L13: 15 Feb : Exam 1 Review.

Housekeeping.

- Exam 1 is Friday (!)
- Optional homework posted – sol'ns to be posted tonight
- No written homework for Friday (just study...)
- Week 2 reading summary + discussion due Monday 11:59 p.m.
- ※ No class on Monday.

Study Guide.

- Population / sample
- Parameter / Statistic
- Variables / data
- Descriptive / inferential statistics
- Qualitative data
  vs.
  Quantitative data < cts. / vs. / discrete.
- Dot plots
- Freq. tables
- Histograms
- Stem & leaf diags.
- What makes a good infographic?
  - Nightingale — military deaths in Crimea
  - Minard — Napoleon march.

- Sampling
- Simple random
- Stratified
- Cluster
- Opportunity
- Systematic

} technical problems – (e.g.,) how to select proportionally the sample candidates from each stratum

- Plotting data in pie charts
- ——— a ——— bar charts

} – Sorted pie charts vs. unsorted
– Bar charts vs. Pareto charts.

- Using random # generator.
- Infographic: Premier League Results (Financial Times)
- Ethical issues w/ sample selec'n.
- When to use (or not) certain sampling techniques
  - Costly, controversial, impossible situations to create sampling frame
  - Cluster sampling is <u>cheaper</u> than stratified sampling
  - Problems w/ opportunity (a.k.a. convenience) sampling.
- Infographic: John Snow Cholera map & Voronoi diagram.

- Sampling error
- Non sampling error
- Sample bias

} real-world examples.

- Sampling { with vs. without } replacement.

- Mean, median, mode : measures of center of data
  - Outliers : the mean is much more sensitive to outliers than the median is.
  - Distinc'n btwn. measures of ctr. ⁉ meas. of spread.
- Creating a freq. table or histogram — determining class bdries.
- Infographics : scatter plot of temps. (NYT)
  time series of global temps. (xkcd.com)

- Descriptive statistics

Sec'ns (from O.S. book):

2.1
2.2
2.3
2.4
2.5
2.6
2.7

- Box plots (box & whisker plots)
- Quartiles & IQR
- Std. deviation & variance } opt'l HW assignment
- z-scores

- Infogr. (not totally relevant): corn / milho.

S.d. & z-scores.

Measures of center – mean, median, mode
– tells the "typical" value of a pt. in a data set

Measures of spread – IQR, s.d., variance
– tells how spread out the data is about its central value.

IQR measures how spread out the data is abt. the median;
S.D. & variance ——→ mean.

Naïve idea / First try: Find the average distance btwn. data points & the mean:

- Compute the mean, call it $\bar{x}$.
- For each data pt. in the set, compute the distance btwn. $\bar{x}$ & that point.
  i.e., if x_i is a pt. in the set, compute $\bar{x} - x_i$.
- Find the mean of those distances.

Try this algorithm out for the set $\{3, 6, 4, 3\}$

- $\bar{x} = \frac{3+6+4+3}{4} = \frac{16}{4} = 4$.
- Distances $= \{4-3, 4-6, 4-4, 4-3\} = \{1, -2, 0, 1\}$
- "Spread" $= \frac{1 + (-2) + 0 + 1}{4} = \frac{0}{4} = 0$.

Slight modification:

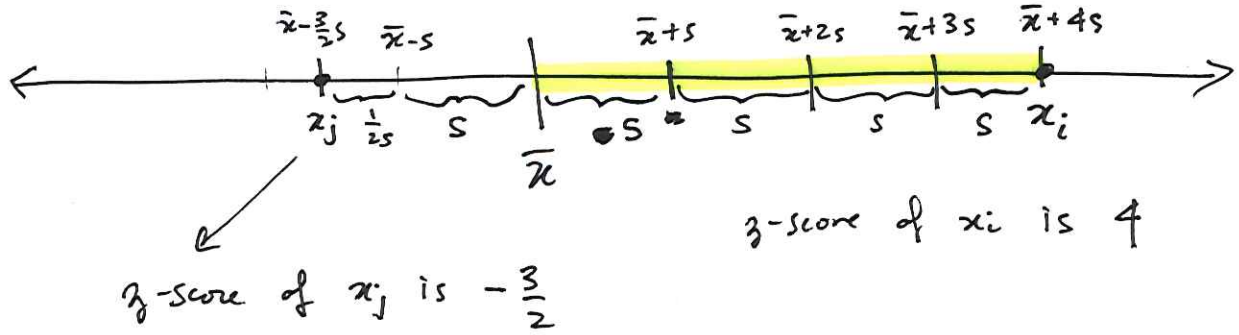
- Compute $\bar{x}$
- For ea. data pt., compute $(\bar{x}-x_i)^2$.
- Find the mean of those, and call it the variance, denote it as s^2.
- Take the sqrt of the variance, call it the s.d., s.

For technical reasons, when computing this mean, divide by $n-1$, not n, where n is the # of data pts. in the set.

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

- s.d. is a property of a data set.
- z-score —— " —— point in a set.

z-score of a pt. tells you how many s.d.'s above the mean that pt. is.



z-score of x_j is $-\frac{3}{2}$

z-score of x_i is 4

the z-score of x_i in a set w/ mean $\bar{x}$ and s.d. s is:

$$z := \frac{x_i - \bar{x}}{s} .$$