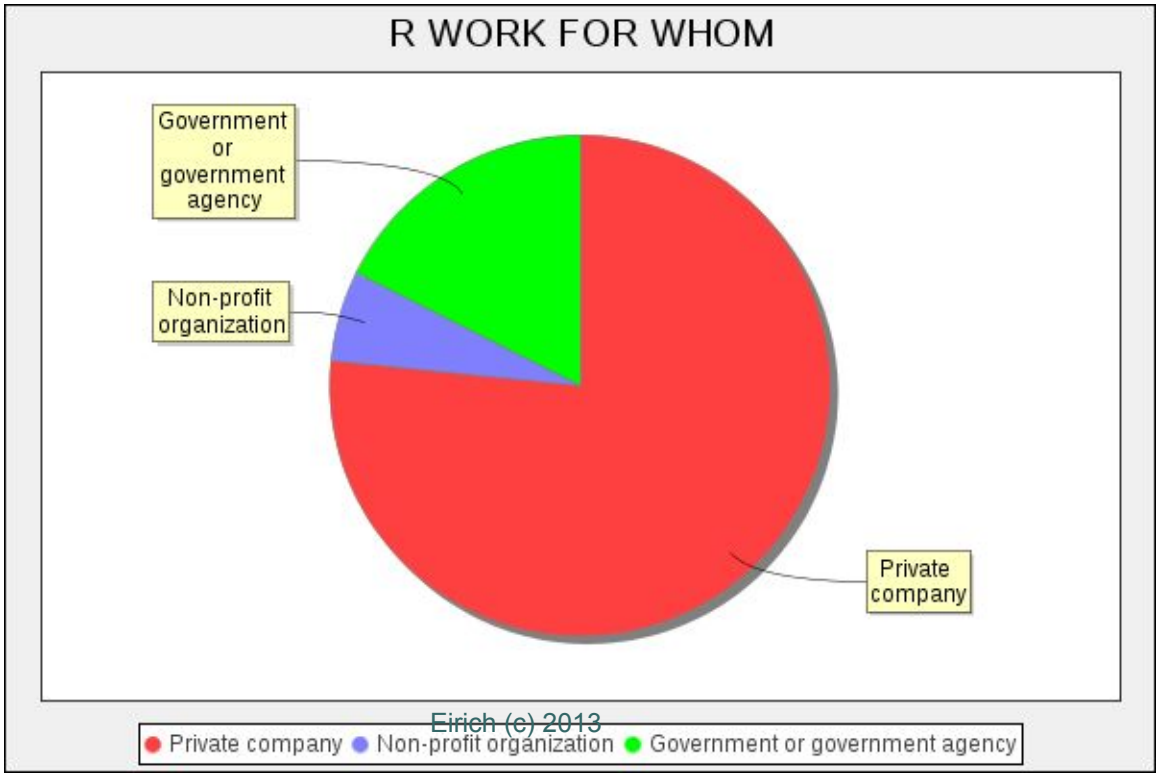
- **Skewness -** Skewness measures the degree and direction of asymmetry. A symmetric distribution such as a normal distribution has a skewness of 0, and a distribution that is skewed to the left, e.g., when the mean is less than the median, has a negative skewness.

# Skewed distributions

- **Kurtosis** - Kurtosis is a measure of the heaviness of the tails of a distribution.
- A normal distribution has a kurtosis of 3.
- Heavy tailed distributions will have kurtosis greater than 3
- Light tailed distributions will have kurtosis less than 3.

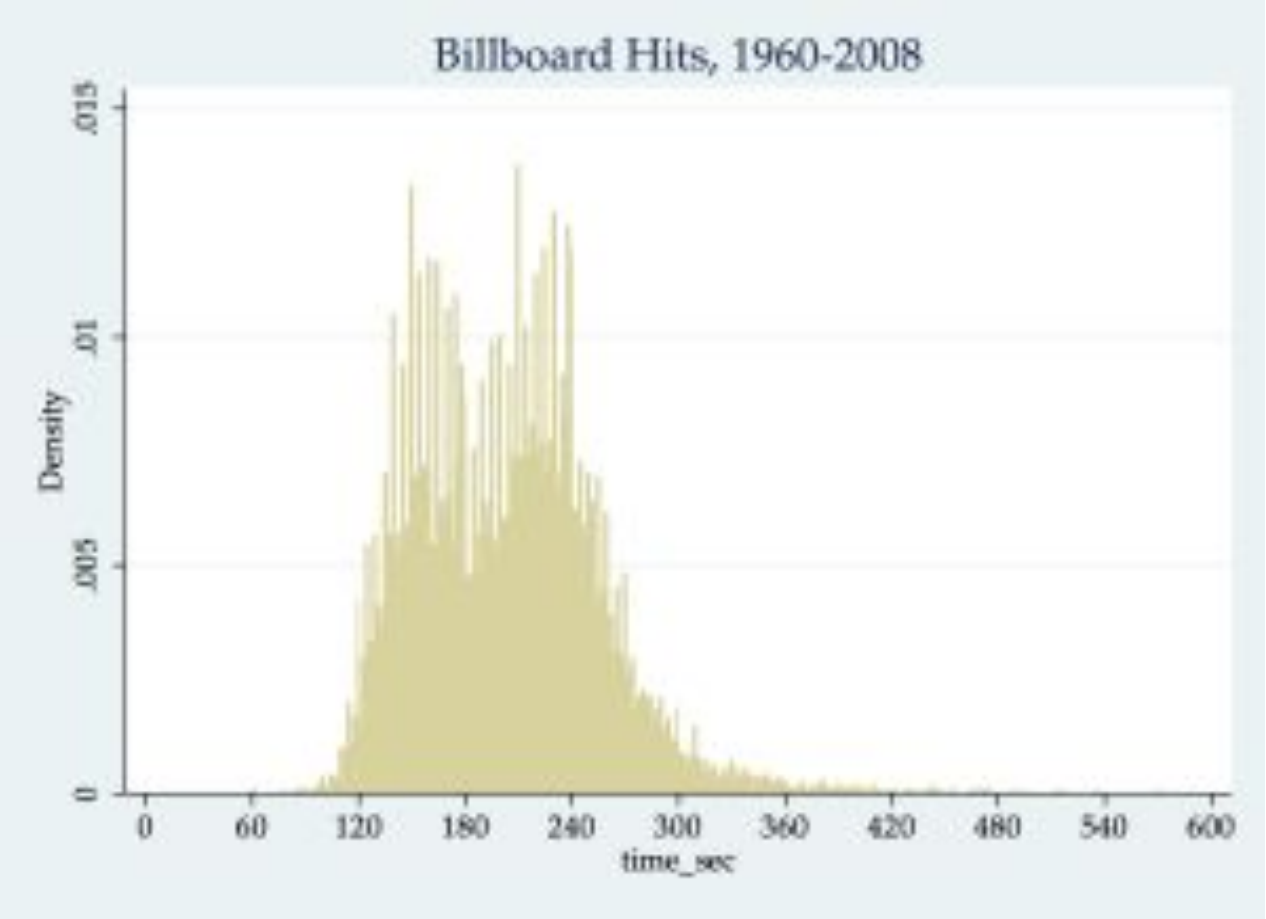
# Mode

- Only appropriate measure with categorical variables



# Bi-modal distribution

- Length of hit songs



# Dispersion in our data

- Range
- IQR
- Standard deviation

# Range

- Maximum - minimum

Range of text messages sent over 4 month period, college students (n=70)

