Lecture #8
6/3/11

Admin

- HW #2 out, due Thurs
  If not in by Friday 5pm ⇒ 0
- I will not be here Wed/Thurs
- Trying to get midterm review room Sun 8-10PM

Plan

- Statistics (cases)
- visualization tools
- data tables
- types of data
  ch 2, 3, 4
- boxplots????
- Demos......

Data: realizations from r.v.'s. The # of data pts, or "observations" or "cases" is n (or sometimes other cases "N")

Statistics: function of data ← stat notation

Purposes ① to estimate stuff you care about ② to summarize a bunch of the data

Summary statistics

- sample avg
- sample var
- sample std dev

  (formulas ↓)
- mode ⟶ most likely # could have multiple modes
- percentiles
  - order the data
    the 10%ile is where 10% data falls below
    the 85%ile is where 85% falls below

imagine $n=100$

$X_{(1)}, \ldots, X_{(10)}, X_{(11)}, \ldots, X_{(100)}$
↑ 10%

- median: 50th %ile

  50% data below, 50% data above

  ... how to calculate? Order them up all up, take middle or...

- quantiles 25th %ile, 75th %ile hyperest: 50% of the data in the middle

- IQR: 75th %ile - 25th %ile

  interquartile range

- range: largest – smallest

- 5 # summary: (smallest, 25th %ile, median, 75th %ile, largest)

- coefficient of variation $C = \frac{s}{\bar{x}}$ (not covered) don't worry about it..

All the summary stats

What is a data table? data set? data frame?

← p → variables, objects, features, attributes, regressors

Stat cols

are cols

rows obs's

n

Numeric data has a unit usually. Sometimes a ratio or %.

Categorical - ordinal (has an order)

Numeric — interval (int)
— ratio (?)

Categorical — nominal (no order)
— ordinal (has order)

Can't be - binary (categorical - nominal

Sometimes data is recoded

ch 5/6
ch 12, ~~13, 14~~, 19

H → 80°
M → 65°
L → 50°

Bar

Best → 5
Good → 3
Fair → 1
Poor → 0

order present
up to you how to recode

We are doing univariate statistics:

PICK 1 Variable ... look at distribution
incl summary statistics

Patience for 2 var's ... (bivariate) later...

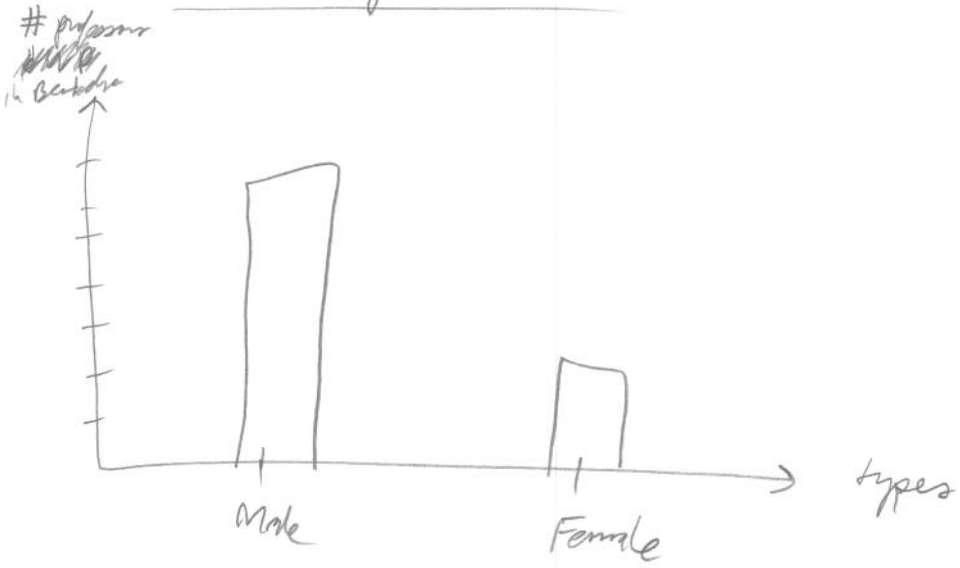
→ How do we do this for Quant data?

Hist
Freq table
Bar Graph / Pics ✓
Box / whisker — 1.5 IQR ✓
Hist ≠ Bar Chart ✓
~~Pareto Chart~~
Archers ✓

Post Stage
Symmetry
multimodal
bell
uniform
tails
L, R skew — redun / avg

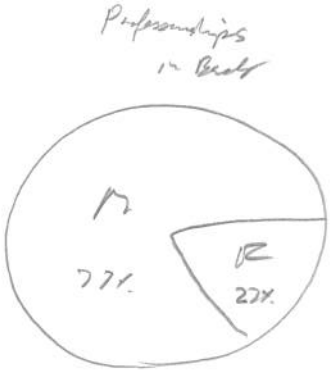
PEMDS

Store w/ categorial data



Freq table

Gender	# professorships
M	100
F	30

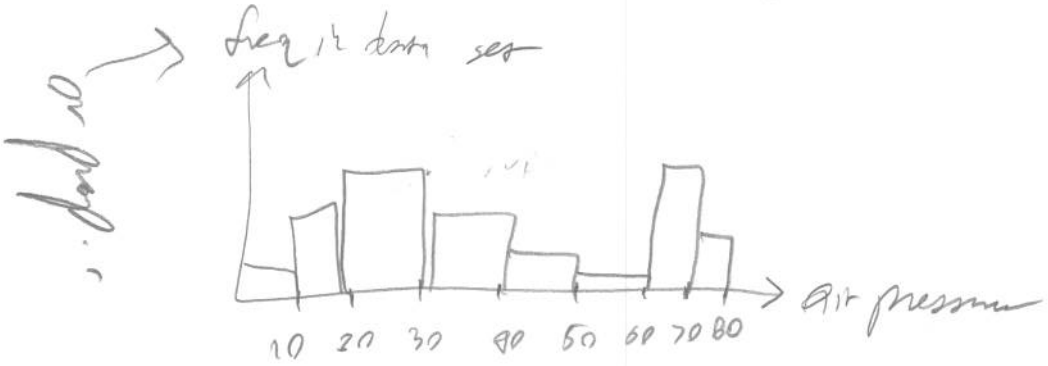


Bar plot (vertical)

Pareto Chart — Ordered by freq... easier to read

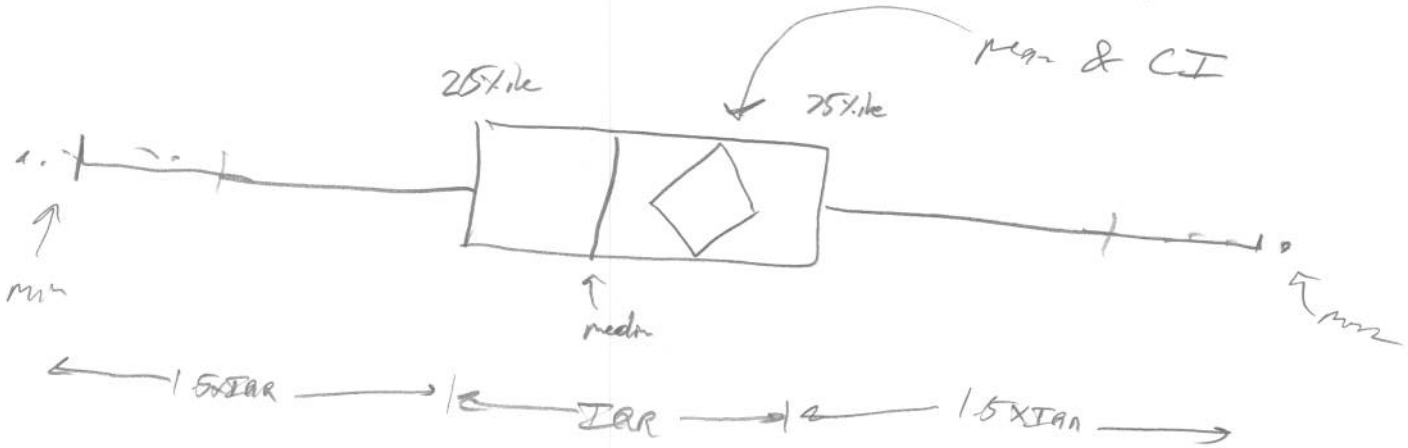
Numerical Data 57.5, 60.1, 32.7

The most normal thing to do is bin the data. Use some rules for binning.



NOT A BAR CHART!

Box and whisker plots 5 + # summary!

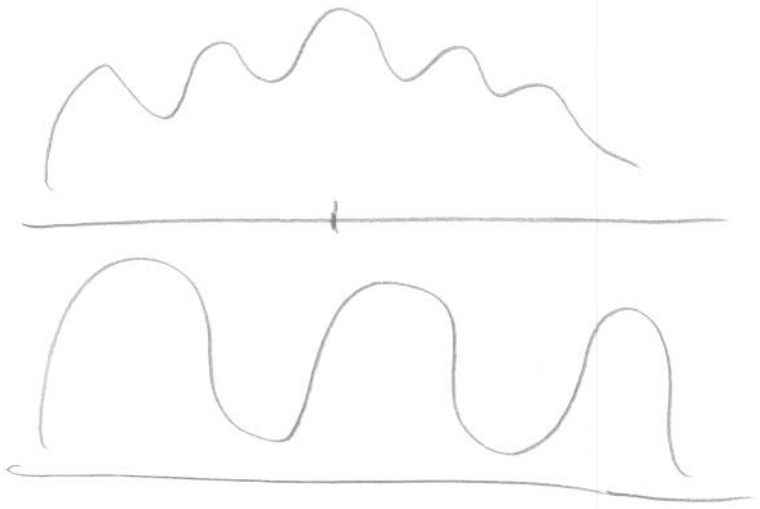


Anything outside the whiskers ⇒ outlier

These plots are great to ~~spot~~ visualize outliers

JMP: hist + box & whisker = good combination

Types of distr's

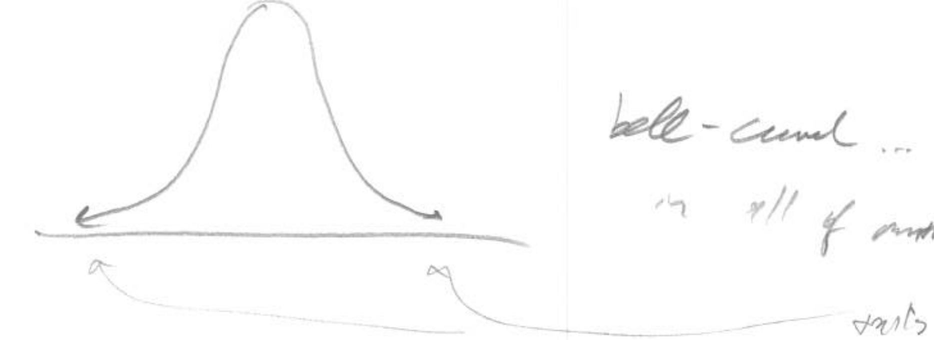


Symmetric

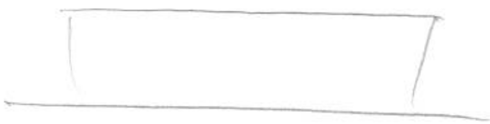
mirror image

mean = median

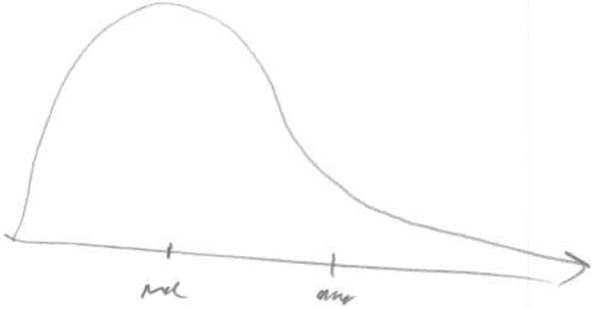
Multimodal



bell-curved ... most common shape in all of nature

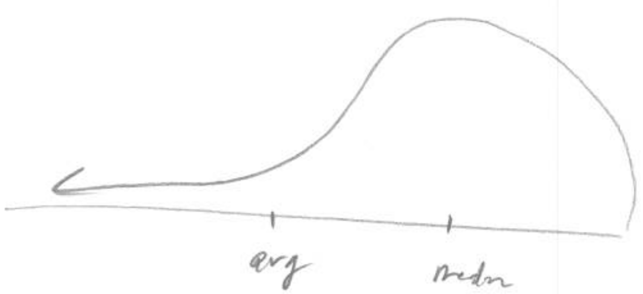


Uniform



right skewed

avg > med



left skewed avg < med

"long tail"

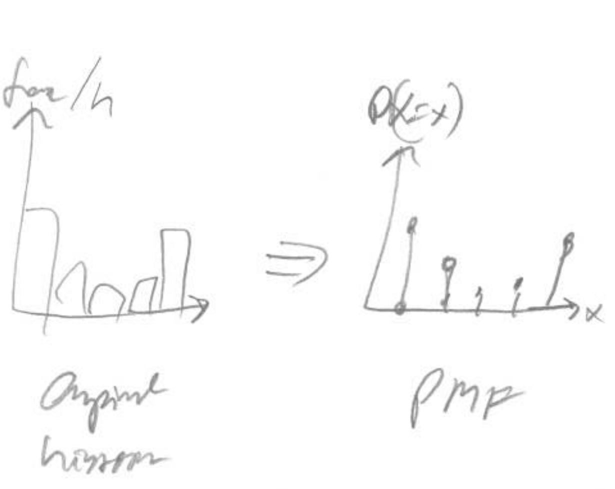
Data transformations

~~Skewness~~

$$\text{skew} = \mathbb{E}(X-\mu)^3$$

$$\text{sample skew} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^3$$

measures lack of symmetry



Revenue

Berkeley Advisors

All Comp: Net Sales ...