# descriptive statistics 1

Adam Okulicz-Kozaryn

adam.okulicz.kozaryn@gmail.com

this version: Tuesday 13th February, 2018 09:55

# outline

basic concepts

summarizing one variable (Wheelan, 2013, ch2): central tendency and dispersion

summarizing more than one variable: crosstabs and correlation, (Wheelan, 2013, ch3,4)

application: income

# Doddle

- https://beta.doodle.com/poll/agf7b9eg4476iexy#table

## interested in working with local non-profit?

- Michael D'Italia: mjd429@camden.rutgers.edu
- again, extra credit for civic engagement!
- again, see syllabus for elaboration

# edu data (edu is most common interest this year)

- US educ data:
  `https://nces.ed.gov/`
  `https://www2.ed.gov/rschstat/landing.jhtml?src=pn`
- compare test scores across countries:
  `http://www.oecd.org/pisa/`
- diversity and disparities:
  `https://s4.ad.brown.edu/Projects/Diversity/Researcher/LTBDDload/DataList.aspx`
- what is college worth:

  `http://www.payscale.com/college-education-value-2013`

## misc

- looking ahead: a lot of material today
- practicing next week
- then one tough class on probability
- and we will relax in second half of the course
- How's Wheelan and Trochim?
- as we discuss topics, let's discuss examples from Wheelan!!

# outline

basic concepts

summarizing one variable (Wheelan, 2013, ch2): central tendency and dispersion

summarizing more than one variable: crosstabs and correlation, (Wheelan, 2013, ch3,4)

application: income

## basic definitions

- observation (U/A) v variable (property, attribute of U/A; eg age, price)
  - extra credit: say I study your grades, what's U/A?
- variable (varies) v constant (constant)
- central tendency v dispersion
  - eg [1,3] v [0,4]: same $\mu$, different $\sigma$
- representativness/external validity: population (students) v sample (this class)
- data: observational (hard (eg gdp) v survey (eg happiness)) v experimental (eg drug trial) [elaborate later in res_des.pdf]
- causation v correlation: `http://www.tylervigen.com/`

## correlation $\neq$ causality is important!

- Perhaps, the most fundamental piece of knowledge here is the understanding that correlation is not causation. It is both important at policy drafting stage–it is easy to mistake correlation for causation and draft unnesessary or wrong policies; and it is important at evaluation stage–it is easy to see positive effect of policy, while there is none. In addition to typical research design/ statistical discussion, I do caution students from evolutionary/behavioral perspective: humans tend to see more causes than there actually are.

# level of measurement

- real continuous: interval/ratio (price, weight, temperature)
- continous/categorical: ordinal (rank of faculty, grades)
- real categorical: nominal (many) or binary (two)
  (eg mode of transportation, gender)
- extra credit: education variable?

# howto describe data?

- numbers
- graphs (always better unless very few data, say $<5$)
  humans recognizes patterns in graphs better and faster
- break it up into subsets/subsamples! dig deeper!
  - say see hist/tab for males and females separately
  - say corr or crosstab for low and hi val separately
    that's a quick way to see nonlinear relationship!
    eg it may first rise and then fall

# outline

basic concepts

summarizing one variable (Wheelan, 2013, ch2): central tendency and dispersion

summarizing more than one variable: crosstabs and correlation, (Wheelan, 2013, ch3,4)

application: income

# definitions of basic summary stats

- start with central tendency, not dispersion:
  - mean $\frac{1+2+2+3+12}{5}=4$ (affected by extremes)
  - median: middle value: 2 (if even take the mean of the middle two)
  - mode: most frequent value: 2

- 1, 2, 2, 3, 12 is right skewed (dispersion, draw)
  - Wheelan had example with few middle class guys at a bar
  - then comes Bill Gates and skewes income distribution

## dispersion or distributions

- draw both freq tab or tabulations and histograms:
  - grades in this class (bimodal)
  - incomes of Hilary, Donald, Bernie, Ted (right skewed)
- can also have class interval or bin:

| | |
|---|---|
| under 35 | 9% |
| 36-45 | 41% |
| 46-64 | 30% |
| above 65 | 20% |

