SJSU SAN JOSÉ STATE UNIVERSITY

Introduction to

Descriptive Statistics

1

SJSU SAN JOSÉ STATE UNIVERSITY

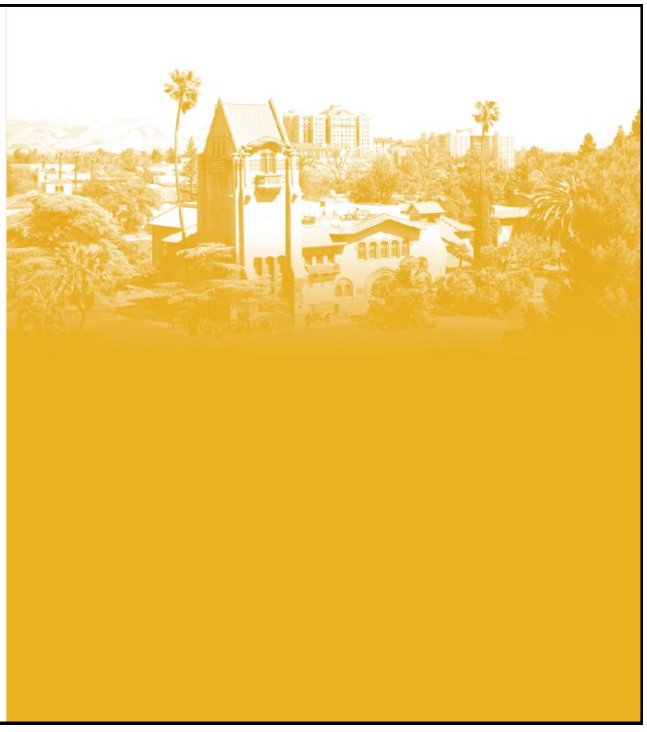
Agenda

- Overview of Data
- What is Statistics?
- Measures of Data:
  - Central Tendency
  - Measures of Dispersions
  - Five Number Summaries

2

SJSU SAN JOSÉ STATE UNIVERSITY

Overview of Data



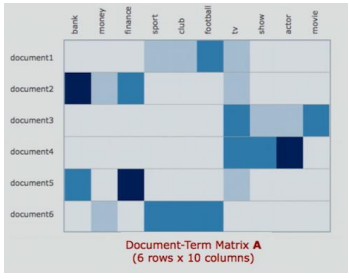
3

SJSU SAN JOSÉ STATE UNIVERSITY

Record or Tabular Dataset

Sale ID	Time	Customer	Product ID	Quantity
S00001	12/1/2012 9:00:00 AM	C0001	P025	1
S00002	12/1/2012 9:05:58 AM	C0025	P025	3
S00003	12/1/2012 9:11:33 AM	C0010	P001	2
S00004	12/1/2012 9:17:16 AM	C0017	P023	4
S00005	12/1/2012 9:23:04 AM	C0018	P016	5
S00006	12/1/2012 9:28:43 AM	C0011	P018	4
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Record or Transaction Data



Document-Term Matrix

Variables

Cases

sepal length	sepal width	petal length	petal width	class
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3	1.4	0.2	Iris-setosa
6.5	3.2	5.1	2	Iris-virginica
6.4	2.7	5.3	1.9	Iris-virginica
6.8	3	5.5	2.1	Iris-virginica
6.7	3.1	4.4	1.4	Iris-versicolor
5.6	3	4.5	1.5	Iris-versicolor
5.8	2.7	4.1	1	Iris-versicolor

Data Matrix

0	0	0	0	9	0
0	8	0	0	0	0
4	0	0	2	0	0
0	0	0	0	0	5
0	0	2	0	0	0

Rows	Columns	Values
5	6	6
0	4	9
1	1	8
2	0	4
2	3	2
3	5	5
4	2	2

Sparse Data Matrix

Images adopted from various internet pages

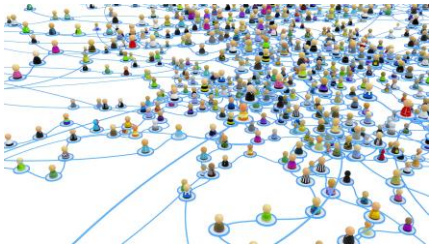
4

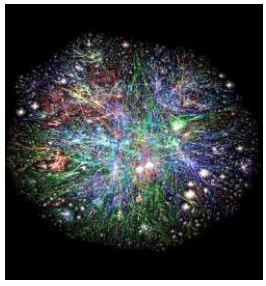
4

SJSU SAN JOSÉ STATE UNIVERSITY

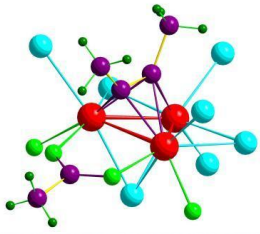
Graph-Based Dataset

- Transportation network
- World Wide Web





Data with Relationships among Objects



Data with Objects that are Graphs

- Molecular Structures
- Social or information networks

5

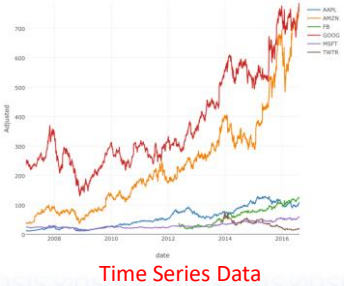
SJSU SAN JOSÉ STATE UNIVERSITY

Ordered Dataset

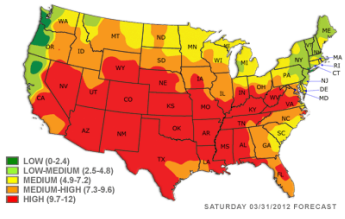
Time	Customer	Items Purchased
t1	C1	A, B
t2	C3	A, C
t2	C1	C, D
t3	C2	A, D
t4	C2	E
t5	C1	A, E

