Categorical (qualitative):
variables take categories as their values such as “yes”, “no”, or “blue”, “brown”, “green”.
Numerical (quantitative):
variables have values that represent a counted or measured quantity.
Discrete variables: arise from a counting process.
Continuous variables: arise from a measuring process.
An ordinal scale classifies data into distinct categories in which ranking is implied.
A **population** contains all of the items or individuals of interest that you seek to study.(i.e.All the items or individuals about which you want to draw conclusion(s).)
A **sample** contains only a portion of a population of interest (i .e .A portion of the population of items or individuals)
A **population parameter** summarizes the value of a specific variable for a population.
A **sample statistic** summarizes the value of a specific variable for sample data
Primary Sources: The data collector is the one using the data for analysis: Data from a political survey. Data collected from an experiment. Observed data.
Secondary Sources: The person performing data analysis is not the data collector: Analyzing census data. Examining data from print journals or data published on the internet.
In a nonprobability sample, items included are chosen without regard to their probability of occurrence. In convenience sampling, items are selected based only on the fact that they are easy, inexpensive, or convenient to sample. In a judgment sample, you get the opinions of preselected experts in the subject matter.
In **convenience sampling**, items are selected based only on the fact that they are easy, inexpensive, or convenient to sample.
In a **judgment sample,** you get the opinions of preselected experts in the subject matter.
In a **probability sample**, items in the sample are chosen on the basis of known probabilities.
Every individual or item from the frame has an equal chance of being selected. Selection may be with replacement (selected individual is returned to frame for possible reselection) or without replacement (selected individual isn’t returned to the frame). Samples obtained from table of random numbers or computer random number generators.
Parameter: – A measure computed from the entire population – As long as the population does not change, the value of the parameter will not change.
• **Statistic:** – A measure computed from a sample that has been selected from a population – The value of the statistic will depend on which sample is selected.
Measuring the Center – Data for a variable of interest forms a distribution – We need to be able to describe the center using a numerical measure.
• **Measuring the Spread** – Values for a variable of interest will take on different values –the data are spread out around the center. – We need to be able to measure the spread.
Mean – The average of the data • **Median** – The midpoint of the data • **Mode** – The data value occurring most frequently
Median:
The median is a center value that divides a data array into two halves (Md). • **Data Array** – Data that have been arranged in numerical order • **Median Index** – i = The index of the point in the data set corresponding to the median value – n = Sample size Formula: $\mathbf{i} = \frac{1}{2}$ n round up if not integer
In an ordered array (lowest to highest), the median is the “middle” number, i.e., the number that splits the distribution in half numerically. – 50% of the data is above the median, 50% is below – Represented as Md • The median is not affected by extreme values.
Mode:
• The value in a data set that occurs most frequently • Is not affected by extreme values. • Can be used for both quantitative and qualitative data. • Can have more than one mode, or no mode. • Distribution with two modes - bimodal
Population Variance:
The average of the squared distances of the data values from the mean. $\sigma^2 = \frac{\Sigma (Xi-\mu)^2}{n}$ n= amount of variables X=variables
Population Standard Deviation:
The most commonly used measure of variation • The positive square root of the variance • Has the same units as the original data $\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\Sigma (Xi-\mu)^2}{n}}$
Sample Variance: $s^2 = \frac{\Sigma (Xi-\bar{x})^2}{n-1}$ **Sample Standard Deviation:** $s = \sqrt{\frac{\Sigma (Xi-\bar{x})^2}{n-1}}$
$\bar{x} = \frac{\Sigma x}{n}$ x , x = sum of variables
Coefficient of Variation:
Is used to compare two or more sets of data measured in different units
Population CV $CV = \frac{\sigma}{\mu}$ *(100)%*
Sample CV $CV = \frac{\sigma}{x}$ *(100)%*
The Empirical Rule:
If the data distribution is bell shaped, then the interval μ ± 1σ contains approximately 68% of the values. μ ± 2σ contains approximately 95% of the values. μ ± 3σ contains virtually all of the data values.

