SJSU SAN JOSÉ STATE UNIVERSITY

Descriptive Statistics
(part 2)

1

SJSU SAN JOSÉ STATE UNIVERSITY

Agenda

- Graphical Summaries of Data
- Correlation and Covariance

2

SJSU SAN JOSÉ STATE UNIVERSITY

Review of Previous Lecture

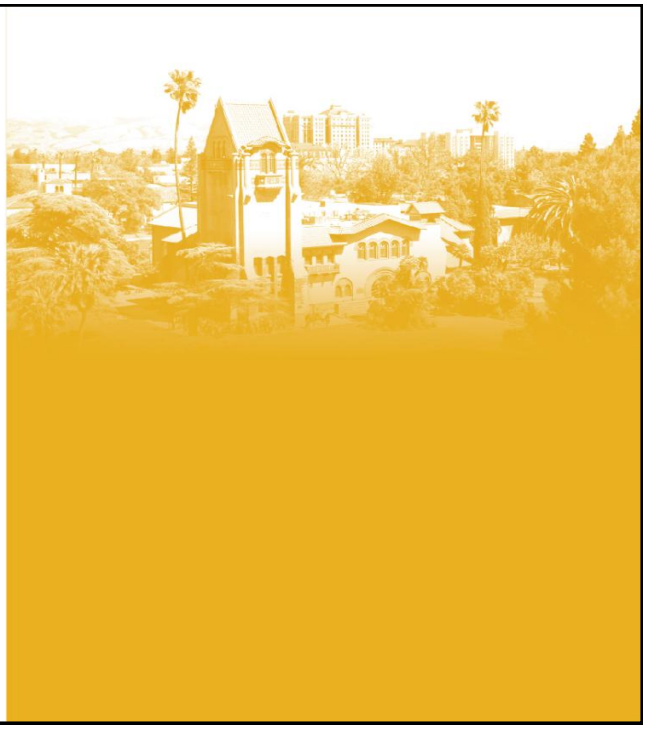
- What is Statistics? Descriptive vs Inferential
- Population, Samples, etc; Sampling Methods
- Data Types and Variable Types
- Measurements of Data
  - Central Tendency
  - Dispersion, Variability or Spread
  - Five Number Summaries

3

3

SJSU SAN JOSÉ STATE UNIVERSITY

Graphic Summaries of Data



4

SJSU SAN JOSÉ STATE UNIVERSITY

Graphical Summaries of data

Frequency table

- **Frequency** – a summary of counts for each category of the data
- **Relative frequency** – ratio between frequency of a category and sum of all frequencies
  - All relative frequencies should add up to 1 or very close to 1

```
1 df['Embarked'].value_counts()
```

```
S    644
C    168
Q     77
Name: Embarked, dtype: int64
```

```
1 # Relative Frequency
2 df['Embarked'].value_counts()/len(df)
```

```
S    0.722783
C    0.188552
Q    0.086420
Name: Embarked, dtype: float64
```

5

SJSU SAN JOSÉ STATE UNIVERSITY

Graphic Displays of Data

- Basic plots: bar and pie charts
- Boxplots: graphic display of five-number summary
- Histograms: x-axis are values, y-axis represents frequencies
- Scatter plots: each pair of values is a pair of coordinates and plotted as points in the plane
- Contingency tables: data summary for two categorical variables
- Segmented Bar and Mosaic plots:

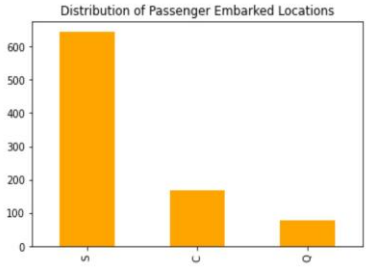
6

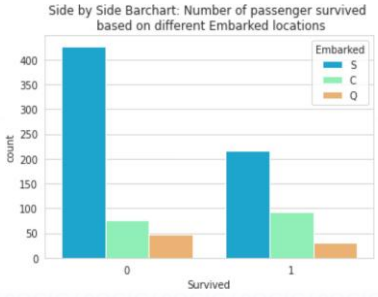
6

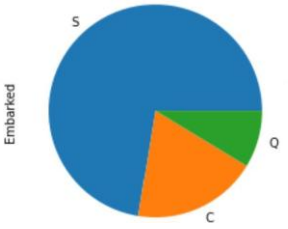
SJSU SAN JOSÉ STATE UNIVERSITY

Graphic Summaries of Data: Basic Charts

- Bar chart - display a single categorical variable
  - Pareto Chart – descending / ascending bar chart
  - Side by side bar chart
- Pie chart







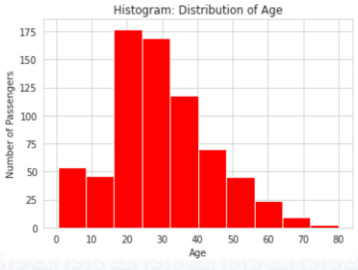
7

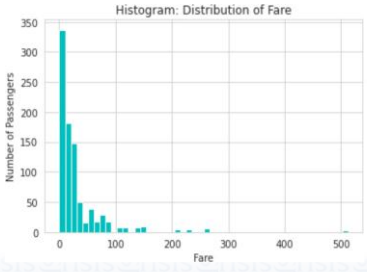
SJSU SAN JOSÉ STATE UNIVERSITY

Graphic Summaries of Data: Histograms

- Histogram
  - Visualization of distribution of continuous variable
  - Frequency distribution of continuous variable by creating classes (groups/bins)
  - All data falls into one of the groups
  - Bins: Same size, No overlap, & No gaps

Class	Frequency
10-19	
20-29	
30-39	
40-49	



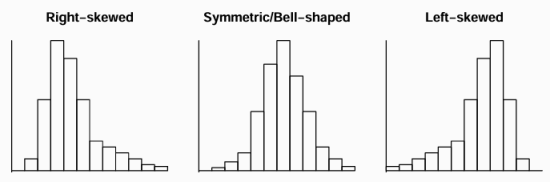


8

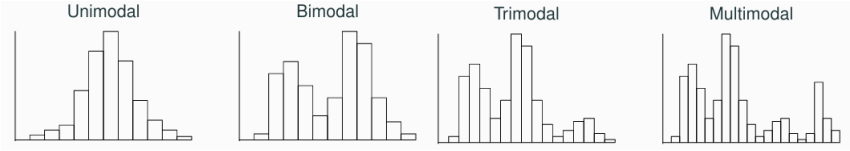
SJSU SAN JOSÉ STATE UNIVERSITY

Skewness and Modes of Histograms

- Skewness



- Mode



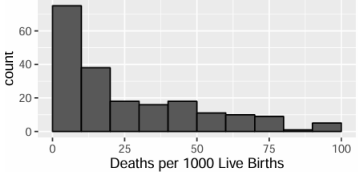
9

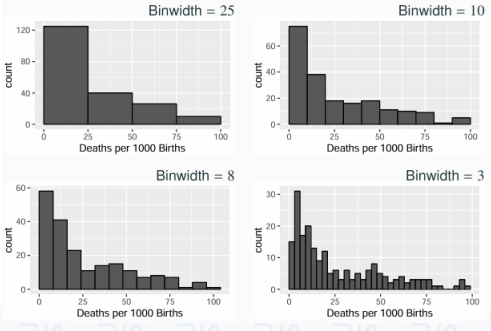
SJSU SAN JOSÉ STATE UNIVERSITY

Histograms

- Data: Infant mortality rates (number of deaths under one year of age per 1000 live births) of 201 countries/regions in 2010-2015.

Interval	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Count	75	38	18	16	18	11	10	9	1	5





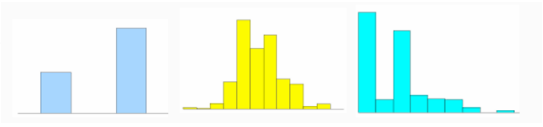
10

SJSU SAN JOSÉ STATE UNIVERSITY

Histogram Exercise

Match the following variables with the histograms and bar graphs given below. Suppose the data represent DATA202 students:

- Height of students
- Gender breakdown of students
- # of pets students have



11

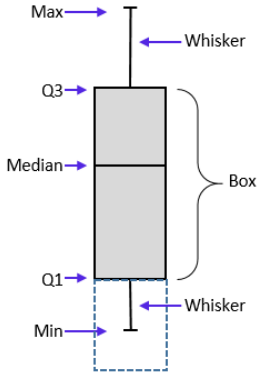
11

SJSU SAN JOSÉ STATE UNIVERSITY

Measuring the Dispersion of Data

Quartiles, **Outliners** and **Boxplots**

- Quartiles: Q_1 (25th percentile), Q_3 (75th percentile)
- Inter-quartile range: IQR = $Q_3 - Q_1$
- 5-number summary: min, Q_1, median, Q_3, max