Customer	Time and Items Purchased
C1	(t1: A,B) (t2:C,D) (t5:A,E)
C2	(t3: A, D) (t4: E)
C3	(t2: A, C)

Sequential Data



Time Series Data



Spatial Data

Sequence Data

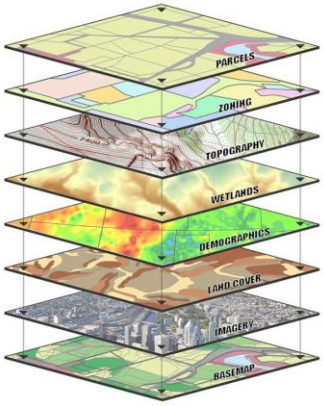
Images adopted from various sources.

6

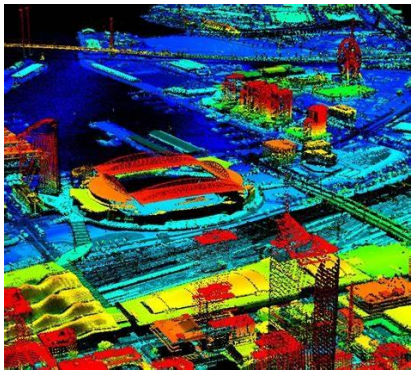
6

SJSU SAN JOSÉ STATE UNIVERSITY

Other Ordered Dataset



GIS Data



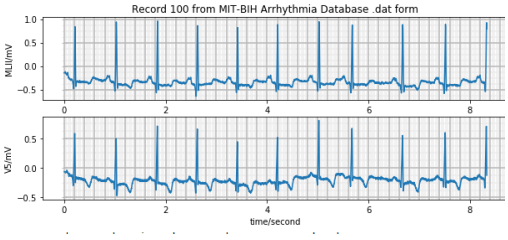
LiDAR Data

Satellite Data

Images adopted from various sources.

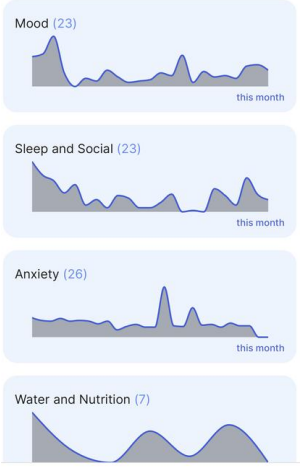
SJSU SAN JOSÉ STATE UNIVERSITY

More Ordered Dataset



ECG Data

Video Data

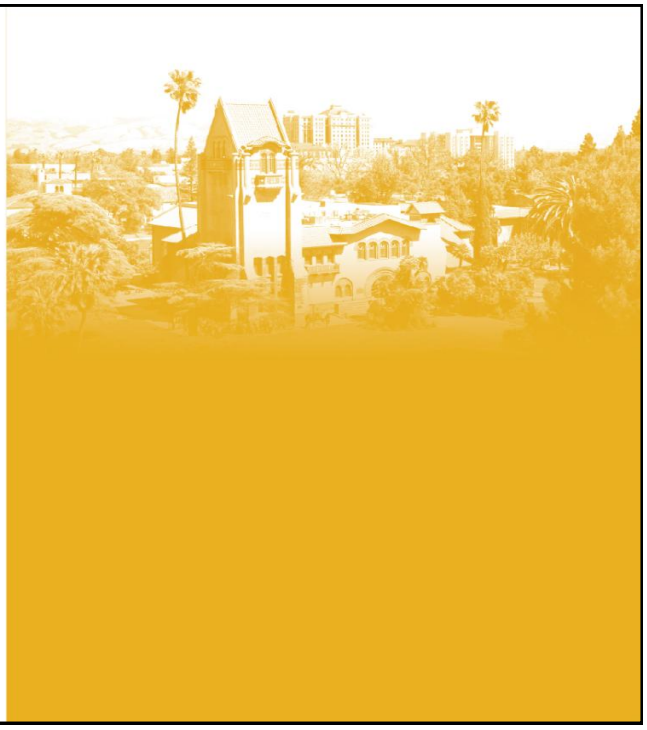


Wellness Data

Images adopted from various sources.

SJSU SAN JOSÉ STATE UNIVERSITY

What is Statistics?



9

SJSU SAN JOSÉ STATE UNIVERSITY

What is Statistics?

