descriptive statistics 1

Adam Okulicz-Kozaryn

`adam.okulicz.kozaryn@gmail.com`

this version: Wednesday 9th September, 2020   14:10

<u>outline</u>

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

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

◇ 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!

$\diamond$ http://www.tylervigen.com/

$\diamond$ a fundamental piece of knowledge: correlation $\neq$
causation

$\diamond$ at policy drafting stage–easy to mistake correlation for
causation and draft unnesessary or wrong policies

$\diamond$ at evaluation stage–easy to see positive effect of policy
(sunk cost, groupthink,etc) while there is none!

$\diamond$

$\diamond$ from evol/beh persp: humans see causes where there
are none

level of measurement

$\diamond$ real continuous: interval/ratio (price, weight, temp)

$\diamond$ continous/categorical: ordinal (rank of faculty, grades)

$\diamond$ real categorical: nominal (many) or binary (two)
 (eg mode of transportation, gender)

$\diamond$ 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

$\diamond$ 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

·

$\diamond$ 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

$\diamond$ uniform

$\diamond$ normal symmetrical unimodal

$\diamond$ left skewed

$\diamond$ right skewed (income)

$\diamond$ bimodal

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

◇

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

◇

mode (high point)
median
mean
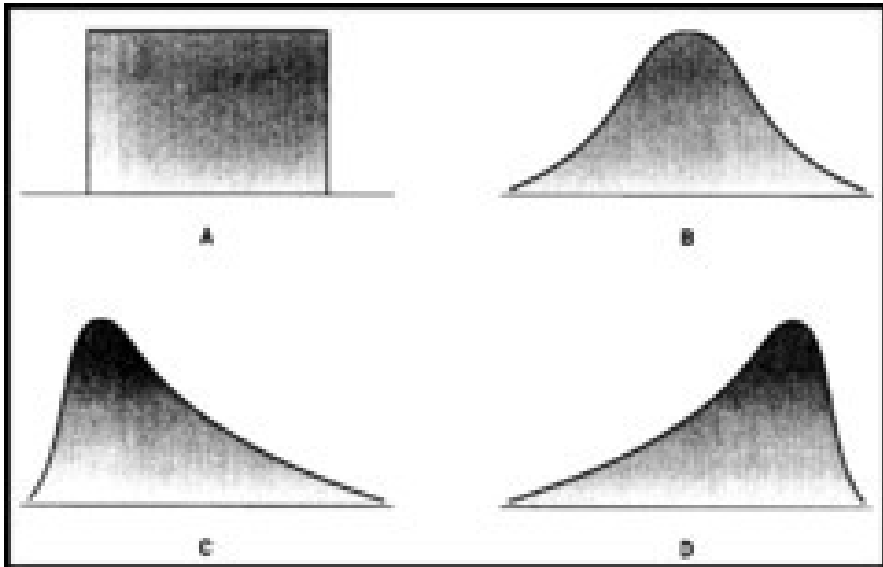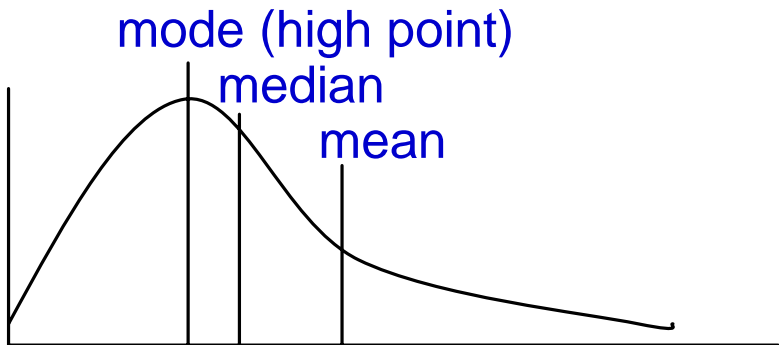
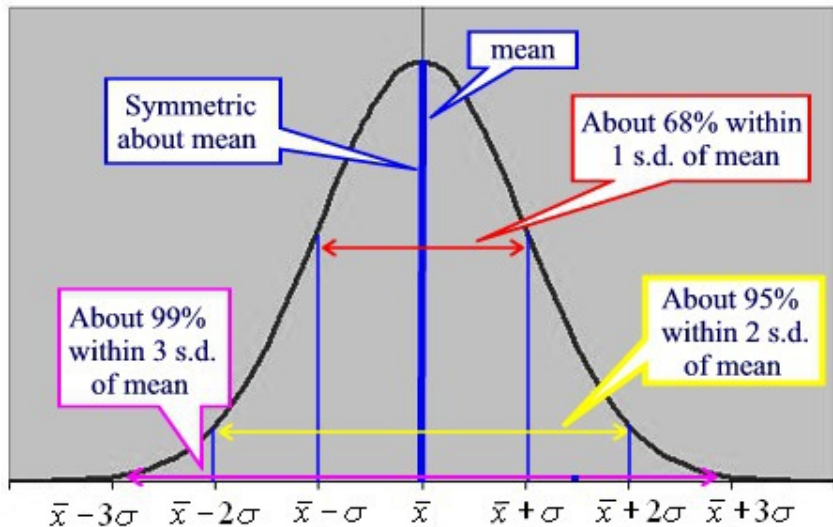# $\mu < M$: left skew (y-axis: density or freq or %)

