

descriptive statistics 1

Adam Okulicz-Kozaryn

`adam.okulicz.kozaryn@gmail.com`

this version: Monday 13th September, 2021 13:59

outline

basic concepts

summarizing one variable (Wheelan, 2013, ch2): central tendency and dispersion [2 vars next week]

application: income

edu data (edu is most common interest this year)

- US <https://nces.ed.gov/>
- NJ <https://www.nj.gov/education/data/>
- compare test scores across countries:
<http://www.oecd.org/pisa/>
- diversity and disparities:
<https://s4.ad.brown.edu/projects/diversity/index.htm>
- what is college worth:
<https://www.bls.gov/ooh/>
<http://www.payscale.com/college-education-value-2013>

misc

- looking ahead: some stats today and next wk
- practicing in 2 wks
- then one tough class on probability
- and relax in second half of the course
- How's Wheelan and Trochim?
- as we cover concepts,
let's discuss ex from Wheelan! 20% participation!

outline

basic concepts

summarizing one variable (Wheelan, 2013, ch2): central tendency and dispersion [2 vars next week]

application: income

if can't measure it, then it's not science

- When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind: it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science, whatever the matter may be. [Lord Kelvin 1824-1907]
- don't say large, increased etc, give numbers!
- but just because got a number, doesn't mean it's right!
- GIGO, triangulate, bias, validity, etc

basic definitions

- observation (U/A) v variable
(property, attribute of U/A; eg age, price)
- extCre: 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
- representativeness/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!

- <http://www.tylervigen.com/>
- a fundamental piece of knowledge: correlation \neq causation
- at policy drafting stage—easy to mistake correlation for causation and draft unnecessary or wrong policies
- at evaluation stage—easy to see positive effect of policy (sunk cost, groupthink, etc) while there is none!
-
- from evol/beh persp: humans see causes where there are none

level of measurement

- important because it determines what you can do with the variable, eg mean v mode
- real continuous: interval/ratio (price, weight, temp)
- continuous/categorical: ordinal (rank of faculty, grades)
- real categorical: nominal (many) or binary (two)
(eg mode of transportation, gender)
- **extCre**: education variable?

outline

basic concepts

summarizing one variable (Wheelan, 2013, ch2): central tendency and dispersion [2 vars next week]

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: ex 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%

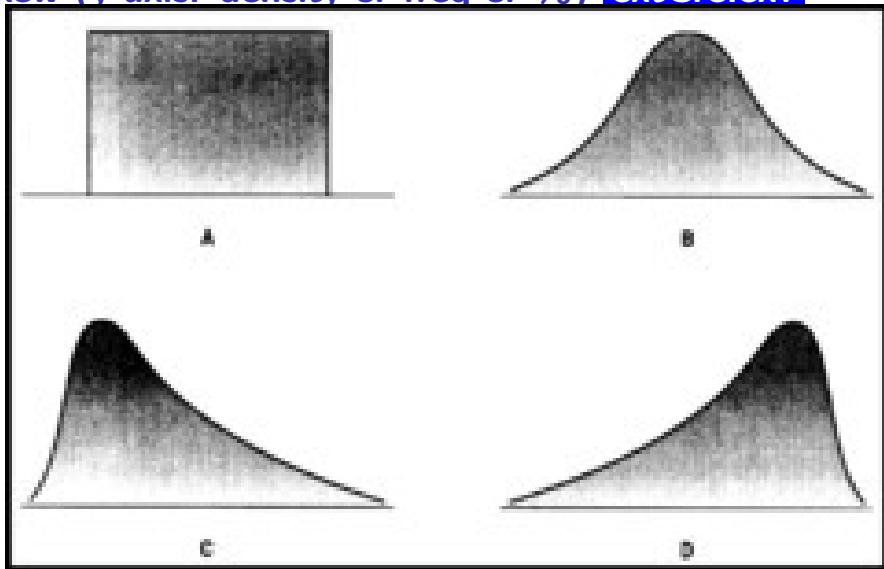
- <http://www.socialresearchmethods.net/kb/statdesc.php>: tab1, fig1

also (Wheelan, 2013, p20-21)

