Lecture 1

Statistics

- Discipline of collecting, organizing, analyzing, presenting, and interpreting data
- The art of learning from data
- Statistics allow for intelligent judgements and informed decisions to be made in the presence of uncertainty and variation

Branches of Statistics

1. Descriptive Statistics- Summarize and describe important features of dta
    - Visually
    - Numerically
2. Inferential Statistics - use information contained in a set of data to generalize to a larger group
    - 40% of Covid infection are asymptomatic

Terminology

Variable - any measure of interested collected on an object whose value may change over time or from one object to another, generally denoted with a capital letter near the end of the alphabet

Types of Variables

Qualitative - measures a quality or characteristic, can be coded numerically. Includes numerical measures where mathematical operations don't apply (categorical)

- Programing language, Phone Number, Serial Number, Material

Quantitative - measures a quantity/numerical value where mathematical operations apply

- Heights, Temperature

Types of Quantitative Variables

Discrete- finite list of possible values, counting

- Heart rate, Shoe size

Continuous - infinite list of possible values within a specified interval, measuring

- Time until train arrives, temperature, length

Experimental Unit - object from which measurements are collected

- Variable: Time until train arrives
- Experimental Unit: Train
- Variable: Number of seat belts on airplane
- Experimental Unit: Airplane

Data - collection of measurements

Types of Data

Univariate - one variable measured on each experimental unit

- X = weight

Bivariate - two variables measured on each experimental unit

- X = weight
- Y = width

Multivariate - more than 2 variables measured on each experimental unit

- X = weight
- Y = width
- Z = Color (x,y,z)

Population - entire group of objects of interest

- All Boeing airplanes in service, All iPhone 11 manufactured

Census - study of all objects in the population

Parameter - numerical measure associated with the population, fixed value, generally unknown, generally denoted by a greek letter

- Average years boeing airplanes have been in service
- Proportion of defective iphone 11

Sample - subset of objects from the population

- 100 selected airplanes in service

Statistic - numerical measure associated with the sample, value varies from sample to sample, estimates the population parameter

- Average number of years for 100 selected airplanes
- Proportion of defective iPhone 11's from 800 tested

Collecting Data (Sampling)

Simple Random Sample (SRS)- all objects from the population are equally likely to be selected

Stratified Samples - separating the objects in the population into non-overlapping groups and sampling from each group

Convenience Sample - selecting objects from the population by convenience under the assumption that they will represent a random sample

Experimental Design - performing an experiment in a systematic way to observe the effect of a factor on a response

Visual Descriptive Statistics

Visual displays provide method for conveying and interpreting information contained in a set of data (table and graphs)

- Qualitative
- Quantitative

Categories of a qualitative variable may or may not have a natural ordering

- Programming Language : Java, C++ (nominal qualitative variable, no ordering)
- Size: x-small, small, medium (ordinal qualitative variable)

Objective: observe and interpret patterns

Frequency - number of observations for each value of the variable, denoted f

Relative frequency - Proportion of observations for each value of the variable relative to the total number of observations collected, denoted f/n where n total number of observations collected
Percentage - Relative frequency * 100%

Types of Tables
Frequency distribution - table identifying the unique values of variable and their corresponding frequencies
Relative frequency distribution - table identifying the unique value of a variable and their corresponding relative frequencies (Sum = 1)
Percentage frequency distribution - table identifying the unique values of a variable and their corresponding percentages (Sum = 100%)

Charts
- Bar
- Pie

Pie Chart - circle is divided into slices that are proportional in size to the frequency, relative frequency, or percentage of each unique value of the variable
Advantages: Easy to interpret when few unique values
Disadvantages - When slices are similar in size, if too many slices hard to read

Bar Chart - 2 dimensional plot with unique values of variable on the horizontal axis and frequency, relative frequency, or percentage on the vertical axis
Advantages: Quickly and easily compare heights of bars, Easy to compare multiple bar charts
Disadvantages: Rearranging values on x axis changes pattern

Visual Displays for Quantitative Variables
- Shape (symmetric, skewed right, skewed left)
- Center
- Spread
- Outliers (extreme observations)

For 2 or more variables: relationships, Outliers

Shapes: Symmetric
Left and right half of the data are mirror images of each other

Skewed Left: Tail of the data is on the left, most data is on the right
Skewed Right: Tail on the data is on the right, most data is on the left