Min=7 … Max=5304

Range=5297

* *portal.acm.org/ft_gateway.cfm?id=1851638&type=pdf*

# Interquartile range

- $75^{th}$ percentile – $25^{th}$ percentile

HOW MUCH MONEY CONTRIBUTED TO RELIGIOUS ORG

**$75^{th}$ Percentile = $1000**

**$25^{th}$ Percentile = $100**

**IQR = $900**

# Standard deviation

- Typical distance an individual is away from the mean value

HOW MUCH MONEY CONTRIBUTED TO RELIGIOUS ORG

**St. Dev. = $6,130.50**

# Standard deviation - formula

S.D. =

$$\sqrt{\frac{\sum_{Xi}^{Xn}(X-M)^2}{N-1}}$$

# Standard deviation - more

$$\text{S.D.} = \sqrt{\frac{\sum_{Xi}^{Xn}(X - M)^2}{N - 1}}$$

1. Find the mean for the variable, M
2. Subtract the mean from each observation, X
3. Square each difference
4. Sum all the squared differences
5. Divide by the number of observations minus 1, N-1 (This gives you the variance.)
6. Take the square root of that quantity*

# Standard deviation – calculation

| ID | Age(X) | X-*M* | X-*M* | (X-*M*)^2 |
|---|---|---|---|---|
| A | 32 | 32-27.3 = | 4.7 | 21.8 |
| B | 23 | 23-27.3 = | -4.3 | 18.8 |
| C | 27 | 27-27.3 = | -0.3 | 0.1 |
| | | | | |
| *M*= | *27.3* | | Sum of (X-*M*)^2 = | 40.7 |
| | | | (Sum of (X-*M*)^2)/N-1 = | 20.35 |
| N=3 | | | Square Root of (Sum of (X-*M*)^2)/N -1= | 4.51 |