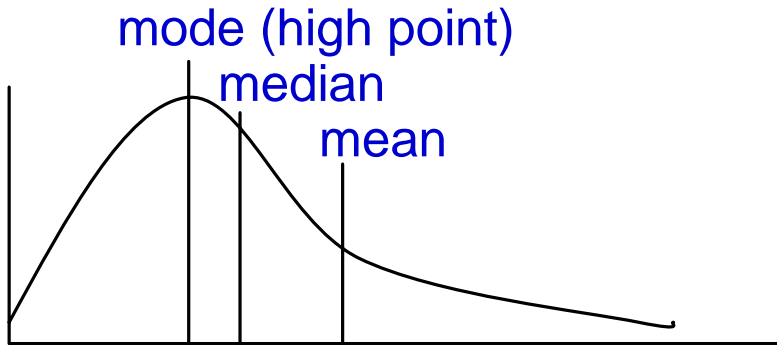
distribution types

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

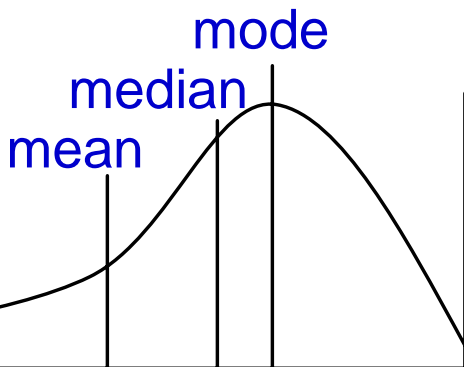
skew (y-axis: density or freq or %) **extCre:ex?**



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



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

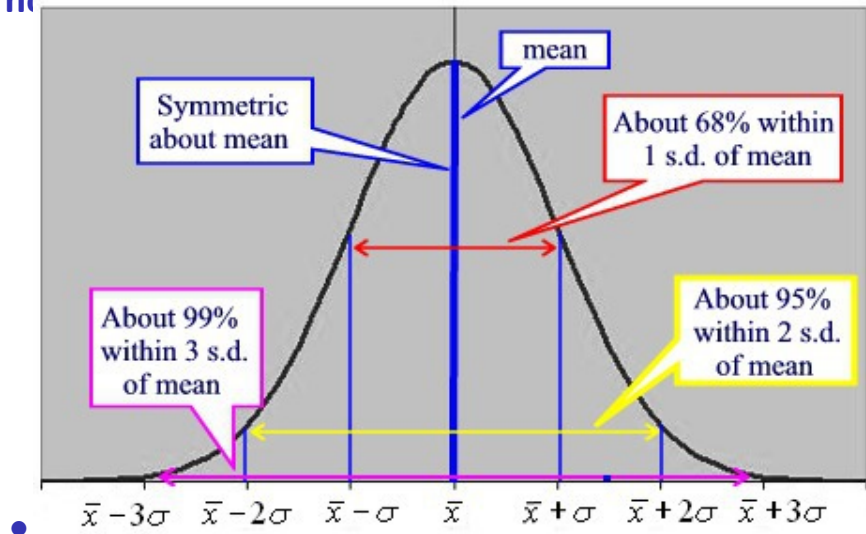


variability

- $range = max - min$
- 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

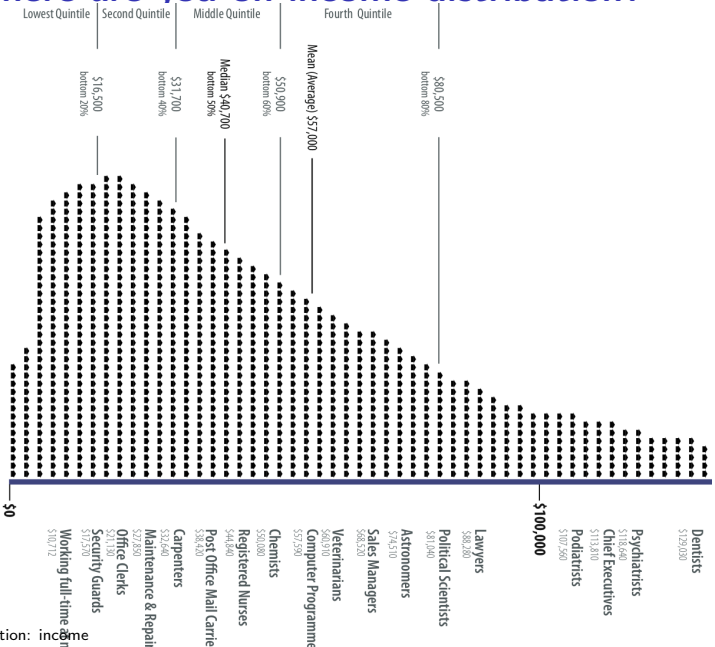
outline

basic concepts

summarizing one variable (Wheelan, 2013, ch2): central tendency and dispersion [2 vars next week]

