STAT206

ANDREW CODISPOTI

1. Descriptive and Inferential Statistics Terminology

(1) Statistics: Science of conducting studies to collect, organize, summarize, analyze and draw conclusions from data.
(2) variable:attribute can be diff values
(3) Categorical variables(Gender), Quantitative variables (age weight)
(4) Discrete var: values that can be counted
(5) continuous var: assume all values between any two specific values, measured
(6) random: vals determined by chance
(7) Data set:collection of data values
(8) descriptive statistics: describe situation in data (collect, organize summarize, present data)
(9) inferential statistics: make inference from samples to population
(10) population:all indviduals under study
(11) sample group selected
(12) statistical hyp: claim
(13) hypothesis testing: decision making process for evaluating a claim about population from samples

study samples of pop and use inferential stats to make inference about pop

1.1. **Freq distribution and graphs.** raw data in original form organized in some way (i.e. freq distribution). can then be presented in graphs or charts

freq. dist organizes raw data into classes and frequencies.

range of data is large (lower class limit, upper class limit, class boundaries, class width, class midpoint, open-ended distribution)

1.2. **General rules for grouped freq. dist.** 5-20 classes, class width should be odd to ensure midpoints are integers, mutually exclusive(non overlapping), continuos, exhaustive, equal in width

1.3. **general procedure.**

(1) determine classes
    (a) find highest(H) and lowest(L) values,
    (b) range = H - L
    (c) number(N) of classes desired
    (d) width = R/N (round to whole)

(e) lowest class limit, convenient value less than or equal to the smallest data value.
(f) add W to get lowr limit of next class
(g) class boundaries
(2) tally data
(3) find freq. of tallies
(4) cumulative freq

1.4. **Histogram.** histogram uses vertical bars to rep. freq of classes
different shapes bell shaped, uniform, j-shaped reverse j-shaped, right and left skewed, bimodal, u-shaped

1.5. **Stem and leaf plots.** uses part of data value as stem and part as leaf to form groups
more informative because shows values
first arrange data in order, separate into classes then plot
look for peaks and gaps, shape of distribution, variability by looking at spread
can put back to back to compare with other data set

1.6. **Descriptive Measures.** describe data with numerical measures
characteristic or measure form population is a parameter, from sample data is called a statistic
Measures of central tendency(mean median, mode) **don't round until done**

1.6.1. *Average.* sum divided by total

$$\mu = \bar{X} = \frac{\sum X}{N}$$

rounding rule: final answer should be rounded to one or more decimal place than original

$$\bar{X} = \frac{\sum f\dot{X}_m}{n} \tag{1}$$

1.6.2. *Median.* ordered so called data array
half smaller half larger, position $\frac{n+1}{2}$
when even number of values, b/w two given data values

1.6.3. *Mode.* value that appears the most frequently
data set can have multiple or no modes

1.6.4. *properties of central tendency.*
(1) mean needs all values of data, varies less than media and mode, unique, cannot be computed for open-ended freq. distribution, highly affected by outliers
(2) median find middle, whether data falls into upper or lower half of the distribution, can be calc for open data dist. less affected by outlier
(3) most typical case req., easier to compute, can be computed for categorical or nominal data, not unique

2. Variation

2.1. **Measure of Variation.** range is simple measure of variability

Variance is measure of variability that uses all the data points, avg. deviation of values from mean.

population variance σ^2 is avg. of squared deviations of values from population mean μ

$$\sigma^2 = \frac{\sum (X-\mu)^2}{N} \tag{2}$$

where N represents total no. of values in pop

sample variance denoted by s^2 sum of squared deviations of the values from sample mean $\bar{X}$ divided by $(n-1)$

$$s^2 = \frac{\sum (X-\bar{X})^2}{n-1} \tag{3}$$

where n is no. values in sample

divided by n-1 b/c unbiased estimate of σ^2. standard deviation is square root of variance

poulation standard deviation

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (X-\bar{\mu})^2}{n-1}} \tag{4}$$

sample standard deviation

$$\begin{aligned} s &= \sqrt{s^2} = \sqrt{\frac{\sum (X-\bar{X})^2}{n-1}} \\ s^2 &= \frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n-1} \\ s^2 &= \frac{\sum f\dot{X}_m^2 - \frac{(\sum f\dot{X}_m)^2}{n}}{n-1} \end{aligned} \tag{5}$$

2.2. **Note.**

(1) value of s^2 always +
(2) sum of (X - $\bar{X}$) is always 0
(3) larger value of s^2 or s, larger variability of data (same with σ)

3. Empirical(Normal) Rule

distro. bell shaped

some % of vals fall within 1 standard deviation of mean (X-nS, X+nS)

4. Chebyshev's Theorem

in an dist. regardl of shape, proportion of value that fall within k standard deviations will be at least $1 - \frac{1}{k^2}$where $k > 1$

subtract average from larger value, divide diff by standard deviation to get k, use cheb to get %

5. Measure of Position

standard score / z-score subtract mean from observation and divide result by standard deviation $z = \frac{X - \bar{X}}{s} or \frac{X - \mu}{\sigma}$

number of standard deviations falls above or below mean

5.1. **Partition Values.**

5.1.1. *percentile.* 100 equal groups

5.1.2. *decile.* 10 equal groups

5.1.3. *quartiles.* 4 equal groups

$$Percentile = \frac{number of data value below X + 0.5}{n} * 100\% \tag{6}$$