TEXAS A&M UNIVERSITY
Engineering

ECEN 758 Data Mining and Analysis: Lecture 3, Data and Attributes II

Joshua Peeples, Ph.D.
Assistant Professor
Department of Electrical and Computer Engineering

Announcements

- Assignment #1 will be released this Wednesday (08/28)
  - Due Friday (11:59 PM), 09/06
- Please reach out if you need assistance
  - Responsive to email between 8 AM and 8 PM (Weekdays)
  - Office hours: MW 4 – 5 PM, WEB 212E; T 4 – 5 PM (virtual, Section 700 priority)
- Additional resources
  - https://dataminingbook.info/resources/
  - Josh Stamer's StatQuest

Last Lecture

- Numerical attributes
  - Analysis, Statistical Measures, Normalization

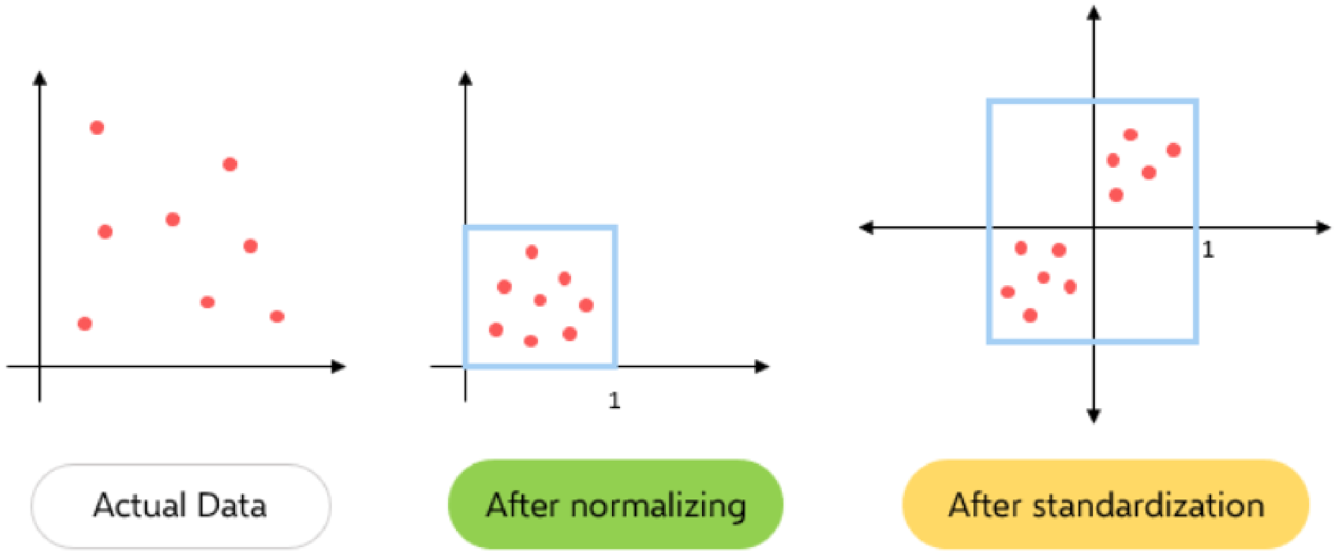


Image from: Someka, How to Normalize Data in Excel.

Today

TEXAS A&M UNIVERSITY Engineering

- Data and attributes
  - **Numerical**
    - Normal distribution
  - **Categorical**
- Reading: ZM Chapters 2 and 3

Review of Last Lecture

Data Representations

- Numeric measurements, observations, settings, counts, time intervals, etc. (binary, integer, fixed-point, floating point)
- Text (characters, words, strings, documents)
- Signals (continuous numeric values)
- Time Series (sequence of discrete-time data points often from sensors, communication signals)
- Image and Video (pixel data, series of image data, voxel data, point-clouds)

Data Types We Will Use

- Data used in Data Mining is generally of two types: Numeric Data and Categorical Data

- Numeric – quantitative, measurable; values are numbers. e.g. 0, 42, 3.1415, 1.602x10^-19

- Categorical – qualitative, recognizable; values are restricted to the possible values in a category and can be represented by a text value or a number. e.g., Tuesday, Medium Rare, Hawaii

Numerical Attributes

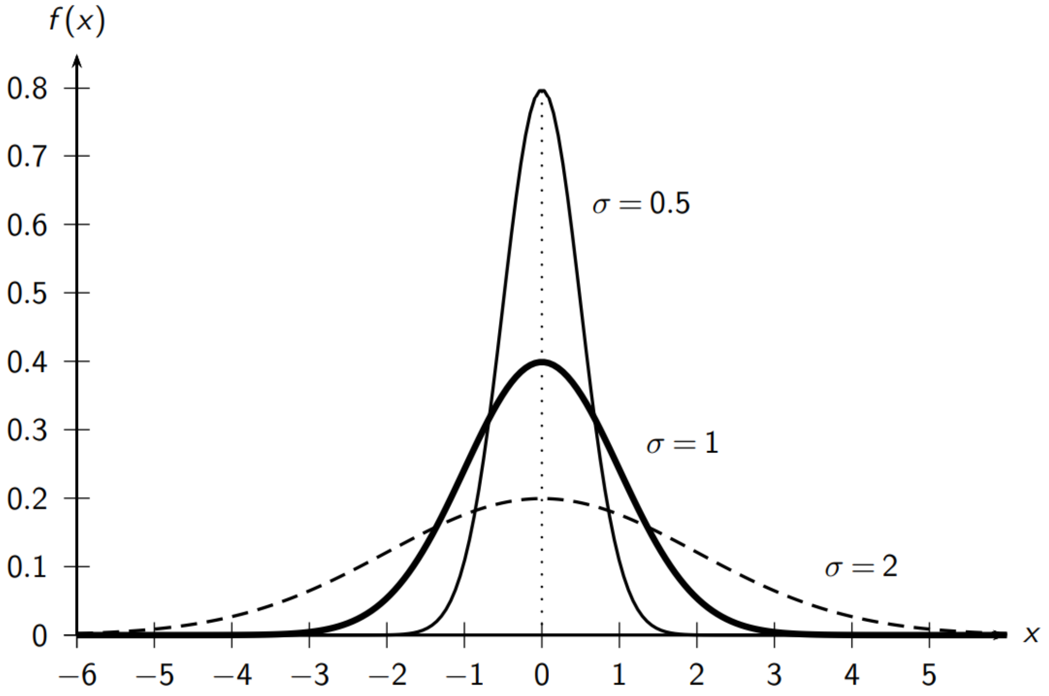
TEXAS A&M UNIVERSITY Engineering

- Univariate
- Bivariate
- Multivariate
- Measures of central tendency
  - Mean, Median, Mode
- Measures of dispersion
  - Range, Interquartile Range, Variance, Standard Deviation
- Normalization

Univariate Normal Distribution

Univariate Normal Distribution

- Two parameters, mean (μ) and variance (σ^2)
- Probability density decreases exponentially as a function of the distance from mean
- Maximum value when x = μ

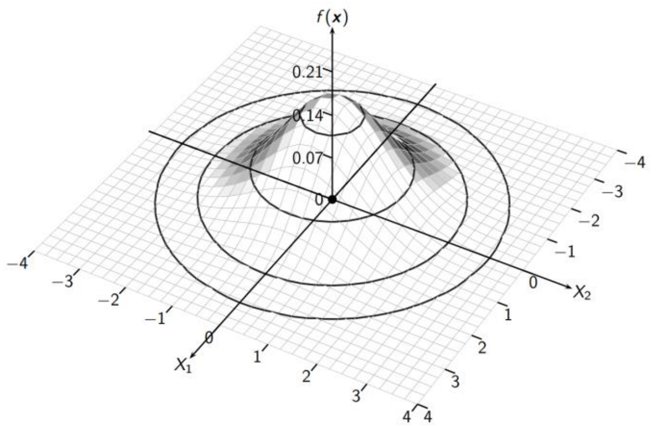


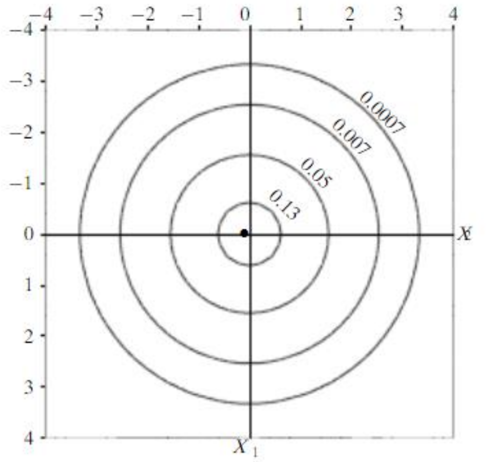
$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