- Statistics is the science of collecting, analyzing, interpreting, presenting, and organizing data.
- It's a way of turning raw numbers into meaningful information that can help us understand patterns, trends, and relationships in various fields such as science, economics, and social sciences.
- There are two main branches of statistics:
  - Descriptive Statistics: summarizing and describing the features of a data set. Common tools include measures of central tendency and measures of variability etc.
  - Inferential Statistics: This involves making predictions or inferences about a population based on a sample of data. Techniques include hypothesis testing, confidence intervals, and regression analysis etc.

10

10

SJSU SAN JOSÉ STATE UNIVERSITY

Population vs Sample

A set of data points is a sample from a population:

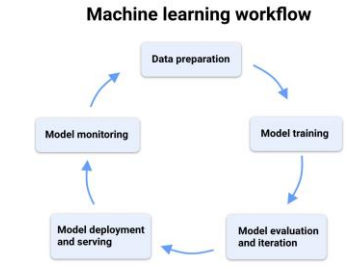
- A population is the entire set of objects or events under study.
  e.g., population can be hypothetical "all students" or all students in this class.
  e.g., population can be all the houses in a region

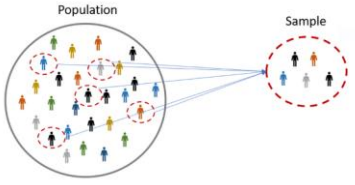
- A sample is a "representative" subset of the objects or events under study. This is needed because it's impossible or intractable to obtain or compute with population data.

SJSU SAN JOSÉ STATE UNIVERSITY

Sampling

- Sampling is an integral part of machine learning (ML) workflow
  - Sampling from all possible real-world data:
    - to create training data
    - to create splits: training, validation and testing data
    - for monitoring purposes
- Not accessible to all real-world data – use a subset of real world data (by sampling) for training model
- Infeasible to process all data available – too much time, computing power and money
- Allows to accomplish a task faster and cheaper
  - e.g. Perform a quick experiment with a subset of the data before running model on all the entire data

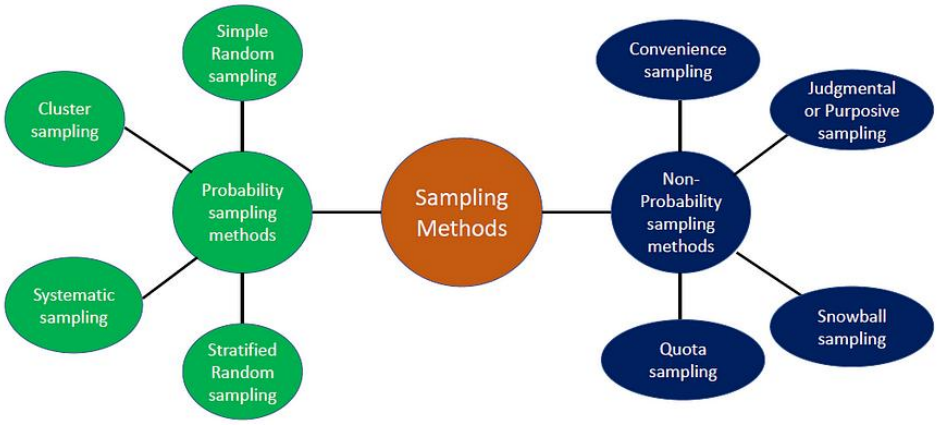




Designing Machine Learning Systems (Chip Huyen, O'Reilly 2022)
https://medium.com/analytics-vidhya/sampling-statistical-approach-in-machine-learning-4903c4Debf86
https://cloud.google.com/vertex-ai/docs/start/introduction-unified-platform

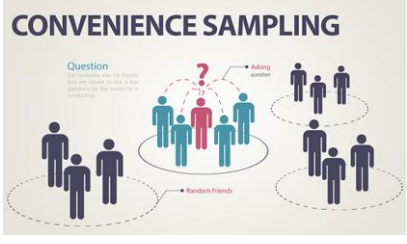
SJSU SAN JOSÉ STATE UNIVERSITY

Introduction to Statistics – Sampling

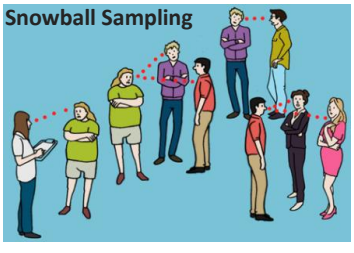


SJSU SAN JOSÉ STATE UNIVERSITY

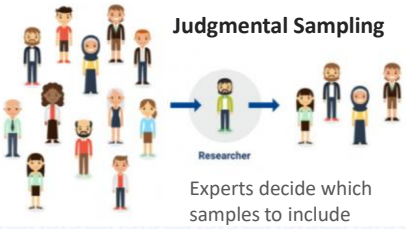
Non Probability Sampling



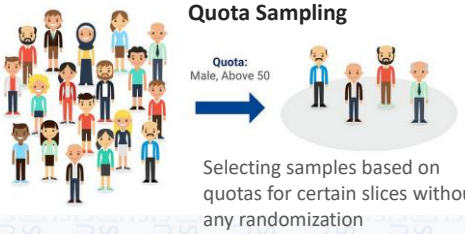
Samples are selected based on availability