◇ .

mode

median

mean
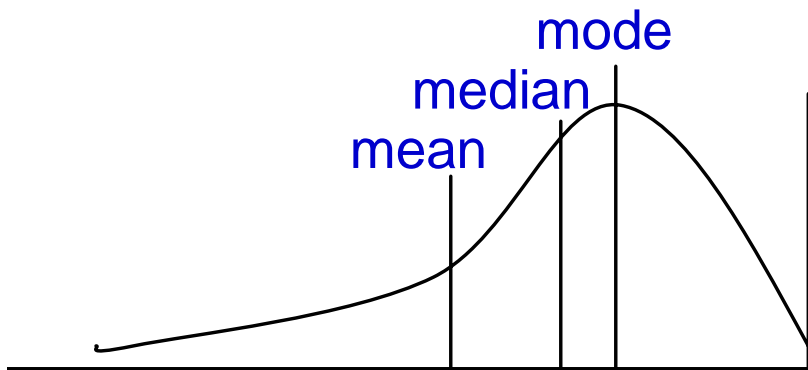
variability

◇ $range = max - min$

◇ 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)

◇

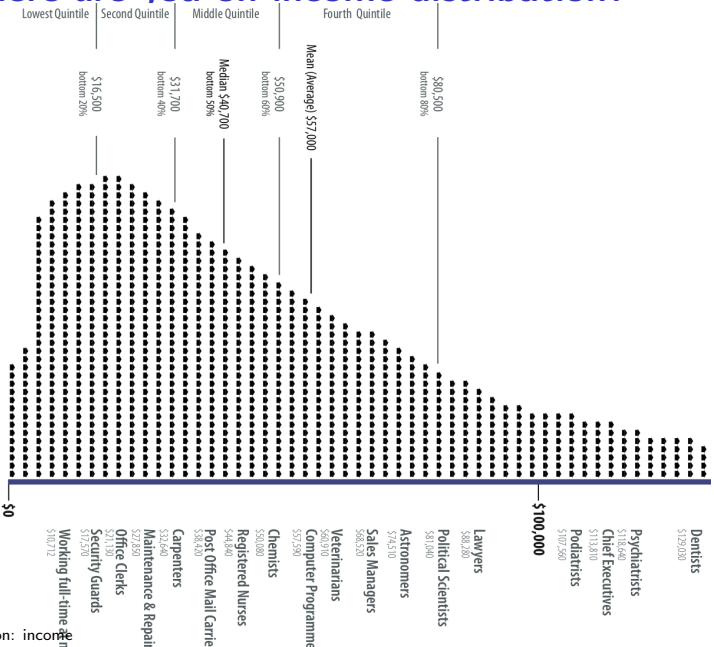· asymptotically, any variable is normally distributed

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?