Multivariate Normal Distribution

Multivariate Normal Distribution

- Parameters: mean vector (**μ**) and covariance matrix (Σ)
- |Σ| determinant of covariance matrix
- Numerator in exponential referred to as **Mahalanobis distance**
- "Standard multivariate normal distribution"
  - Zero mean vector and identity covariance





$$f(\boldsymbol{x}|\boldsymbol{\mu},\Sigma) = \frac{1}{(\sqrt{2\pi})^d \sqrt{|\Sigma|}} \exp\left\{-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}{2}\right\}$$

Geometry of Multivariate Normal

- Mean vector translates center of distribution
- Covariance matrix scales and rotates
- Can use Eigendecomposition to express covariance matrix

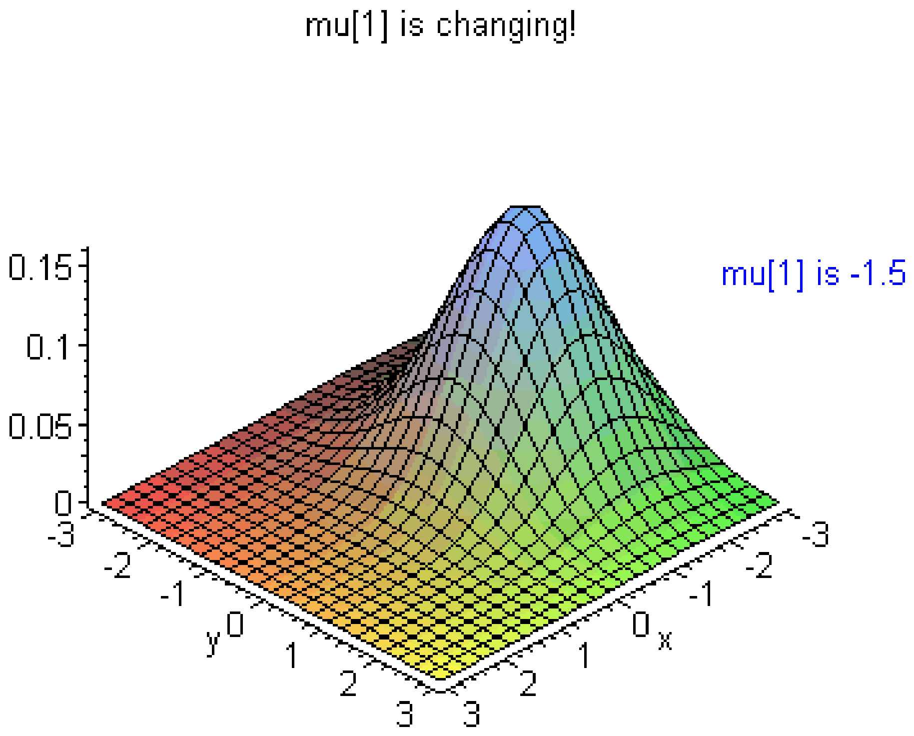
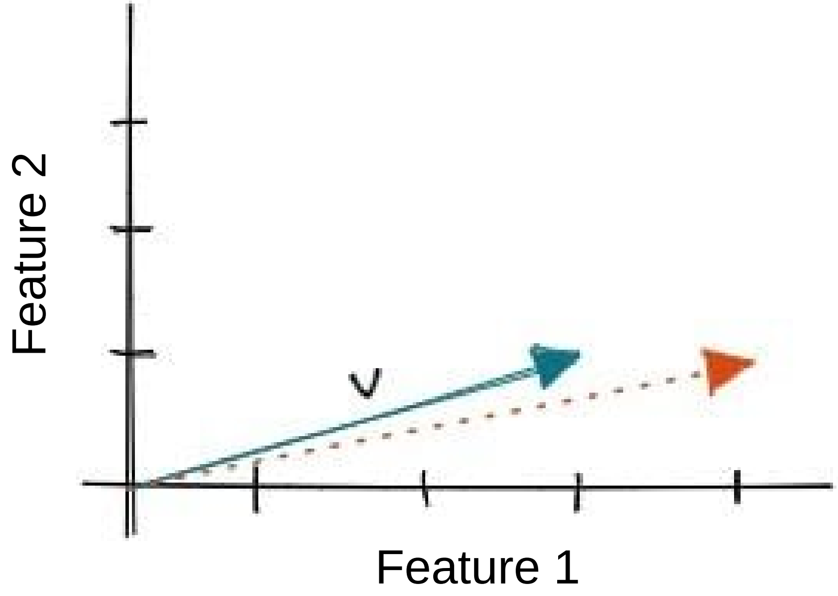


Image from: Statistics How To: Bivariate Normal Distribution

Eigenvectors and Eigenvalues

- Take a vector and apply linear transformation
- Identify vector(s) whose direction will not be changed after transformation
- Only magnitude will be scaled up or down



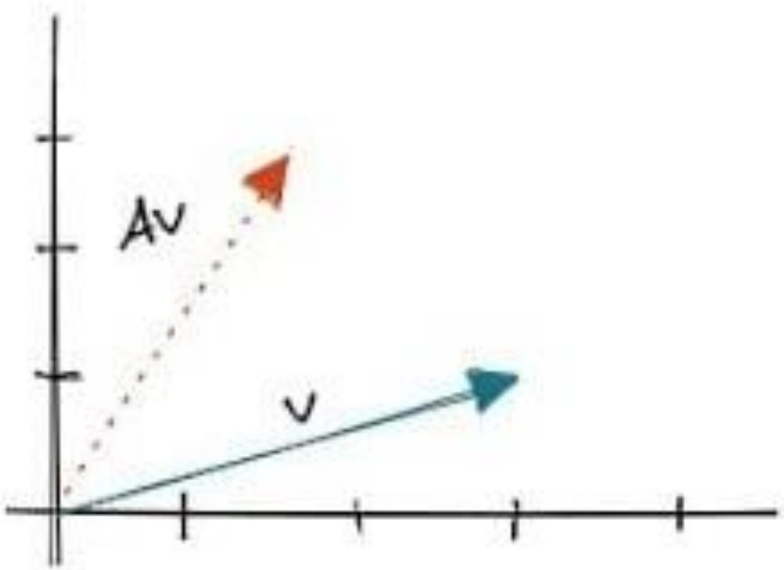
$$Av = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \boxed{\begin{bmatrix} 3 \\ 1 \end{bmatrix}} = \boxed{\begin{bmatrix} 4 \\ 1 \end{bmatrix}}$$

Image modified from: M. Lanhenke, Eigendecomposition Explained

Eigenvectors and Eigenvalues

Not an eigenvector

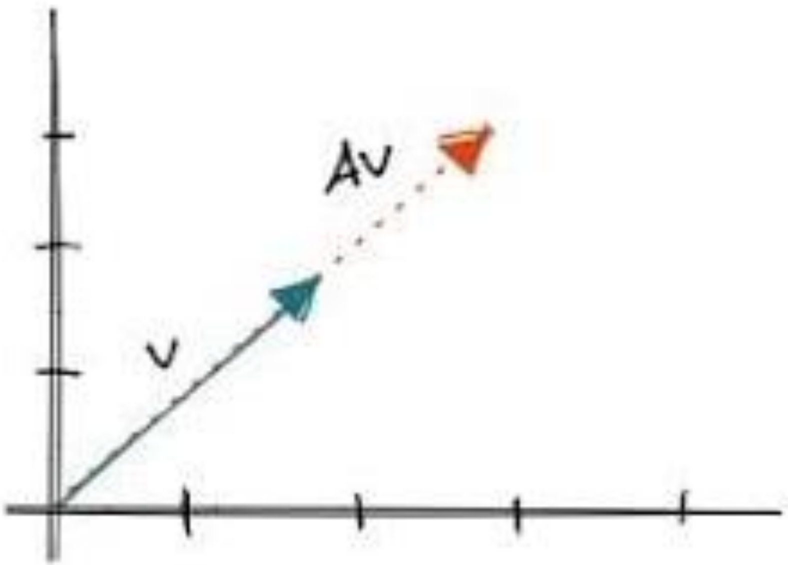
Feature 2



Feature 1

Eigenvector

Feature 2

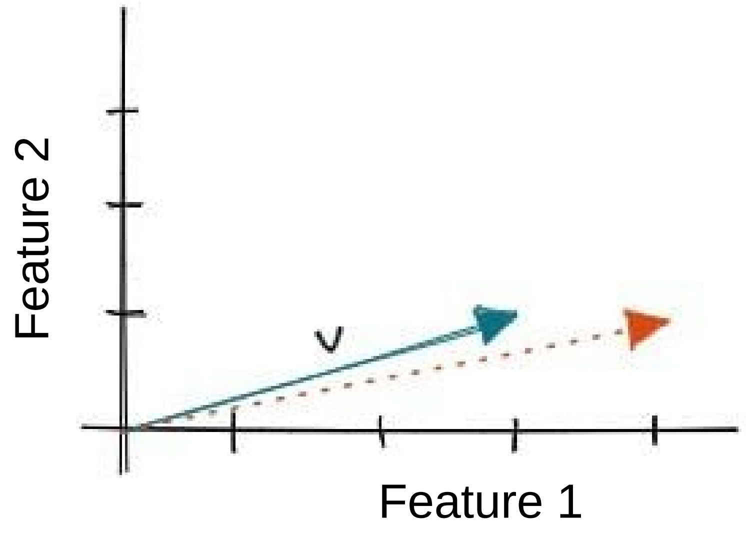


Feature 1

Image modified from: M. Lanhenke, Eigendecomposition Explained

Eigenvectors and Eigenvalues

- Matrix multiplication has same effect as scaler
- Matrix (A) is composed of eigenvectors
- Scaler values are called eigenvalues (λ)
- Eigendecomposition equation sets determinant of A minus λ*I equal to 0
  - “Area” = 0 (2D case)



