SJSU SAN JOSÉ STATE UNIVERSITY

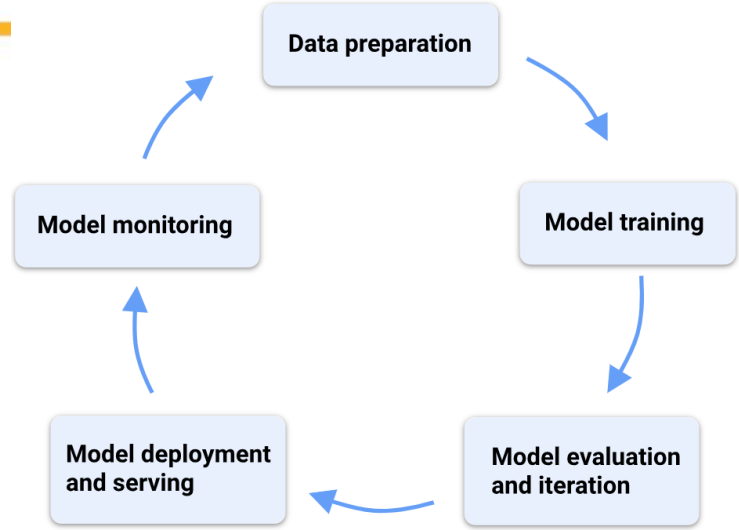
DATA 220
Mathematical Methods for Data Analytics

Dr. Mohammad Masum

Introduction to Statistics

- What is Statistics?
  - Descriptive Statistics
  - Inferential Statistics
- What is Data?
- Sample vs. Population
  - Statistic vs parameter

Introduction to Statistics – Sampling



- Sampling is Integral part of 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 that you have access to – too much time, too much computing power, too much money
- Allows to accomplish a task faster and cheaper
  - For instance, performing 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-4903c40ebf86
https://cloud.google.com/vertex-ai/docs/start/introduction-unified-platform

Sampling

Non-Probability Sampling

- Convenience Sampling
- Snowball Sampling
- Judgement Sampling
- Quota Sampling

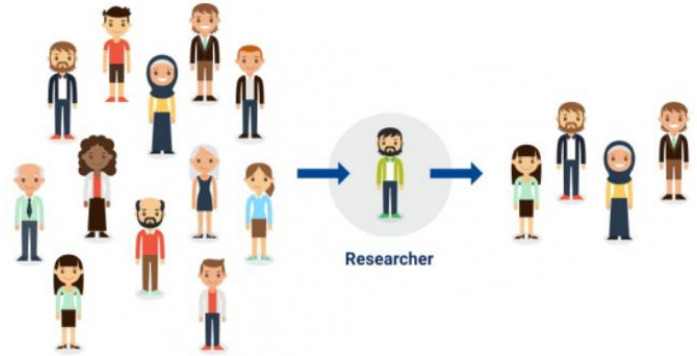
Probability Sampling

- Simple Random Sampling
- Stratified Sampling
- Cluster Sampling
- Systematic Sampling

https://medium.com/analytics-vidhya/sampling-statistical-approach-in-machine-learning-4903c40ebf86

Non-Probability Sampling

- **Convenience Sampling** – samples are selected based on availability
- **Snowball Sampling** – Future samples are selected based on existing samples
- **Judgement Sampling** – experts decide what samples to include
- **Quota Sampling** – selecting samples based on quotas for certain slices without any randomization



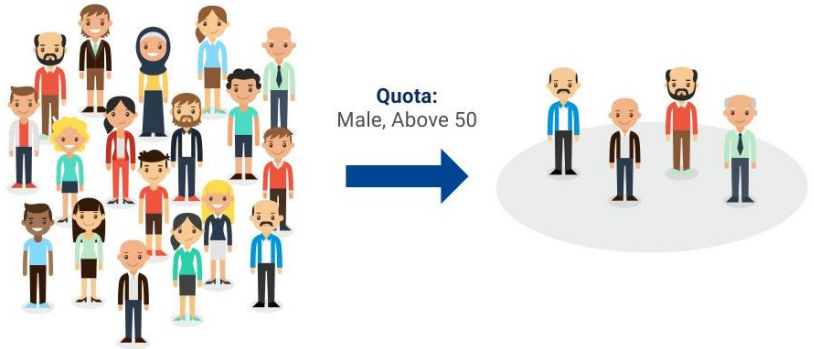
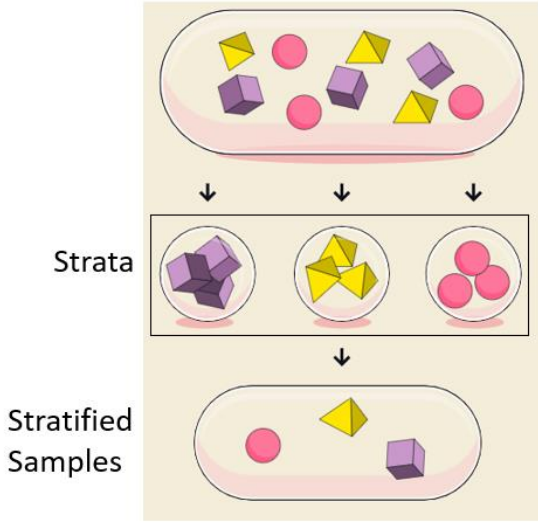


Image source - https://medium.com/analytics-vidhya/sampling-statistical-approach-in-machine-learning-4903c40ebf86

Probability Sampling

- **Simple Random Sampling:** Randomly selecting individuals from a population, where each has an equal chance to be selected
- **Stratified Sampling:** Dividing the population into distinct groups and then randomly selecting samples from each group.
- **Cluster Sampling:** Dividing the population into clusters or groups, then randomly choosing some clusters
- **Systematic Sampling:** Selecting every nth individual from a population after randomly choosing a starting point