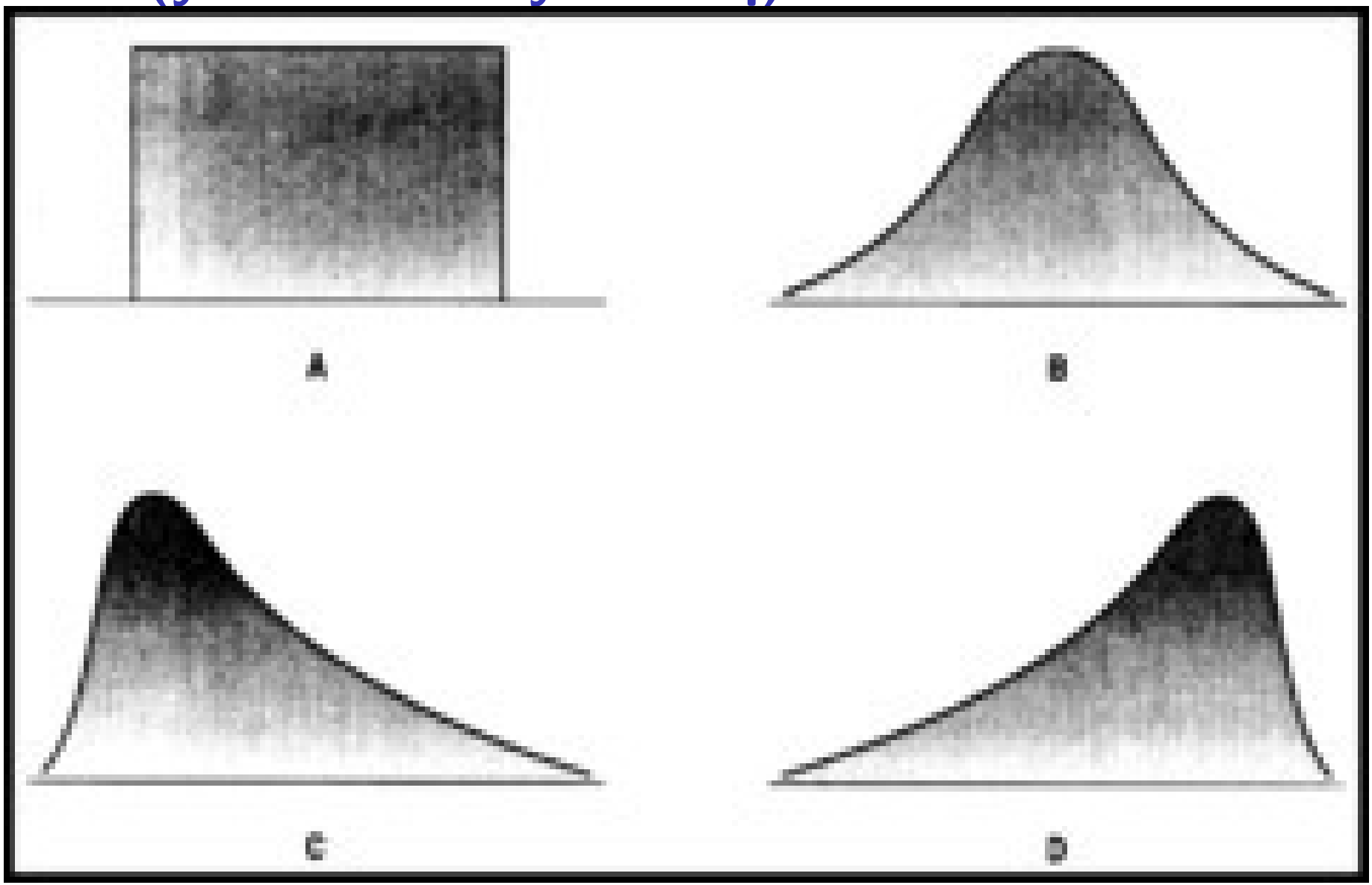
- `http://www.socialresearchmethods.net/kb/statdesc.php`: tab1, fig1

  also (Wheelan, 2013, p20-21)

# distribution types

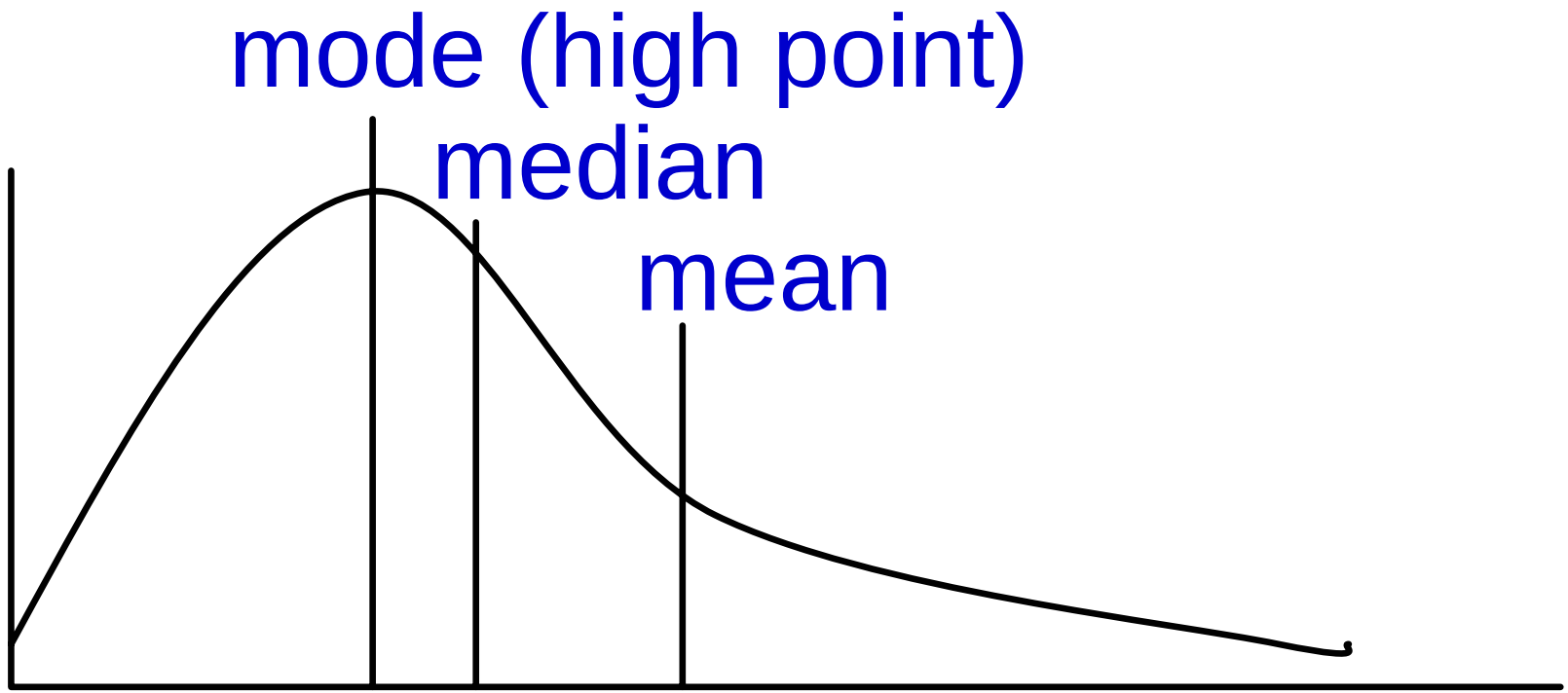
- uniform
- normal symmetrical unimodal
- left skewed
- right skewed (income)
- bimodal