Lowest Quintile | Second Quintile | Middle Quintile | Fourth Quintile

$16,500 bottom 20%
$31,700 bottom 40%
Median $40,700 bottom 50%
$50,900 bottom 60%
Mean (Average) $57,000
$80,500 bottom 80%

$0

$100,000

Working full-time $10,712
Security Guards $17,570
Office Clerks $21,130
Maintenance & Repair $27,850
Carpenters $32,640
Post Office Mail Carriers $44,840
Registered Nurses $54,210
Chemists $50,080
Computer Programmers $57,350
Veterinarians $60,910
Sales Managers $68,520
Astronomers $74,510
Political Scientists $81,040
Lawyers $80,260
Podiatrists $107,560
Psychiatrists $118,640
Chief Executives $113,810
Dentists $129,030

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))...
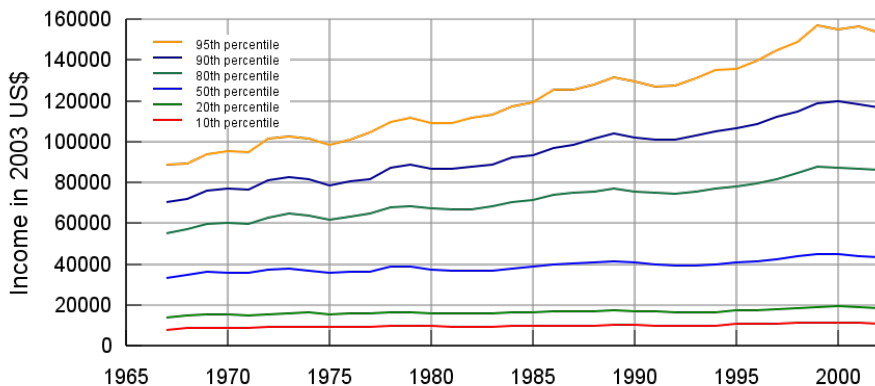
Long-term real growth in US GDP per capita 1871–20...

GDP per capita adjusted for inflation using 2005 dollars
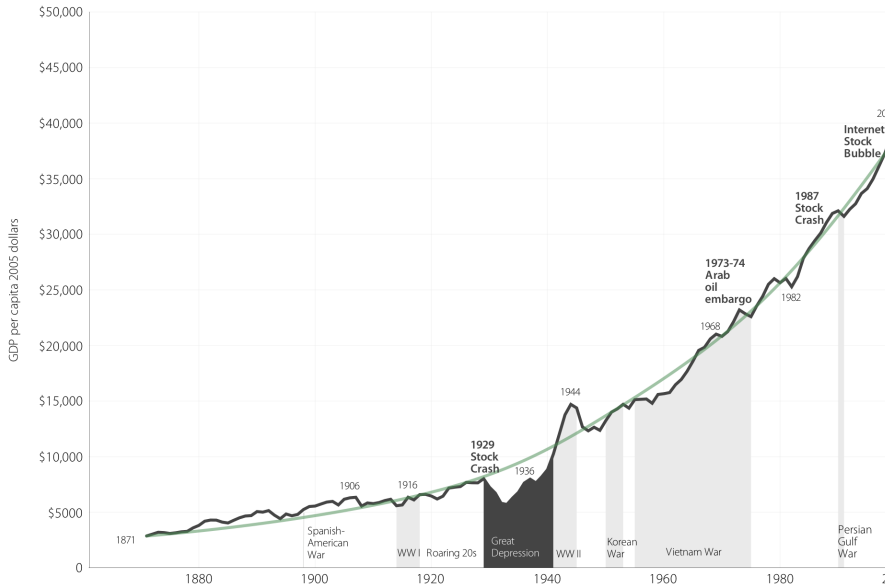
Data from MeasuringWorth.com

application: income   21/25

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)

· can you translate this to plain English? extCre

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

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.