What is Statistics?

* Statistic: number calculated from data (median income, unemployment rate)
* As a field, science to design studies, collecting data & analyzing, and quantifying uncertainty
* To extract info & knowledge from data
* To understand the world around us

Statistical Methods

* Description: summarizing & analyzing data as well as exploring patterns in data (min/max, avg, tables, plots)
* Descriptive statistics
* Inference: decision-making or predictions of population based on data from smaller group (sample)
* Inferential statistics
* Design research: planning how to obtain data efficiently, how to extract useful data, & how to derive meaningful outcome from data

Descriptive statistics include reaching out to all JMU students about dining

Inferential statistics include emailing opinion survey to student sample

What are Data

* Data are the info we collected
* To answer our research question
* Cases/subject we’re interested
* Case/subject: something or someone we are interested
* Variables: features or characteristics of each case/subject
* Data usually vary & are usually some value with context

Types of Data – Variables

* Categorical variables: label, category, identifier (descriptive)
* gender, class
* Quantitative variables: measurable, quantity, magnitude
* Age, salary, height

Level of Measurement

* Nominal: “name” the attribute uniquely
* Ordinal: can be rank-ordered
* Interval: distance is meaningful or interpretable (but ratio doesn’t make sense) body temp
* Ratio: absolute zero, ratio fraction is meaningful (age)

Discrete variable

* Counting, only integers
* Number of students, results of rolling dice

Continuous Variable

* Measuring, any value (within a range), fractions
* Weight, time, length, age

Metadata, Independent and Dependent Variables

**Central Tendency**

Ways to present data

* Present (store) in a table
* Summarizing collected data (descriptive statistics)
* Charts
* Move in the following 2 weeks
* Numbers (statistics)
* Include units & proper precisions

What to know first about data

* Center
* Mean, median, mode
* Shape
* Skewed/tail
* Spread
* Range & variation

Marginal vs. Conditional

Relationship between 2 categorical variables

* Contingency tables
* Frequencies and/or relative frequencies of observations; which are 2 categorical variables
* Two dimensions: one category across columns, the other across rows
* Display distribution along each variable, contingent on the value of the other variable
* Mutually exclusive

Marginal distribution: distribution of 1 or 2 variables displayed on the margins of the contingency table

Conditional distribution: distribution of how frequently one of the categorical variable is observed for just those cases that satisfy certain condition on another variable

Conditional probability: probability of how frequently one of the categorical variable is observed for just those cases that satisfy certain condition on another variable

Categorical data

* Frequency tables: shows list of categories & how often each category occurs
* Relative frequency tables: categories but with the proportion or percent, rather than count of the data in each category
* Bar charts: display the distribution of a categorical variable
* Frequency or relative frequency
* Length of each bar should be proportional to the value it presents

Pie Charts

* Shows whole group of observations as a circle
* Divide the circle into sectors or slices
* Each slice should be proportional to the fraction of the whole

Spread & shape

Shape – shape of the distribution of data

Symmetrical or not

* Symmetrical & asymmetrical
* Skewed: distribution is skewed if one tail is longer than other
* Right skew (positive skew): mode < median < mean (pg.5 notebook)
* Left skew (negative skew): mean < median < mode

Type of data and best measure of central tendency (pg. 5)

Spread: range & variation

* Range: diff between max and min
* Measuring variability: how far a data value is from the mean
* When variance increases, range increases

2 categorical variables:

Keywords for conditional distributions & probabilities

* Given that, who, which, that, consider only, only, …etc.

Segmented Bar chart: use rectangular segments to present categorical data

Bessel’s correction (4): why we divide by (n-1)

Rationale for correction: smaller the sample, the less likely you are to capture the variance, unbiased is better

Descriptive statistics:

* Min, 1st quantile, median, mean, 3rd quantile, max
* Stat.desc: # of observations, # of null observations, # of “not available” observations, range, sum, standard error, mean of population, width, variance, standard deviation, coefficient of variance

2.1 bar charts/bar plots

2.3 segmented bar charts

2.6 pie & waffle charts

2.10 contingency tables/marginal & conditional distributions

Present the relationship between 2 categoricals: contingency table & segmented bar chart

Not used to display categorical: histogram

Statistics used to describe spread of dataset: variance, min/max, range

Commands calculate stand.dev named “mydata”: var(mydata), sd(mydata)

Operator used to assign a value to a variable: <-

Function to create a vector: c

**Quantitative data**

Area principal: area occupied by a part of the graph should correspond to the magnitude of the data value it’s representing

How to describe & display quantitative variables?

* Plots to show the distribution of data: histogram
* Plots to show important statistics: box plot

How to describe distribution of a quantitative variable

* Central tendency, shape, spread

Shape – modality (7)

* Unimodal
* Bimodal
* Uniform
* Multimodal

Bins are groups used to generation frequencies in a histogram

* Bins are equal size and each has a quant lower bound & a quant upper bound

Interquartile range (iqr) = Q3 – Q1

Outliers (>upper fence or <lower fence) would be true if over upper lim or under lower lim

Upper limit = Q3 +1.5 IQR

Lower limit = Q1 –1.5 IQR

2.2 Histogram

2.4 box plots

2.9 scatterplots, covariance, and correlation

6.1 simple linear regression

Fivenum provides min, max, q3, median, q1

Box plot shows upper lim, min, lower lim, max, median, q3, q1

Summary provides median, mean, max, q3, q1, min

**Correlation, covariance, correlation coefficient r**

Scatterplot

* 2 continuous quantitative variables
* Direction: pos or neg
* Form: straight (linear), curved, other or no pattern
* Strength: tightly clustered or not
* Outliers?
* X: explanatory variable, IV, predictor variable
* Y: response variable, DV, predicted variable

Correlation: standardized or normalized version of covariance

Covariance: measure of the joint of 2 quantitative variables

Correlation coefficient

* Range (-1,1)

Conditions & properties of correlation

Correlation vs association

Correlation vs causation

Linear regression model & r^2

Least squares method

* Residual = actual value – predicted value

Linear regression calculation

R squared tells the proportion of variability in the data that is explained by the linear fit or linear model

Relationship between quantitatives are dependent (have a relationship) or independent

Negative relationship R<0, zero linear model (r=0), positive relationship (r>0)

A simple linear regression tries to fit a linear model (that is, a line) to represent the relationship between two continuous quantitative variables.

1.7 PDFs & CDFs

1.8 Z-score problems with the Normal Model

2.8 QQ Plots and tests for Normality