# skew (y-axis: density or freq)



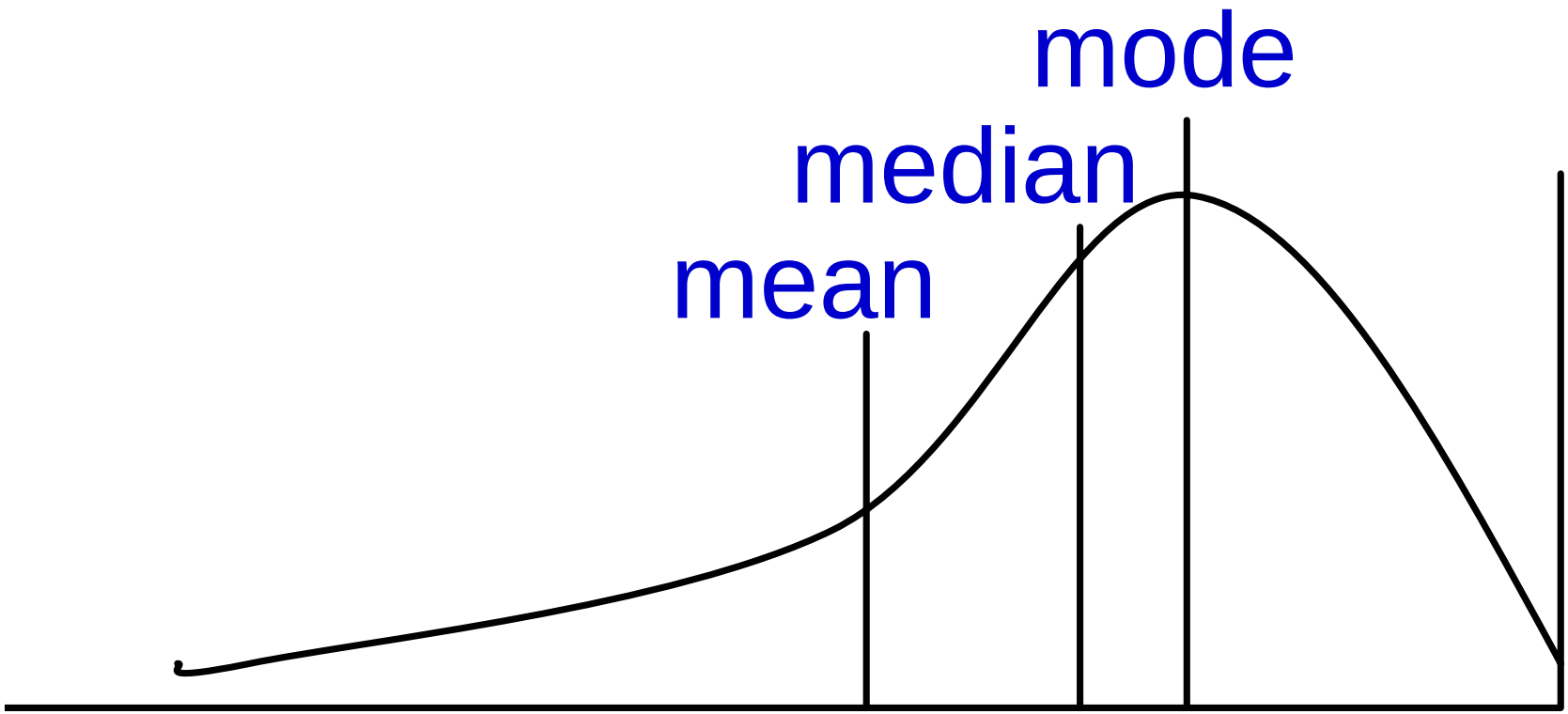

◇

# $\mu > M$: right skew (y-axis: density or freq)




◇

# $\mu < M$: left skew (y-axis: density or freq)



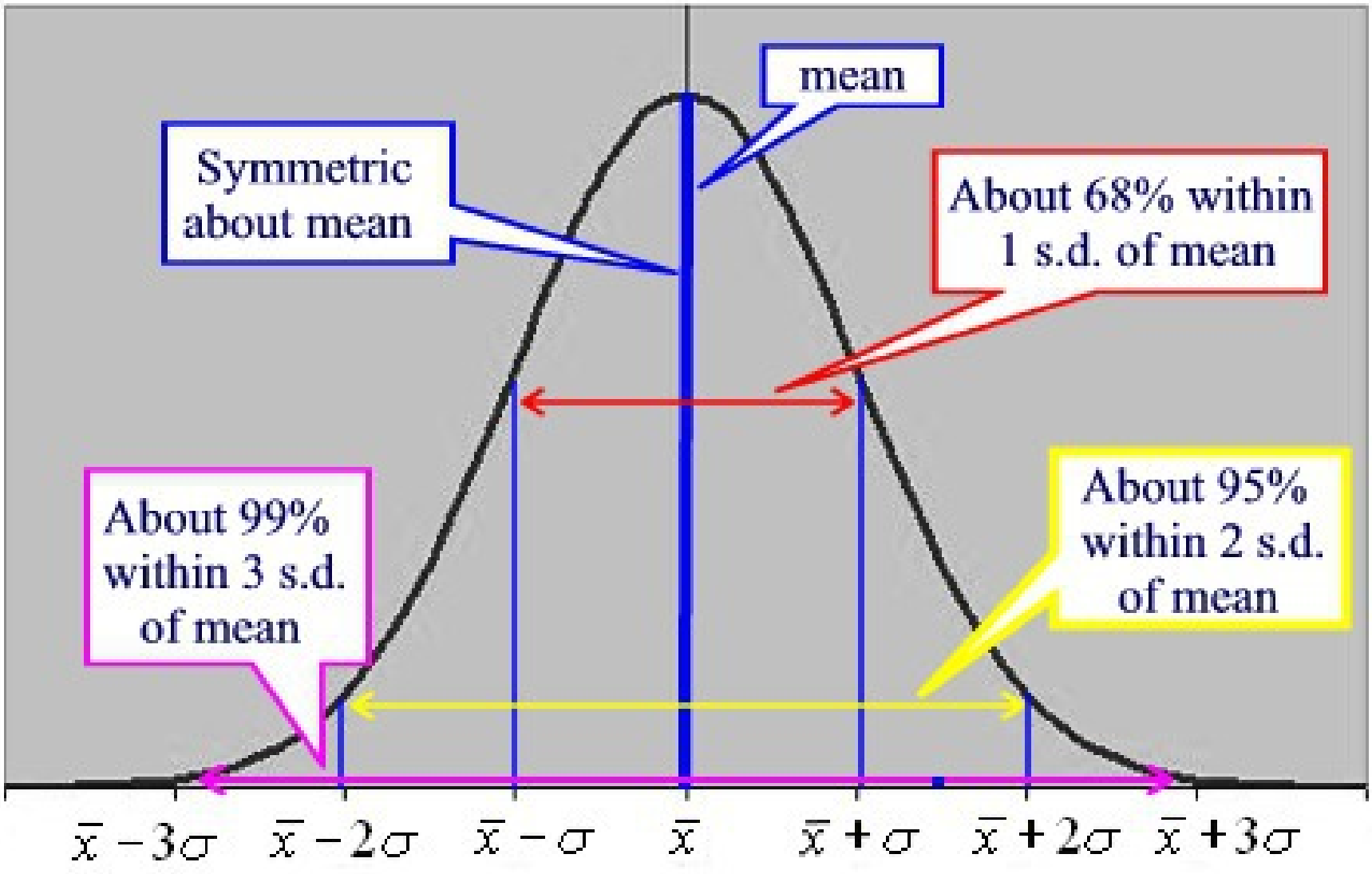

◇

# variability

- $range = max - min$
- p-th percentile: p % are below it; eg 75th percentile of income distribution : 75% of people are poorer than me
- quartile =25 %
- decile = 10%
- median = 2nd quartile = 5th decile = 50th percentile

`http://en.wikipedia.org/wiki/Household_income_in_the_United_States`

# normal distribution (Wheelan, 2013, fig on p26)




◇

- asymptotically, any variable is normally distributed

# outline

basic concepts

summarizing one variable (Wheelan, 2013, ch2): central tendency and dispersion

summarizing more than one variable: crosstabs and correlation, (Wheelan, 2013, ch3,4)

application: income

## few categories / categorical

- ◇ use contingency table / cross-tabs
  (because you cross-tabulate data)
- ◇ use percents, not counts: then usually it's clear
- : so what's the relationship: age and being a student?

| What is your age? | Are you a student? | | | Total |
|---|---|---|---|---|
| | Yes - Full Time | Yes - Part Time | No | |
| 15 and under | 88% | 12% | - | 8 |
| 16 - 18 | 95% | - | 5% | 42 |
| 19 - 23 | 68% | 12% | 20% | 205 |
| 24 - 29 | 16% | 10% | 74% | 353 |
| 30 - 35 | 5% | 9% | 86% | 192 |
| 36 - 45 | 4% | 8% | 88% | 165 |
| over 45 | 1% | 7% | 92% | 129 |