Future samples are selected based on existing sample



Experts decide which samples to include

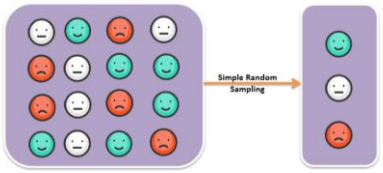


Selecting samples based on quotas for certain slices without any randomization

SJSU SAN JOSÉ STATE UNIVERSITY

Probability Sampling

Simple Random Sampling

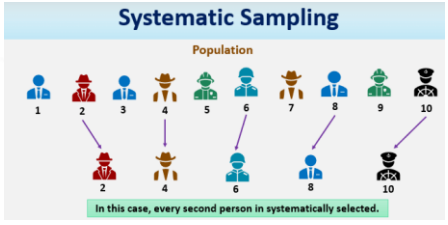


Randomly selecting individuals from a population

Cluster Sampling

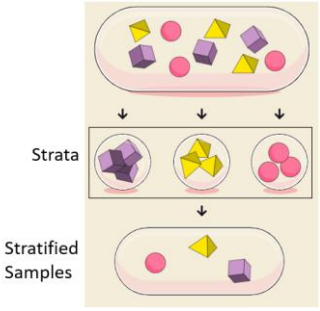
Divide the population into clusters or groups, then randomly choose some clusters

Systematic Sampling



Selecting every n^{th} individual from a population after randomly choosing a starting point

Stratified Random Sampling



Divide the population into distinct groups & randomly select samples from each group

15

SJSU SAN JOSÉ STATE UNIVERSITY

Introduction to Statistics – Data Sampling Errors

Selection Bias – Data (samples) are NOT selected in a way that is reflective or representative of the real-world distribution (entire population).

- Coverage Bias – Data is not selected in a representative fashion.
- Non-Response Bias – Data ends up unrepresentative due to participation gaps in the data collection process.
- Sampling Bias – Proper randomization is not used during data collection.

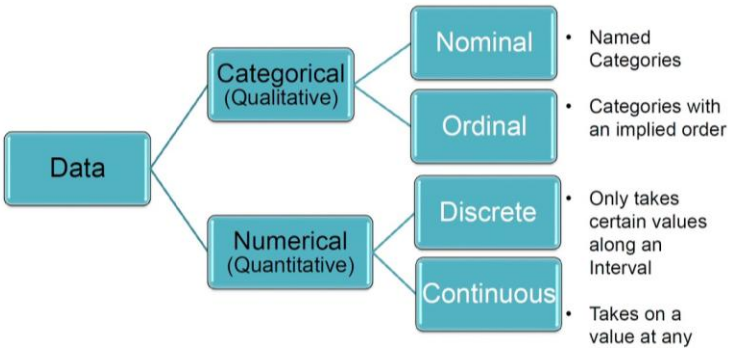
https://developers.google.com/machine-learning/crash-course/fairness/
https://www.qualtrics.com/experience-management/research/selection-bias/

16

SJSU SAN JOSÉ STATE UNIVERSITY

Variable Types

Recognizing variable types is important in choosing the appropriate statistical methods, visualization techniques, and modeling approaches.



17

17

SJSU SAN JOSÉ STATE UNIVERSITY

Questions

What are the data attribute types (nominal or ordinal) of the following data types?

- Course letter grades
- Gender
- Customer satisfaction level
- Marital status

18

18

SJSU SAN JOSÉ STATE UNIVERSITY

Question

Classify the following as Categorical (nominal or ordinal) or Numerical (discrete or continuous).

- Time in terms of AM or PM.
- Brightness as measured by a light meter.
- Brightness as measured by people's judgments.
- Angles as measured in degrees between 0 and 360.
- Bronze, Silver, and Gold medals as awarded at the Olympics.
- Height above sea level.
- Number of patients in a hospital.
- Military rank.

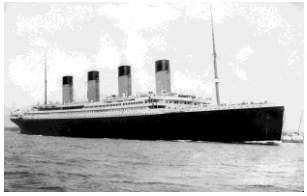
19

SJSU SAN JOSÉ STATE UNIVERSITY

Variable Type Example – Titanic Dataset

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

- **Pclass** - The passenger class (1 = first class, 2 = second class, 3 = third class)
- **SibSp** - The number of siblings or spouses the passenger had on board.
- **Parch** - The number of parents or children the passenger had on board.
- **Embarked** - The port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