**The result:**

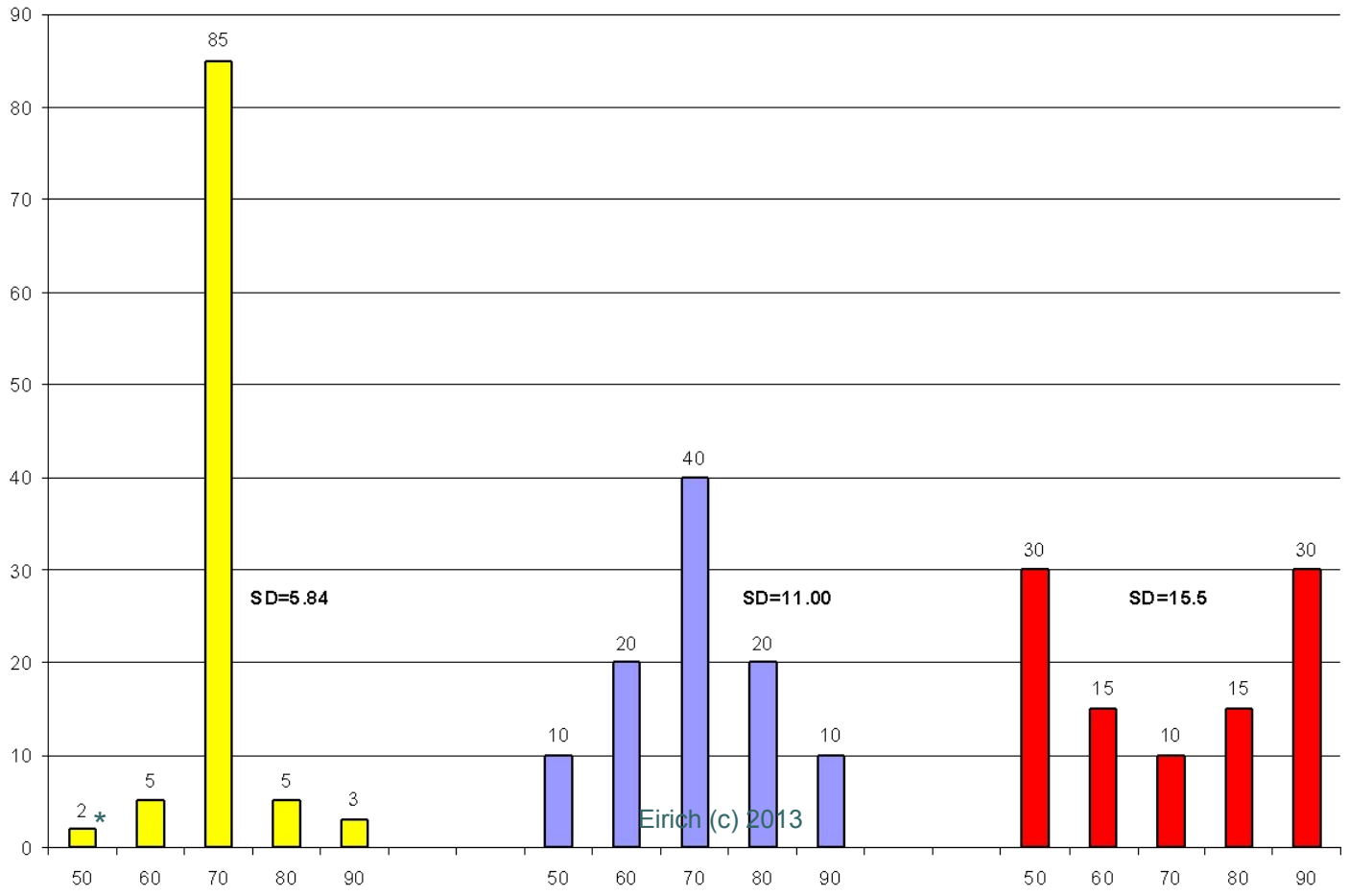
**s= 4.51**

- The standard deviation of age for these 3 observations is 4.51

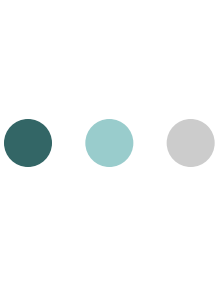
# A little more on standard deviation

# St. dev. and variable distributions

Three distributions of grades, all with mean=70

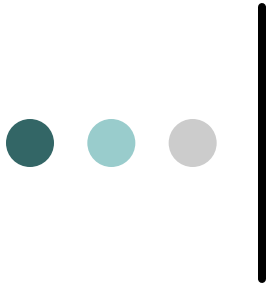


# Coefficient of variation = st. dev/mean



# 4. Probability

# Probability = long-run proportion of times an event will happen



# Probability = ranges from 0 to 1

# Probability rule #1
# Pr(A and B) = Pr(A) * Pr(B), if A and B are independent

# Probability rule #1

Probability of getting 2 heads in a row =

= 0.5 (chance first time) * 0.5 (chance second time)

= 0.25

# Probability rule #2
# Pr(A and B) = P(A) * P(B, given A),
# If A and B overlap

# Conditional probability

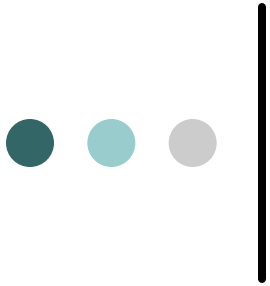
## What is the probability (in 2018) of finding a strong Democrat who is okay with a police officer ever striking a citizen? How about a strong Republican?

# Probability rule #2

| Probability of being ... | | |
|---|---|---|
| | | |
| … strong Democrat | 0.14 x | |
| … ok police hitting, given you're strong Dem | 0.55 = | |
| … both Dem and ok police hitting | **0.066** | |
| | | |
| … strong Republican | | 0.11 x |
| … ok police hitting, given you're strong Rep | | 0.80 = |
| … both Rep and ok police hitting | | **0.088** |

# Conditional probability

The ratio of Republicans who are ok with police hitting citizens under some circumstance to Democrats is 1.33 to 1, while the ratio of Republicans to Democrats overall (unconditionally) in the population is 0.78 to 1.



# Probability distributions

# Probability distribution = a histogram of the likelihood of various outcomes for a variable

The expected value of a probability distribution (say, Y) is just the mean/average

E(Y) = mean of Y

# Expected value

$E(Y) = \sum yP(y)$, which means that the mean of Y is the value of each outcome times its probability of occurring, for discrete variables

# Expected value example

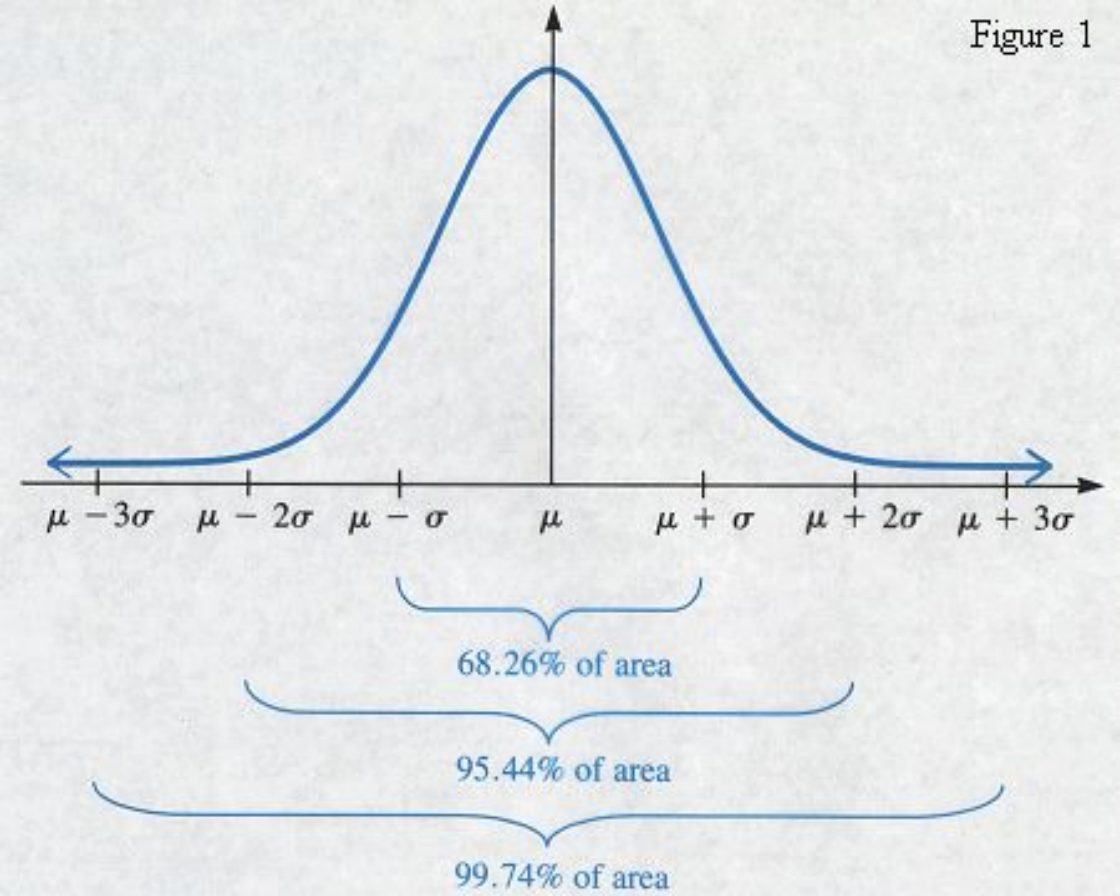
A lottery ticket costs $1. With probability of 0.0000001, you win $1M and with probability of 0.9999999, you win nothing. What is the expected value of the lottery ticket?