$$Av = \lambda v$$

$$(A - \lambda I)v = 0$$

$$det(A - \lambda I) = 0$$

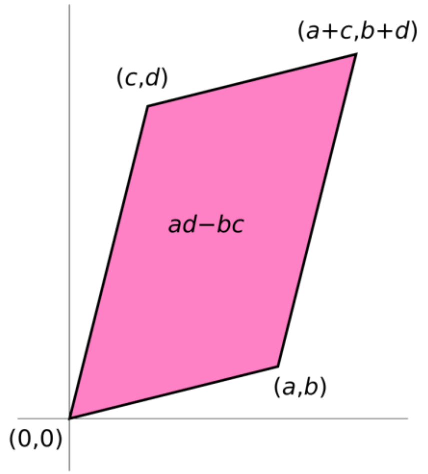


Image modified from: M. Lanhenke, Eigendecomposition Explained; Image of parallelogram: J. Niesen

Eigendecomposition

- Covariance matrix is positive semidefinite
- Diagonal matrix, Λ, is used to record eigenvalues
- Eigenvectors with “orthonormal” column vectors

$$\mathbf{U} = \begin{pmatrix} | & | & & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_d \\ | & | & & | \end{pmatrix} \qquad \mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_d \end{pmatrix}$$

Normalized → $\mathbf{u}_i^T\mathbf{u}_i = 1 \quad \text{for all } i$

Orthogonal → $\mathbf{u}_i^T\mathbf{u}_j = 0 \quad \text{for all } i \neq j$

$$\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$$

Image modified from: M. Lanhenke, Eigendecomposition Explained

Iris Sepal Length and Sepal Width

- X_1: Sepal Length
- X_2: Sepal Width
- U: Eigenvectors
- Λ: Eigenvalues

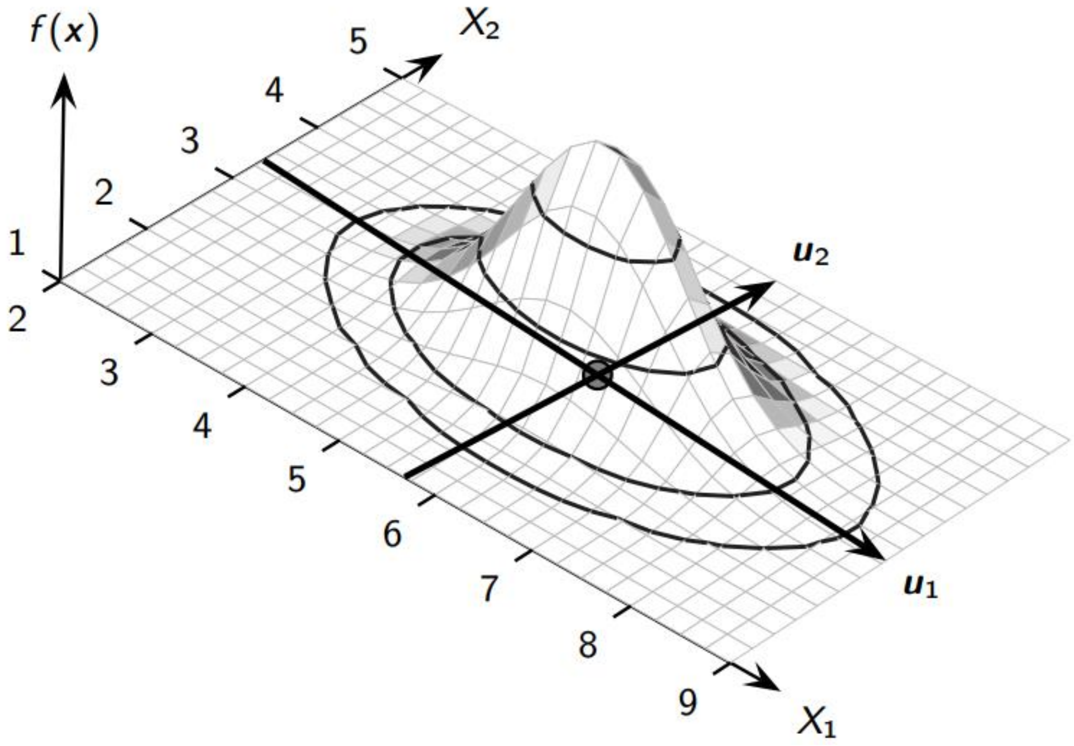
$$\hat{\mu} = \begin{pmatrix} 5.843 \\ 3.054 \end{pmatrix}$$

$$\hat{\Sigma} = \begin{pmatrix} 0.681 & -0.039 \\ -0.039 & 0.187 \end{pmatrix}$$

$$\hat{\Sigma} = U \Lambda U^T$$

$$U = \begin{pmatrix} -0.997 & -0.078 \\ 0.078 & -0.997 \end{pmatrix}$$

$$\Lambda = \begin{pmatrix} 0.684 & 0 \\ 0 & 0.184 \end{pmatrix}$$

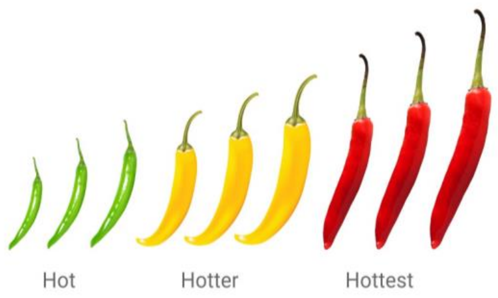


Categorical Data

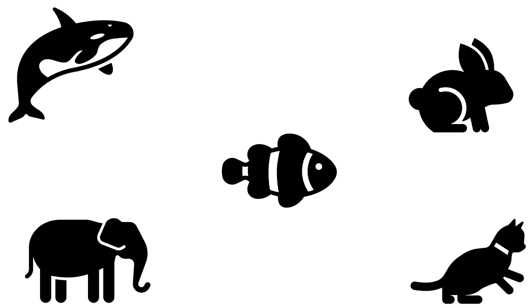
Types of Categorical Data

- Ordinal – values have an underlying, natural order. E.g., Monday, February, C-, 3rd-gear, Medium Rare, above average.

- Nominal – there is no underlying order in values. E.g., snake, brown, Fiat 500



Ordinal



Nominal

Image from: QuestionPro

TEXAS A&M UNIVERSITY Engineering

Univariate Analysis

Univariate Categorical Data

TEXAS A&M UNIVERSITY Engineering

- Focused on single attribute (e.g., feature)
- Data represented as matrix, **D**
- Each row is a sample and column is an attribute
- X is a random variable
- Domain of X is comprised of *m* symbolic values

$$\mathbf{D} = \begin{pmatrix} X \\ \hline x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