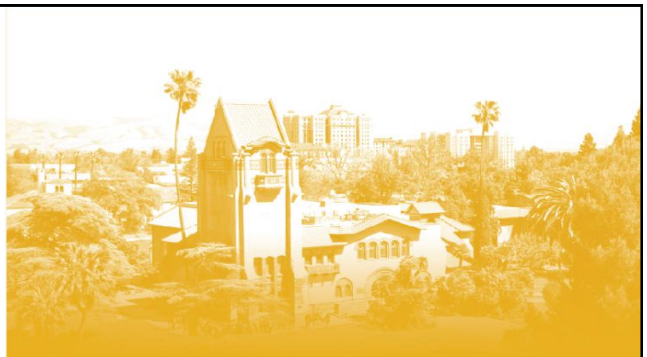


https://en.wikipedia.org/wiki/Titanic/

20

SJSU SAN JOSÉ STATE UNIVERSITY

Measurements of Data



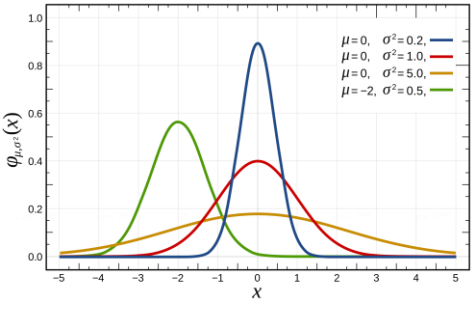
21

SJSU SAN JOSÉ STATE UNIVERSITY

Important Measurements of Data

To better understand the data, here are some important measures:

- Central Tendency
- Dispersion
- Graphic Summaries of Data
- Covariance and Correlation Analysis (later)



22

SJSU SAN JOSÉ STATE UNIVERSITY

Measurements of Central Tendency

23

SJSU SAN JOSÉ STATE UNIVERSITY

Measures of Dispersion

Here are some common measures of central tendency:

- Mean
- Median
- Mode

24

SJSU SAN JOSÉ STATE UNIVERSITY

Mean

- The mean is the arithmetic average of a set of values.
- To find the mean, all the values are summed up and then divided by the # of values.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{\sum x}{N}$$

sample population

Pros:

- Simple to calculate.
- Utilizes all data points.

$$\bar{x} = \frac{\Sigma x}{n} = \frac{22 + 22 + 26 + 24 + 23}{5} = \frac{117}{5} = 23.4$$

Cons:

- Sensitive to outliers (extreme values can skew the mean).

SJSU SAN JOSÉ STATE UNIVERSITY

Median

- The median is the middle value in a dataset when the values are arranged in ascending or descending order.
- To find the median:
  - Arrange data in order.
  - Find the middle position: n+1 / 2.
  - For even # of data, the median is the average of the two middle numbers.

22 22 **23** 24 26

Pros:

- Not affected by outliers.
- Represents the midpoint effectively.

22 22 **23** **24** 26 27

$$\text{Median} = \frac{23 + 24}{2} = \frac{47}{2} = 23.5$$

Cons:

- Does not utilize all data points (only the middle one).

SJSU SAN JOSÉ STATE UNIVERSITY

Mode

- The mode is the most frequently occurring value in a dataset.
- A dataset may have one mode (unimodal), more than one mode (bimodal or multimodal), or no mode if no value repeats.

Pros:

- Can be used with categorical data.
- Represents the most common value.

Cons:

- Might not be unique (or may not exist).

22	22	**23**	**24**	26	27

What is the mode?

27

27

SJSU SAN JOSÉ STATE UNIVERSITY

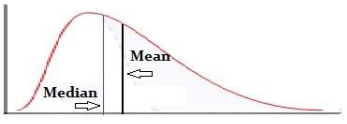
Summary of Mean, Median & Mode

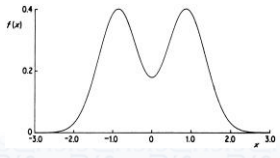
- **Mean** (algebraic measure) (sample vs. population):

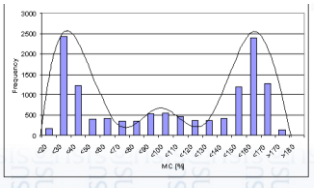
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{\sum x}{N} \qquad \bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

sample vs population weighted mean

- **Median**: middle value (odd # of values) or average of the middle 2 values (otherwise)
- **Mode**: Value that occurs most frequently in the data
  - unimodal
  - bimodal
  - trimodal







28

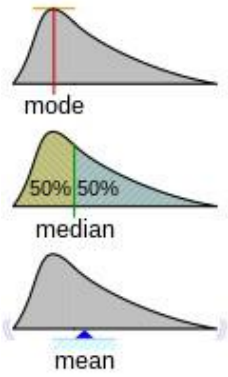
28

SJSU SAN JOSÉ STATE UNIVERSITY

Mean vs Median vs Mode

Comparison of common averages of values { 1, 2, 2, 3, 4, 7, 9 }

Type	Description	Example	Result
Arithmetic mean	*Sum of values of a data set divided by number of values*	(1+2+2+3+4+7+9) / 7	**4**
Median	*Middle value separating the greater and lesser halves of a data set*	1, 2, 2, **3**, 4, 7, 9	**3**
Mode	*Most frequent value in a data set*	1, **2**, **2**, 3, 4, 7, 9	**2**



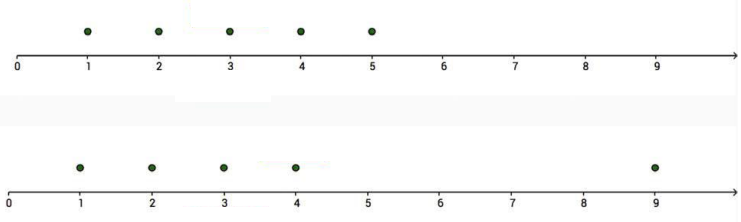
- Mean is best for datasets without outliers and symmetric distributions.
- Median is useful for skewed distributions or when outliers are present.
- Mode is ideal for categorical data or when you're interested in the most common value.