E(Ticket) = (Winning Amount * the Probability of Winning) + (Losing Amount * the Probability of Losing)

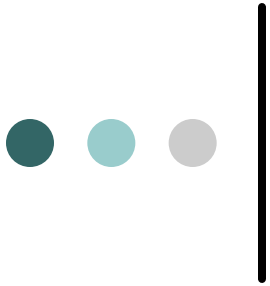
E(Ticket) = 1,000,000*0.0000001 + 0*0.99999999
E(Ticket) = 0.10 + 0
**E(Ticket) = $0.10 = 10¢**

# There is a special kind of probability distribution …

# The standard normal curve



# More on that distribution later …



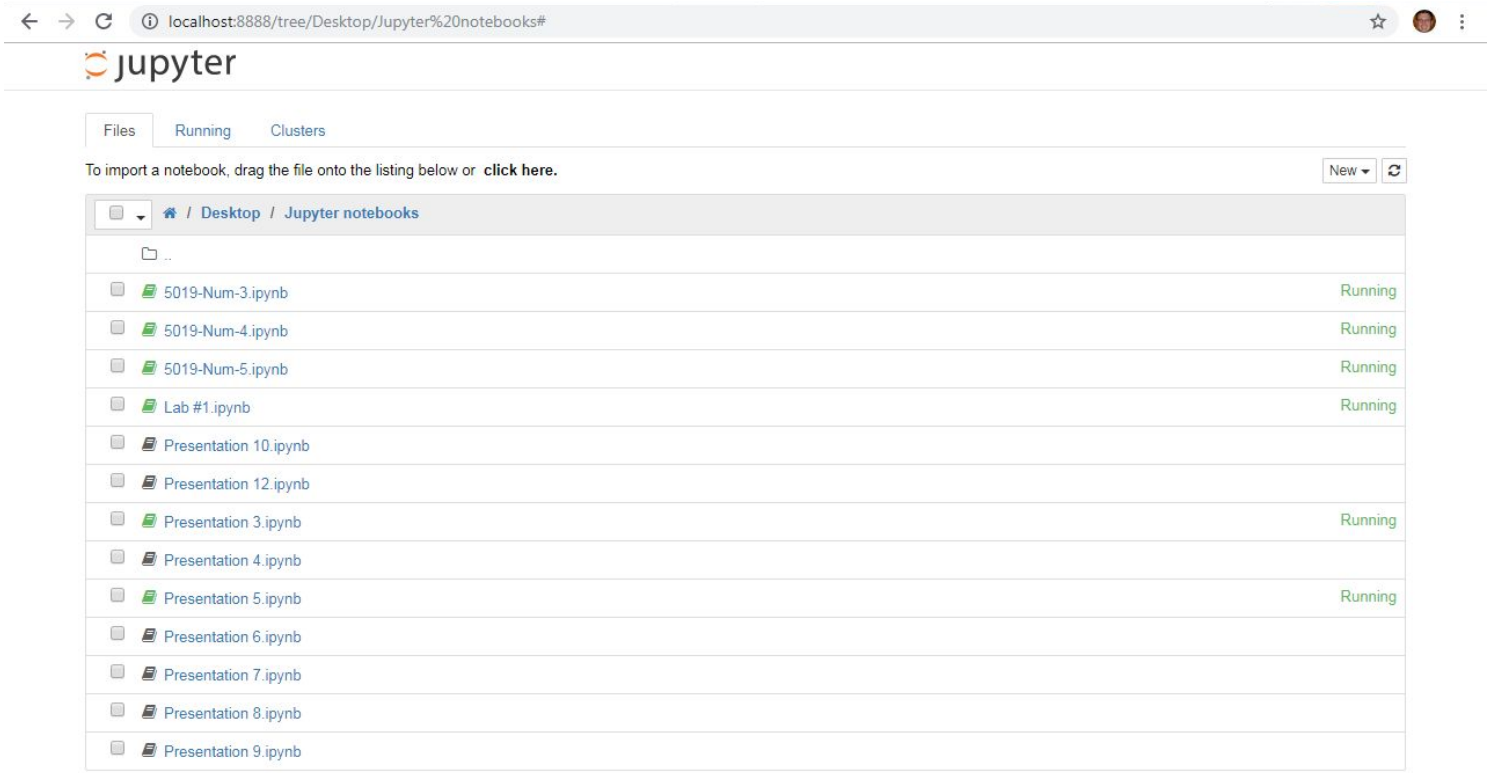
# Our first lab

# Step 1: Download Python

Go here and follow the instructions (choose Python 3.7)

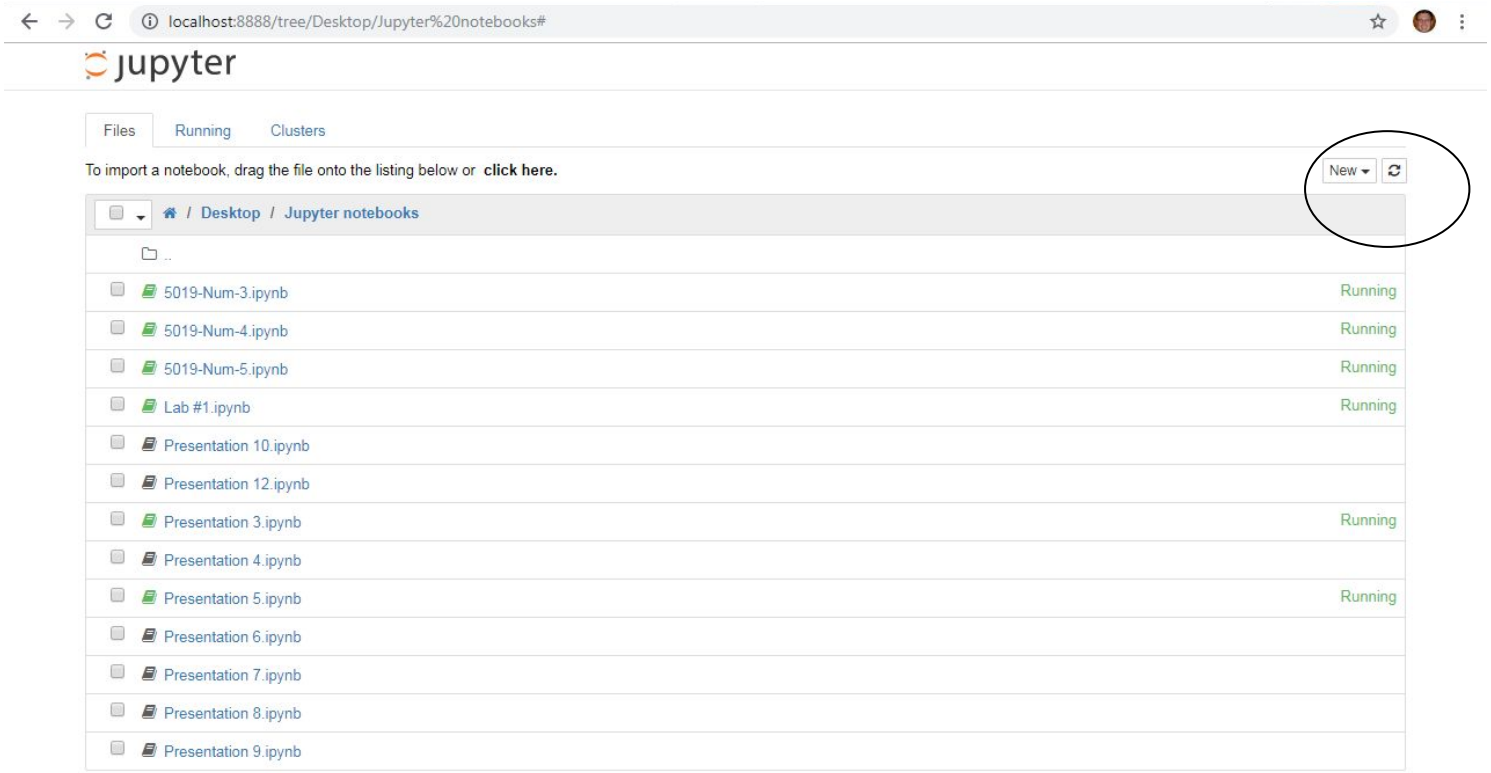
# Step 2a: Open IPython Notebook program