Standardized Data Values • *The number of standard deviations a value is from the mean* • *Standardized data values are also referred to as z scores.*
– *Population z score* $\frac{x-\mu}{\sigma} = z$
– *Sample z score* $\frac{x-\bar{x}}{s} = z$ x – data value μ - population mean σ – population standard deviation x – sample mean s – sample standard deviation
Relative Frequency
• Step 3: The proportion of total observations that are in a given category. **Relative Frequency** = $\frac{fi}{n}$
fi - Frequency of the “i”th value of the discete variables. n= Total number of obersvations. k The number of different values for the discrete variable
• **Continuous data** – data whose possible values are uncountable and that may assume any value in an interval (weight, length, time) • Discrete data with many possible outcomes (income, stock prices, GPA’s) • Summarized in a grouped data frequency distribution • Data are organized in classes. Classes must be mutually exclusive. – Classes do not overlap. • Classes must be all-inclusive. – A set of classes contains all possible data values. • Classes should be of equal width, if possible. – The distance between the lowest and the highest possible values in each class is equal for all classes. • Empty classes should be avoided.
Developing Frequency Distribution for Continuous Data • Step 1: Determine the number of classes. • Step 2: Establish the class width. • Step 3: Determine the class boundaries for each class. – the upper and lower values of each class • Step 4: Determine the class frequency for each class. – number of data points in each class. **Minimum Class Width =** $\frac{\text{Largest value} - \text{smallest value}}{\text{number of classes}}$
Cumulative Frequency Distribution – a summary of a set of data that displays the number of observations with values less than or equal to the upper limit of each of its classes • Cumulative Relative Frequency Distribution – a summary of a set of data that displays the proportion of observations with values less than or equal to the upper limit of each of its classes
Bar Charts • A graphical representation of a categorical data set in which a rectangle or bar is drawn over each category or class • The length or height of each bar represents the frequency or percentage of observations or some other measure associated with the category. • The bars may be vertical or
Pie Charts • A graph in the shape of a circle. • Use to show visually the parts of a total • The circle is divided into “slices” corresponding to the categories or classes to be displayed. • The size of each slice is proportional to the magnitude of the displayed variable associated with each category or class.horizontal.
Probability – The chance that a particular event will occur – The probability value will be in the range 0 to 1. • **Experiment** – A process that produces a single outcome whose result cannot be predicted with certainty • **Sample Space** – The collection of all outcomes that can result from a selection, decision, or experiment
Types of Events • **Mutually Exclusive Events** – Two events are mutually exclusive if the occurrence of one event precludes the occurrence of the other event. • **Independent Events** – Two events are independent if the

occurrence of one event in no way influences the probability of the occurrence of the other event. • **Dependent Events** – Two events are dependent if the occurrence of one event impacts the probability of the other event occurring.

Addition Rule for Individual Outcomes The probability of an event Ei is equal to the sum of the probabilities of the individual outcomes forming Ei . For example, if Ei = {e1 , e2 , e3 } then P(Ei) = P(e1) + P(e2) + P(e3).

Complement Rule • The complement of an event E is the collection of all possible outcomes not contained in event E.

The probability of the complement of event E is 1 minus the probability of event E. P(E) = 1 - P(E). line on top of first e

This rule is corollary to Probability Rules 1 and 2

Rule 4: Addition Rule for Any Two Events – Keyword “or” Means Addition

P(E1 or E2) = P(E1) + P(E2) - P(E1 and E2)

Rule 8: Multiplication Rule for Any Two Events: we do not know the joint relative frequencies, the multiplication rule for two events can be used. For two events E1 and E2 P(E1 and E2) = P(E1) P(E2 |E1)

|Rule 9: Multiplication Rule for Independent Events: The joint probability of two independent events is simply the product of the probabilities of the two events. For independent events E1 , E2 P(E1 and E2) = P(E1) P(E2)

Bayes' Theorem • A special application of conditional probability • A way to formally incorporate the new information • Probability assessment for events of interest may be based on relative frequency or subjectivity p(e/b)= $\frac{p(ei)p(b/ei)}{p(e1)p(b/e1)+p(e2)p(b/e2)+\ldots.+p(ek)p(b/ek)}$ ei= the “i”th event of interest of the k possible events. b=event that has occurred that might impact P(ei)

Events of ei,ei,,,,,ei are mutually exclusive and collectively exhaustive

Introduction to Discrete Probability Distributions • **Random Variable** – Takes on different numerical values based on chance • **Discrete Random Variable** – Can only assume a finite number of values or an infinite sequence of values such as 0,1,2,…Many possible outcomes: – number of complaints per day – number of TVs in a household – number of rings before the phone is answered • Only two possible outcomes: – gender: male or female – defective item: yes or no – game result: won or lost • **Continuous Random Variable** – Can assume an uncountable infinite number of values

Standard Deviation of a Discrete Random Variable • **Standard deviation** measures the spread, or dispersion, in a set of data. • The standard deviation also measures the spread in the values of a random variable. $\sqrt{\Sigma[(x-\mu)^2 \times P(x)]}$ = σ x E(x) = Expected value of x x = Values of the discrete random variable P(x) = Probability of the random variable taking on the value x

The Binomial Probability Distribution • *A distribution that gives the probability of x successes in n trials in a process that meets the following conditions:* 1. A trial has only two possible outcomes: a success or a failure. 2. There is a fixed number, n, of identical trials. 3. The trials of the experiment are independent of each other. This means that if one outcome is a success, this does not influence the chance of another outcome being a success. 4. The process must be consistent in generating successes and failures. That is, the probability, p, associated with a success remains constant from trial to trial. 5. If p represents the probability of a success, then (1 – p) = q is the probability of a failure.

Counting Rule for Combinations • A method for counting the number of ways binomial events can occur

$C_x^n = \frac{n!}{x!(n-x)!}$ n! = n(n - 1)(n - 2) . . . (2)(1) x! = x(x - 1)(x - 2) . . . (2)(1) 0! = 1 by definition

Binomial Formula: $P(x) = \frac{n!}{x!(n-x)!} p^x q^{n-x}$ P(x) - probability of x successes in n trials, with probability of success p on each trial x - number of successes in sample, (x = 0, 1, 2, ..., n) p - probability of a success q - probability of a failure q = 1 – p n - number of trials (sample size)

Finding an Exact Probability Using the Cumulative Binomial Table: P(x=2)={x<=2)-P(x<=1)

Mean and Standard Deviation for a Binomial Distribution: μ=E(x) =np, σ = $\sqrt{npq}$

Other Discrete Probability Distributions • **Poisson Distribution** – Describes a process that extends over space, time, or any well defined segment/interval or unit of inspection in which the outcomes of interest occur at random and the number of outcomes that occur in any given interval are counted – Is used when the total number of possible outcomes cannot be determined.**Characteristics of the Poisson** Distribution • The outcomes of interest are rare relative to the possible outcomes. • The average number of outcomes of interest per segment interval is λ . • The number of outcomes of interest are random, and the occurrence of one outcome does not influence the chances of another outcome of interest. • The probability that an outcome of interest occurs in a given segment is the same for all segments.

The Poisson Distribution: Mean or expected value: μ_x =E(x) =λt. Standard deviation σ x= $\sqrt{\lambda t}$ .Probability distribution: P(x) = $\frac{(\lambda t)^x e^{-\lambda t}}{x!}$ or poisson distribution formula. t - Number of segments of interest x - Number of successes in t segments λ - Expected number of successes in one segment e - Base of the natural logarithm system (2.71828)

Hypergeometric Distribution • n trials in a sample taken from a finite population of size N • Sample taken without replacement • Trials are dependent. • The probability changes from trial to trial. • Concerned with finding the probability of x successes in the sample where there are X successes in the population. Two possible outcomes per trial: $P(x) = \frac{C_{n-x}^{N-x} C_x^X}{C_n^N}$ N - Population size X - Number of successes in the population n - Sample size x - Number of successes in the sample n – x - Number of failures in the sample

The Normal Probability Distribution • Is a bell-shaped distribution with the following properties: 1. It is **unimoda**l; that is, the normal distribution peaks at a single value. 2. It is **symmetrica**l; this means that the two areas under the curve between the mean and any two points equidistant on either side of the mean are identical. 3. The mean, median, and mode are equal.

Normal Probability Density Function: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$ x= any value of the continuous random variable. σ Population standard deviation.

Mean and Standard Deviation of a Uniform Distribution: Mean or Expected Value: $E(x) = \mu = \frac{a+b}{2}$

Standard Deviation $\sigma = \sqrt{\frac{(b-a)^2}{12}}$ a - The smallest value assumed by the uniform random variable of interest b - The largest value assumed by the uniform random variable of interest.

Exponential Density Function:

A continuous random variable that is exponentially distributed has the probability density function

$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0$ 1/λ Means time between events (λ>0)

Exponential Probability: The probability that a value will fall within an interval is the area under the graph between the two points defining the interval $P(0 \leq x \leq a) = 1 - e^{-\lambda a}$ a= the value of interest. 1/λ= the mean.