Bernoulli Variable

- Special case when m=2

$$X(v) = \begin{cases} 1 & \text{if } v = a_1 \\ 0 & \text{if } v = a_2 \end{cases}$$

$$dom(X) = \{0, 1\}$$

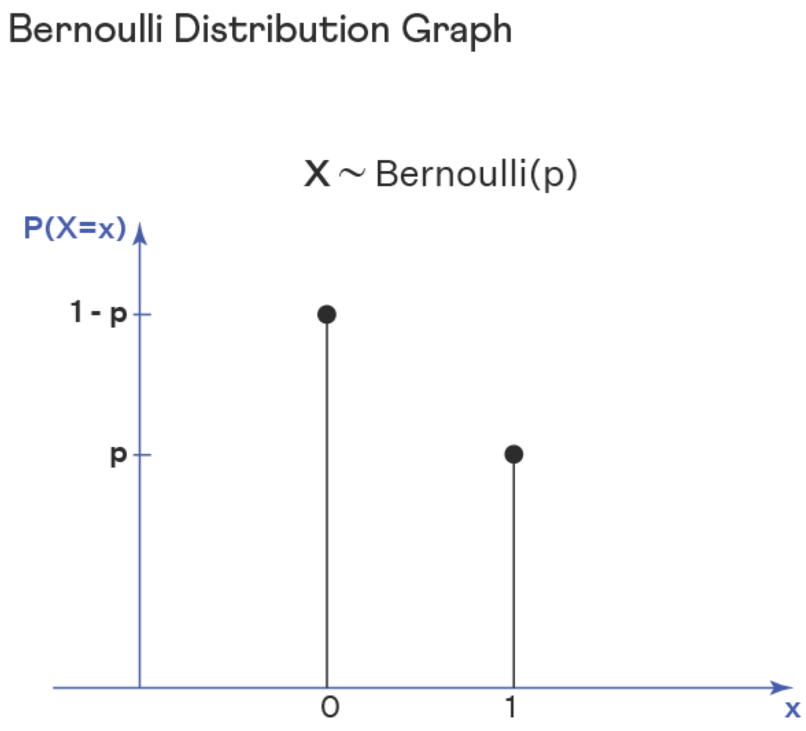


Image from: Cue Math, Bernoulli Distribution

Bernoulli Variable: PMF

$$P(X = x) = f(x) = p^x (1 - p)^{1-x}$$

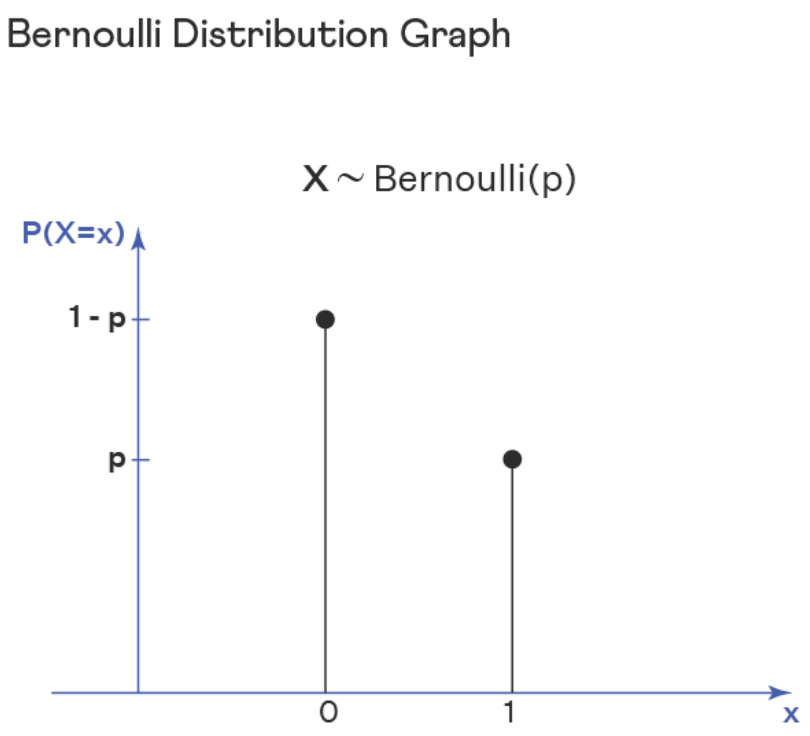


Image from: Cue Math, Bernoulli Distribution

Bernoulli Variable: Mean and Variance

TEXAS A&M UNIVERSITY Engineering

- Expected value

$$\mu = E[X] = 1 \cdot p + 0 \cdot (1 - p) = p$$

- Variance

$$\sigma^2 = var(X) = p(1 - p)$$

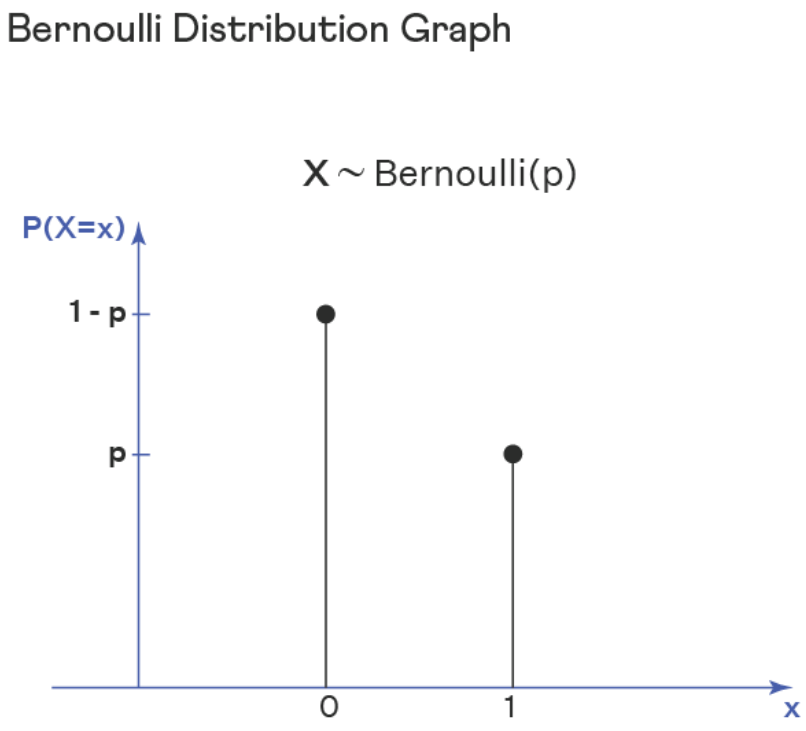


Image from: Cue Math, Bernoulli Distribution

Binomial Distribution

- Multiple trials
- PMF

$$f(N = n_1 | n, p) = \binom{n}{n_1} p^{n_1} (1-p)^{n-n_1}$$

- *N* is the sum of *n* independent Bernoulli random variables

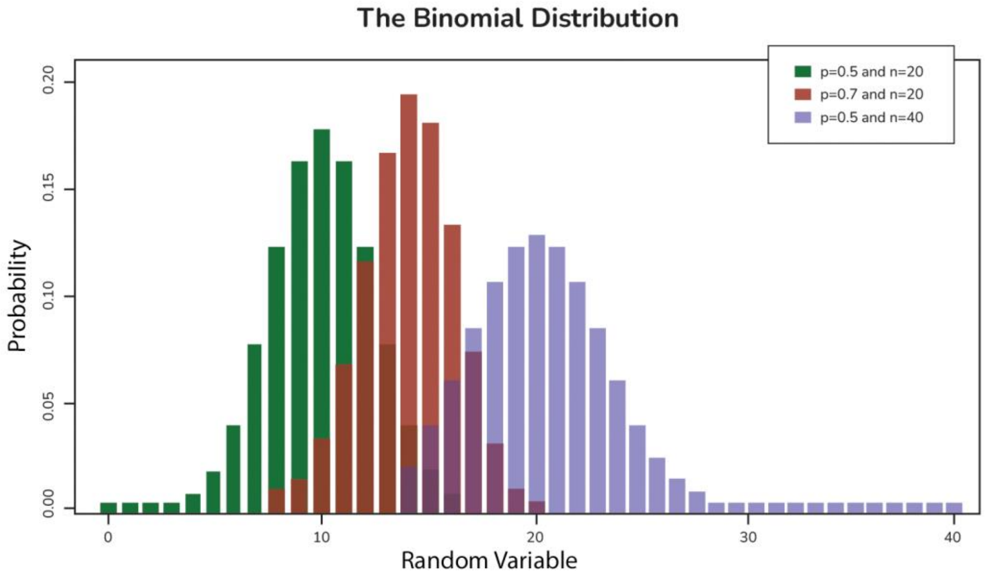


Image from: A. Jain, All about Binomial Distribution

Binomial Distribution

TEXAS A&M UNIVERSITY Engineering

- Mean

$$\mu_N = E[N] = E\left[\sum_{i=1}^{n} x_i\right] = \sum_{i=1}^{n} E[x_i] = \sum_{i=1}^{n} p = np$$

- Variance

$$\sigma_N^2 = var(N) = \sum_{i=1}^{n} var(x_i) = \sum_{i=1}^{n} p(1-p) = np(1-p)$$