# crosstabs: row percents v col percents

Sort: Cols ▾ Rows ▾ | Count | All % | Row % | Col %

Number of Employees at Company

| Job Satisfaction | 1-25 | 26-100 | 101-999 | 1,000-3,000 | > 3000 | Total |
|---|---|---|---|---|---|---|
| Hate my job | 24.4% | 14.1% | 26.9% | 12.8% | 21.8% | 100% |
| I'm not happy in my job | 31.6% | 21.3% | 19.2% | 6.3% | 21.5% | 100% |
| It's a paycheck | 27.6% | 20.4% | 22.6% | 7.7% | 21.8% | 100% |
| I enjoy going to work | 32.3% | 21.8% | 21.3% | 7.0% | 17.6% | 100% |
| Love my job | 47.8% | 17.2% | 17.0% | 5.0% | 13.0% | 100% |

Sort: Cols ▾ Rows ▾ | Count | All % | Row % | Col %

Number of Employees at Company

| Job Satisfaction | 1-25 | 26-100 | 101-999 | 1,000-3,000 | > 3000 |
|---|---|---|---|---|---|
| Hate my job | 0.8% | 0.8% | 1.5% | 2.2% | 1.5% |
| I'm not happy in my job | 6.6% | 7.9% | 7.1% | 7.2% | 9.3% |
| It's a paycheck | 12.6% | 16.4% | 18.1% | 18.9% | 20.4% |
| I enjoy going to work | 43.3% | 51.6% | 50.3% | 50.8% | 48.4% |
| Love my job | 36.7% | 23.2% | 23.0% | 20.9% | 20.5% |
| Total | 100% | 100% | 100% | 100% | 100% |

## percentage change v percentage point change

- say good school's dropout rate increases from 2% to 4%
  - percentage point increase is $4 - 2 = 2$
  - percentage increase is $(\frac{4-2}{2}) * 100 = 100$

- say bad school's dropout rate increases from 50% to 75%
  - percentage point increase is $75 - 50 = 25$
  - percentage increase is $(\frac{75-50}{50}) * 100 = 50$

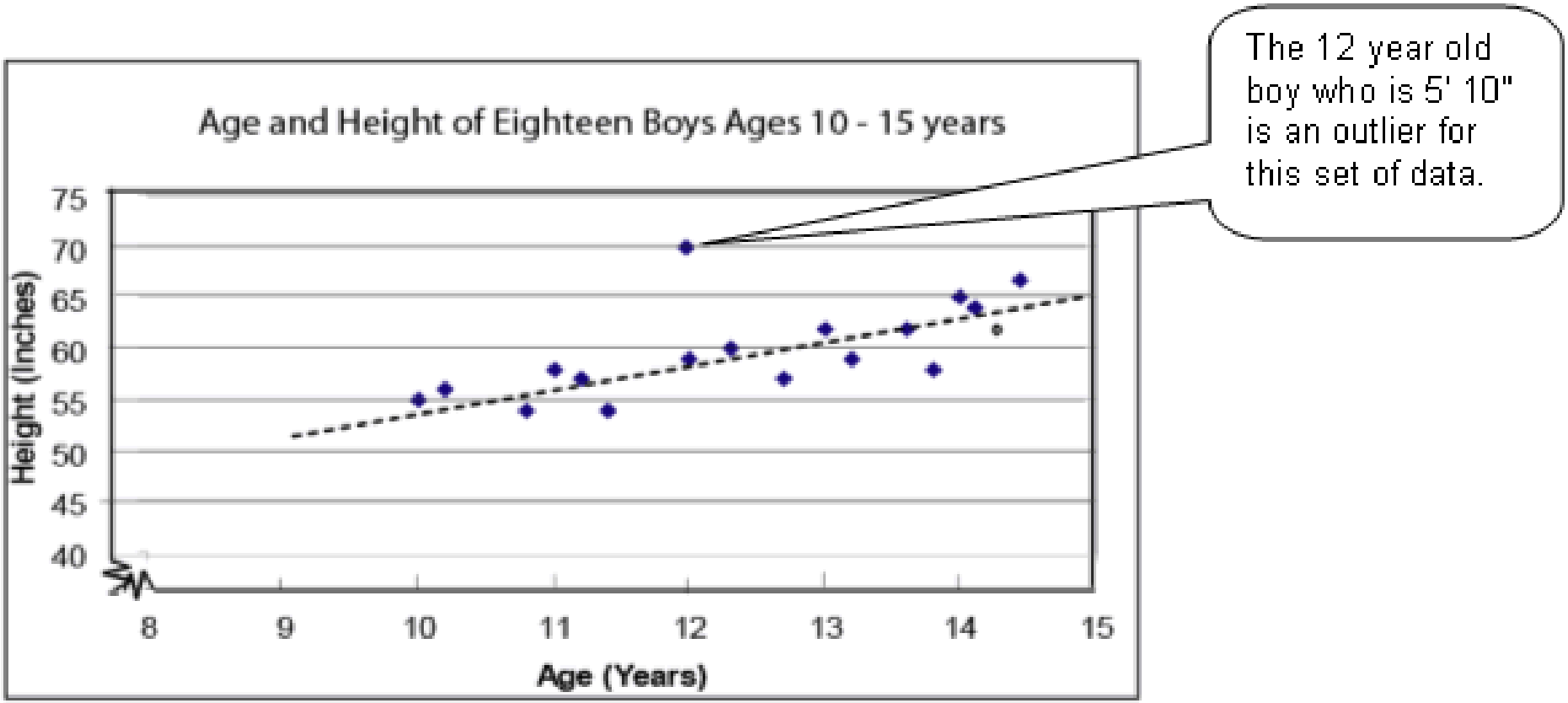
- if you start from low base (eg 2), then small percentage point increase is huge percent increase!

## many categories / continuous data

- use correlation and scatterplots
  - just plot them in scatterplot; identify outliers!
  - blackboard: examples with outliers
  - correlation ranges between -1 and 1
  - $< |4|$ low
  - $|.4 - .6|$ moderate
  - $> |.7|$ strong
- again, keep in mind causation v correlation

TODO: just insert here one of these corr coef graphs showng strength of relationship based on look

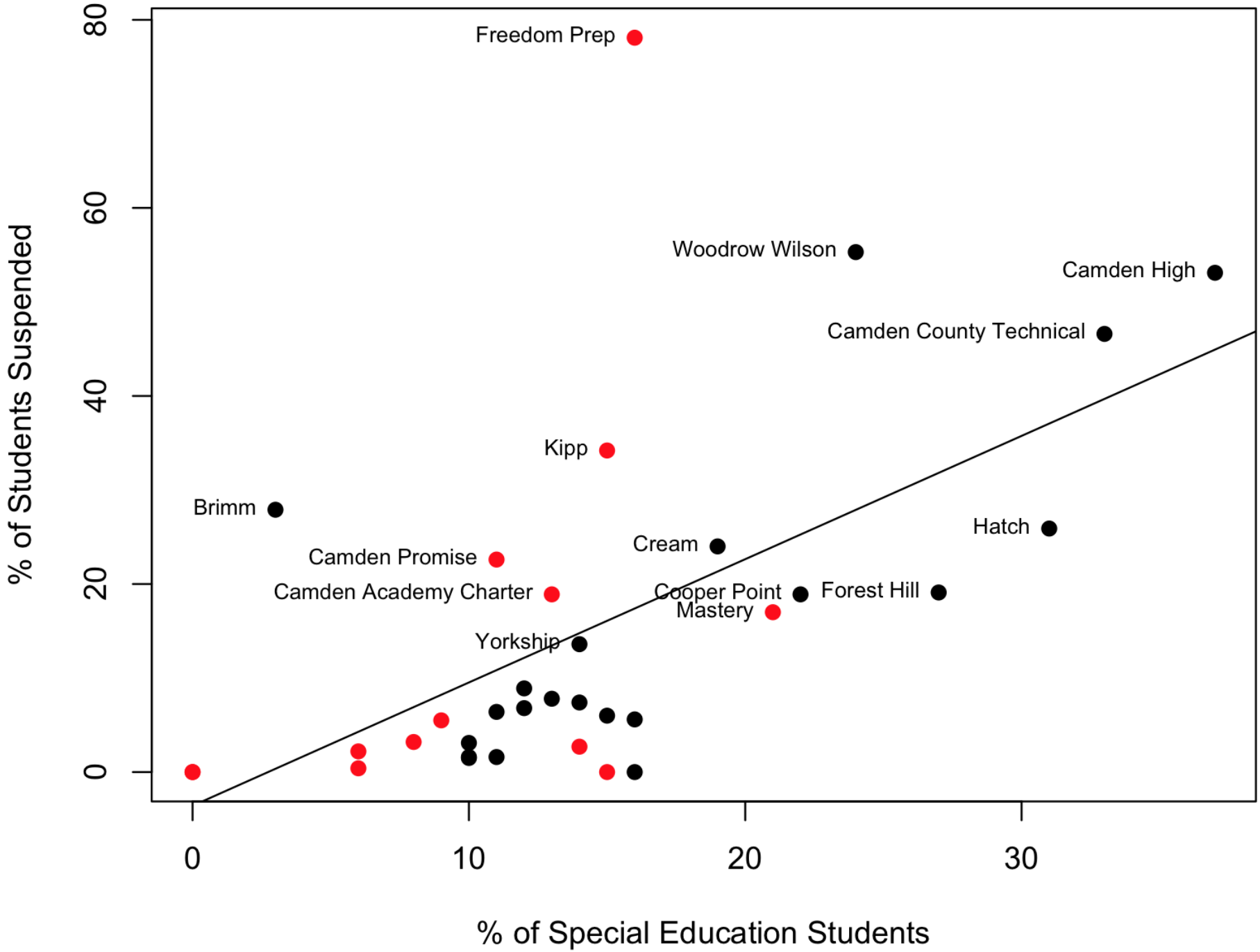
# scatterplot




- also see `http://www.socialresearchmethods.net/kb/statcorr.php`

- next slide: `https://danley.camden.rutgers.edu/2017/04/13/who-suspends-the-highest-percentage-of-camden-students-freedom-prep/`

- red: charter/renaissance; black: Camden schools


Freedom Prep
Woodrow Wilson
Camden High
Camden County Technical
Kipp
Brimm
Hatch
Cream
Camden Promise
Camden Academy Charter
Cooper Point
Forest Hill
Mastery
Yorkship
0
20
40
60
80
% of Students Suspended
0
10
20
30
% of Special Education Students

# do scatterplots

- it is useful to produce a scatterplot
- you'd see outliers–
- and whether the relationship is due to them
- blackboard: relationships biased due to outliers
- say marriage rate and divorce rate and Nevada

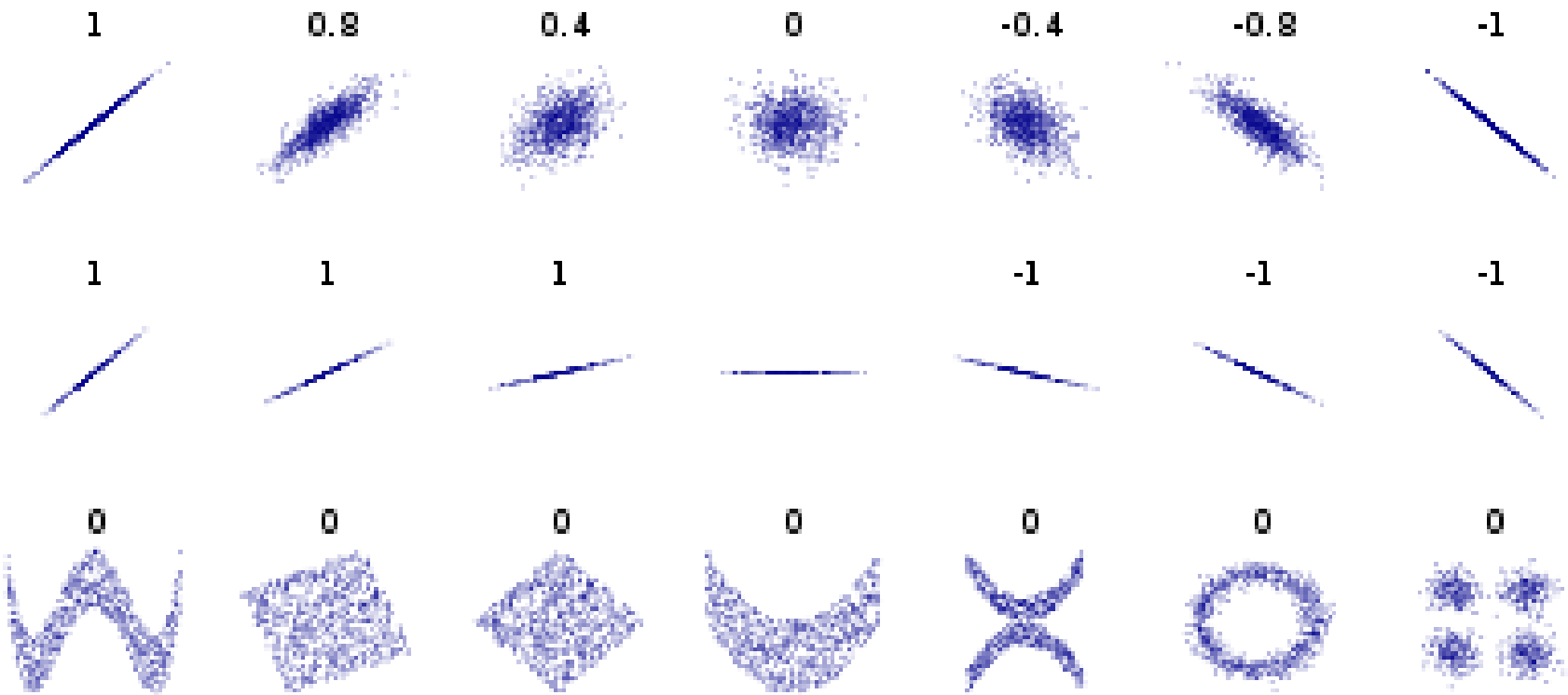
# Wheelan in ch11 mentions Whitehall studies

- fascinating stuff!
- high status causes better health!
  - great book 'Status Syndrome' `http://a.co/jaUuwT7`
- say nobel prize or oscar boosts one's health and longevity
  - these successful folks live longer and in better health
  - than exact same people (income, lifestyle, etc) but without status

# closer look at status syndrome

- `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2566175/`
- see Table 2A for correlations
  - especially 'Decision latitude'
  - conclusions? extra credit

