Test 1 Review

January-13-15
5:03 PM

Population	Set of units
Variable	Characteristic of Population

- Can carry out **measurements** to give **values** to variables
- If you want to measure every population unit, you do a **population of measurements**
  - **Census** is examining all population measurements
  - Can only do for small populations

- **Sample** is a subset of the population
- **Sample of measurements** is measuring the SAMPLE

Descriptive statistics	Describing important aspects of a set of measurements
Statistical inference	Using a **sample** of measurements to make generalizations about the whole **population** of measurements

Random Samples	A sample, where, whenever you choose a new unit, the units that remain have the same chance of being chosen

- Can do **With Replacement** or **Without Replacement**
- Can do with a **Random Number Table** or a **Random Number Generator**

- List of units is called a **frame**, therefore, we already have the list
- If you don't have the list, you can do a **systematic sample,** which is close to a random sample

- Can sample a process, or studying what could potentially happen
  - Process is the sequence from raw goods to finished product

Finite Population	Measuring what has or will be produced
Infinite Population	Measuring what could potentially be produced

- When you a measure a process and it has no unusual process variations, it is in **statistical control**
  - **Constant amount of variation**
    - This variation is around a **constant** level, or a horizontal line
  - Predictable, can make inferences about it
- Can do a **runs plot** or **times series plot**
  - Process measurement vs. Time

- **Qualitative**
  - **Nominal/Nominative**
    - Gender, Colour of a car
  - **Ordinal**
    - When you ask someone to **rank** something
- **Quantitative**
  - **Interval**
    - Distances within points are fixed and meaningful
    - Zero is arbitrary
    - Ex: A 1-7 scale measuring satisfaction
  - **Ratio**

        - Meaningful zero
        - Money, grades
    - **Continuous**
    - **Discrete**

- **Making a survey**
    - Order of questions matters

- **Sampling Designs** are ways of taking a sample and making a **sample survey**
- **Stratified Random Sampling**
    - Divide the population into non-overlapping groups of similar units called **strata,** and then choose a random sample from each stratum
- **Multistage cluster sampling**
    - If you don't have a list for the sample you want to take, you can cluster the units and narrow down
        - EX: Choose a sample of counties
            - Sample of townships
                - Sample of households in each township
- **Can combine both**
    - EX: Divide country into strata (provinces) and do multistage sampling on these strata

- **Systematic Sampling**
    - Randomly choose first unit
    - Choose second unit by choosing the next 50th (depends on interval) unit
        - Keep choosing until you have the sample size you want
    - Number in total/Sample size you want=Interval

- **Undercoverage**
    - Not including some population units
        - IE people who don't have phones
- **Nonresponse**
    - When people don't participate or cannot be contacted
    - The people who DO participate will be biased
- **Response**
    - Structuring the questions in a bad way
    - Wording questions badly

- Stem and Leaf
    - Can construct histogram
        - Number of classes= 2^(x) where 2^(x) is greater than the units in total
        - Class length = (Largest unit-Smallest unit) /Number of classes
    - Can construct frequency distribution
        - Relative frequency
        - Percent frequency

- Normal curve
    - Positively skewed (right)
    - Negatively (left)

- Can use dot plot as another graphing method
    - Use number line
    - Use dots above each other to show frequency

- **Mean** is one measure of the set's **central tendency**

    - Mean is a population parameter, it describes an aspect of the population

- Can estimate a population parameter with a point estimate
    - One number estimate of the value of a parameter
    - Can use a sample statistic from sample measurements that describes an aspect of the sample
        - Sample mean (x bar or M)
            - Sample mean is (Sum of all the units/number of units)

- **Median** is the middle number in a population of measurements
    - If odd, is the middle measurement
    - If even, is the average of two middle measurements

- **Mode** is the most frequently occurring measurement
    - Bimodal=2 modes
    - Multimodal=Many modes
    - Most occurring **class=**modal class

- When comparing median and mode
    - Median is resistant to large units because it doesn't care about how big the units are
        - Median doesn't care about **exact size**
    - Mean is changed by extreme values, median is not

- **Range** is largest measurement-smallest measurement
- **Population Variance (σ^2)**
    - Average of ---> The (distance) deviation of each unit from the mean squared
        - EX (Distance 1^2+Distance2^2)/2
- **Population standard deviation is σ**
    - Square root of variance

- Can use sample variance and sample standard deviation
- **Sample Variance s^2**= (Sum of all the squared deviations/number of units-1)
    - This equals the **point estimate** of population variance **σ^2**
- **Sample Standard Deviation**
    - Square root of **S^2** is **point estimate** of population SD

-

Tolerance interval	An interval that contains a certain percentage of the population
Three-sigma interval	Interval that contains almost all of the population a normally distributed population

- Companies can provide **specifications** that specify what certain **individual measurements** should be, and if the process can meet these measurements, it is **capable**

- If we want to assume that a process is **consistently capable**, the measurements must be within the three-sigma interval (if it is normally distributed) of what the specifications are
    - Therefore, all measurements are in the criteria provided, satisfying the criteria

- Can use Chebyshev's theorem if you don't think that the empirical rule will work or for double mound distributions
    - Set K, greater than 1
    - $100(1-1/k^2)$ = How much of the population lies within k standard deviations of the mean
    - Can solve for answer

- Z scores
    - z= (x-mean)/SD

- z= how many standard deviations X is from the mean
- Positive=greater than mean

- Coefficient of variation=SD/Mean x100
  - Tells us how much, **relative to the mean,** the variation is higher or lower
  - Can be used to assess risk (higher=more risk)

- Pth percentile is when p percent of units equal to or less than p
- To calculate
  - Order in ascending order
  - Find i=(p/100)n
  - i=the number that is equal to that percentile (if a decimal, round up)
  - Find ith number, all scores equal to or less than this number are in the percentile

- Can use percentiles to describe very skewed populations
- Can also divide population into four parts, 25% each, called quartiles
- **First quartile=**25th percentile
- **Second=**50th percentile
- **Third=**75th percentile
- Five number summary= Smallest Unit, First, Second (median), Third, and Largest unit
- IQR= Q3-Q1 to show what is in the BOX=Middle 50% of measurements

- **Box and Whisper display or Box Plot**
  - Find Q1 and Q2 and Q3
  - Draw box From Q1 to Q3
  - Draw line at Q2 or Median
  - Draw inner fences
    - Q1-1.5(IQR)
    - Q3+1.5(IQR)
  - Draw outer fences
    - Q1-3(IQR)
    - Q3+3(IQR)
  - Draw whiskers
    - On left side, it is the smallest measurement between inner fences
    - On right side, it is the largest measurement between inner fences
  - Draw outliers
    - Between inner and outer fence=mild outlier (STAR)
    - Outside outer fence=extreme outlier (O symbol)

- Can use **bar chart** or **pie** chart for qualitative data
  - If you want to estimate a proportion of a category to the population you can use a sample to find the sample proportion, which is a point estimate of the population proportion

- **Pareto Chart**
  - Identify problems
  -

Vital Few	Few defects that make up most of the total
Trivial many	Many defects that make up the remainder of the total

  - Order frequency of defects in descending order, want to deal with most troublesome ones first
  - Line over bars is the **cumulative percentage point**

- **Scatter plots**
  - Use dependent and independent variable

    - Straight (linear) vs curved relationships

- Can use scale break to make data look far more impressive than it actually is
- Can used compressed axis to make data look like it has less of a trend

- **Weighted Mean**
    - Sum of all weights*values/sum of all values

- **Weighted mean for frequency table or grouped data**
    - Multiply class midpoints by frequency and add all of them together
    - Add all frequencies together
    - Sum of mxf/Sum of f=mean
- **Sample Variance for Grouped Data**
    - **Numerator**
        - Find the weighted mean
        - Find the deviation of each midpoint from mean and square it to find D
        - Multiply D by the frequency to find WD
        - Do this for each midpoint
        - Add all the WD`s together to find the sum to find the numerator
    - **Denominator**
        - Add all frequencies together for denominator and subtract one

- **Probability**
    - Classical method, where you assign probabilities if all outcomes are equally likely
    - Long-run relative frequency means that over the long run, probabilities observed will equal actual probabilities
        - Can use relative frequency method
            - If someone says something 14/100 times there is a 14% probability
    - Subjective probability
        - Basing probability on past events, cant recreate

    - Pr= (Desired outcomes/Total outcomes) if the events are equally likely