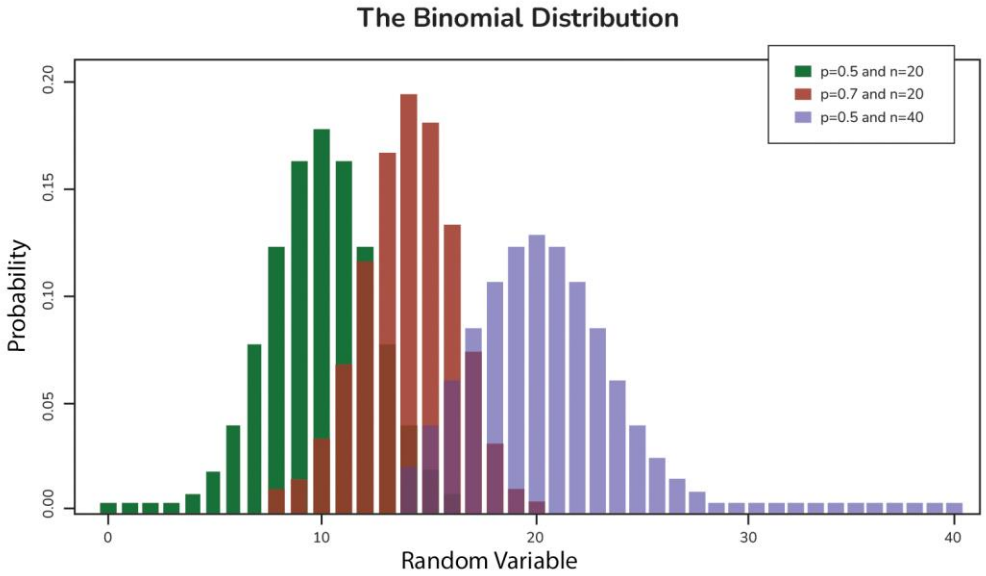


Image from: A. Jain, All about Binomial Distribution

Multivariate Analysis

Multivariate Bernoulli Variable

- Generalize beyond m = 2
- Assume only one of the symbolic values at any one time

$$\mathbf{X}(v) = \mathbf{e}_i \text{ if } v = a_i$$

$$P(\mathbf{X} = \mathbf{e}_i) = f(\mathbf{e}_i) = p_i = \prod_{j=1}^{m} p_j^{e_{ij}}$$

$$\sum_{i=1}^{m} p_i = 1.$$

Multivariate Bernoulli Variable: Mean

$$\mu = E[\boldsymbol{X}] = \sum_{i=1}^{m} \boldsymbol{e}_i f(\boldsymbol{e}_i) = \sum_{i=1}^{m} \boldsymbol{e}_i p_i = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} p_1 + \cdots + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} p_m = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_m \end{pmatrix} = \boldsymbol{p}$$

Mean

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i = \sum_{i=1}^{m} \frac{n_i}{n} \boldsymbol{e}_i = \begin{pmatrix} n_1/n \\ n_2/n \\ \vdots \\ n_m/n \end{pmatrix} = \begin{pmatrix} \hat{p}_1 \\ \hat{p}_2 \\ \vdots \\ \hat{p}_m \end{pmatrix} = \hat{\boldsymbol{p}}$$

Sample mean

Iris Sepal Length

Bins	Domain	Counts
$[4.3, 5.2]$	Very Short (a_1)	$n_1 = 45$
$(5.2, 6.1]$	Short (a_2)	$n_2 = 50$
$(6.1, 7.0]$	Long (a_3)	$n_3 = 43$
$(7.0, 7.9]$	Very Long (a_4)	$n_4 = 12$

We model `sepal length` as a multivariate Bernoulli variable $\boldsymbol{X}$

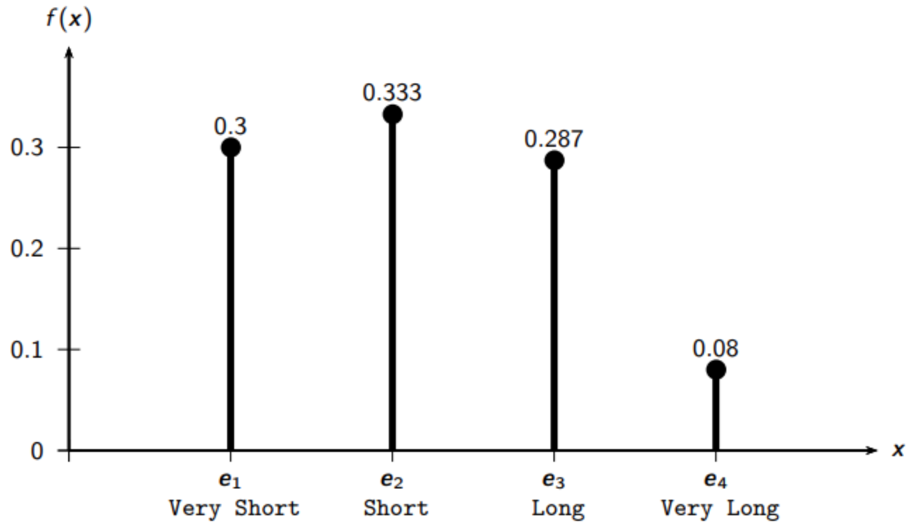
$$\boldsymbol{X}(v) = \begin{cases} \mathbf{e}_1 = (1,0,0,0) & \text{if } v = a_1 \\ \mathbf{e}_2 = (0,1,0,0) & \text{if } v = a_2 \\ \mathbf{e}_3 = (0,0,1,0) & \text{if } v = a_3 \\ \mathbf{e}_4 = (0,0,0,1) & \text{if } v = a_4 \end{cases}$$

For example, the symbolic point $x_1 = \text{Short} = a_2$ is represented as the vector $(0,1,0,0)^T = \mathbf{e}_2$.

Probability Mass Function

The total sample size is $n = 150$; the estimates $\hat{p}_i$ are:

$$\hat{p}_1 = 45/150 = 0.3$$
$$\hat{p}_2 = 50/150 = 0.333$$
$$\hat{p}_3 = 43/150 = 0.287$$
$$\hat{p}_4 = 12/150 = 0.08$$



Multivariate Bernoulli Variable: Covariance

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1m} \\ \sigma_{12} & \sigma_2^2 & \dots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \dots & \sigma_m^2 \end{pmatrix} = \begin{pmatrix} p_1(1-p_1) & -p_1 p_2 & \cdots & -p_1 p_m \\ -p_1 p_2 & p_2(1-p_2) & \cdots & -p_2 p_m \\ \vdots & \vdots & \ddots & \vdots \\ -p_1 p_m & -p_2 p_m & \cdots & p_m(1-p_m) \end{pmatrix}$$

$$\Sigma = diag(\boldsymbol{p}) - \boldsymbol{p} \cdot \boldsymbol{p}^T \text{ where } \boldsymbol{\mu} = \boldsymbol{p} = (p_1, \cdots, p_m)^T$$

Transformed Dataset

- Can encode and center data

	X
x_1	Short
x_2	Short
x_3	Long
x_4	Short
x_5	Long

	A_1	A_2
$\mathbf{x}_1$	0	1
$\mathbf{x}_2$	0	1
$\mathbf{x}_3$	1	0
$\mathbf{x}_4$	0	1
$\mathbf{x}_5$	1	0

	Z_1	Z_2
$\mathbf{z}_1$	-0.4	0.4
$\mathbf{z}_2$	-0.4	0.4
$\mathbf{z}_3$	0.6	-0.6
$\mathbf{z}_4$	-0.4	0.4
$\mathbf{z}_5$	0.6	-0.6

X is the multivariate Bernoulli variable

$$\mathbf{X}(v) = \begin{cases} \mathbf{e}_1 = (1,0)^T & \text{if } v = \texttt{Long}(a_1) \\ \mathbf{e}_2 = (0,1)^T & \text{if } v = \texttt{Short}(a_2) \end{cases}$$

The sample mean and covariance matrix are

$$\hat{\boldsymbol{\mu}} = \hat{\mathbf{p}} = (2/5, 3/5)^T = (0.4, 0.6)^T \qquad \widehat{\Sigma} = diag(\hat{\mathbf{p}}) - \hat{\mathbf{p}}\hat{\mathbf{p}}^T = \begin{pmatrix} 0.24 & -0.24 \\ -0.24 & 0.24 \end{pmatrix}$$

Multinomial Distribution

- PMF

$$f\left(\boldsymbol{N}=(n_1,n_2,\ldots,n_m)\mid \boldsymbol{p}\right)=\binom{n}{n_1\, n_2 \ldots n_m}\prod_{i=1}^{m} p_i^{n_i}$$

- Mean

$$\mu_{\boldsymbol{N}}=E[\boldsymbol{N}]=nE[\boldsymbol{X}]=n\cdot\boldsymbol{\mu}=n\cdot\boldsymbol{p}=\begin{pmatrix} np_1 \\ \vdots \\ np_m \end{pmatrix}$$

- Covariance

$$\Sigma_{\boldsymbol{N}}=n\cdot\left(diag(\boldsymbol{p})-\boldsymbol{p}\boldsymbol{p}^T\right)$$

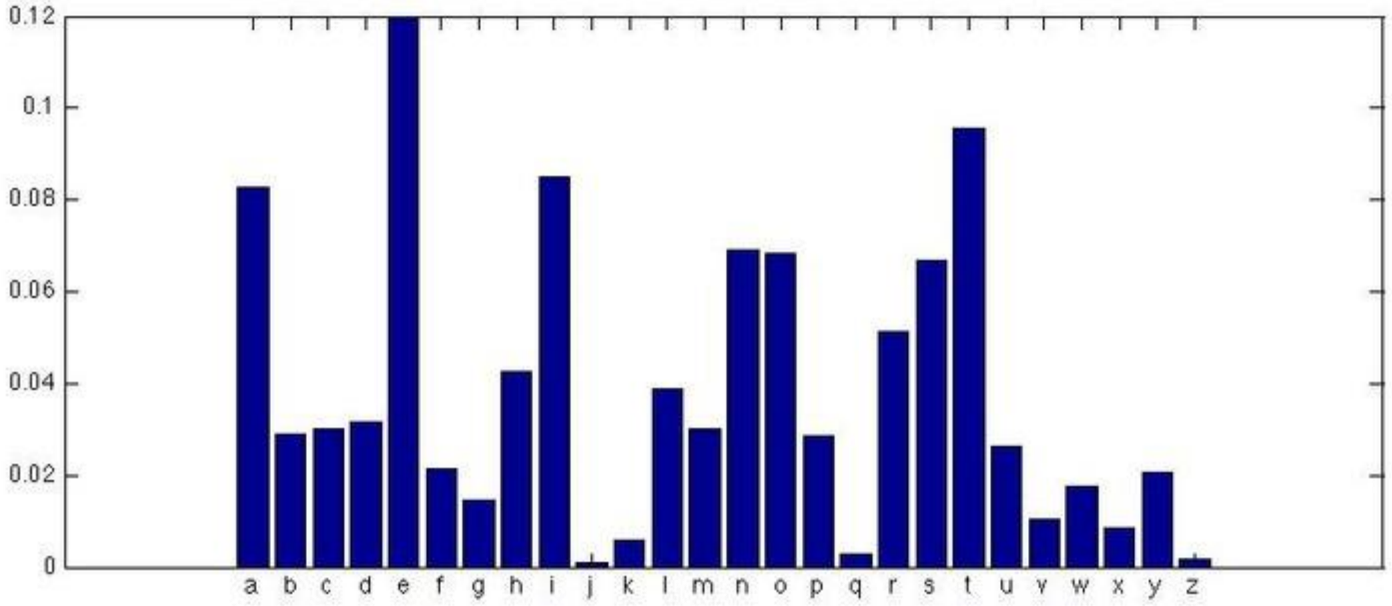


Image from: P. Hall, Introduction to Machine Learning for Computer Graphics

Bivariate Analysis

Bivariate Analysis

TEXAS A&M UNIVERSITY Engineering

- Consider two categorical attributes, X_1 and X_2
- Can model as joint distribution

$$\mathbf{X}\left((v_1, v_2)^T\right) = \begin{pmatrix} X_1(v_1) \\ X_2(v_2) \end{pmatrix} = \begin{pmatrix} \mathbf{e}_{1i} \\ \mathbf{e}_{2j} \end{pmatrix}$$

$$\mathbf{P}_{12} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1m_2} \\ p_{21} & p_{22} & \cdots & p_{2m_2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m_1 1} & p_{m_1 2} & \cdots & p_{m_1 m_2} \end{pmatrix}$$

Bivariate Example

X_1:sepal length

Bins	Domain	Counts
$[4.3, 5.2]$	Very Short (a_1)	$n_1 = 45$
$(5.2, 6.1]$	Short (a_2)	$n_2 = 50$
$(6.1, 7.0]$	Long (a_3)	$n_3 = 43$
$(7.0, 7.9]$	Very Long (a_4)	$n_4 = 12$

X_2:sepal width

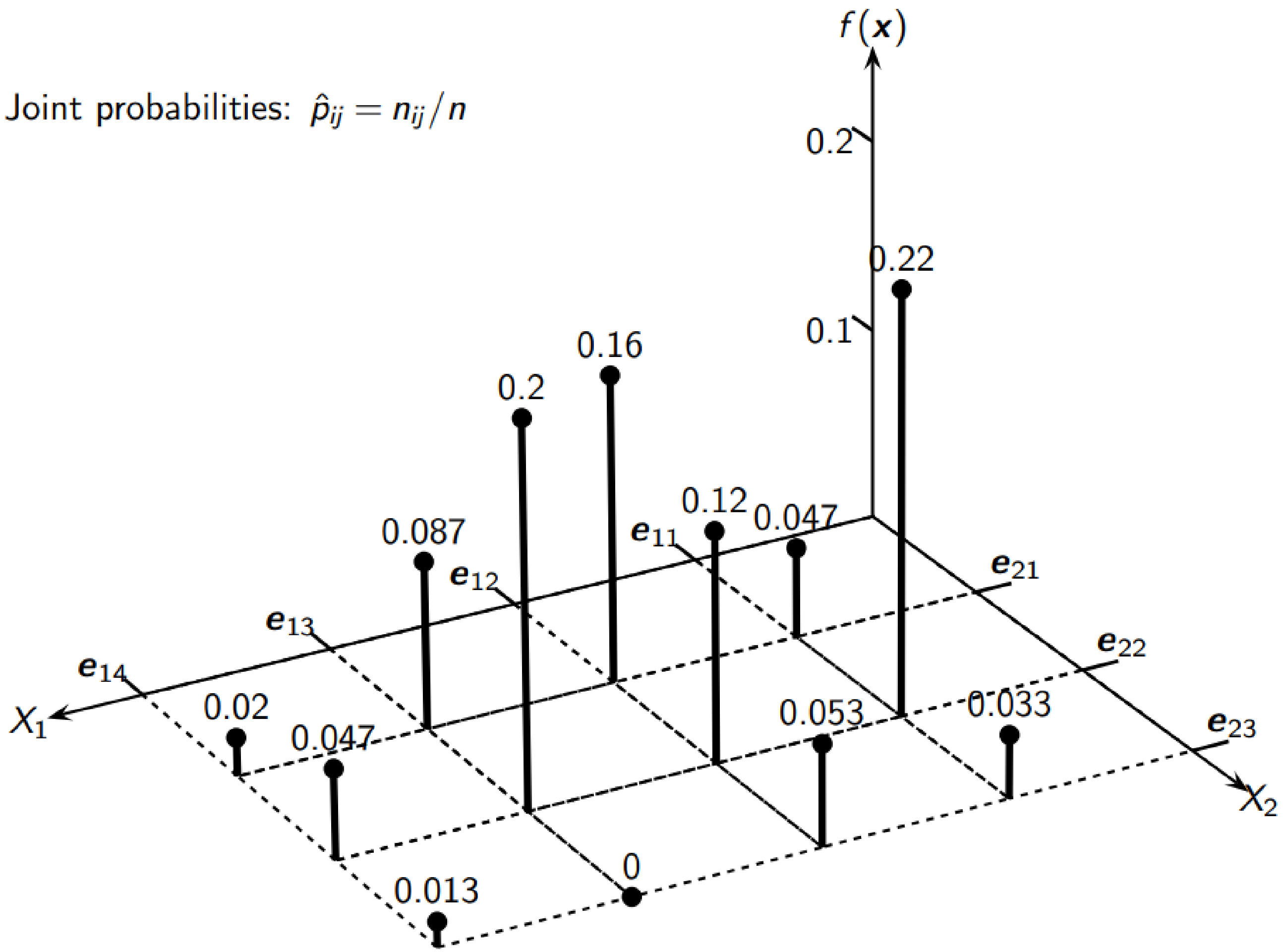
Bins	Domain	Counts
$[2.0, 2.8]$	Short (a_1)	47
$(2.8, 3.6]$	Medium (a_2)	88
$(3.6, 4.4]$	Long (a_3)	15

Observed Counts (n_{ij})

		X_2		
		Short ($\mathbf{e}_{21}$)	Medium ($\mathbf{e}_{22}$)	Long ($\mathbf{e}_{23}$)
X_1	Very Short ($\mathbf{e}_{11}$)	7	33	5
	Short ($\mathbf{e}_{12}$)	24	18	8
	Long ($\mathbf{e}_{13}$)	13	30	0
	Very Long ($\mathbf{e}_{14}$)	3	7	2