29

SJSU SAN JOSÉ STATE UNIVERSITY

Mean vs Median

- Which is more sensitive to extreme values or outliers? Mean or Median?

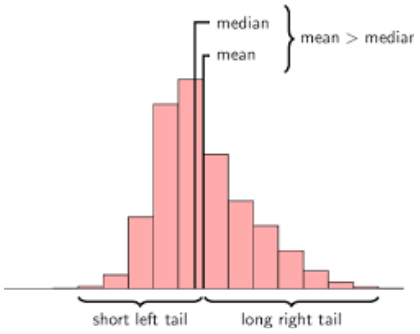


30

SJSU SAN JOSÉ STATE UNIVERSITY

Mean, Median, and Skewness

The following distribution is called right-skewed since the mean is greater than the median.

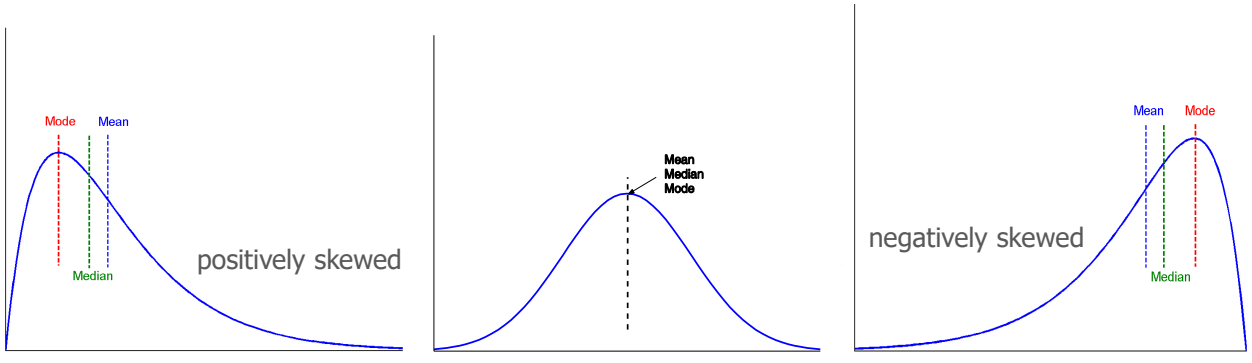


Note: skewness often "follows the longer tail".

31

SJSU SAN JOSÉ STATE UNIVERSITY

Symmetric vs. Skewed Data

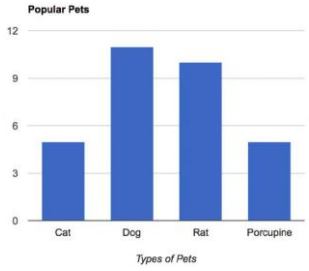


32

SJSU SAN JOSÉ STATE UNIVERSITY

Questions

- Is income positively or negatively skewed?
- For categorical variables, which makes the most sense: mean, median or mode? Why?



33

SJSU SAN JOSÉ STATE UNIVERSITY

Example: Anscombe's Data

- The following four data sets comprise the Anscombe's Quartet; all four sets of data have identical simple summary statistics.

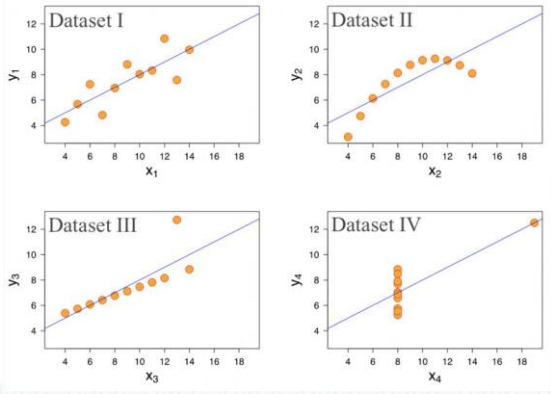
	Dataset I		Dataset II		Dataset III		Dataset IV	
	x	y	x	y	x	y	x	y
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.95	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.74	13	12.74	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.1	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.1	4	5.39	19	12.5
	12	10.84	12	9.13	12	8.15	8	5.56
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
Sum:	99.00	82.51	99.00	82.51	99.00	82.51	99.00	82.51
Avg:	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
Std:	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03

34

SJSU SAN JOSÉ STATE UNIVERSITY

Example: Anscombe's Data

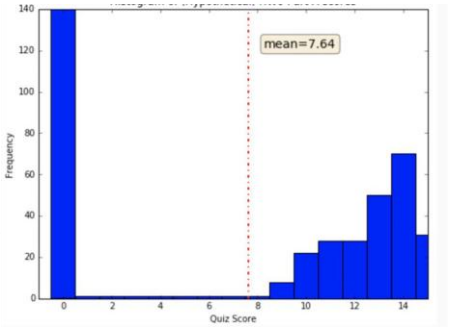
- Summary statistics clearly don't tell the story of how they differ. But a picture can be worth a thousand words:



SJSU SAN JOSÉ STATE UNIVERSITY

More Example

- If the average score for an assignment is: 7.64/15 = 50.9%, what does that suggest?



