* studying data: formalized process
* statistics - describes any bit of information
* 1986
  Challenger - o rings died
  Not seal because of cold
  - plot ranges are very off - only showed best parts
  - resiliency of rubber was impacted by temperature

Population: all potential observations from a distribution of interest

enumerative study:
- tangible populations
- sampling frame: list of population exists and possible
- finite

analytic study:
- ongoing process
- conceptual/hypothetical
- ~infinite ex: rainfall over time
  ex: dripping pop - all the milk coming

Sample:
measurements: ex: height, etc
  ↳ make up dataset

Observational Study
- scientists cannot control variables of interest
- step after controlled studies

- is sample representative of pop?
- census = complete enumeration
  all in computing - every person is measured
* random sample is best
  simple random sample
  "draw names from hat"
  sample variation
  sample tends to be diff from population
  samples of convenience
  - can be biased
  easiest to collect

volunteer bias
- people choose to participate or not
* historical control

variable: characteristic we are interested in

data
- categorical - not/partially additive
- numerical - additive

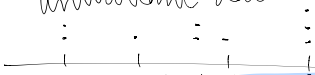
univariant - one imp bit
bivariant - 2 imp bits
tri/quad/multivariant (etc)
ex: X_i = (height$_h$, color$_c$)

descriptive stats:
describes & simplifies data sets

inferential statistics:
attempts to say something about data

Dotplots

univariant data



question: typical datapoint looks like...

Histograms

- bar chart for numbers
- shape describes distribution

Create

1) find range (max − min)
2) Breaks into intervals
   ~ √sample size
   class = range / # classes
3) freq chart
4) column chart

ex: range = 44 − 6

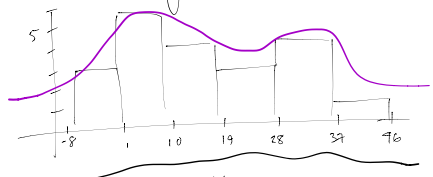
5 classes

class size = $\frac{44-6}{5} = \sim 9$

3) class	freq	relative freq
[-8, 1)	3	0.15
[1, 10)	5	0.25
[10, 19)	4	0.2
[19, 28)	3	0.15
[28, 37)	4	0.2
[37, 46)	1	0.05
	(# of data) = 20	= 1

how do you decide class size?

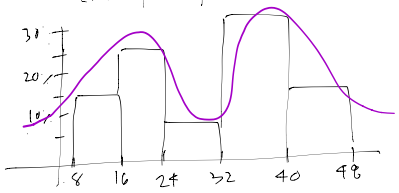
histogram



ex2 6 classes

$\frac{44-6}{6} \sim 8$

class	freq	R freq
[-8, 0)	3	15
[0, 8)	5	25
[8, 16)	1	05
[16, 24)	6	3
[24 32)	2	1
[32, 40)	3	15



Shapes

- dist. left skewed — neg
- symm
- right skewed — positive
- bimodal
- multimodal

*common features are more trustworthy

*shape related to real life events

Density Histograms

use for pre-binned data

Measures of location

- summary of data set
- sample statistic - derived from sample
- sample mean - most common for sample center
- μ - mean from all population, not sample
- mean is greatly effected by outliers

outlier very different from rest, generally wrong data

robust resistant to outliers

(mean $= \bar{X}$ / median $= \tilde{X}$) $X_{(i)}$ = difference, $X_{(1)}$ = smallest, etc

median - order stat, more robust

if n is odd, $\tilde{X} = X_{(\frac{n+1}{2})}$

if n is even, $\tilde{X} = \frac{X_{(\frac{n}{2})} + X_{(\frac{n+2}{2})}}{2}$

$\tilde{\mu}$ = median of population

categorical data - measured in proportions (from population)

$\hat{p} = \frac{x}{n}$

+skew = $\bar{X} > \tilde{X}$

-skew = $\tilde{X} > \bar{X}$

symm = $\tilde{X} = \bar{X}$

variability/spread

Range → susceptible to outliers

deviations from mean

how far from mean for each datapoint

- + & - dev. cancel each other out, = 0

Sample variance: measures spread of data

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1} S_{xx}$$

sample standard variation: $\sqrt{}$ of sample variance

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left[\sum_{i=1}^{n} x_i\right]^2\right]$$

ex: 19 data points

$$\frac{1}{19}\left[\sum_{i=1}^{20} x_i^2 - \frac{1}{20}\left(\sum_{i=1}^{20} x_i\right)^2\right]$$

$s^2 = 186.42$

$\sqrt{186.42} = 13.65 = s$

not robust

~2/3 data falls w/i 1 SD

~95% in 2 devs

~all data in 3 std $(\bar{x} - 3s, \bar{x} + 3s)$

ex $(\bar{x} - s, \bar{x} + s) = (1.71, 29.02)$

$(\bar{x} - 2s, \bar{x} + 2s) = (-11.95, 42.68)$

true for normal distribut.

Ipad died ☹ January 25, 2016

Question 8:
Average = sum of Xi's/10

10(63,000) = sum of Xi's

1/10th of old number subtracted
add another 1/10th of new number, = new total

n-1 in variance – corrects for bias
n in population variance
usually don't know population mean, only have sample

IQR = Q3 - Q1
EX: 72.6 - 21 = 51.6
outliers, small:
21 - 1.5 * 51.6
= 56.4
72.6 + 1.5 * 51.6
150

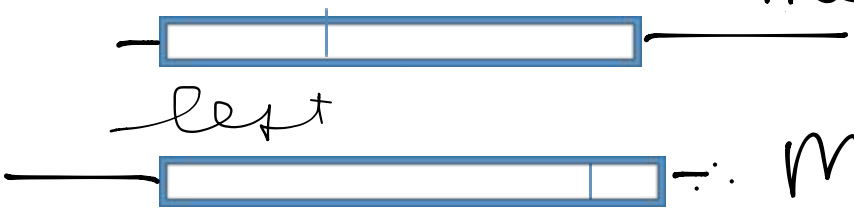
Finding Quartiles
Q1: median of bottom half:
Q2: median
Q3: median of top half:

-robust measure of spread
-interquartile range or fourth spread: Q3-Q1

Boxplot
Plot lower quartile and upper quartile, put box around it and a line where median is
"whiskers" where greatest and least ***last non-outliers*** **exist**
outliers: inner quartile range * 1.5
if no outliers exist, the whiskers sit on top and bottom values

Inner quartile range = Q3-Q1
small outlier cutoff: Q1-1.5 *(IQR)
large outlier cutoff: Q3+1.5*(IQR)
DIFFERENT THAN WHISKER POINTS

Right skewed



∴ mean > median (usually)

left

∴ mean < median

Not clear is skewed or not if whisker is out further, for example:

Histogram vs. Box Plot?
Histogram shares more information, but difficult to compare
Box Plots are easier to compare
Probability:
Dealing with chance, randomness, and uncertainty
Experiment:
Cannot be determined in advance what outcome will be
Sample Space:
Set of all possible outcomes

Randomly pull card?
Cards:
S = {ace of spades, 2 of spades, ace of hearts, etc…}
Event: draw a hearts (subset, A-K hearts)

Component breaks on computer?
S = (0, infinity) or [0, infinity) or [0, 1 million]
Subset = [0, 20]

Event: "interesting" subset of sample space
Empty set – subset of every set

Sample range
highest - lowest
Sample standard deviation:
sum of squares of differences
divided by (n-1) & then √
of final answer !!