Center: Balancing point of the data, most common or likely observation to occur
Spread- Variability of the data, grouping of the data

Graphs
1. Stem and Leaf Plot

2. Dot Plot
3. Histogram

Others

4. Box Plot
5. Bar Chart
6. Line Plot /Time series
7. Scatter Plot (bivariate data)

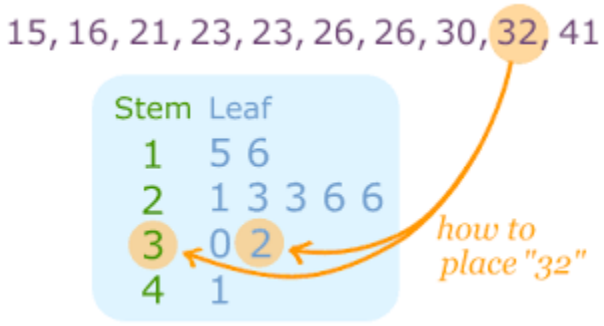
Stem and Leaf Plot

Use: Small to medium size data sets consisting of observations with at least 2 digits

1. Separate each observation into stem and leaf (Leaf = last digit, stem - remaining digits)
2. List stems in order in vertical line
3. Draw vertical line to the right of stem list
4. List leaves in numerical order horizontally
5. Indicate units

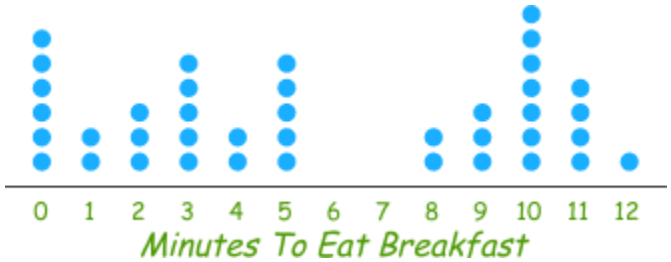
Split Stem and Leaf Plot

- If there are too many leaves corresponding to stems, split stem and leaft plot can be created
- Each stem will be listed twice
- First occurrence of stem is 0-4, second 5-9



Dot Plot

1. Put variable on horizontal axis
2. Put dot for each observation in the data set at location



Histogram

1. Choose the number of intervals.groups to divide your data into
    - Rule of thumb: number of class = sqrt(number of observations)

    - If there are too few intervals, info is lost
    - If there are too many, info could be difficult to interpret
2. Determine length for intervals, same length
    - Range = max-min
3. Create interval by finding boundaries
4. Use data to determine frequencies using left inclusion rule or right inclusion rule
    - Left inclusion rule: Observations that are equal to the left side of a specified interval should be included in the interval group
5. Construct histogram

Density Histogram is created by replacing the vertical axis with a density scale where rectangle height = relative frequency /class width
Density of histogram = 1

Lecture 2

Measures of Center

1. Mean - arithmetic average of a set of data, sensitive to outliers (x1 + x2 + .. + xn) / n where n = sample size
    - Sample mean of population mean is mew *u*
2. Median - Middle observation in an ordered set of data, resistant to outliers
    - Order the data
    - Find the position of the median
    - Find the median, if n is even position of the median will be halfway between n/2 and n/2 + 1 positions, average of those 2 positions
3. Mode - Most frequently occurring observation in a set of data
    - Unimodal: Model, a single moss frequently occurring observation value
    - Bimodal: 2 modes, two observation values that occur equally and more than all other observation values
    - Multimodal: more than two observation values occurring equally and more than all other observation values

Symmetric - Mean = Median

- If the shape of the data is symmetric and unimodal, mean = median = mode
- Skewed right, Mean > median
- Skewed left, mean < median

Trimmed Mean - arithmetic average of a set of data after 100% of the sammelst and 100% of the largest observations are deleted from the data, in other words, delete the smallest observations and the largest observations before calculating the mean
Xtr = trimmed mean
Ex: Find the 10% trimmed mean Xtr(10)
100(alpha)%, alpha = .1
n(alpha) = 50(.1)=5
Remove the 5 smallest and 5 largest observations, then find trimmed mean

Pth percentile - the value of the o servation in an ordered set of data that separates the lower p% of data from the upper (100 - p)% of data