- And what does the graph suggest?

SJSU SAN JOSÉ STATE UNIVERSITY

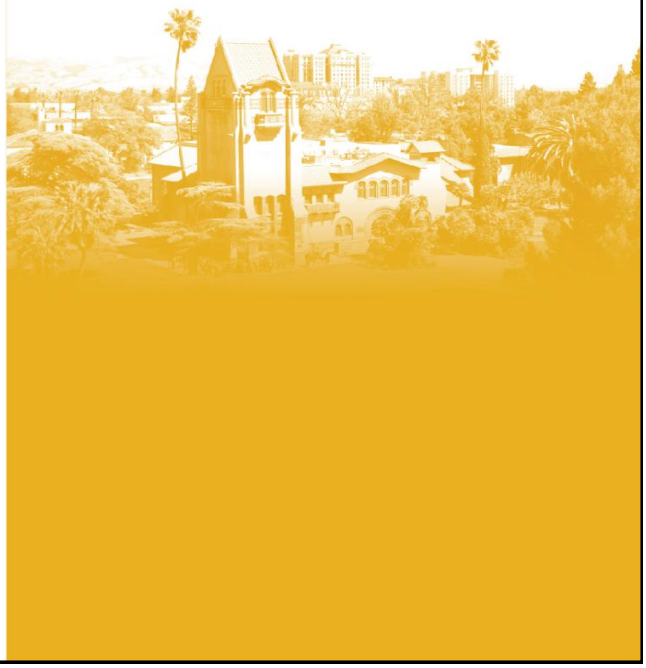
Practical Applications of Mean, Median and Mode

- Business:
  - Mean is used to determine average sales.
  - Median income to gauge typical earnings.
  - Mode for the most common product sold.
- Healthcare:
  - Median survival time in clinical trials provides a clear central measure unaffected by outliers.
- Education:
  - Mode can show the most common score in a test, indicating the most frequent performance level.

37

SJSU SAN JOSÉ STATE UNIVERSITY

Measurements of Dispersion



38

SJSU SAN JOSÉ STATE UNIVERSITY

Measurements of Dispersion

- Dispersion refers to the variations of items among themselves or around an average
- Greater variation ➔ More Dispersion
- Measurements of dispersion is useful for:
  - Determine the reliability of mean etc
  - Compare variability of 2 or more data sets
  - Facilitate the use of other statistical measures

39

39

SJSU SAN JOSÉ STATE UNIVERSITY

Measures of Dispersion

Here are some common measures of dispersion:

- Range
- Variance or Standard Deviation
- Interquartile Range (IQR)
- Coefficient of Variation

40

40

SJSU SAN JOSÉ STATE UNIVERSITY

Range

- The range is the simplest measure of dispersion.
- It is the difference between the highest and lowest values in a dataset.

Range = (maximum data value) – (minimum data value)

- Pros:
  - Easy to calculate and understand.
  - Provides a quick sense of the data spread.
- Cons:
  - Sensitive to outliers.
  - Does not provide information about the distribution of data.

22, 22, 26, 24

$$\text{range} = (\text{maximum value}) - (\text{minimum value}) = 26 - 22 = 4.0$$

SJSU SAN JOSÉ STATE UNIVERSITY

Variance and Standard Deviation

- Variance measures the average squared deviation from the mean.
- Standard deviation is the square root of variance.
- It provides an overall sense of how spread out the values are.

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}} \quad \text{sample standard deviation}$$

$$\text{population standard deviation} \quad \sigma = \sqrt{\frac{\Sigma(x - \mu)^2}{N}}$$

- Pros:
  - Uses all data points.
  - Provides a comprehensive measure of variability.
- Cons:
  - Units are squared (for variance), which may be difficult to interpret.
  - Sensitive to outliers.

22, 22, 26, 24

$n = 4$

$\Sigma x = 94$

$\Sigma x^2 = 2220$

$$s = \sqrt{\frac{n(\Sigma x^2) - (\Sigma x)^2}{n(n - 1)}} = \sqrt{\frac{4(2220) - (94)^2}{4(4 - 1)}} = \sqrt{\frac{44}{12}} = 1.9$$

SJSU SAN JOSÉ STATE UNIVERSITY

Interquartile Range (IQR)

- Interquartile range is the range of the middle 50% of data.
- It is the difference between the third quartile (Q3) and the first quartile (Q1).

$$\text{Interquartile range (or IQR)} = Q_3 - Q_1$$

Pros:

- Not affected by outliers.
- Provides a measure of spread for the central portion of the data.

Cons:

- Does not use all data points.
- Requires data to be ordered.