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 geo, 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/>

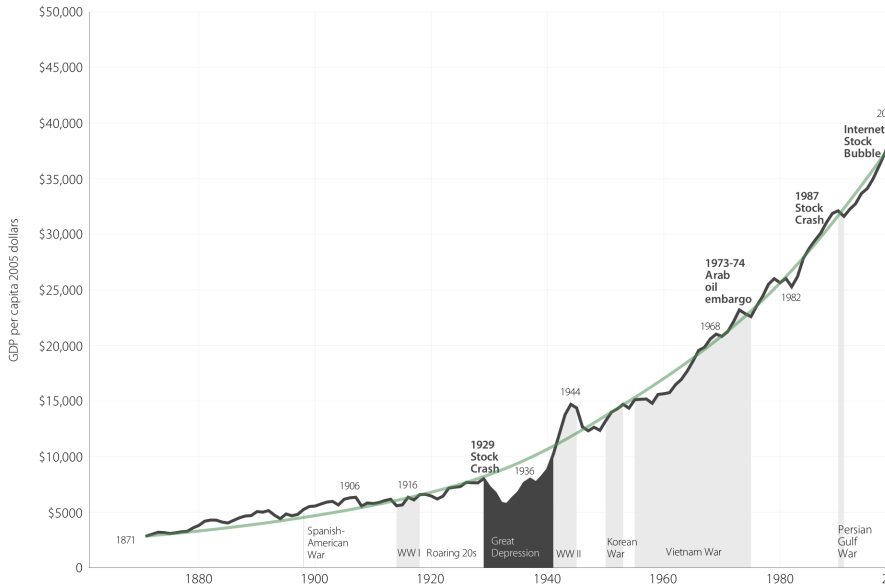
- also see Wheelan (2013, ch2) and http://en.wikipedia.org/wiki/Household_income_in_the_United_States#Household_income

○

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

Long-term real growth in US GDP per capita 1871–2000

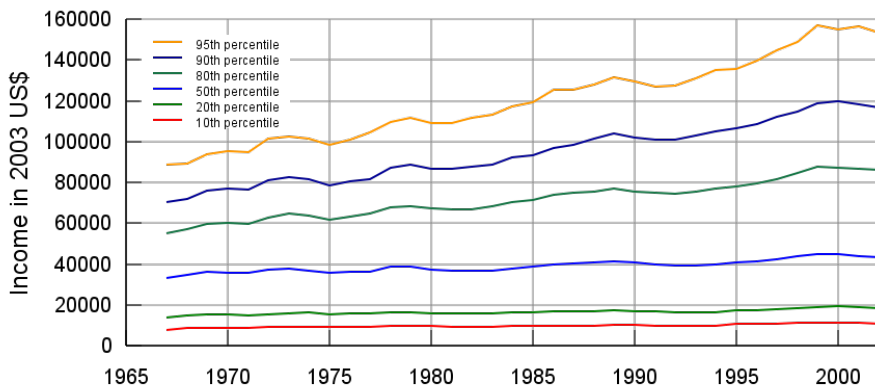
GDP per capita adjusted for inflation using 2005 dollars



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 1st decile has better quality of life than 9th decile 100 years ago (Derek Bok [\(Bok, 2010\)](#))
- can you translate this to plain English? **extCre**

discussion

- US News tweaking their methodology on college rankings
 - <https://www.insidehighered.com/admissions/article/2021/09/13/us-news-tweaks-its-methodology>
- the methodology looks at SAT and ACT scores which a lot of schools shifted away from due to COVID-19, US News needed to adjust the formula
- The new calculations decreased the percent of students who needed to submit their scores benefiting schools with more affluent students whose parents pushed them to take the tests prior to COVID-19.
- Another interesting point in the article is that Forbes changed their methodology which helped increase public

wrap-up

- end every class discussing what we covered and quick look at next week
- end with a review Q&A,
- give some examples (essp in pub pol and pub adm) for concepts covered
- students will discuss concepts from the class
-
- quick look at next class

bibliography I

BOK, D. (2010): The politics of happiness: What government can learn from the new research on well-being, Princeton University Press, Princeton NJ.

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