- Position: p/100(n+1)
- The 99th percentile represents the value in the data set that separates the lower 99% of observations from the top 1% of observations

Quartiles - divide the data into 4 equal parts

- First Quartile - 25th percentile, separates the lower 25% of observations from the upper 75% of observations in an ordered set of data also called the lower quartile (Q1, Qt)
  Formula: 0.25(n+1)
- Second Quartile - 50th percentile, separates the lower 50% of observations from the upper 50% observations in an ordered set of data also called the median (Q2, or m)
  Formula: 0.50(n+1)
- Third Quartile -75th percentile; separates the lower 75% of observations from the upper 25% of observations in an ordered set of data also called the upper quartile (Q3 or Qu)
  Formula: 0.75(n+1)

Proportions

1. Success
2. Failure

X = number of successes
(n - x) = number of failures
Sample proportion of successes (p hat)
p hat = x / n = # of successes / # of observations

Sample proportion of failures = 1 - p hat
Population proportion of successes = p

Measures of Spread

1. Range
2. Interquartile Range
3. Variance
4. Standard Deviation

Range - the difference between the maximum and minimum values in a set of data

- Notation: R where R = max - min

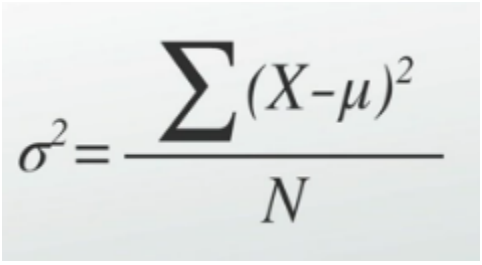
Interquartile Range - difference between the upper and lower quartile, the range of the middle 50% of observations

- Notation: IQR where IQR = Q3 - Q1
- Resistant to outliers

Variance - Average of the squared deviations of each observation from their mean

- The variance is measured in units^2 and is difficult to interpret

- Variance is sensitive to outliers
- Notation: S^2
- (n-1) is called degrees of freedom (df)

Population variance
- Notation : Sigma-squared
- 

Standard Deviation - Positive square root of the variance
- Measured deviation is measured in units
- Bigger standard deviation, more spread out the data is
- Sensitive to outliers
- Notation: S

Population standard deviation
- Notation: Sigma

Properties of Variances
- Location Change
- Scale Change

Box Plot
1. Draw horizontal measurement scale
2. Draw vertical lines at Q1,Q2, Q3. Draw a horizontal line at the top and bottom of the vertical lines to create a box
3. Identify outliers using fences
    - Lower fence = Q1 - 1.5(IQR)
    - Upper fence = Q3 + 1.5(IQR)
4. Draw horizontal line from Q1 to the minimum and from Q3 to the maximum minus outliers, called whiskers

If right whisker is longer than left whisker substantially, data is skewed right
If left whisker is longer than right whisker substantially, data is skewed left

Lecture 3

Probability - study of randomness and uncertainty, represents the likelihood of chance of outcoming occurring

- Probability is used to express confidence in results/ conclusions/ estimates about a population that is from a sample
- Bridge between descriptive statistics and inferential statistics

Experiment - Activity or process whose outcome is subject to a random outcome, can be performed in a controlled environment such as a lab or in nature

Sample space (S) - set of all possible outcomes of an experiment

Event - Collection / Subset of outcomes from the sample space

Simple event - An event consisting of only one outcome from the sample space
Null Event - An even twitch does not contain any outcomes, empty set

Examples:

1. 3 items
    - Variable: Identify each item as acceptable (a) or defective (d)
    - Sample Space: {aaa,aad,ada,daa,add,dad,dda,ddd}
    - Events:
        - D : At least one defective item = {aaa,ada,daa,add,dad,dda,ddd}
        - N : no defective items = {aaa}
        - E : an equal number of defective and acceptable items = {0}

Event Relationships

Complement - Complement of an event A is the set of all elements of the sample space that are not in event A

Union - Union of 2 vents A and B is the event consisting of all outcomes belonging to A or B or both

Intersection - intersection of 2 events A and B is the event consisting of all outcomes shared by A and B

Mutually Exclusive/ Disjoint - 2 Events A and B are said to be mutually exclusive / disjoint if there are no shared outcomes between the 2 events

Production/ Multiplication Rule for Ordered Pairs

- Product Rule: If the first element in an ordered pair can be observed n1 ways and the second element can be observed n2 ways, then the total number of outcomes is equal to n1n2

Permutation Rule:

- Permutation - Ordered subset
- Permutation Rule: $P(n,k) = n!/(n-k)!$

Combination Rule

- Combination - unordered subset

- Combination Rule: n!/(kl!(n-k)!)

Lecture 4

Law of Large Numbers - Theoretical pro ability of an outcome is equal to the proportion of times that outcome would occur in a very long series of repetitions of the xperiment, probability is equal to the long run frequency of the outcome
Limit n - > infinity (f/n)

Probability - P(A), can be expressed as proportion or percentage

P(A) = 0, event will never occur
P(A) = 1, event will always occur
P(A) = sum P(Ei) (countably infinite) = Na/N = # ouf outcomes in A / # of outcomes in S

Axioms of Probability

1. For any event A, P(A) >= 0
2. P(S) = 1
3. If A1, A2, A3 is an infinite collection of disjoint events, then P(A1, A2…) is the sum of all events

Probability Properties
Sample Space - Set of all possible outcomes of an experiment P(S) = 1
Event - Collection / subset of outcomes from the sample space 0 <= P(A) <= 1
Null event - An event which does not contain any outcomes P(0/) = 0
Compliment - Complement of an event A ios the set of all elements of the sample space that are not in event A
Union - The union of 2 or more events in the event consisting consisting of all outcomes belonging to those events Ex: A U B = B U A, P(A U B) = P(A) + P(B) - P(A ∩ B)
Ex: A U B U C , P(A U B U C) = P(A) + P(B) + P(C) - P(A∩B) - P(B∩C)+P(A∩B∩C)
Mutually exclusive / disjoint - events with no shared outcome
(A ∩ B) = ∅
P(∅) = 0
P(A ∩ B) = P(A) + P(B)

Conditional probability - probability of an event A occurring given an event B has occurred
P(A|B)

P(A|B) = P(A∩ B)/ P(B) for P(B) > 0

Probability of S changes depending on Y. Thus S and Y are said to be dependent events

Occurance of one event does not influence the occurrence of another event, the events are independent

Independence - A and B are said to be independent if knowing that one has occurred does not change the probability of the other occurring

- Rule 1: A and B are independent events iff (if and only if) P(A|B) = P(A) or P(B|A) = P(B)
- A and B are independent events iff $P(A \cap B) = P(A)P(B)$

Lecture 5

Multiplication Rule: $P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$

Law of Total PRobability - calculate the weighted average of probabilities for an event B over k mutually exclusive and exhaustive events

Exhaustive events - set of events if at least one of the events must occur

Bayes Theorem - Calculates P(A|B)

Lecture 6

Random Variables - allow outcomes from an experiment to be represented numerically

- Function defined over a sample space that assigns a real number to each outcome (Capital letter)

Types of Random Variables

1. Discrete - finite number of possible values
2. Continuus - infinite number of possible values

Examples:

Discrete: Number of defective items out of three observed, number of empty car spaces in a parking lot, Socket size (inches), Sum of 2 rolled 6 sided dice

Continuous: Length of time until a battery fails, Weight of a steel plate

Examples:

1. Experiment: Observer 3 items
   Variable: Identify each item as acceptable or defective
   Sample Space S = {aaa,aad,ada,daa,add,dad,dda,ddd}
   - Random Variable: X = # of defective items = {0,1,2,3}

Values that are only 0 and 1 is called a Bernoulli random variable or indicator random variable or dummy random variable (Successes or failures)

Discrete random variable - a random variable foir which there exists a discrete set of values with non-zero probability, a random variable whose range space is either finite or countably infinite
Probability distribution for a discrete random variable which is the sum of the probabilities of all outcomes that are assigned the value x also called a probability mass function (pmf)

Rules: p(x) is a pmf if and only if p(x) > 0 and the sum of all probabilities is 1

A probability distribution /pmf is a table or mathematical formula that provides the possible values of random variable and their corresponding probabilities

Can be represented as

1. Line Graphs
2. Probability Histograms

Cumulative distribution function (cdf)
Rules:

1. 0 <= F(x) <= 1
2. F(x) is non decreasing
3. F(a) = 0 for a < xmin where xmin is the minimum value of x
4. F(b) = 1 for b>= max where xmas is the maximum value of x

A cdf can b e represented with a step graph
Note:

- Pmf can be recovered from cdf by calculating jump value
- Cdf can be used to find probabilities of interest

<u>Lecture 7</u>

Expected Value / Mean of a discrete random variable X with a pmf p(x) is given by E[X]

Sum of x times p(x)

Moment about the origin

Variance - Average of the distances squared of each observed value to the mean
Parameter of a distribution - Quantity defined by p(x) such that it can be assigned any one of a number of possible values, with each different value determining a different probability distribution

Family of distribution - the collection of all probability distributions for different values of the parameter

Bernoulli Distribution - Suppose an experiment results in only 2 possible outcomes, “success” and “failure”, with probabilities P(success) = p and P(failure) = 1-p. Experiment is called Bernoulli trial and X is a Bernoulli

Binomial Distribution
Suppose an experiment consists of n independent and identical Bernoulli trails. In other words, the experience satisfies the following 4 assumptions:

1. The experiment consists of n smaller experiments called trials where n is fixed in advance of the experiment
2. Each trial can result in one of the same two possible outcomes which are referred to as a successes or as a failure
3. The trails are independent
4. The probability of success,p, is constant from trial to trial. Note: This can be achieved by sampling with replacement

Hypergeometric Distribution
The hypergeometric distribution is related to the binomial distribution in that both are interested in a dichotomous (2 outcomes) population where the number of successes is of interest. The binomial disruption requires the probability of success p to be constant for every trial which can be achieved by sapling with replacement . I
N the hypergeometric distribution, each trial changes the probability for each subsequent trial because the sampling is done without replacement

1. The population or set to be sample d consists of N individuals, objects, or elements
2. Each individual can be characterized as a success or a failure and there are M successes in the population
3. A sample of n individuals is selected without replacement in such a way that each subset of size n is equally likely to be chosen

Negative Binomial Distribution
Is related to binomial distribution. Binomial is the number of successes when the number n of trials is fixed. Negative binomial distribution fixes the number of successes desired and lets the number of trails be random.

1. Experiment consists of a sequence of independent trails
2. Outcome of each trial can be characterized as a success or failure
3. Probability of success p is constant

4. Experiment continues until total of r successes have been observed

Geometric Distribution
Special case of the negative binomial distribution where r = 1. In other words, the experiment consists of independent and identical Bernoulli trials and is permed until 1 success occurs.

Poisson Distribution
Consider an experiment where it is of interest to count the number of times an event occurs in a given interval (time, space, region, volume, etc). The poisson process satisfies the following

1. The number of outcomes occurring in one time interval is independent of the number that occurs in any other disjoint intervals
2. The probability that a single outcome will occur during a short interval is proportional to the length of the interval and does not depend on anything outside of the interval
3. The probability of more than one occurrence of the event in a very short interval is negligible

Poisson Approximation of Binomial
Can be used to provide an accurate approximation for binomial distribution

Lecture 8

Discrete Random Variable - random variable that takes either a finite number of possible values or at most countably infinite (pmf, p(x), cdf, P(X <= x), E[X], Var [x])

Continuous Random Variable - random variable that takes an infinite number of possible values that is not countable

Example:
X = Time until a component Fails
Y = Temperature at which specific liquid freezes

Probability Density Function

- Analogous pmf for discrete random variable, probability density function descries the probability distribution for a continuous random variable

Probability distribution for a continuous random variable also called probability density function (pdf) or density curve

Cumulative Distribution Function
Cdf for a continuous random variable, X is defined for every number x by F(x) = P(X<=x)

Rules:

1. F(X) -> 0 as x -> negative infinity
2. F(X) -> 1 as x -> infinity
3. F(x) is a monotonic non-decreasing function of x
4. F(x) does not need to be smooth, but it is continuous

Percentiles of a Continuous Distribution
Let p be a number between 0 and 1. The (100p)^th percentile of the distribution of a continuous

Expected Values / Mean

- Expected value/ mean of a continuous random variable is an integral

Variance of a continuous random variable

Lecture 9

Uniform distribution

- A continuous random variable X is said to have a Uniform Distribution on the interval [A,B] in the pdf of X is given by 1/ B- A

1. Uniform
2. Normal
3. Standard Normal
4. Exponential
5. Gamma