# correlations for different scenarios




◇

# outline

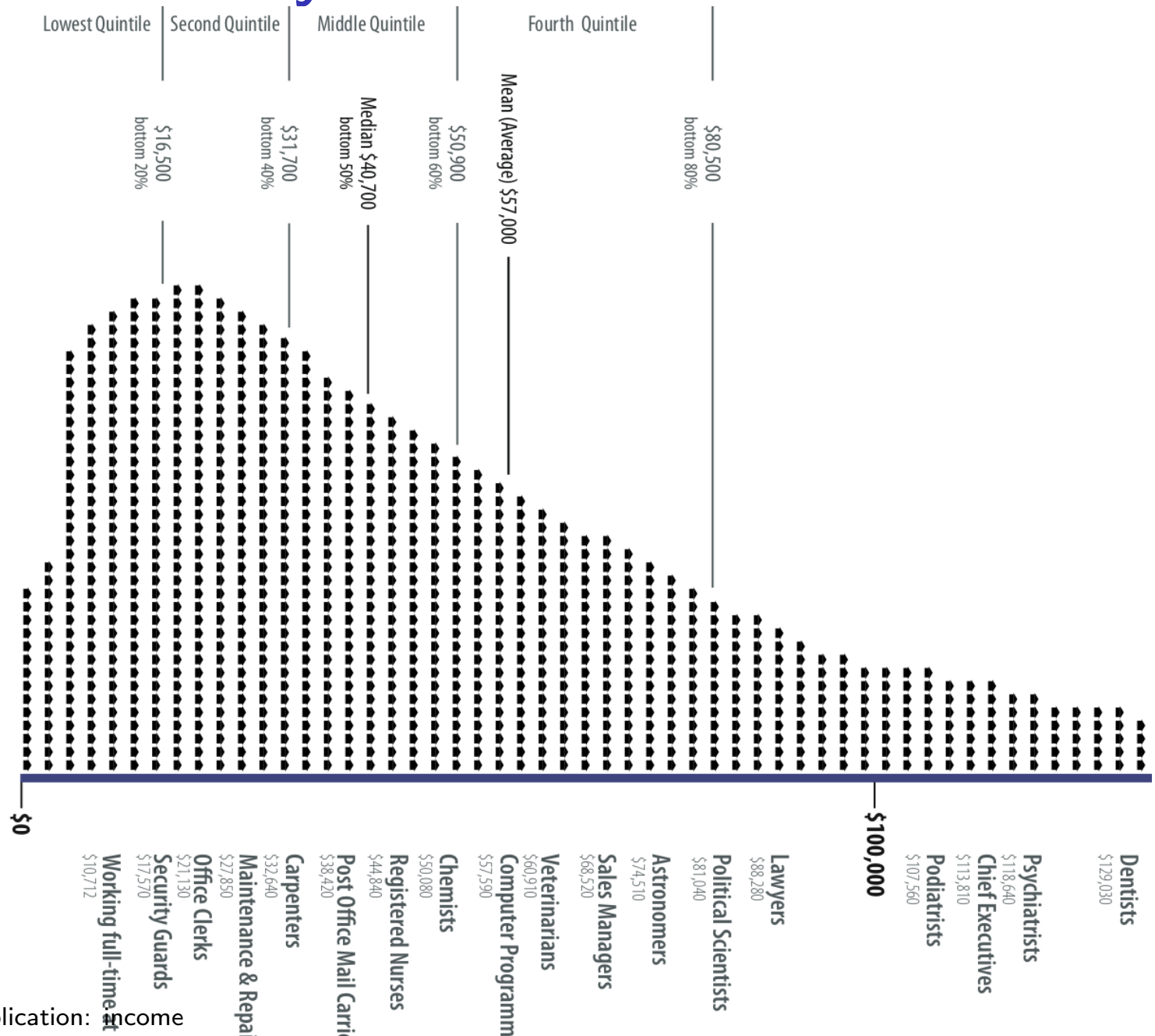
basic concepts

summarizing one variable (Wheelan, 2013, ch2): central tendency and dispersion

summarizing more than one variable: crosstabs and correlation, (Wheelan, 2013, ch3,4)