SJSU SAN JOSÉ STATE UNIVERSITY

Measures of Dispersion – Variance Threshold

Variance threshold – a baseline feature selection method

- Compute the variance of each feature in the dataset.
- Set a threshold value for the variance
  - Remove the features with variance below the threshold (low variability)
  - Retain the features with variance above the threshold for further analysis or modeling

- A feature with higher variance indicates that the data points are more diverse and less clustered around the mean ➔ carries more information or exhibits greater variability
- Conversely, a feature with lower variance indicates that the data points are closer to the mean, indicating less variability and potentially less informative content

SJSU SAN JOSÉ STATE UNIVERSITY

Variance Threshold Example

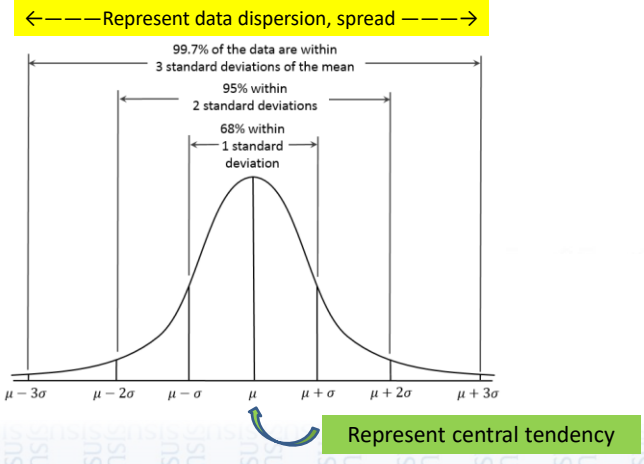
Consider a dataset with the following features: **Age**, **Height**, **Weight**, and **Income**. Use the variance threshold method to select features with a variance above a threshold of 10.

- Compute the variance of each feature:
  - Age: Variance = 10.5
  - Height: Variance = 2.1
  - Weight: Variance = 15.2
  - Income: Variance = 4.8
- Features with variance < 10 (Height and Income) ➔ low variability.
- Remove the features with low variance (Height and Income) from the dataset

SJSU SAN JOSÉ STATE UNIVERSITY

Empirical (or 68–95–99.7) Rule for Data

- For data that's symmetric or has a bell-shape distribution, the empirical rule applies:



SJSU SAN JOSÉ STATE UNIVERSITY

Chebyshev's Theorem

For datasets without with bell-shaped distributions, Chebyshev's theorem can be used:

- At least 3/4 (or 75%) of all values lie within 2 standard deviations of the mean.
- At least 8/9 (or 89%) of all values lie within 3 standard deviations of the mean.

- However, results from Chebyshev's theorem are only approximate.

SJSU SAN JOSÉ STATE UNIVERSITY

Example

Given a dataset with a mean of 100 and a standard deviation of 15, what can we conclude?

- If data distribution is bell shaped, we can apply the empirical rule:
  - 1 std dev ➔ 68% of data points is between 85 and 115
  - 2 std devs ➔ 95% of data points is between 70 and 130
  - 32 std devs ➔ 99.7% of data points is between 55 and 145

- Otherwise, use the Chebyshev's Theorem:
  - At least 3/4 (or 75%) of data points are within 2 std devs of the mean (between 70 and 130)
  - At least 8/9 (or 89%) of data points are within 3 std devs of the mean (between 55 and 145).

SJSU SAN JOSÉ STATE UNIVERSITY

Z-Score

Z score a measure of position which tells us how many standard deviations the original observation falls away from the mean and in which direction.

- It has no units of measurement.
- Z scores for "usual" data points should be between -2 and 2.

Sample	Population
$z = \frac{x - \bar{x}}{s}$ or	$z = \frac{x - \mu}{\sigma}$

Z score has various applications:

- Z score standardization
- Outlier detection

Usual values: $-2 \leq z\text{ score} \leq 2$

Unusual values: $z\text{ score} < -2$ *or* $z\text{ score} > 2$

SJSU SAN JOSÉ STATE UNIVERSITY

Example: Z Score

Suppose heights of this class students approximately bell shaped and symmetrical with mean 65" and std. dev. 1.7".

- What is the z score of a student who is (a) 70" tall, (b) 63" tall?
- Find what is the height of a student with a z score of 1.5?

SJSU SAN JOSÉ STATE UNIVERSITY

5-Number Summaries

- Quartiles: data is divided into four equal parts
- 5-Number Summary: Minimum, Q1(25th percentile), Median, Q3(75th percentile), Maximum

- Interquartile Range (IQR): difference between upper and lower quartile $\boldsymbol{Q3}$ - $\boldsymbol{Q}1$
- Outliers: we can find outliers using IQR ➔ $[\boldsymbol{Q1} - \boldsymbol{1.5} * \boldsymbol{IQR}, \boldsymbol{Q3} + \boldsymbol{1.5} * \boldsymbol{IQR}]$

SJSU SAN JOSÉ STATE UNIVERSITY

Example: Five Number Summary and Boxplot

- Given the dataset as shown:

Table 3-5 *Sorted* Counts of Chocolate Chips in Chips Ahoy (Regular) Cookies

19	19	20	20	20	20	22	22	22	22
23	23	23	23	23	23	23	24	24	24
24	24	25	25	25	25	25	25	25	26
26	26	26	26	26	27	27	28	28	30

- The Five-Number Summary is as follows: