descriptive statistics 1

Adam Okulicz-Kozaryn

adam.okulicz.kozaryn@gmail.com

this version: Wednesday 2$^{\text{nd}}$ September, 2020 12:50

outline

basic concepts

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

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

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 μ, different σ
- 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]

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

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

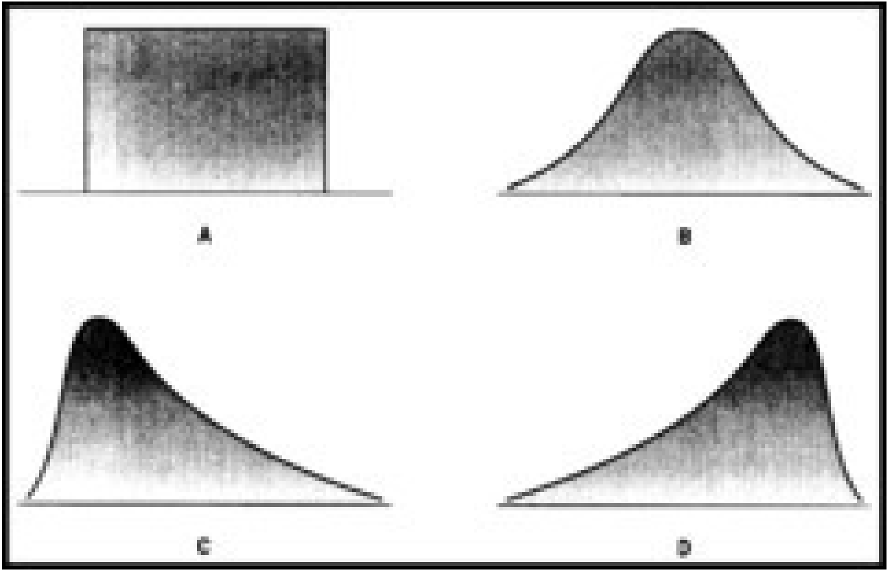
    also (Wheelan, 2013, p20-21)

distribution types

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

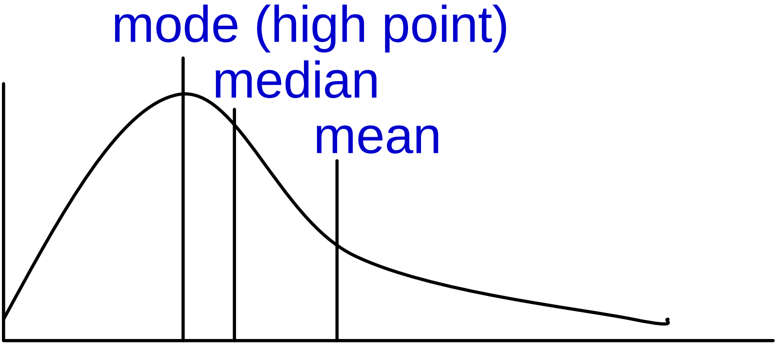
skew (y-axis: density or freq)

◇



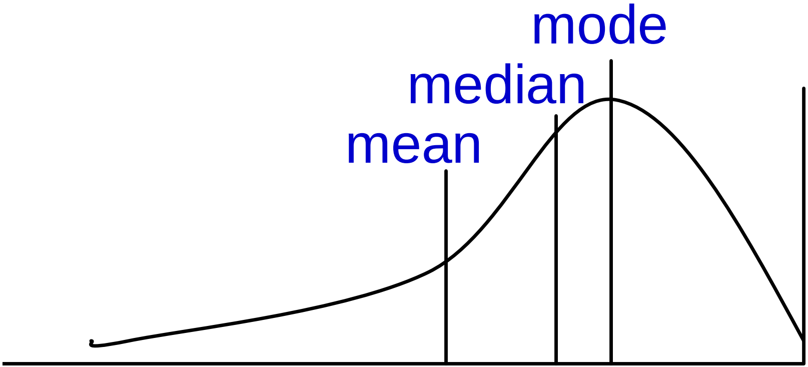
$\mu > M$: right skew (y-axis: density or freq)

◇



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

- ◇

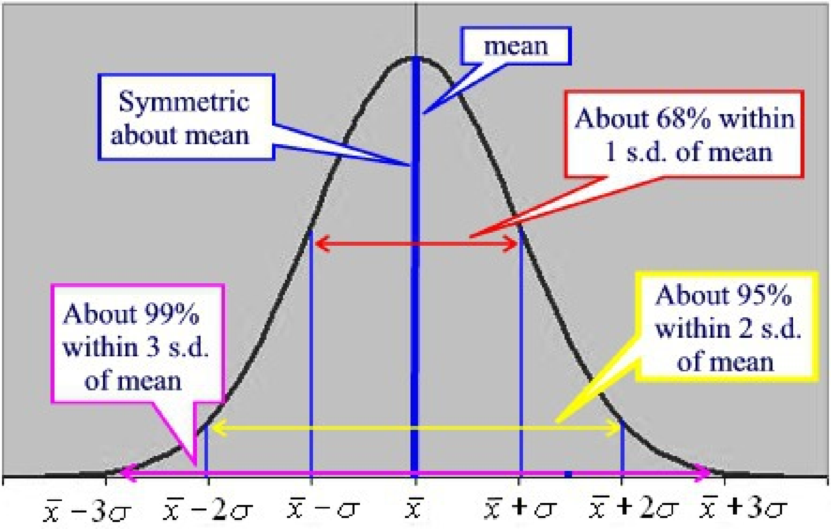


variability

- $range = max - min$
- 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

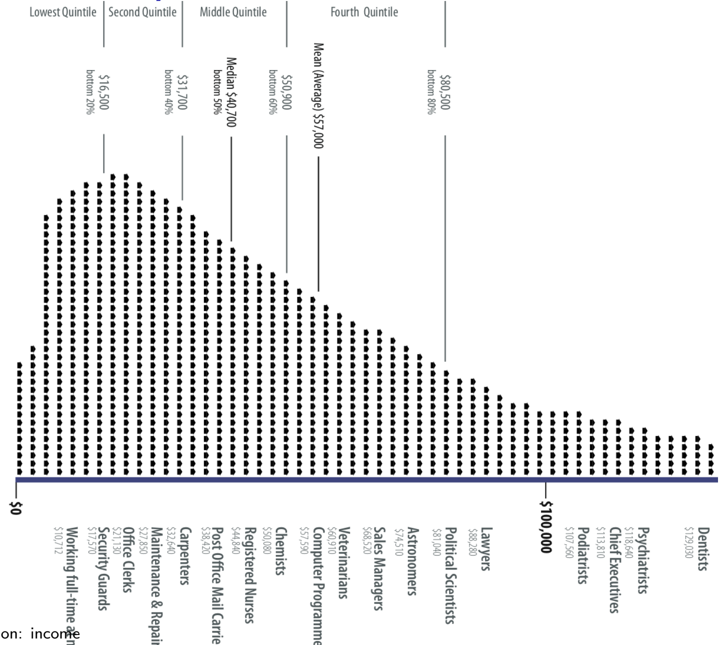
outline

basic concepts

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

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,

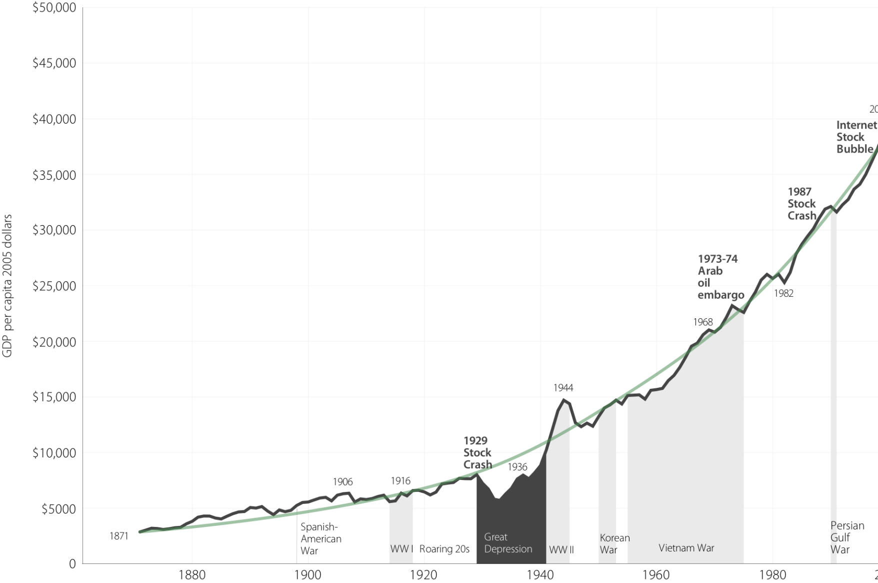
  by geography, and by both

- 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))...

Long-term real growth in US GDP per capita 1871–200

GDP per capita adjusted for inflation using 2005 dollars

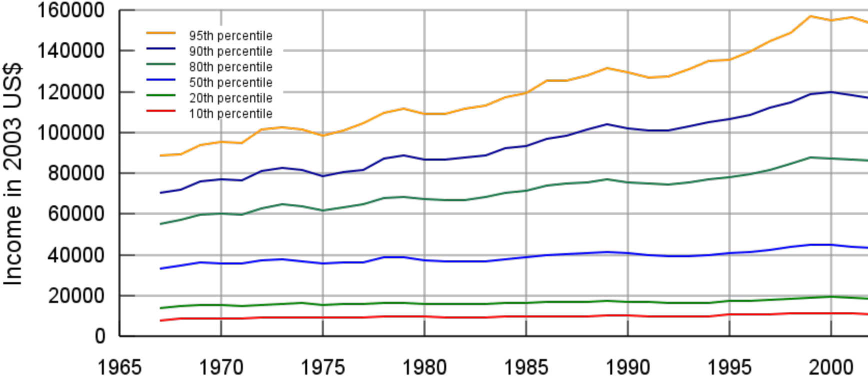


Data from MeasuringWorth.com

VisualizingEco

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.