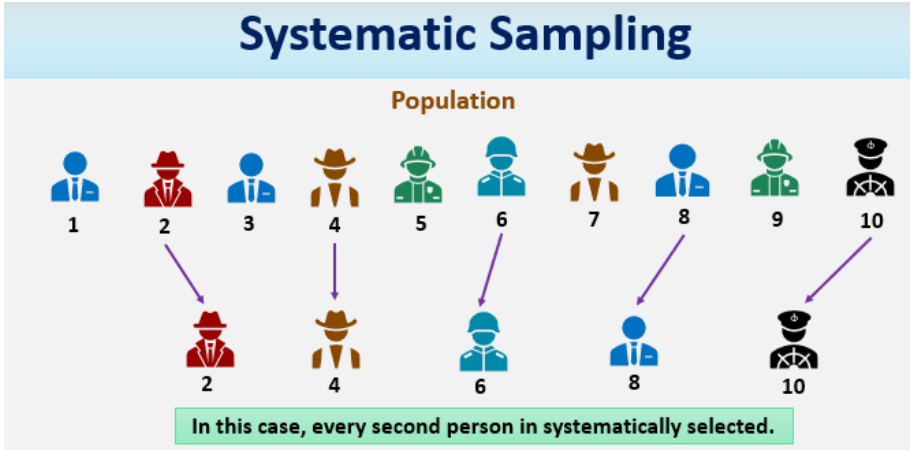


Image source - https://medium.com/analytics-vidhya/sampling-statistical-approach-in-machine-learning-4903c40ebf86

Introduction to Statistics – Sampling

A. We take a list of all students enrolled and identify every 100^{th} student for our sample
B. We stand in the commons area and stop students as they walk by to get their opinion.
C. We randomly select 50 classes then interview everyone in those classes.
D. We take a list of all students that is numbered. We use a random number generator to select 500 numbers. The students corresponding to those numbers on the list are selected for the sample.
E. We recognize that women may have a different opinion than men so we first divide the population by gender then randomly select male and then randomly select female students for our sample.

Introduction to Statistics – Selection Bias

- **Selection Bias** – Data (samples) are selected in a way that is not reflective / 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/

Introduction to Statistics - Variable types

- Recognizing variable types is important - choosing appropriate statistical methods, visualization techniques, and modeling approaches
- **Qualitative or categorical** - represent qualities or categories that cannot be measured numerically
  - **Ordinal:** categories with a meaningful order
  - **Nominal:** without a specific order among categories
- **Quantitative** - represent numerical measurements that have meaningful magnitudes and differences
  - **Discrete:** distinct, separate values and often involve counting
  - **Continuous:** can take on any value within a certain range

Introduction to Statistics - Variable types – 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/

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
```

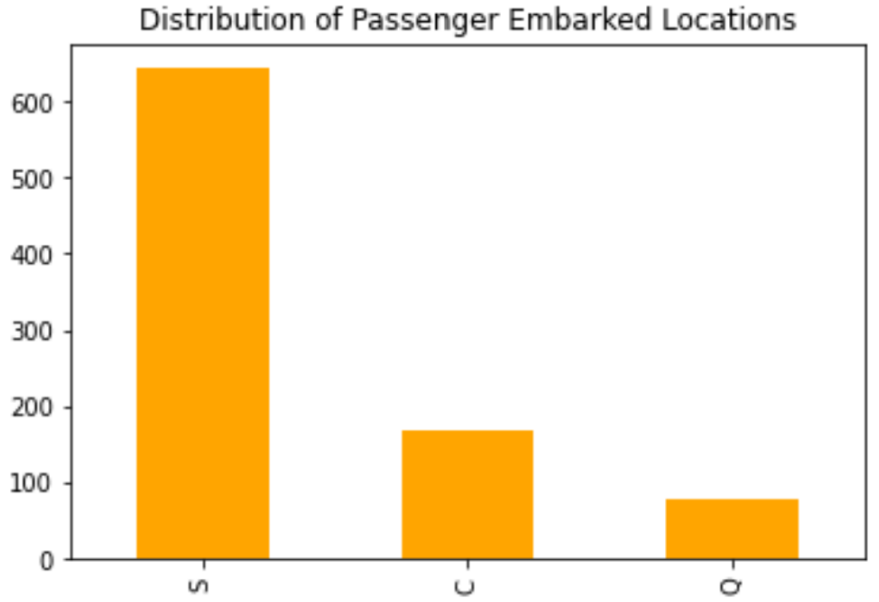
```
1 # Relative Frequency
2 df['Embarked'].value_counts()/len(df)
```

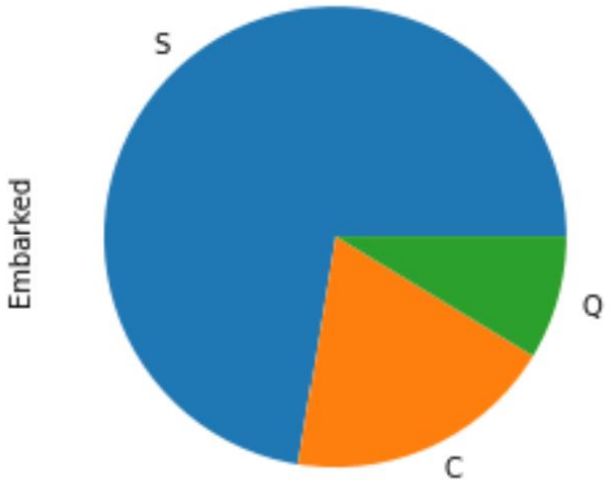
```
S    0.722783
C    0.188552
Q    0.086420
Name: Embarked, dtype: float64
```

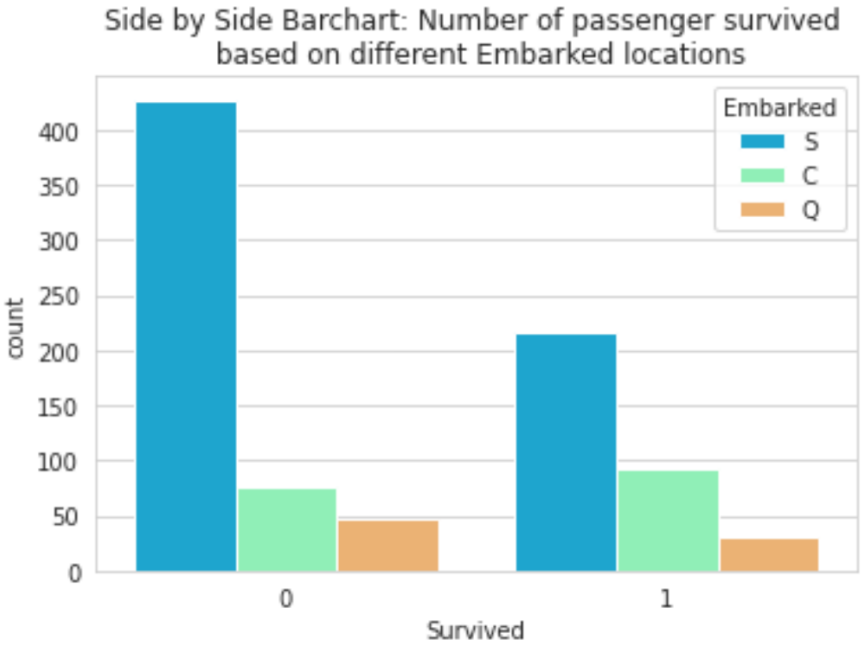
SJSU SAN JOSÉ STATE UNIVERSITY

Graphical Summaries of data

- Bar chart
  - Pareto Chart – descending / ascending bar chart
  - Side by side bar chart
- Pie chart



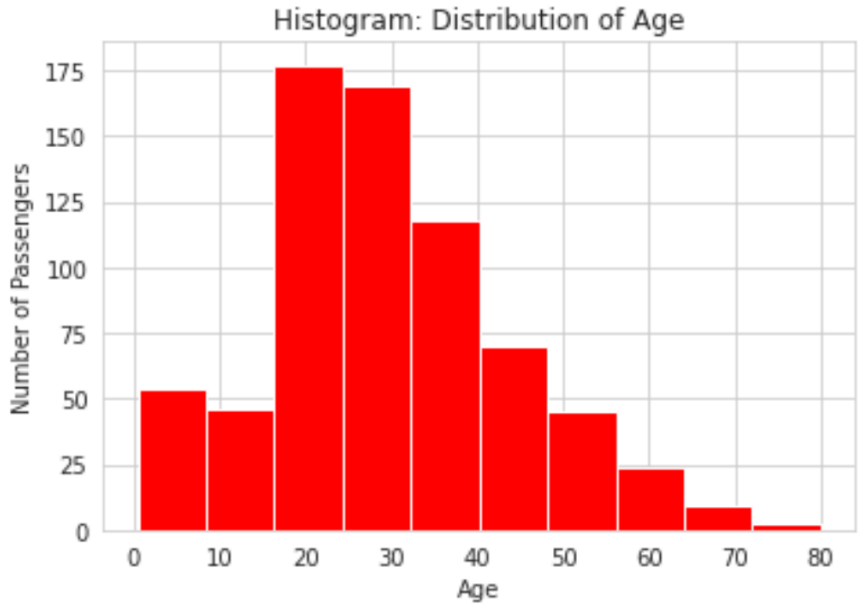


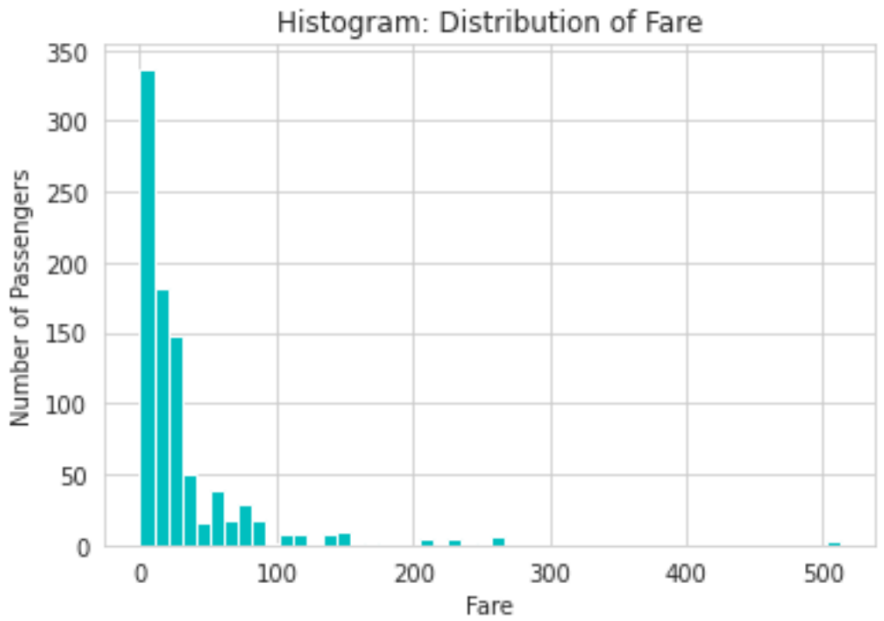


Graphical Summaries of data

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

- Histogram
  - Can be used for visualize 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

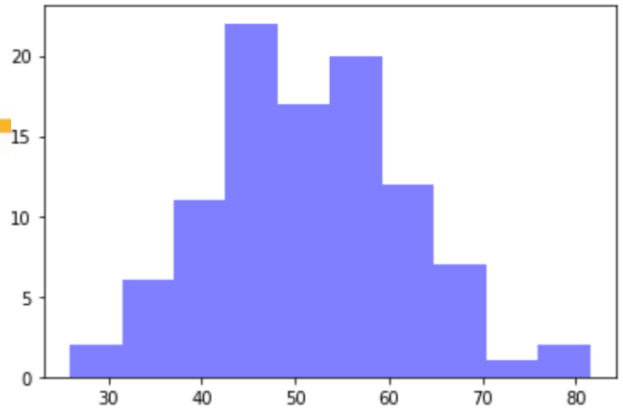


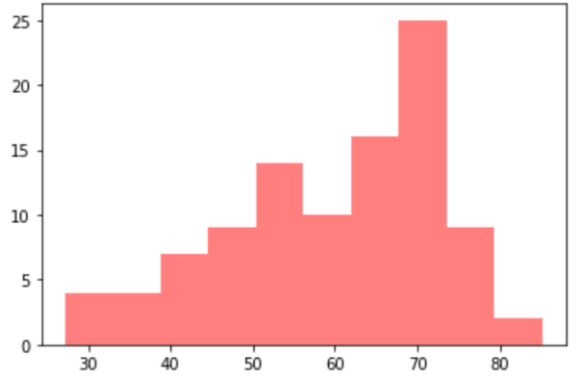


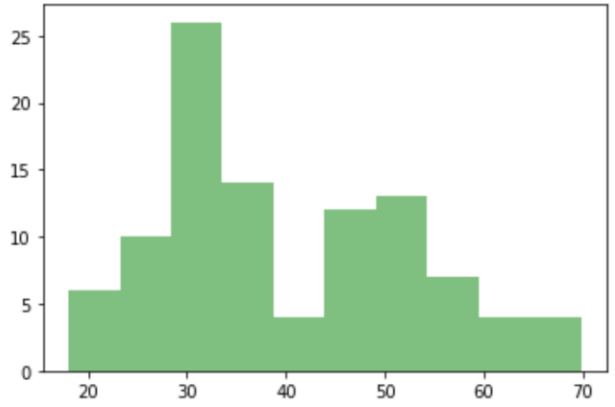
SJSU SAN JOSÉ STATE UNIVERSITY

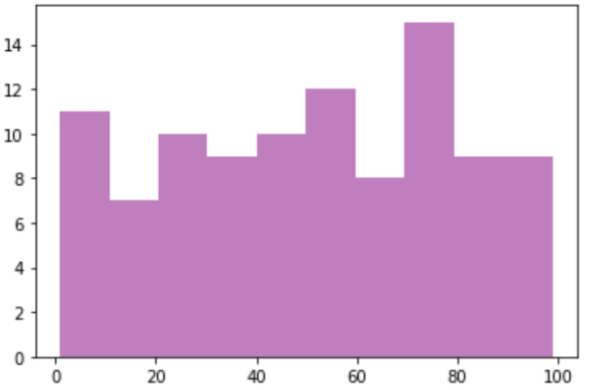
Graphical Summaries of data

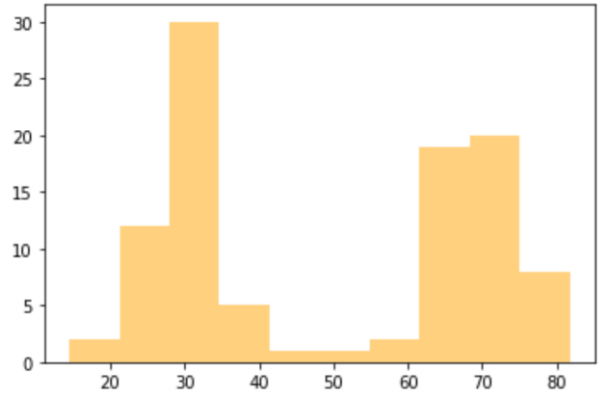
- Graphical Summaries of data
  - Histogram
    - Can be used for visualize distribution of continuous variable
    - Shape
      - Symmetric
      - Skewed right
      - Skewed left
      - Unimodal
      - Bimodal
      - Uniform











Graphical Summaries of data

- Determining Bins for Histogram
  - Square root rule: Approximate Square root of sample size
    - Simple & easy to apply
    - May not provide optimal number of bins
  - Scott's normal reference rule: $\boldsymbol{bins = 3.5\,\sigma\, n^{-\frac{1}{3}}}$
    - Consider both sample size and standard deviation of the data
    - Assumes symmetric distribution (which may not be the case)
  - Freedman-Diaconis Rule: $\boldsymbol{bins = \frac{IQR}{2}\ n^{-\frac{1}{3}}}$
    - Consider sample size and variability of the data
    - Can be sensitive to outliers

Numerical Summaries of data

- Numerical Summaries of data
  - Mean
  - Median
- Comparing the mean and median
  - Which one is best measures of center?
  - Which one is resistant to outliers?

Numerical Summaries of data

Accessing individual data may not be feasible for all cases -

- **Data Privacy:** For confidentiality, data might be presented as frequency distributions to protect individual data points.
- **Data Collection Challenges:** When precise data collection is difficult, intervals or categories are used for reporting.

Calculate means for grouped data

1. Compute midpoint of each class in the frequency distribution
2. For each class: midpoint * frequency
3. Find sum of all the values in step 2 and then divide by total # of observations

	Class	Frequency	Midpoint	Midpoint_Frequency
0	$0-$49	25	24.5	612.5
1	$50-$99	45	74.5	3352.5
2	$100-$149	30	124.5	3735.0

```
[4]   1 total_midpoint_frequency
```

```
7700.0
```

```
[5]   1 total_frequency
```

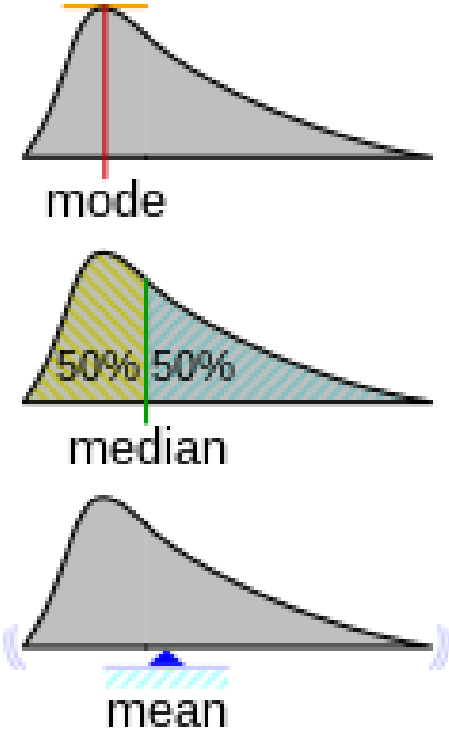
```
100
```