application: income

# where are you on income distribution?

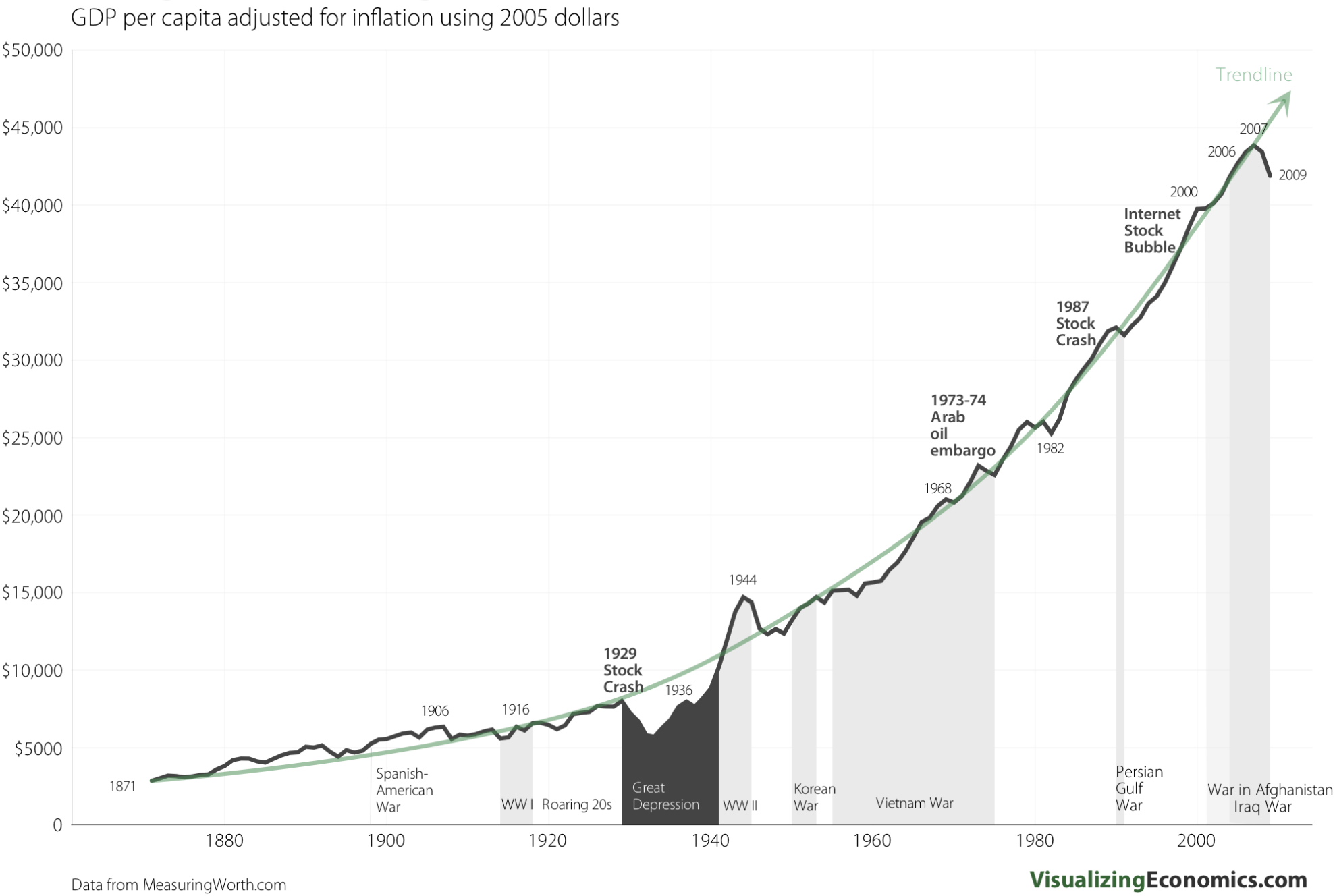


## idea for a project: what you can do

- it would be interesting to break income down by sociodemographics,
  by geography, and by both

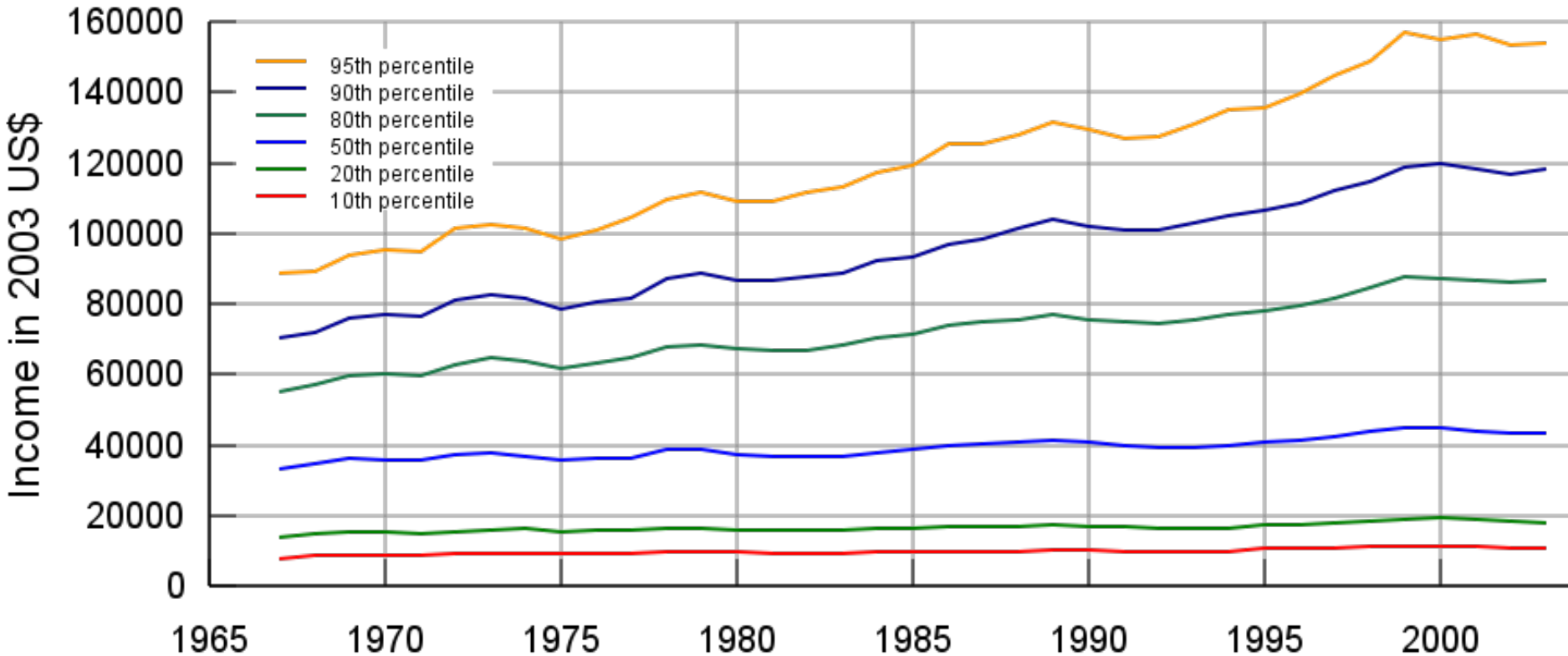
- get data and do it yourself, eg:
  `http://visualizingeconomics.com/cool-data/`
- and lots of nice visualizations here `http://www.gapminder.org/`
- also see Wheelan (2013, ch2) and `http://en.wikipedia.org/wiki/Household_income_in_the_United_States#Household_income`

- and now let's plot income over time (also see (Wheelan, 2013, p16))...


GDP per capita adjusted for inflation using 2005 dollars
$50,000
$45,000
$40,000
$35,000
$30,000
$25,000
$20,000
$15,000
$10,000
$5000
0
1880
1900
1920
1940
1960
1980
2000
Trendline
1871
1906
1916
1929
Stock
Crash
1936
1944
1968
1973-74
Arab
oil
embargo
1982
1987
Stock
Crash
2000
Internet
Stock
Bubble
2006
2007
2009
Spanish-
American
War
WW I
Roaring 20s
Great
Depression
WW II
Korean
War
Vietnam War
Persian
Gulf
War
War in Afghanistan
Iraq War
Data from MeasuringWorth.com
VisualizingEconomics.com

# but median income has not been growing much

◇



# how about income distribution over time?

- another interesting thing is to look over time at income distribution
- today's bottom decile has better quality of life than 9th decile 100 years ago (Derek Bok)
- can you translate this to plain English? extra credit

# next week

- we will always end the class by having a quick look at the next class

# bibliography I

Okulicz-Kozaryn, A. and J. M. Mazelis (2016): "More Unequal In Income, More Unequal in Wellbeing," Social Indicators Research.

Wheelan, C. (2013): Naked statistics: stripping the dread from the data, WW Norton & Company.