It pops up on your browser!



# Step 2b: Open IPython Notebook

Open a new notebook (otherwise called Jupyter)



# Step 2c: Run some commands

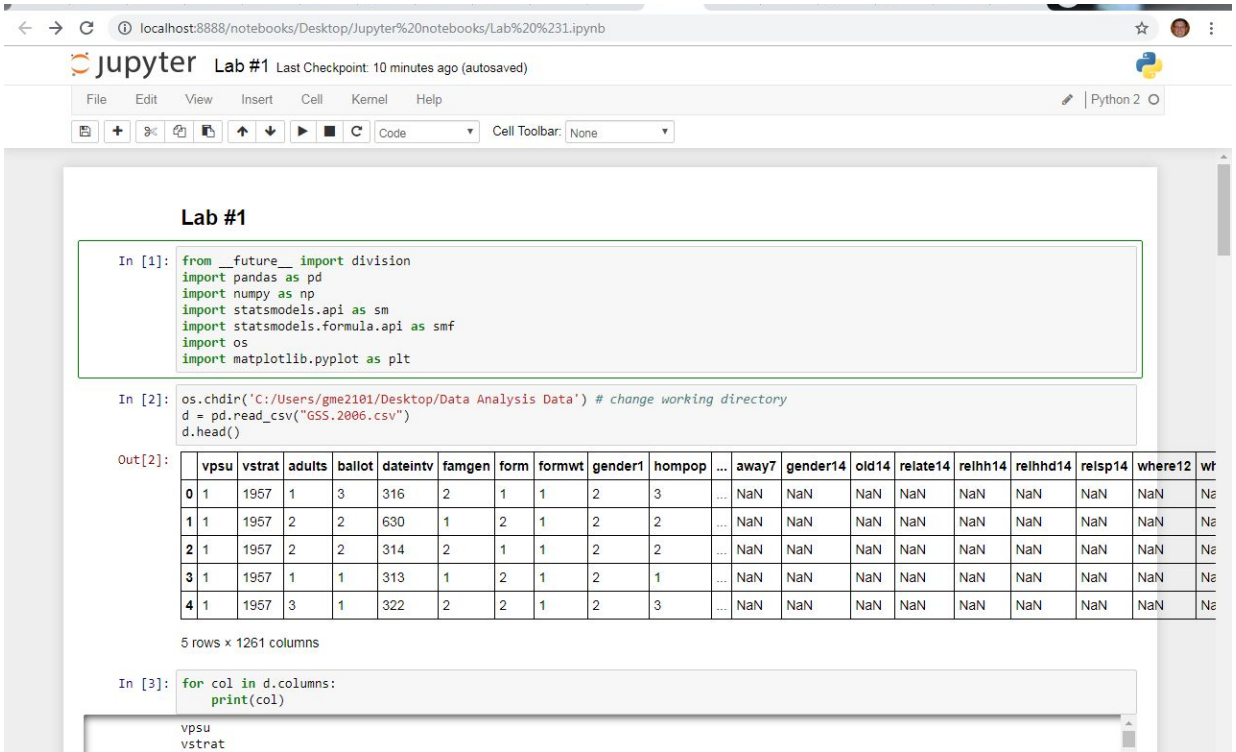
```
from __future__ import division
import pandas as pd
import numpy as np
import statsmodels.api as sm
import statsmodels.formula.api as smf
import os
import matplotlib.pyplot as plt
```

# Step 3a: Get your data

Get the GSS.2006.csv from me, from Courseworks

# Step 3b: Load our data

```
os.chdir('C:/Users/gme2101/Desktop/Data Analysis Data') # change working directory

d = pd.read_csv("GSS.2006.csv")
```



localhost:8888/notebooks/Desktop/Jupyter%20notebooks/Lab%20%231.ipynb

jupyter Lab #1 Last Checkpoint: 10 minutes ago (autosaved)

File Edit View Insert Cell Kernel Help Python 2

Code Cell Toolbar: None

**Lab #1**

In [1]:
```
from __future__ import division
import pandas as pd
import numpy as np
import statsmodels.api as sm
import statsmodels.formula.api as smf
import os
import matplotlib.pyplot as plt
```

In [2]:
```
os.chdir('C:/Users/gme2101/Desktop/Data Analysis Data') # change working directory
d = pd.read_csv("GSS.2006.csv")
d.head()
```

Out[2]:

| | vpsu | vstrat | adults | ballot | dateintv | famgen | form | formwt | gender1 | hompop | ... | away7 | gender14 | old14 | relate14 | relhh14 | relhhd14 | relsp14 | where12 | wh |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 1957 | 1 | 3 | 316 | 2 | 1 | 1 | 2 | 3 | ... | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | Na |
| 1 | 1 | 1957 | 2 | 2 | 630 | 1 | 2 | 1 | 2 | 2 | ... | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | Na |
| 2 | 1 | 1957 | 2 | 2 | 314 | 2 | 1 | 1 | 2 | 2 | ... | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | Na |
| 3 | 1 | 1957 | 1 | 1 | 313 | 1 | 2 | 1 | 2 | 1 | ... | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | Na |
| 4 | 1 | 1957 | 3 | 1 | 322 | 2 | 2 | 1 | 2 | 3 | ... | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | Na |