```
[7]   1 group_mean
```

```
77.0
```

Introduction to Statistics

- Mode
  - values that occurs most frequently in a set of data (categorical variables)
    - understanding dominant categories and making informed decisions based on the frequency of occurrence
  - Values corresponding to the peaks of the distribution (Continuous Variables)
    - detecting potential clusters or patterns, guiding decisions like price points, or understanding key trends in a dataset



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

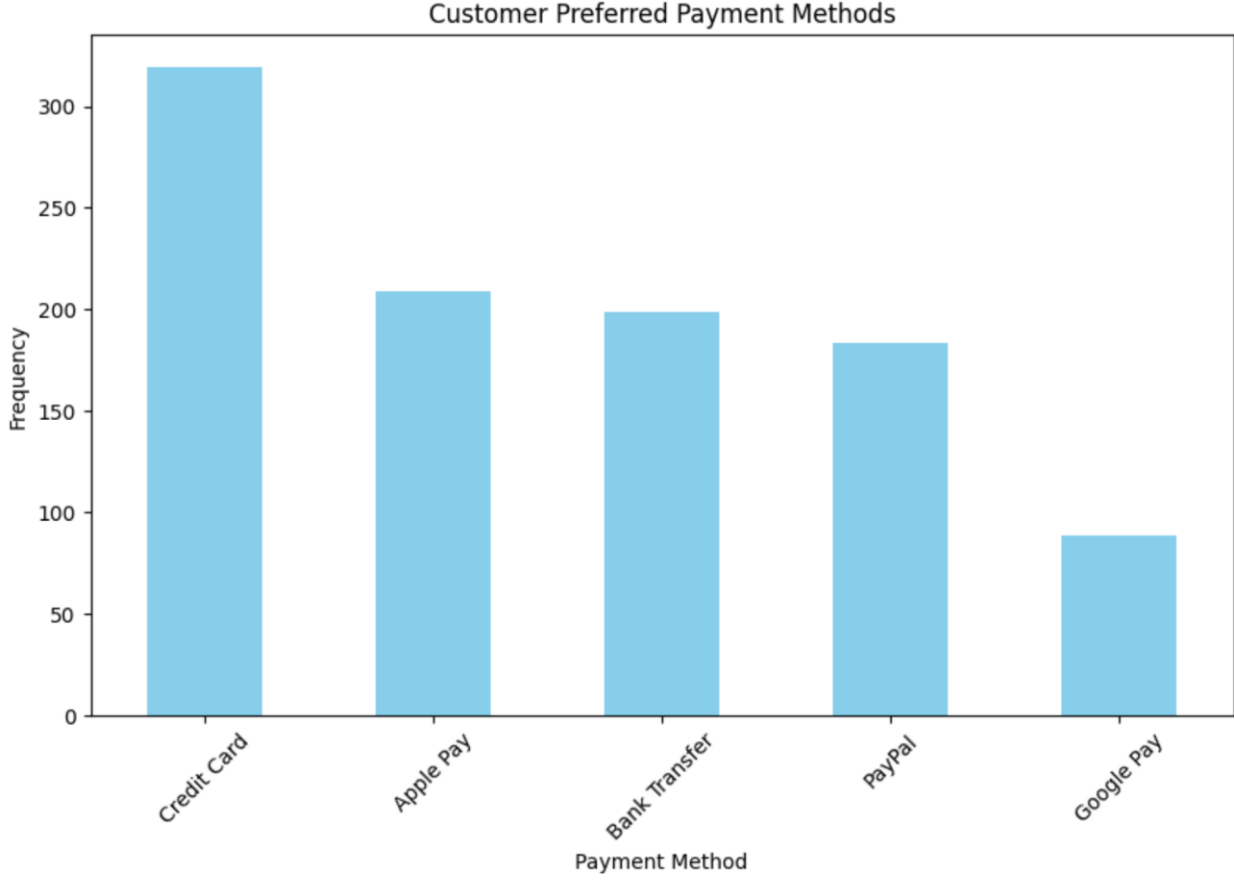
Type	Description	Example	Result
Arithmetic mean	*Sum of values of a data set divided by number of values*	(1+2+2+3+4+7+9) / 7	**4**
Median	*Middle value separating the greater and lesser halves of a data set*	1, 2, 2, **3**, 4, 7, 9	**3**
Mode	*Most frequent value in a data set*	1, **2**, **2**, 3, 4,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

https://en.wikipedia.org/wiki/Mode_(statistics)

SJSU SAN JOSÉ STATE UNIVERSITY

Introduction to Statistics – Modes

- Explore 1,000 customer payment preferences for an online retail platform
- Use the mode to make a decision