Bivariate PMF

Joint probabilities: $\hat{p}_{ij} = n_{ij}/n$



Contingency Analysis

- Observed counts for each attribute and symbolic values
- Multinomial distribution

$$N_{12} = n \cdot \widehat{P}_{12} = \begin{pmatrix} n_{11} & n_{12} & \cdots & n_{1m_2} \\ n_{21} & n_{22} & \cdots & n_{2m_2} \\ \vdots & \vdots & \ddots & \vdots \\ n_{m_1 1} & n_{m_1 2} & \cdots & n_{m_1 m_2} \end{pmatrix}$$

Contingency Table Example

Sepal length (X_1)	Sepal width (X_2)				
		Short a_{21}	Medium a_{22}	Long a_{23}	Row Counts
	Very Short (a_{11})	7	33	5	$n_1^1 = 45$
	Short (a_{12})	24	18	8	$n_2^1 = 50$
	Long (a_{13})	13	30	0	$n_3^1 = 43$
	Very Long (a_{14})	3	7	2	$n_4^1 = 12$
	Column Counts	$n_1^2 = 47$	$n_2^2 = 88$	$n_3^2 = 15$	$n = 150$

Independence Test

Chi-Squared Test

- Assume two attributes are independent
- Chi-squared quantifies difference between observed and expected counts

$$\hat{p}_{ij} = \hat{p}_i^1 \cdot \hat{p}_j^2$$

$$e_{ij} = n \cdot \hat{p}_{ij} = n \cdot \hat{p}_i^1 \cdot \hat{p}_j^2 = n \cdot \frac{n_i^1}{n} \cdot \frac{n_j^2}{n} = \frac{n_i^1 n_j^2}{n}$$

$$\chi^2 = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

Chi-squared Density Function

TEXAS A&M UNIVERSITY Engineering

- Sampling distribution for statistic follows density function
- *q* is degrees of freedom

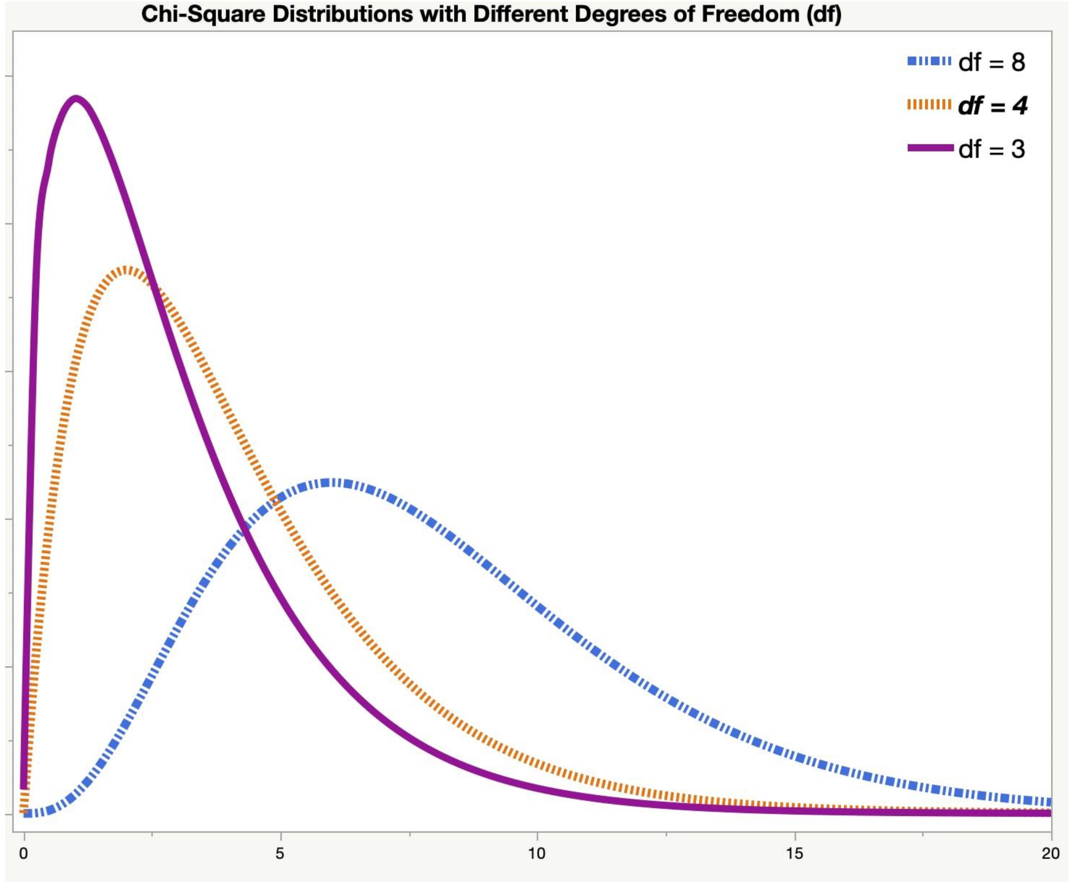


Image from: JMP

Chi-Squared Test Example

	Expected Counts	X_2		
		Short (a_{21})	Medium (a_{22})	Short (a_{23})
X_1	Very Short (a_{11})	14.1	26.4	4.5
	Short (a_{12})	15.67	29.33	5.0
	Long (a_{13})	13.47	25.23	4.3
	Very Long (a_{14})	3.76	7.04	1.2

	Observed Counts	X_2		
		Short (a_{21})	Medium (a_{22})	Long (a_{23})
	Very Short (a_{11})	7	33	5
	Short (a_{12})	24	18	8
	Long (a_{13})	13	30	0
	Very Long (a_{14})	3	7	2

The chi-squared statistic value is $\chi^2 = 21.8$.

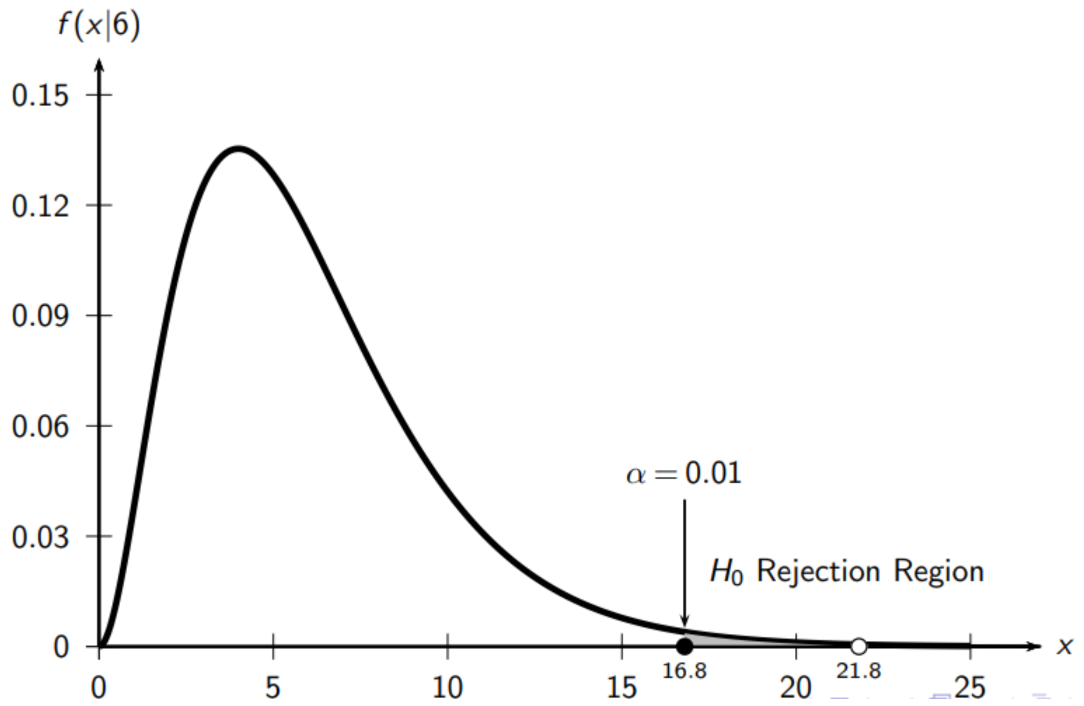
The number of degrees of freedom are

$$q = (m_1 - 1) \cdot (m_2 - 1) = 3 \cdot 2 = 6$$

Chi-Squared Distribution

TEXAS A&M UNIVERSITY Engineering

- p-value is probability of obtaining value at least as extreme as observed value
- Null hypothesis: independent
- Rejected if p-value less than alpha (e.g., 0.01)
- p-value of 21.8 = 0.0013



Distance and Angle Measures

Distance and Angle

- Can compute distance or angle between data points
- Rely on matching/mismatching of values across attributes
- s is number of matches
- Compute number of mismatches as d - s

$$\mathbf{x}_i = \begin{pmatrix} \mathbf{e}_{1i_1} \\ \vdots \\ \mathbf{e}_{d\,i_d} \end{pmatrix} \qquad \mathbf{x}_j = \begin{pmatrix} \mathbf{e}_{1j_1} \\ \vdots \\ \mathbf{e}_{d\,j_d} \end{pmatrix}$$

$$s = \mathbf{x}_i^T \mathbf{x}_j = \sum_{k=1}^{d} (\mathbf{e}_{ki_k})^T \mathbf{e}_{kj_k}$$

Common Distance Measures

The *Euclidean distance* between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is given as

$$\delta(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\| = \sqrt{\boldsymbol{x}_i^T \boldsymbol{x}_i - 2\boldsymbol{x}_i \boldsymbol{x}_j + \boldsymbol{x}_j^T \boldsymbol{x}_j} = \sqrt{2(d-s)}$$

The *Hamming distance* is given as

$$\delta_H(\boldsymbol{x}_i, \boldsymbol{x}_j) = d - s$$

Cosine Similarity: The cosine of the angle is given as

$$\cos\theta = \frac{\boldsymbol{x}_i^T \boldsymbol{x}_j}{\|\boldsymbol{x}_i\| \cdot \|\boldsymbol{x}_j\|} = \frac{s}{d}$$

The *Jaccard Coefficient* is given as

$$J(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{s}{2(d-s)+s} = \frac{s}{2d-s}$$

Discretization

- Converts numeric attributes into categorical attributes
- K is number of intervals
- Equal-width intervals partitions data evenly
- Equal-frequency intervals partitions data into equal number of data points

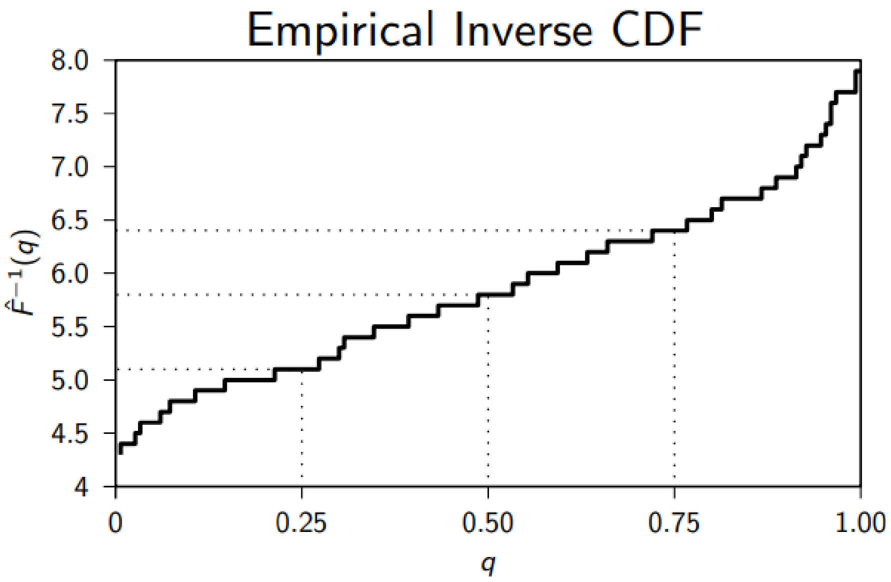
Equal-width:

$$w = \frac{x_{\max} - x_{\min}}{k}$$

Equal-frequency:

$$\hat{F}^{-1}(q) = \min\{x \mid P(X \leq x) \geq q\}$$

Equal-Frequency Discretization: Sepal Length



Quartile values:

$$\hat{F}^{-}1(0.25) = 5.1$$

$$\hat{F}^{-}1(0.5) = 5.8$$

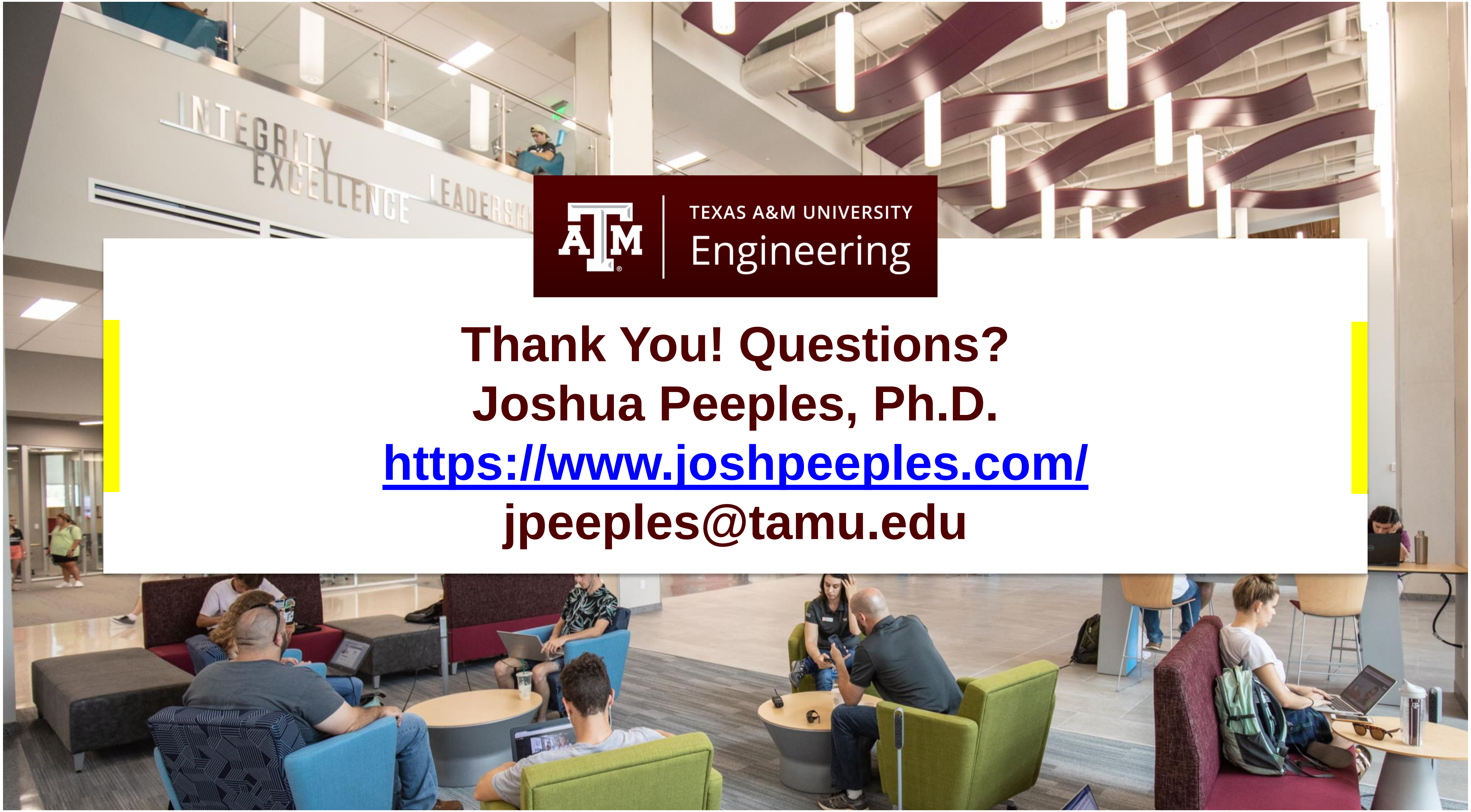
$$\hat{F}^{-}1(0.75) = 6.4$$

Range: $[4.3, 7.9]$

Bin	Width	Count
$[4.3, 5.1]$	0.8	$n_1 = 41$
$(5.1, 5.8]$	0.7	$n_2 = 39$
$(5.8, 6.4]$	0.6	$n_3 = 35$
$(6.4, 7.9]$	1.5	$n_4 = 35$

Next class

- Dimensionality reduction



TEXAS A&M UNIVERSITY Engineering

Thank You! Questions?

Joshua Peeples, Ph.D.

https://www.joshpeeples.com/

jpeeples@tamu.edu

TEXAS A&M UNIVERSITY
Engineering

Supplemental Slides

Additional Resources

TEXAS A&M UNIVERSITY Engineering

- StatQuest
  - Intuitive explanations of concepts covered in course
  - Probability Distributions
  - Normal Distribution
  - Binomial Distribution
- Eigendecomposition Explained