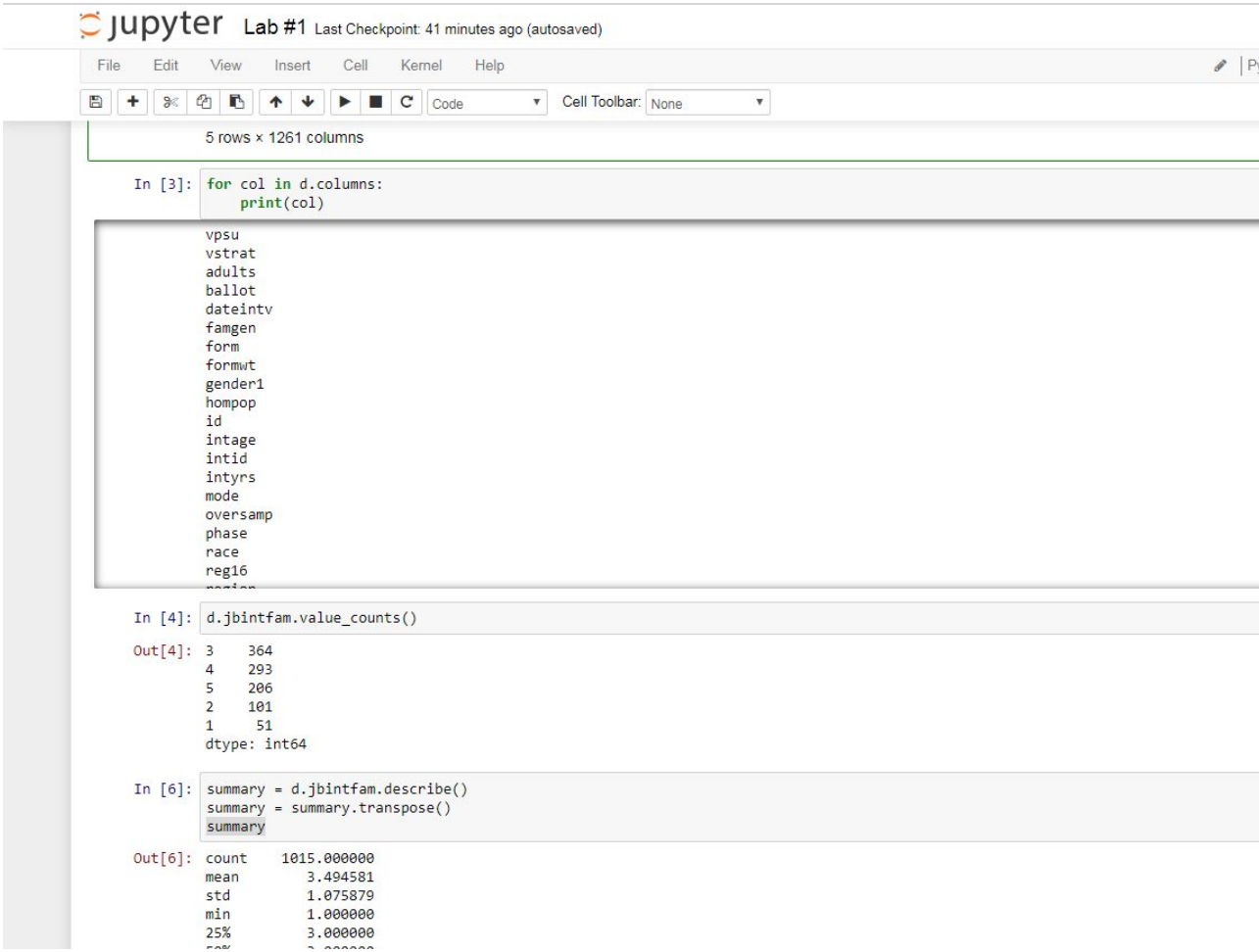
5 rows × 1261 columns

In [3]:
```
for col in d.columns:
    print(col)
```

```
vpsu
vstrat
```

# Step 4: Inspect your data

```
for col in d.columns:
    print(col)
```




*

# Step 5: View your data

```
d.head()
```

```
d.head()
```

Out[2]:

| | vpsu | vstrat | adults | ballot | dateintv | famgen | form | formwt | gender1 | hompop | ... | away7 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 1957 | 1 | 3 | 316 | 2 | 1 | 1 | 2 | 3 | ... | NaN |
| **1** | 1 | 1957 | 2 | 2 | 630 | 1 | 2 | 1 | 2 | 2 | ... | NaN |
| **2** | 1 | 1957 | 2 | 2 | 314 | 2 | 1 | 1 | 2 | 2 | ... | NaN |
| **3** | 1 | 1957 | 1 | 1 | 313 | 1 | 2 | 1 | 2 | 1 | ... | NaN |
| **4** | 1 | 1957 | 3 | 1 | 322 | 2 | 2 | 1 | 2 | 3 | ... | NaN |

5 rows × 1261 columns

# Your first lab is ...

Python Lab Assignment #1

Answer these **three** questions as best you can:

1-- Choose one variable, look at its distribution, and plot a histogram of it. Explain what you take away from looking at the variable.

# Your first lab is ...

Python Lab Assignment #1 (due on June 5th) -- *continued*

2-- Choose some continuous-ish variable, and calculate its mean and standard deviation by some grouping variable. Explain what conclusion you draw from this analysis.

# Your first lab is ...

Python Lab Assignment #1 (due on June 5th) -- *continued*

3-- Choose two categorical-ish variables, and cross-tabulate them. Explain what conclusion you draw from this analysis.

# Step 6a: Go to website and find a variable … What website? Here.

# Step 6b: Go to website and find a variable

1008) How often do you feel that the demands of your job interfere with your family life? (JBINTFAM)

| | TOTAL | % |
|---|---|---|
| 1) Always | 51 | 5.0 |
| 2) Often | 101 | 9.9 |
| 3) Sometimes | 364 | 35.8 |
| 4) Hardly ever | 293 | 28.8 |
| 5) Never | 206 | 20.2 |
| 8) Can't choose | 1 | 0.1 |
| 9) No answer | 2 | 0.2 |
| Missing | 3492 | |
| **TOTAL** | 1018 | 100.0 |

[ Analyze results | Add to Question Bank | View Question Bank ]

# Step 6c: Go to website and find another variable

3) Last week were you working full-time, part-time, going to school, keeping house, or what? (WRKSTAT)

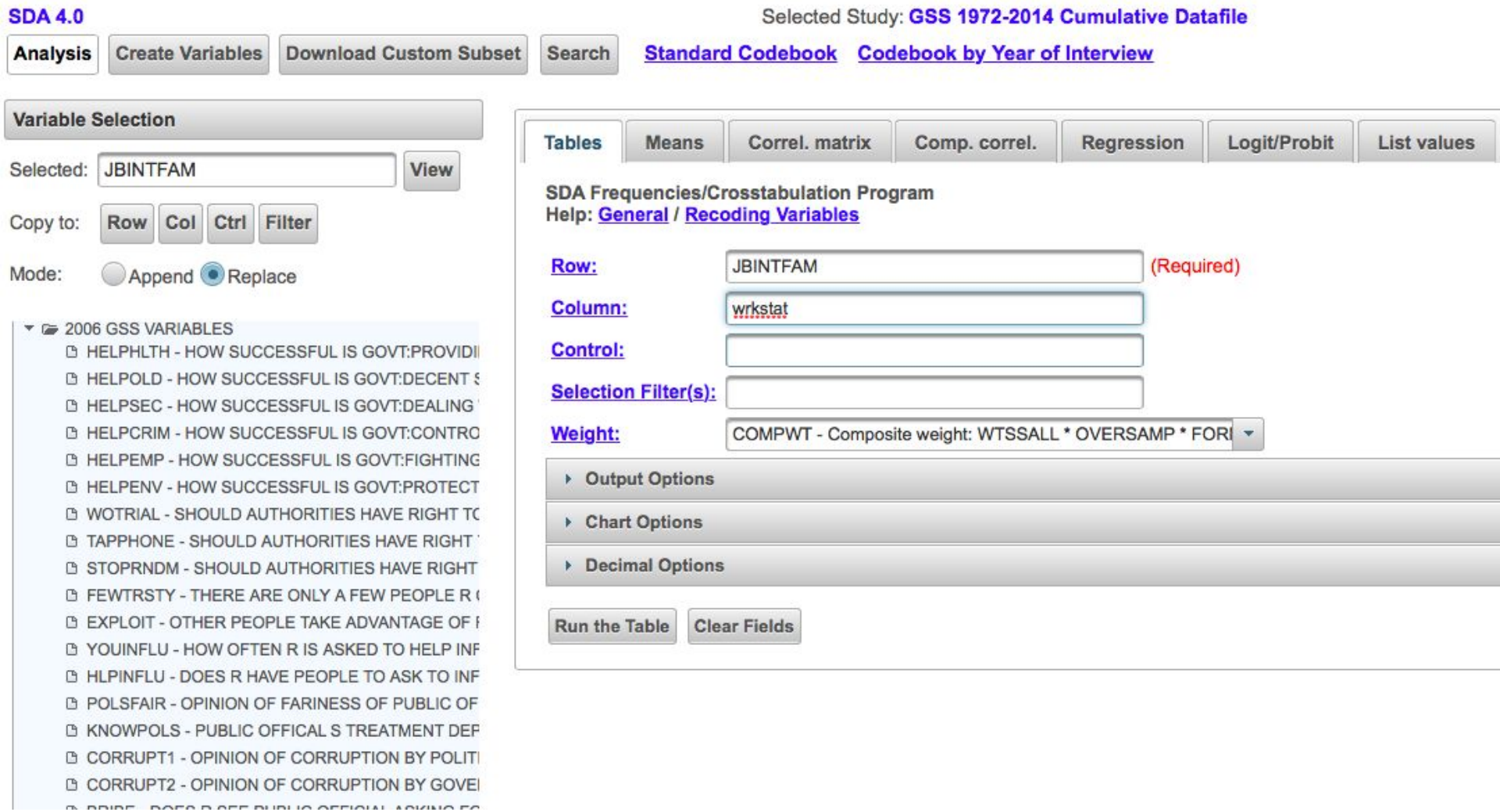
| | TOTAL | % |
|---|---|---|
| 1) Working full-time | 2321 | 51.5 |
| 2) Working part-time | 440 | 9.8 |
| 3) With a job, but not at work because of temporary illness, vacation, strike | 84 | 1.9 |
| 4) Unemployed, laid off, looking for work | 143 | 3.2 |
| 5) Retired | 712 | 15.8 |
| 6) In school | 139 | 3.1 |
| 7) Keeping house | 490 | 10.9 |
| 8) Other | 177 | 3.9 |
| 9) No answer | 4 | 0.1 |
| **TOTAL** | 4510 | 100.0 |

# Step 6d: Go to website and find one more variable

4) IF WORKING, FULL OR PART TIME: How many hours did you work last week, at all jobs? (HRS1)

| | N | Mean | Std.Deviation |
|---|---|---|---|
| **TOTAL** | 2765 | 42.608 | 15 |

# Step 6e: Or check out this website too ...



# Step 7a: Let's look at the data now

```
summary = d.jbintfam.describe()
summary = summary.transpose()
summary

d.jbintfam.value_counts()

my_tab = pd.crosstab(index=d["jbintfam"],  # Make a
crosstab
                                columns="count")
def compute_percentage(x):
      pct = float(x/my_tab['count'].sum()) * 100
      return round(pct, 2)

my_tab['percentage'] = my_tab.apply(compute_percentage,
axis=1)

my_tab
```

# Step 7a: Let's look at the data now

In [4]:
```
d.jbintfam.value_counts()
```

Out[4]:
```
3    364
4    293
5    206
2    101
1     51
dtype: int64
```

In [6]:
```
summary = d.jbintfam.describe()
summary = summary.transpose()
summary
```

Out[6]:
```
count    1015.000000
mean        3.494581
std         1.075879
min         1.000000
25%         3.000000
50%         3.000000
75%         4.000000
max         5.000000
Name: jbintfam, dtype: float64
```

In [46]:
```
my_tab = pd.crosstab(index=d["jbintfam"],  # Make a crosstab
                     columns="count")
def compute_percentage(x):
      pct = float(x/my_tab['count'].sum()) * 100
      return round(pct, 2)

my_tab['percentage'] = my_tab.apply(compute_percentage, axis=1)

my_tab
```

Out[46]:

| col_0 | count | percentage |
|---|---|---|
| jbintfam | | |
| 1 | 51 | 5.02 |
| 2 | 101 | 9.95 |
| 3 | 364 | 35.86 |
| 4 | 293 | 28.87 |