```
Most common payment method: Credit Card
Decision: Promote Credit Card offers to boost sales.
```

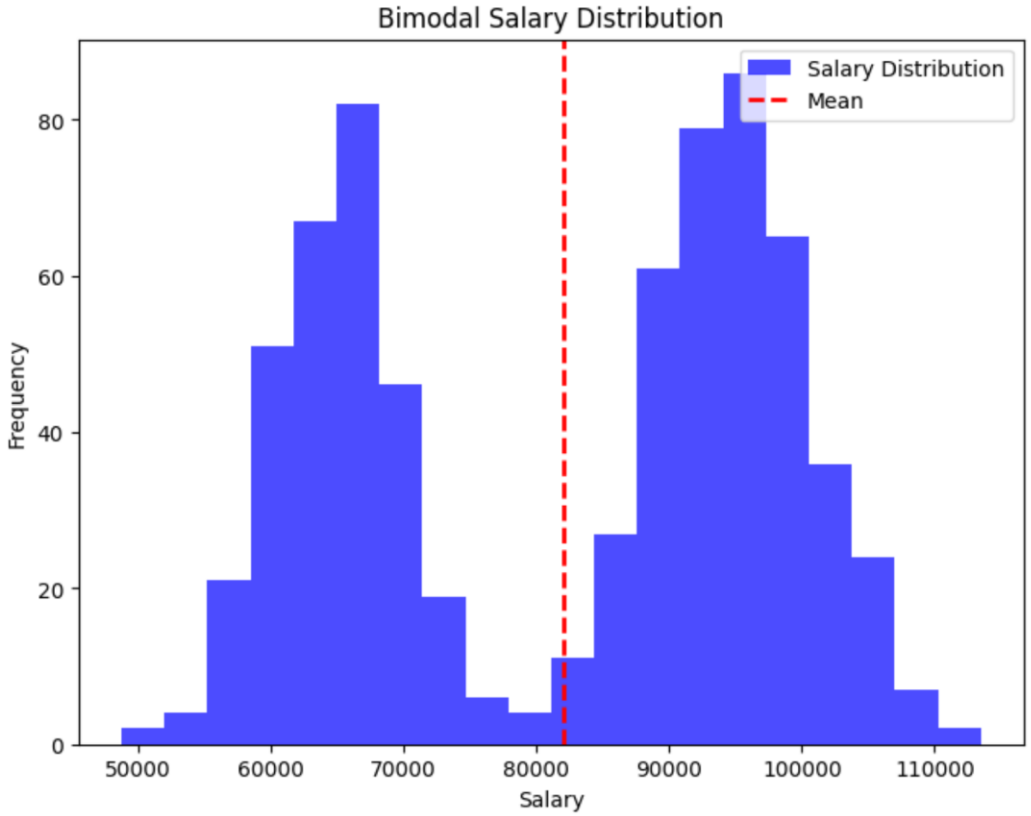
```python
# Make a decision based on the mode
if mode_payment == 'Credit Card':
    decision = "Promote Credit Card offers to boost sales."
elif mode_payment == 'PayPal':
    decision = "Enhance PayPal integration for customer convenience."
else:
    decision = "Explore incentives for using the most common payment method."

print("Most common payment method:", mode_payment)
print("Decision:", decision)
```

SJSU SAN JOSÉ STATE UNIVERSITY

Introduction to Statistics – Modes

- Consider a dataset of employee salaries - contains two distinct groups: junior employees and senior employees



Identify the Modes:

- First mode: Corresponds to junior employee salaries occurs at ..
- Second mode: Corresponds to senior employee salaries occurs at ..

Decision: Adjust Compensation Packages:

- Different compensation packages for junior and senior employees based on the modes.
  - Junior employees: Competitive entry-level salaries, growth opportunities.
  - Senior employees: Compensation reflects experience and responsibilities.

Introduction to Statistics

- Measures of Spread
  - Range: largest data value – lowest data value
  - Variance: measures how the data points are spread from the mean
    - Standard deviation: square root of the variance
  - Practice:

$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

S^2 = sample variance

x_i = the value of the one observation

$\bar{x}$ = the mean value of all observations

n = the number of observations

Introduction to Statistics

- Variance threshold – a baseline feature selection method
  - A feature with higher variance indicates that the data points are more diverse and less clustered around the mean
    - feature 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

Introduction to Statistics

- Variance threshold – a baseline feature selection method
    1. Compute the variance of each feature in the dataset.
    2. Set a threshold value for the variance
        - Remove the features with variance below the threshold: Features with a variance below this threshold are considered to have low variability
        - Retain the features with variance above the threshold for further analysis or modeling

Introduction to Statistics

- Variance threshold – a baseline feature selection method
  - Example: Consider a dataset with the following features: Age, Height, Weight, and Income. We want to 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 below the threshold (Height and Income) are considered to have low variability.
  - Remove the features with low variance (Height and Income) from the dataset

https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.VarianceThreshold.html

Introduction to Statistics

- The empirical Rule (68-95-99.7 rule)
  - Many histogram have a bell shaped or symmetric distribution, in that case we can apply empirical rule
- Practice
  - Suppose we have a bell-shaped and symmetrical distribution data with mean 100 and std of 5. Based on the empirical rule:
    - What are the cut off values for middle 68%
    - What percentage of the values are greater than 110?
    - What percentage of the values are less than 95?
    - What percentage of the values are between 85 and 115?
    - What percentage of the values are between 90 and 105?

Introduction to Statistics

- Chebyshev's inequality
  - If the distribution is unknown
    - At least 75% data within 2 std
    - At least 88.9% data within 3 std
- Suppose that we have a data distribution with mean 50 and std 3:
  - Assume the data distribution is bell-shaped: what are the cut-off values for middle 95%
  - If we can not assume the data distribution is bell-shaped, at least what percentage of values are between 41 and 59?

Introduction to Statistics

$$Z = \frac{x - \mu}{\sigma}$$

Z = standard score

x = observed value

μ = mean of the sample

σ = standard deviation of the sample

- Measure of position
  - Z score: tells us how many standard deviations the original observation falls away from the mean and in which direction
  - Example: suppose heights of this class students approximately bell-shaped and symmetrical with mean 65 inches and std 1.7 inches.

1. What is the z-score of a student who is (a) 70 inches tall (b) 63 inches tall?
2. Find what is the height of a student with a z-score of -1.5?

feature_value	z_score
8	0.592187
3	-1.258396
8	0.592187
9	0.962303
4	-0.888280