- Outliner: a value higher/lower than 1.5x IQR of Q_1 or Q_3

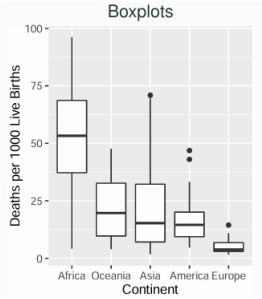
12

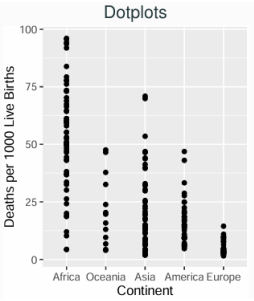
12

SJSU SAN JOSÉ STATE UNIVERSITY

Boxplot Examples

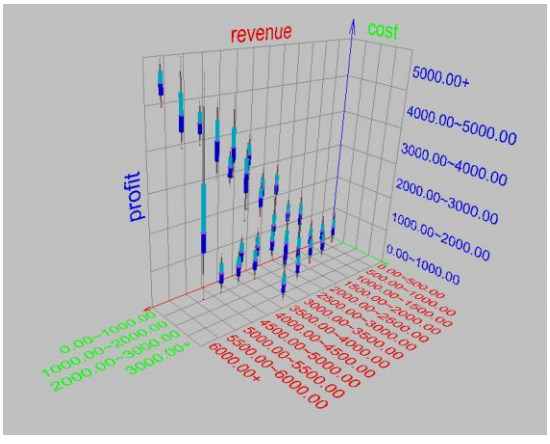
- Like histograms, boxplots of related distributions are often placed side-by-side for comparison.





SJSU SAN JOSÉ STATE UNIVERSITY

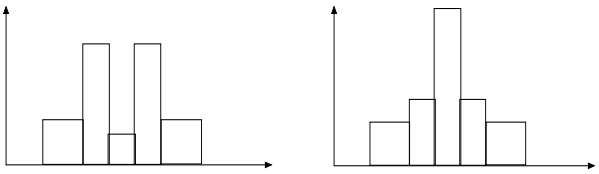
Visualization of Data Dispersion



SJSU SAN JOSÉ STATE UNIVERSITY

Histograms Often Tell More than Boxplots

- Consider the following histograms:



- These may have the same boxplot representation:
  - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions.

15

15

SJSU SAN JOSÉ STATE UNIVERSITY

Graphic Summaries of Data: Scatter Plots

Scatter Plots

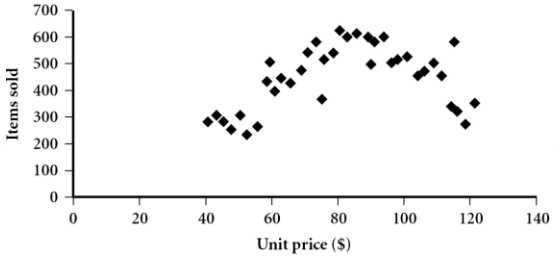
- Allows to visualize association between two variables
- Predictor (or explanatory) variables and response variable
- Can be used for examining association between variables
  - Positive association
  - Negative association
  - No association
- Form can be linear and non-linear

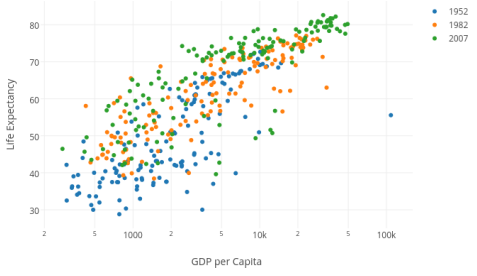
16

SJSU SAN JOSÉ STATE UNIVERSITY

Scatter Plots

- Provides a first look at bivariate data to see clusters of points, outliers, etc.
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane

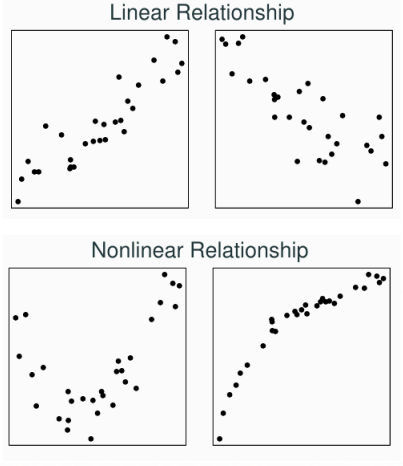


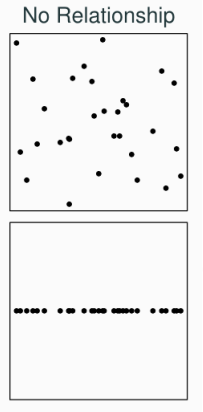


17

SJSU SAN JOSÉ STATE UNIVERSITY

Different Relationships from Scatter Plots

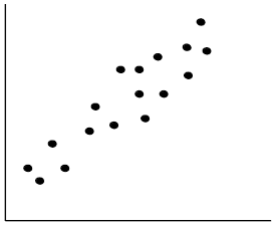




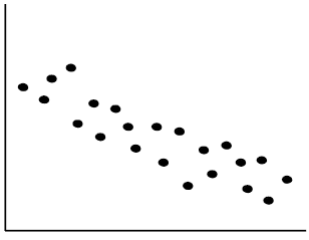
18

SJSU SAN JOSÉ STATE UNIVERSITY

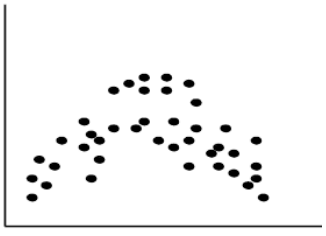
Positively and Negatively Correlated Data



Positively correlated



Negatively correlated



The left half is positively correlated
The right half is negatively correlated

19

SJSU SAN JOSÉ STATE UNIVERSITY

Contingency Tables

- Contingency Table: A table that summarizes data for two categorical variables
- Example: Breakdown of passengers on Titanic by class and survival status

		Died	Survived	Total
Class	1st	122	203	325
	2nd	167	118	285
	3rd	528	178	706
	Crew	673	212	885
	Sum	1490	711	2201

20

SJSU SAN JOSÉ STATE UNIVERSITY

Contingency Tables: Overall Proportions

- From a contingency table, we can divide each cell by the overall total to get the proportions of observations in the different combinations:

		Survived		
		No	Yes	Total
Class	1st	122/2201 ≈ 0.06	203/2201 ≈ 0.09	325/2201 ≈ 0.15
	2nd	167/2201 ≈ 0.08	118/2201 ≈ 0.05	285/2201 ≈ 0.13
	3rd	528/2201 ≈ 0.24	178/2201 ≈ 0.08	706/2201 ≈ 0.32
	Crew	673/2201 ≈ 0.31	212/2201 ≈ 0.10	885/2201 ≈ 0.40
	Sum	1490/2201 ≈ 0.68	711/2201 ≈ 0.32	1

SJSU SAN JOSÉ STATE UNIVERSITY

Contingency Tables: Row Proportions

- From a contingency table, we can divide each cell by the corresponding row totals to get the proportions of passengers survived in the four classes (rows):

		Survived		
		No	Yes	Total
Class	1st	122/325 ≈ 0.38	203/325 ≈ 0.62	1
	2nd	167/285 ≈ 0.59	118/285 ≈ 0.41	1
	3rd	528/706 ≈ 0.75	178/706 ≈ 0.25	1
	Crew	673/885 ≈ 0.76	212/885 ≈ 0.24	1

SJSU SAN JOSÉ STATE UNIVERSITY

Contingency Tables: Column Proportions

- From a contingency table, we can divide each cell by the corresponding column totals to get the proportions of passengers survived in each of the four classes:

		Survived	
		No	Yes
Class	1st	122/1490 ≈ 0.08	203/711 ≈ 0.29
	2nd	167/1490 ≈ 0.11	118/711 ≈ 0.17
	3rd	528/1490 ≈ 0.35	178/711 ≈ 0.25
	Crew	673/1490 ≈ 0.45	212/711 ≈ 0.30
	Sum	1	1

SJSU SAN JOSÉ STATE UNIVERSITY

Independence of Two Categorical Variables

- If the row proportions do not change from row to row, the two categorical variables are **independent**. Otherwise, the two categorical variables are **associated**.
- In the Titanic example, the survival of passengers is associated with the class they were in because the survival rates differ substantially from class to class:

		Survived		
		No	Yes	Total
Class	1st	122/2201 ≈ 0.06	203/2201 ≈ 0.09	325/2201 ≈ 0.15
	2nd	167/2201 ≈ 0.08	118/2201 ≈ 0.05	285/2201 ≈ 0.13
	3rd	528/2201 ≈ 0.24	178/2201 ≈ 0.08	706/2201 ≈ 0.32
	Crew	673/2201 ≈ 0.31	212/2201 ≈ 0.10	885/2201 ≈ 0.40
	Sum	1490/2201 ≈ 0.68	711/2201 ≈ 0.32	1

SJSU SAN JOSÉ STATE UNIVERSITY

Independence of Two Categorical Variables

- We can also define two categorical variables to be **independent** if the column proportions do not vary from column to column.
- The two conditions are equivalent… Why?

		Survived	
		No	Yes
Class	1st	122/1490 ≈ 0.08	203/711 ≈ 0.29
	2nd	167/1490 ≈ 0.11	118/711 ≈ 0.17
	3rd	528/1490 ≈ 0.35	178/711 ≈ 0.25
	Crew	673/1490 ≈ 0.45	212/711 ≈ 0.30
	Sum	1	1

SJSU SAN JOSÉ STATE UNIVERSITY

Exercise

The table below shows the breakdown of cases of injuries in the U.S in a certain year. by circumstance and gender. Counts are in millions.

	Circumstance			
Gender	Work	Home	Other	Total
Male	8.0	9.8	17.8	35.6
Female	1.3	11.6	12.9	25.8
Total	9.3	21.4	30.7	61.4

- What proportion of injury cases occurred at work?
- What proportion of injury cases occurred at work and on women?
- Among all injury cases occurred on women, what proportion occurred at work?
- Among all injury cases occurred at work, what proportion occurred on women?
- Is the circumstance of injury cases independent of the gender of the victims?

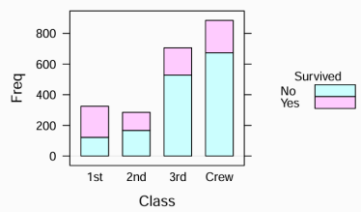
Source: Vital and Health Statistics published by the National Center for Health Statistics

SJSU SAN JOSÉ STATE UNIVERSITY

Segmented Bar Plots

- **Segmented bar plots** (stacked bar plots) are a type of bar chart that displays multiple categorical variables in one bar.
- Each bar is divided into segments that represent the different categories or groups within the variable.
- The length of each segment corresponds to the proportion or count of that category in relation to the total.

		Died	Survived	Total
Class	1st	122	203	325
	2nd	167	118	285
	3rd	528	178	706
	Crew	673	212	885
	Sum	1490	711	2201

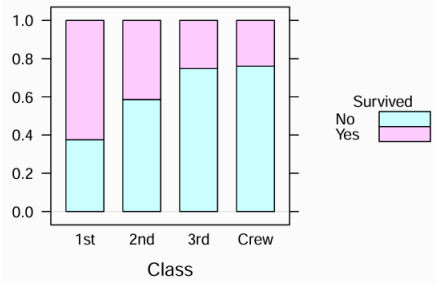


SJSU SAN JOSÉ STATE UNIVERSITY

Standardized Segmented Bar Plots

- **Standardized segmented bar plots** are generated on the row proportions.
- They are convenient for comparing row proportions, and determining whether the two variables are independent

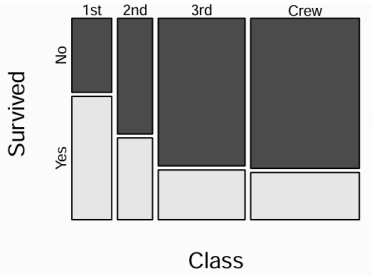
		Survived		
		No	Yes	Total
Class	1st	122/325 ≈ 0.38	203/325 ≈ 0.62	1
	2nd	167/285 ≈ 0.59	118/285 ≈ 0.41	1
	3rd	528/706 ≈ 0.75	178/706 ≈ 0.25	1
	Crew	673/885 ≈ 0.76	212/885 ≈ 0.24	1



SJSU SAN JOSÉ STATE UNIVERSITY

Mosaic Plots

- **Mosaic plots** are graphical representations of multivariate categorical data. They use tiles to represent the proportions of combinations of categories.
- Each tile's size is proportional to the frequency or count of the corresponding category combination.

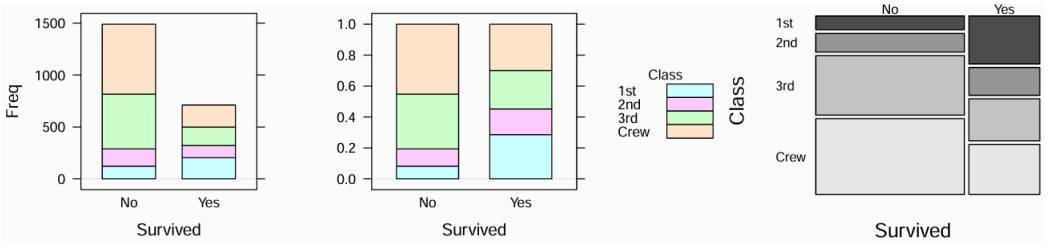


- bar widths = row totals
- segment lengths within a bar = row proportions

$$\boxed{\text{segment area}} = (\text{barwidth}) \times (\text{segment length})$$
$$= \text{row total} \times (\text{row proportion})$$
$$= \text{row total} \times \frac{\text{cell count}}{\text{row total}} = \boxed{\text{cell count}}$$

SJSU SAN JOSÉ STATE UNIVERSITY

Segmented Bar vs Mosaic Plots



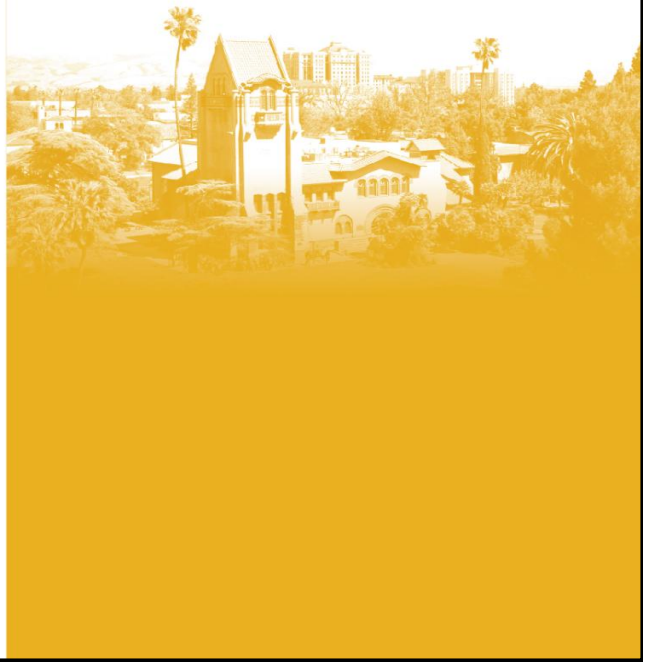
SJSU SAN JOSÉ STATE UNIVERSITY

Ways to Inspect Relationships Between Variables

- Numerical vs Numerical
  - scatterplots
- Categorical vs Categorical
  - contingency tables
  - segmented bar plots, standardized segmented bar plots, mosaic plots
- Categorical vs Numerical
  - side-by-side boxplots
  - histograms by group on the same horizontal axis

SJSU SAN JOSÉ STATE UNIVERSITY

Correlation and Covariance



SJSU SAN JOSÉ STATE UNIVERSITY

Covariance

- Covariance is a measure of how much two random variables change together. It indicates the direction of the linear relationship between variables.

$$Cov(X,Y) = \frac{\Sigma(X-\mu_X)(Y-\mu_Y)}{N}$$

- It's the average of the product of the deviations of each pair of data points from their respective means.
- Covariance values can range from negative to positive infinity.
  - Positive covariance indicates that the variables tend to increase together.
  - Negative covariance indicates that as one variable increases, the other tends to decrease.
  - Zero covariance suggests no linear relationship.
- Covariance provides a sense of the direction of the relationship but not its strength.

33

SJSU SAN JOSÉ STATE UNIVERSITY

Correlation

- Correlation is a statistical measure that describes the strength and direction of a relationship between two variables. It shows how one variable changes in relation to another.

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

- Correlation values range from -1 to 1:
  - +1: Perfect positive correlation (as one variable increases, the other also increases).
  - 0: No correlation (no linear relationship between the variables).
  - -1: Perfect negative correlation (as one variable increases, the other decreases).
- Pearson Correlation Coefficient: Measures linear relationships.
- Spearman's Rank Correlation: Measures monotonic relationships (not necessarily linear).

34

SJSU SAN JOSÉ STATE UNIVERSITY

Example: Covariance and Correlation of Two Variables

- Calculate the Mean of X and Y
- Calculate the Deviations from the Mean
- Calculate Covariance
- Calculate the Standard Deviations of X and Y
- Calculate Correlation

	x	y
1	0	1
2	2	2
3	1	1
4	1	0

35

35