# Step 8a: Let's look at the data now

```
d.groupby(['jbintfam'])['hrs1'].mean()
d.groupby(['jbintfam'])['hrs1'].std()
```

```
In [19]: d.groupby(['jbintfam'])['hrs1'].mean()

Out[19]: jbintfam
         1           49.955556
         2           46.336957
         3           44.540936
         4           39.894161
         5           38.345550
         Name: hrs1, dtype: float64
```

On average, people who say they **always** have work/family conflict (=1) work 49.95 hours, while people who say they **never** have work/family conflict (=5) work 38.34 hours

# Step 8b: Let's look at the data this way too

```
In [19]: d.groupby(['jbintfam'])['hrs1'].mean()

Out[19]: jbintfam
         1           49.955556
         2           46.336957
         3           44.540936
         4           39.894161
         5           38.345550
         Name: hrs1, dtype: float64
```

```
In [49]: d.groupby(['jbintfam'])['hrs1'].std()

Out[49]: jbintfam
         1           18.320326
         2           10.937663
         3           13.054693
         4           11.514910
         5           13.148783
         Name: hrs1, dtype: float64
```

# Step 8b: Let's look at the data this way too (BTW, by sex and jbintfam)

```
d.groupby(['jbintfam', 'sex'])['hrs1'].mean()
```

```
In [20]: d.groupby(['jbintfam', 'sex'])['hrs1'].mean()

Out[20]: jbintfam  sex
         1         1      57.724138
                   2      35.875000
         2         1      47.627119
                   2      44.030303
         3         1      46.558011
                   2      42.273292
         4         1      41.007194
                   2      38.748148
         5         1      40.141304
                   2      36.676768
         Name: hrs1, dtype: float64
```

# Step 8b: Let's look at the data this way too (BTW, by sex and jbintfam)

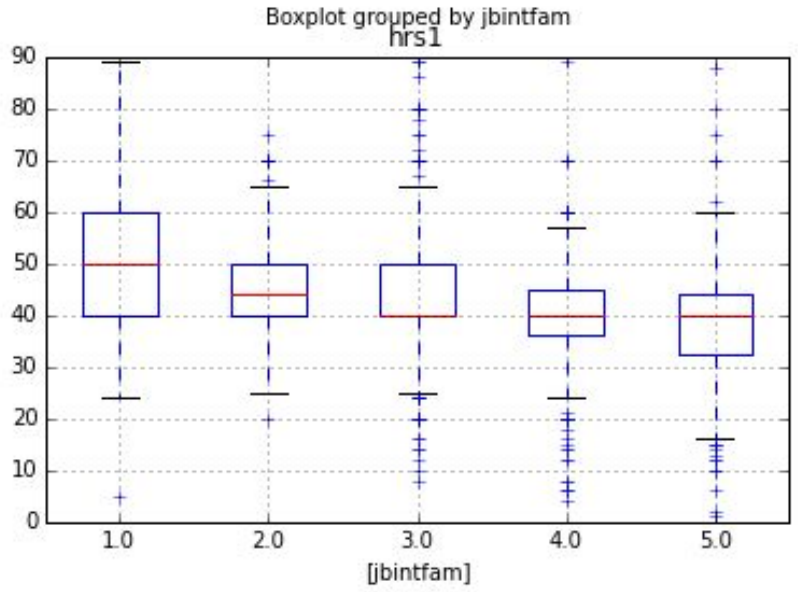
```
d.groupby(['jbintfam', 'sex'])['hrs1'].count()
```

```
In [57]: d.groupby(['jbintfam', 'sex'])['hrs1'].count()

Out[57]: jbintfam  sex
         1         1      29
                   2      16
         2         1      59
                   2      33
         3         1     181
                   2     161
         4         1     139
                   2     135
         5         1      92
                   2      99
         Name: hrs1, dtype: int64
```

# Step 8c: Draw a picture

```
%matplotlib inline
d.boxplot(column='hrs1', by=['jbintfam'])
```

```
In [29]: %matplotlib inline
         d.boxplot(column='hrs1', by=['jbintfam'])

Out[29]: <matplotlib.axes._subplots.AxesSubplot at 0x211d9e48>
```



# Step 8f: Let's look at the data now like this

```
res = pd.crosstab(d.jbintfam, d.wrkstat)
res.astype('float').div(res.sum(axis=0), axis=1)
```

[45]:
```
res = pd.crosstab(d.jbintfam, d.wrkstat)
res.astype('float').div(res.sum(axis=0), axis=1)
```

t[45]:

| wrkstat | 1.0 | 2.0 | 3.0 | 4.0 | 5.0 | 6.0 | 7.0 | 8.0 |
|---|---|---|---|---|---|---|---|---|
| jbintfam | | | | | | | | |
| 1 | 0.049140 | 0.051471 | 0.068966 | 0.000000 | 0 | 0.111111 | 0.0625 | 0.000000 |
| 2 | 0.109337 | 0.036765 | 0.103448 | 0.333333 | 0 | 0.111111 | 0.1250 | 0.000000 |
| 3 | 0.373464 | 0.279412 | 0.379310 | 0.333333 | 0 | 0.444444 | 0.2500 | 0.285714 |
| 4 | 0.285012 | 0.316176 | 0.379310 | 0.000000 | 0 | 0.222222 | 0.1875 | 0.285714 |
| 5 | 0.183047 | 0.316176 | 0.068966 | 0.333333 | 1 | 0.111111 | 0.3750 | 0.428571 |

# Step 8g: Let's look at the data now -- Zoom in

```
res = pd.crosstab(d.jbintfam, d.wrkstat)
res.astype('float').div(res.sum(axis=0), axis=1)
```

```
In [45]: res = pd.crosstab(d.jbintfam,
         res.astype('float').div(res.s
```

Out[45]:

| wrkstat | 1.0 | 2.0 | 3.0 |
|---|---|---|---|
| jbintfam | | | |
| 1 | 0.049140 | 0.051471 | 0.0 |
| 2 | 0.109337 | 0.036765 | 0.1 |
| 3 | 0.373464 | 0.279412 | 0.3 |
| 4 | 0.285012 | 0.316176 | 0.3 |
| 5 | 0.183047 | 0.316176 | 0.0 |

Only 18.30% of people who work full time (wrkstat=1) say they **never** have work/family conflict (=5), while 31.61% of people who work part time (wrkstat=2) say they **never** have work/family conflict (=5).

# Find a full another example of what I want here too

```
Files > Lab 1 > Lab #1-Example.html

--or--

Files > Lab 1 > Lab #1-Example.ipynb
```