```
Mean of feature_value:  6.4
Standard deviation of feature_value:  2.701851217221259
```

Introduction to Statistics

- Z score and the empirical rule
  - Draw the curve and mark the cut-off values for z-scores
- How comparing values in two different distributions?
  - Example: The mean length of one-year-old spotted flounder is 126 mm with standard deviation of 18 mm and the mean length of two-year-old spotted flounder is 162 mm with a standard deviation of 28 mm. The distribution of flounder lengths is approximately bell-shaped.
    1. Anna caught a one-year-old flounder that was 150 mm in length. What is the z score for this weight?
    2. Luis caught a two-year old flounder that was 190 mm in length. What is the z-score for this length?
    3. Whose fish is longer relative to fish the same age?

Example taken from Dr. Linda Galloways' note

Introduction to Statistics

- Application of Z-score:
  - Z- score standardization
  - Z score for outlier detection
    - Assumes normality
    - Sensitive to outliers
    - Fixed threshold
      - Can be solved using Quantiles (or, dynamic thresholding)
  - Modified Z-score for outlier detection: $\boldsymbol{modified\ Z-score} = \frac{\mathbf{0.6745}\ (\boldsymbol{x-median})}{\boldsymbol{MAD}}$
    - Median Absolute Deviation, $\mathrm{MAD} = median\ (|x - median|)$
    - Robust to non-normal distribution
    - Less sensitive to extreme outliers

Introduction to Statistics

- Modified Z-score for outlier detection: $\boldsymbol{modified\ Z-score} = \frac{\mathbf{0.6745\ (x-median)}}{\boldsymbol{MAD}}$
  - Median Absolute Deviation, $\text{MAD} = median\ (|x - median|)$

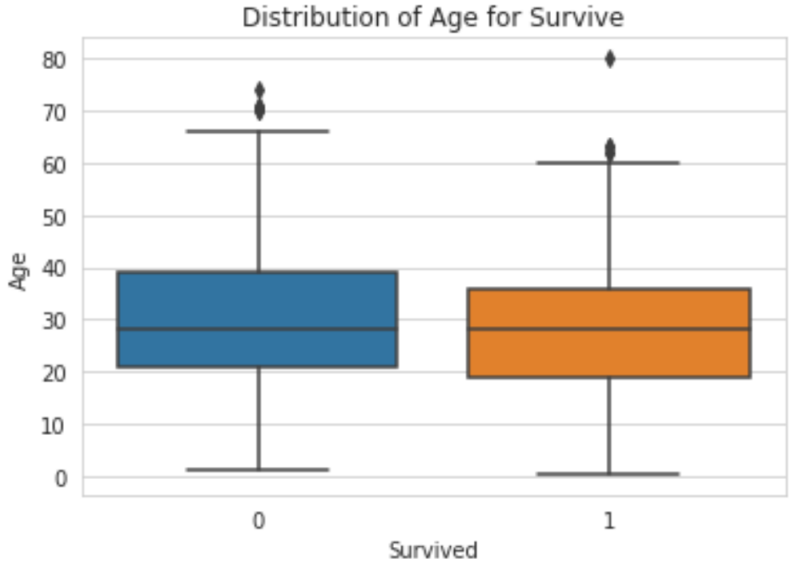
feature_value	z_score	abs_deviation	modified_z_score
8	0.592187	0.0	0.0000
3	-1.258396	5.0	-3.3725
8	0.592187	0.0	0.0000
9	0.962303	1.0	0.6745
4	-0.888280	4.0	-2.6980

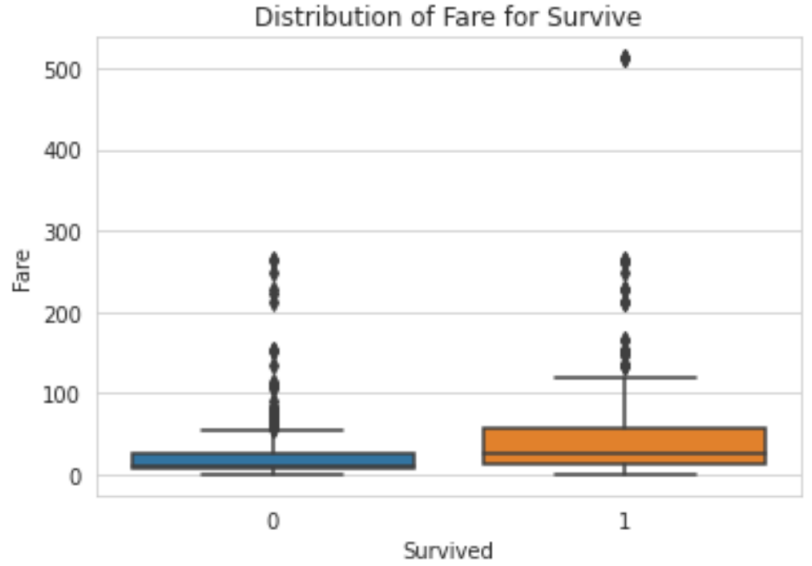
Introduction to Statistics

- Quartiles: divide the data into four equal parts
- Five Number Summary: minimum, Q1, median, Q3, maximum
- Interquartile Range (IQR): Q3-Q1
- Outliers: we can find outliers using IQR ➔ $[\boldsymbol{Q1 - 1.5 * IQR, Q3 + 1.5 * IQR}]$
- Practice: Is there any outlier in the table?

Introduction to Statistics

- Boxplots: a graph that represent the five number summary
  - We can also include outliers if any
  - Can be used to understand the distribution of the data
- Comparing boxplots





Which distribution has the largest IQR?
Which distribution has the largest maximum value?
Which distribution has a lower Q1?
Which distribution has a lower median?
Which distribution is symmetric?

Introduction to Statistics

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

Introduction to Statistics

- Strength of scatter plots
  - Correlation coefficient (*r*) – only calculate how strong **a linear relation** is
  - Calculate correlation coefficient
  - Values of *r* ranges between **-1 and 1**
- Examples

Introduction to Statistics

- Correlation matrix
- Covariance – a single number that measures the linear relationship between two variables
  - Correlation is the scaled version of the covariance
  - Covariance value can be **between** $-x \ \ and + x$ while correlation coefficient can be between -1 and 1

Introduction to Statistics

- Kendall Correlation for Ordinal Data

https://personal.utdallas.edu/~herve/Abdi-KendallCorrelation2007-pretty.pdf

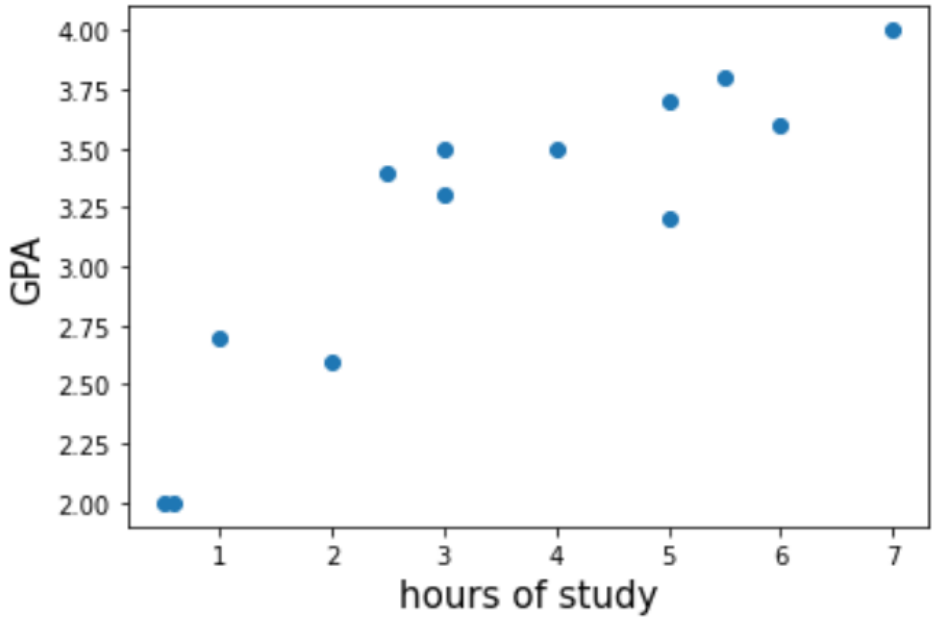
Introduction to Statistics

- Linear Regression
- What is extrapolation?
- What is residual?
- How to interpret the slope?

SJSU SAN JOSÉ STATE UNIVERSITY

Introduction to Statistics

- Practice –Describe the scatterplot. Remember that you must mention direction and shape.
  - Now obtain the regression equation



	hours_of_study	gpa
0	5.0	3.7
1	3.0	3.5
2	7.0	4.0
3	1.0	2.7
4	0.5	2.0
5	4.0	3.5
6	3.0	3.3
7	5.5	3.8
8	2.5	3.4
9	2.0	2.6
10	0.6	2.0
11	6.0	3.6
12	5.0	3.2

```
1 # The coefficients
2 print("Coefficients: \n", regr.coef_)
```

```
Coefficients:
 [0.26937586]
```

```
[30] 1 # The intercept
     2 print("Intercepts: \n", regr.intercept_)
```

```
Intercepts:
 2.2423960474085756
```

Introduction to Statistics

- Using the regression equation, what would you predict that the GPA would be for a student who studies **4.2 hours**?
- Using the regression equation, what would you predict that the GPA would be for a student who studies **8 hours**?
- Calculate the residual for student **index 5**
- What is the slope? (give units)
- Interpret the slope.

Probability

- Probability
  - A fair die is rolled: find the probability that an even number comes up.
- Sample space
  - Find the sample space when a coin is tossed (a) once (b) twice and (c) thrice

Probability

- Probability model
  - The probabilities should be legitimate between 0 and 1
  - Sum of all the probabilities in the sample space should equal 1

- Probability rules
  - The probability of an event is always in between 0 and 1
  - If a event A cannot occur then P(A) = 0
  - If a event is certain to occur then P(A) = 1

Probability

- Compound event
  - is formed by combining two or more events
  - can be represented using contingency table (frequency table that represents two variables)
- Practice: A collection of color pencils are listed below in the table.

1. What is the probability that it is green?
2. What is the probability that it does not have an eraser?
3. What is the probability that it is green and eraser?

	Red	Green	Blue	Total
Eraser	10	15	5	
No Eraser	3	5	2	
Total				

Probability

- What is the probability that it is green and has no eraser?
- What is the probability that it either red or has an eraser?
- What is the probability that it is green or has no eraser?

	Red	Green	Blue	Total
Eraser	10	15	5	
No Eraser	3	5	2	
Total				

Probability

- General Addition Rule: P(A or B) =
- Mutually exclusive events: no intersection or overlap of two events
- Addition rule for mutually exclusive events: P(A or B) =
- Practice: given- P(A) = 0.32, P(B) = 0.12, P(C) = 0.15
    1. If the P(A or B) is 0.44, are A and B mutually exclusive? Why or why not?
    2. if the P(B or C) is 0.2, are B and C mutually exclusive? Why or why not?
- Complements:

Probability

- Independence
  - Multiplication rule for two independent event: if A and B are independent → P(A and B) =

- Example: Two dice are rolled, and each produces a number between 1 and 6. Let A represent the event in which the number on the first die is even and B represent the event in which the number on the second die is 6.
  - Explain why events A and B are independent
  - Find P(A), P(B), and P(A and B)

Probability

- Sampling with or without replacement
- Practice:

Probability

- Practice: A jar contains 10 balls: 7 red and 3 blue:
  - First list the possible outcomes for drawing two balls with replacement then find the probability of obtaining each outcome
    - what is the probability of getting red both times?
    - what is the probability of getting blue exactly once?
    - what is the probability of getting blue at least once?

Probability

- Practice: Taz's tree cutting service sends a survey to its customers after each job. It was found that 26% of the customers from the past year were dissatisfied with the service given. Let D represent dissatisfied, and S represent satisfied.
  - Give a probability model for the satisfaction level when randomly selecting three customers. (hint: find the sample space for randomly selecting three customers using S and D.)
  - What is the probability that when three customers are randomly selected, at least one is dissatisfied?
  - What is the probability that when four customers are randomly selected, at least one is dissatisfied?
- Practice: A fair coin is tossed five times. What is the probability that it comes up heads at least once?

Example taken from Dr. Linda Galloways' note

Probability

- A **random variable** is a numerical outcome of a probability experiment
  - Discrete random variable
  - Continuous random variable
- Probability distribution
  - Criteria for probability distribution is same as probability model
    - Probability should be [0,1] and sum of the probabilities equal 1
  - Example:

Probability

- Practice: