

Foundation of Data Science and Analytics

1. Introduction

Arun K. Timalsina, PhD

Setopati Projection: Balen Shah to be elected Kathmandu Mayor

All mayoral candidates apart from Shah, Singh and Sthapit set to lose deposit



Forecast was TWO Weeks before Result Announcement!

Setopati Projection: Balen Shah to be elected Kathmandu Mayor

All mayoral candidates apart from Shah, Singh and Sthapit set to

Suhang Nembang of CPN-UML emerges victorious in Ilam-2

Defeats Khadka of Nepali Congress by a margin of 5,830.



THE KATHMANDU POST

Published at : April 30, 2024

Updated at : April 30, 2024 08:12

१८ वैशाख २०८२, मंगलवार

April 30, 2024

इलाम-२ को निर्वाचनबारे सेतोपाटी विश्लेषण

सेतोपाटी टिमले पछिल्लो साता इलाम-२ का १०९६ मतदातासँग कुरा गरेको थियो

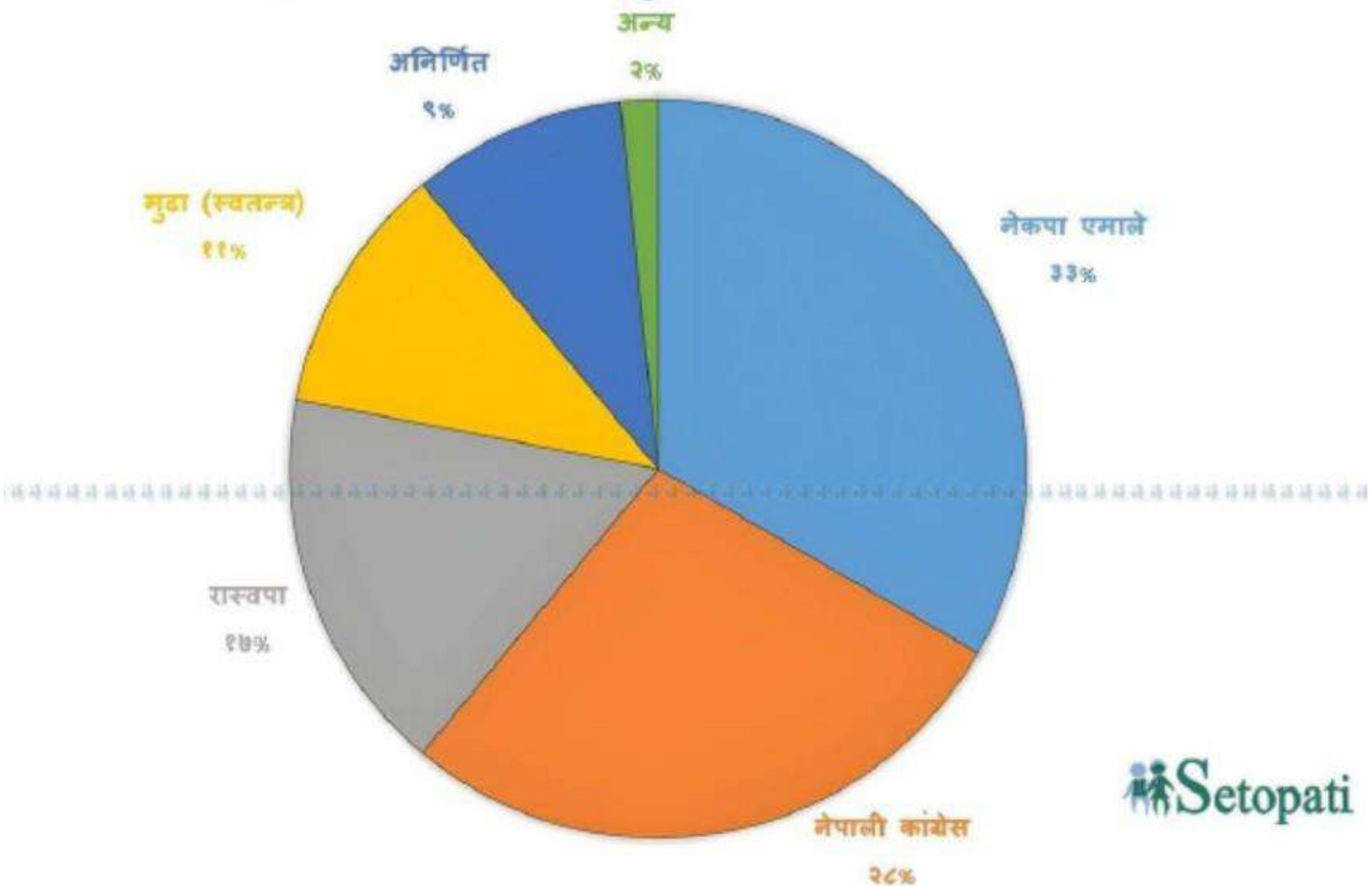
मनोज/प्रशन्न/राजु/सुदीप

इलाम, वैशाख ११

<https://www.setopati.com/exclusive/premium-story/327822>



इलाम- २ को चुनावी विश्लेषण



<https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did>

Feb 16, 2012 11:02am EST

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did



Kashmir Hill Former Staff

Tech

Welcome to The Not-So Private Parts where technology & privacy...

This article is more than 8 years old.



TARGET

Target has got you in its aim

Every time you go shopping, you share intimate details about your consumption

Flash Sale: Less than \$1/week

Sign

≡ Forbes

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

By Kashmir Hill, Former Staff. Welcome to The Not-So Private Parts where technology & privacy...

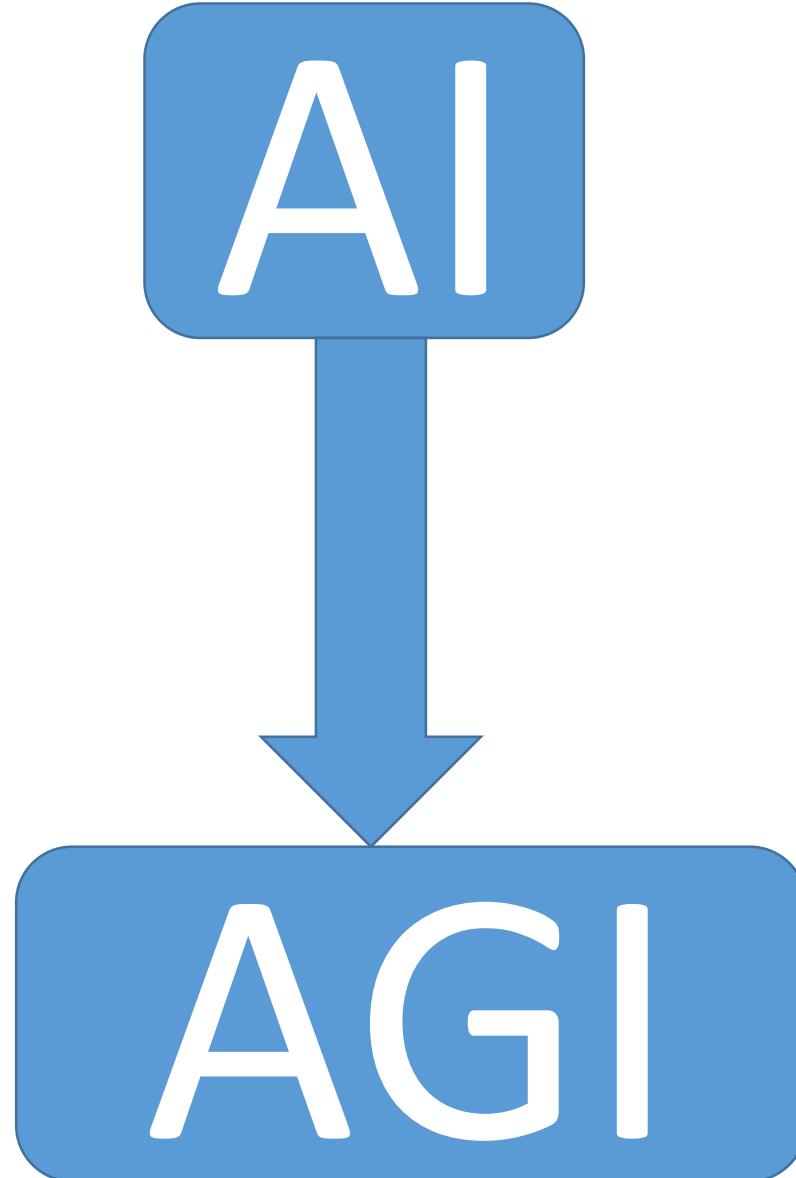
Feb 16, 2012, 11:02am EST

Plausible Effectiveness of Deep Learning

2012 Imagenet challenge:

Classify 1 million
images into 1000
classes.

	horseshoe crab 0.99%	African elephant 0.99%	mongoose 0.94%	Indian elephant 0.88%	dingo 0.87%	
Cliff dwelling L2 11.0% - Mah. 99.9%	cliff 0.07%	dam 0.00%	stone wall 0.00%	brick 0.00%	castle 0.00%	
Gondola L2 4.4% - Mah. 99.7%	shopping cart 1.07%	unicycle 0.84%	covered wagon 0.83%	garbage truck 0.79%	forklift 0.78%	
	dock 0.11%	canoe 0.03%	fishing rod 0.01%	bridge 0.01%	boathouse 0.01%	
Palm L2 6.4% - Mah. 98.1%	crane 0.87%	stupa 0.83%	roller coaster 0.79%	bell cote 0.78%	flagpole 0.75%	
	cabbage tree 0.81%	pine 0.30%	pandanus 0.14%	iron tree 0.07%	logwood 0.06%	



What the course FDSA is about?

Foundation of Data Science & Analytics

- Overall Summary of Data Science & Analytics
- Mathematics of Data Analysis
 - Basic Statistics , Regression, Matrix factorization
- Data Wrangling /Cleaning (EDA)
- Model and Evaluation specifics and setups
- OLTP/OLAP – NoSQL Specifics
- Related Research Trends

Course Contents

1. Introduction to Data Science (3 Hrs)
Data Science Hype, Why data science, Getting Past the Hype, The Current Landscape, Role of Data Scientist
2. Data Types and Data Science Processes (7 Hrs)
 - 2.1. Facets of data: Structured data, Unstructured data, Natural language, Machine-generated data, Graph-based or network data, Audio, image, and video, Streaming data
 - 2.2. Process Overview, Defining goals, Retrieving data, Data preparation, Exploratory Data Analysis, Data Wrangling & Cleaning, Data Integration and Transformation, Data Reduction, Data modeling and Result Presentation
3. Mathematical Foundation for Data Science (20 Hrs)
 - 3.1. Introduction and Descriptive Statistics : An overview of probability and statistics, Pictorial and tabular methods in descriptive statistics, Measures of central tendency, dispersion, and direction, Joint and conditional probabilities, Central limit theorem (4 Hrs)
 - 3.2. Random Variables and Probability Distributions: Random variables, Probability distributions for random variables, Expected values of discrete random variables and continuous distributions, The binomial probability distribution, Hypothesis testing using the binomial distribution, The Poisson probability distribution (4 Hrs)

Course Contents

3.3. Hypothesis Testing Procedures: Tests about the mean of a normal population, The t-test, Z-tests for differences between two populations means, The two-sample t-test, A confidence interval for the mean of a normal population (4 Hrs)

4. Regression and associated Models (8 Hrs)

4.1 Empirical Models, Simple Linear Regression, MLE and Least Square Estimator, Logistic Regression, Hypothesis tests in simple linear regression, t-tests and ANOVA, Confidence intervals, Residual Analysis, Coefficient of Determination, Correlation

4.2 Multiple Linear Regression, Matrix approach to Multiple Linear Regression, Hypothesis tests, Polynomial Regression Models, Categorical Regressors and Indicator variables, Selection of variables and Model building

4.3 Matrix Factorization, Probabilistic Matrix Factorization, Non-Negative MF, Applications (2 Hrs)

Course Contents

5. Modeling and validation processes for Machine Learning Techniques (8 Hrs)
 - 5.1. Supervised learning algorithms & Unsupervised learning algorithms.
 - 5.2. Modeling Process, Training model, Validating model, Cross Validation methods, Predicting new observations - Interpretation
 - 5.3. Measures for Model Performance and Evaluation: Classification accuracy, Confusion matrix, Sensitivity and specificity, Recall and precision, F-score, ROC curve, Clustering performance measures, other measures
6. Association and Other types of Analysis (12 Hrs)
 - 6.1. Market Basket Analysis using frequent itemset, Association rules generation from transactional dataset, Apriori and other algorithms, Correlation analysis
 - 6.2. Outlier Analysis, Trend analysis, Time series analysis, Social network analysis
7. Database and Datawarehousing (6 Hrs)
DBMS fundamentals, Relational Algebra and SQL, OLTP, Datawarehouse, Multidimensional data model, Data Cubes, NoSQL, OLAP Operations |
8. Ethics and Recent Trends (4 Hrs)
Data Science Ethics, Doing good data science, Owners of the data, Privacy aspects, Social impact, Getting informed consent, The Five Cs, Future Trends.

References

1. Introducing Data Science: Big Data. Machine Learning and More, Using Python Tools. Cielen D, Meysman AD, Ali M. Manning, 2016
2. An Introduction to Statistical Learning: with Applications in R, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Springer, 1st edition, 2013
3. Applied Statistics and Probability for Engineers, Doglas C. Montgomery, Goerge C Runger, Wiley, 2014
4. Ethics and Data Science, D J Patil, Hilary Mason, Mike Loukides, O' Reilly, 2018
5. Applied Data Science with Python and Jupyter: Galea A., Packt Publishing Ltd; 2018.
6. Adhikari A, DeNero J. Computational and Inferential Thinking: The Foundations of Data Science., 2017

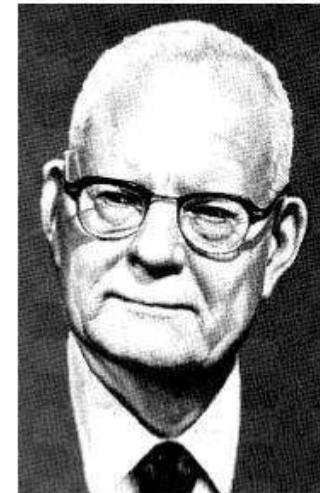
Data Analysis : Timeline

1935: “The Design of Experiments”

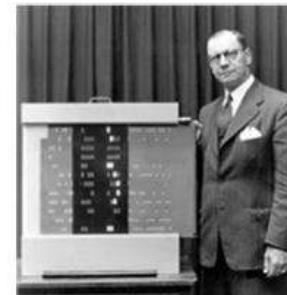
R.A. Fisher



1939: “Quality Control”



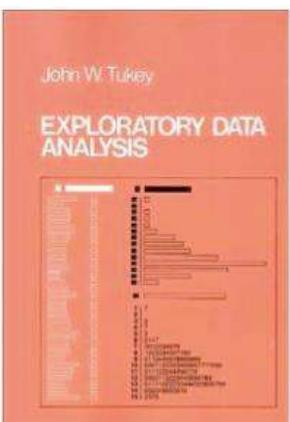
1958: “A Business Intelligence System”



Peter Luhn

W.E. Demming

1977: “Exploratory Data Analysis”



1989: “Business Intelligence”



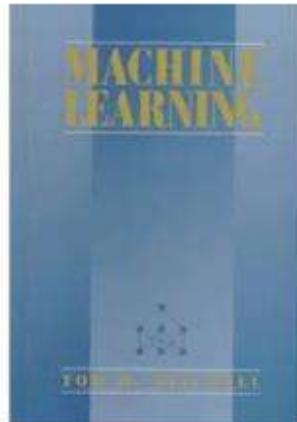
Howard
Dresner

Data Analysis : Timeline

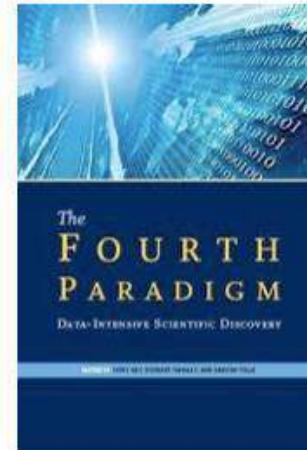
1996: Google



1997: "Machine Learning"



2007: "The Fourth Paradigm"



First 3 paradigms of science :
Empirical,
Theoretical and
Simulation.
4th Data Driven
Science

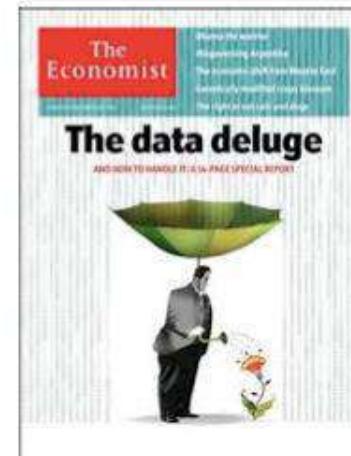
2009: "The Unreasonable Effectiveness of Data"



Peter Norvig :
Simple Model
+ Voluminous Data
→ Complex Model

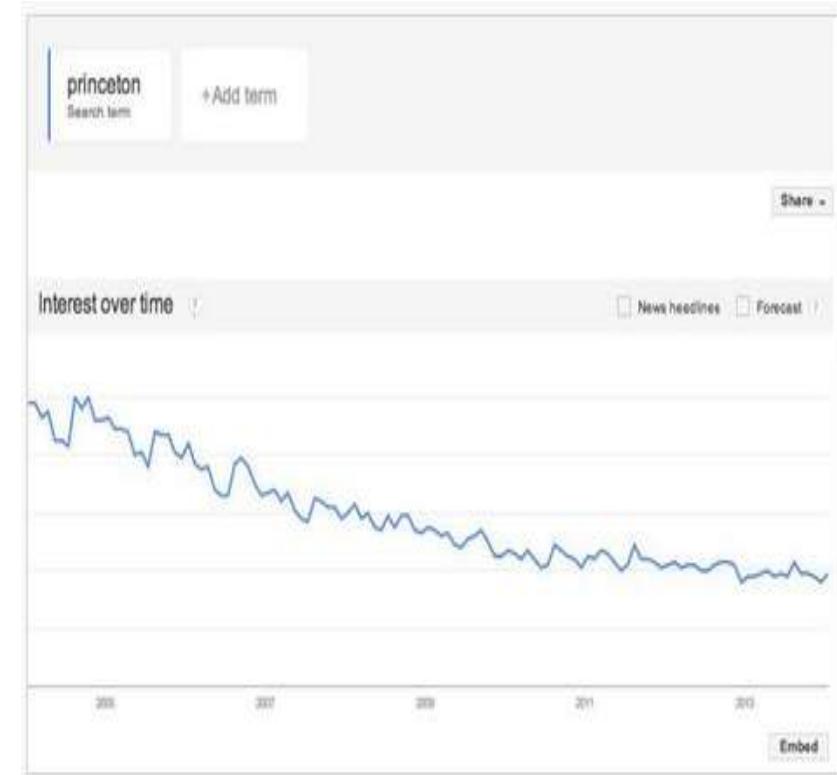
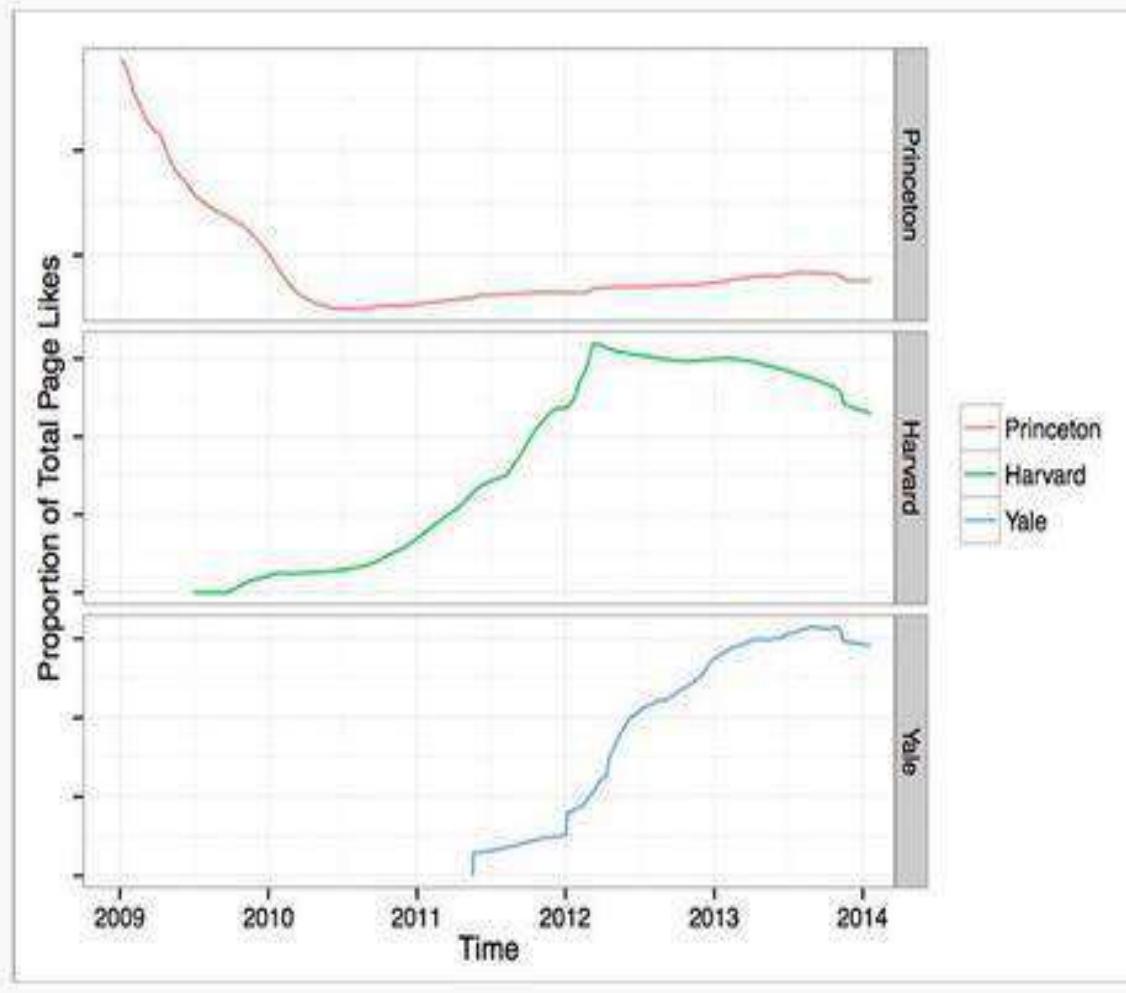
n Timalsina

2010: "The Data Deluge"



Data Makes Everything Clearer

In keeping with the scientific principle “correlation equals causation,” our research unequivocally demonstrated that Princeton may be in danger of disappearing entirely. Looking at page likes on Facebook, we find the following alarming trend:



and based on Princeton search trends:

“This trend suggests that Princeton will have only half its current enrollment by 2018, and by 2021 it will have no students at all,...

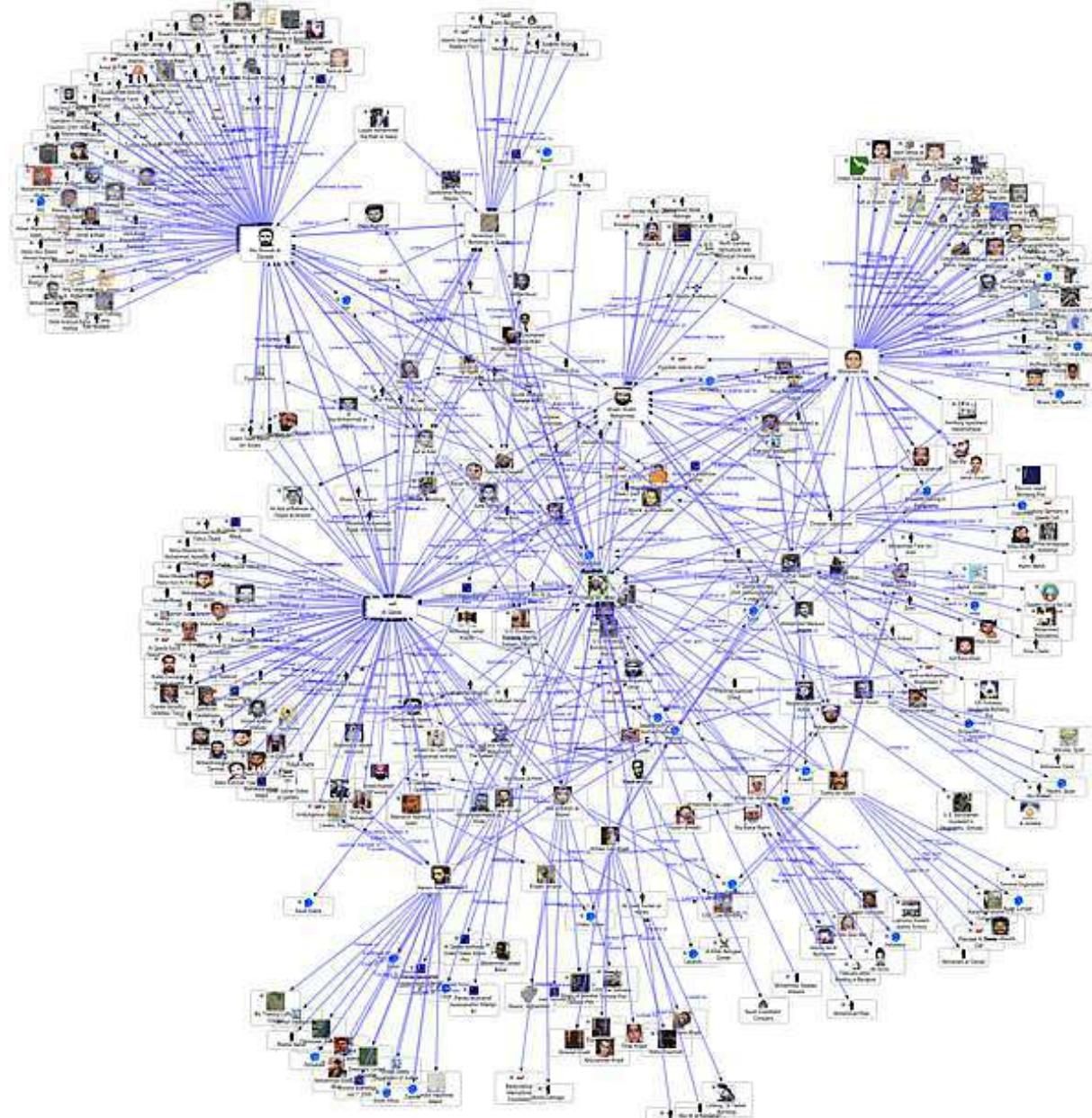
<http://techcrunch.com/2014/01/23/facebook-losing-users-princeton-losing-credibility/>

Graph Data

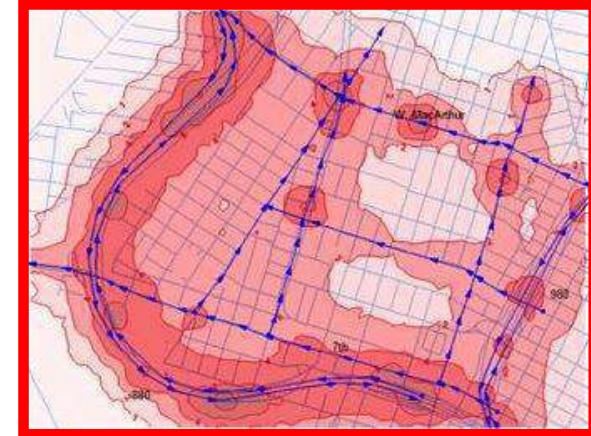
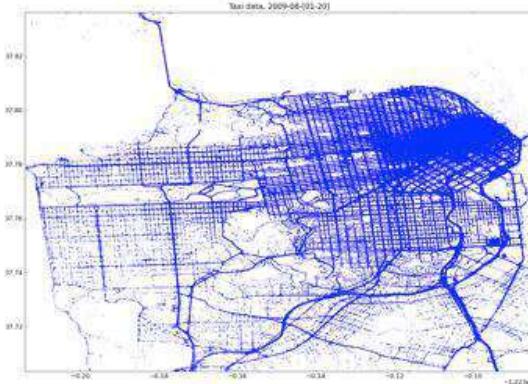
Lots of interesting data has a graph structure:

- Social networks
- Communication networks
- Computer Networks
- Road networks
- Citations
- Collaborations/Relationships
- ...

Some of these graphs can get quite large (e.g., Facebook* user graph)

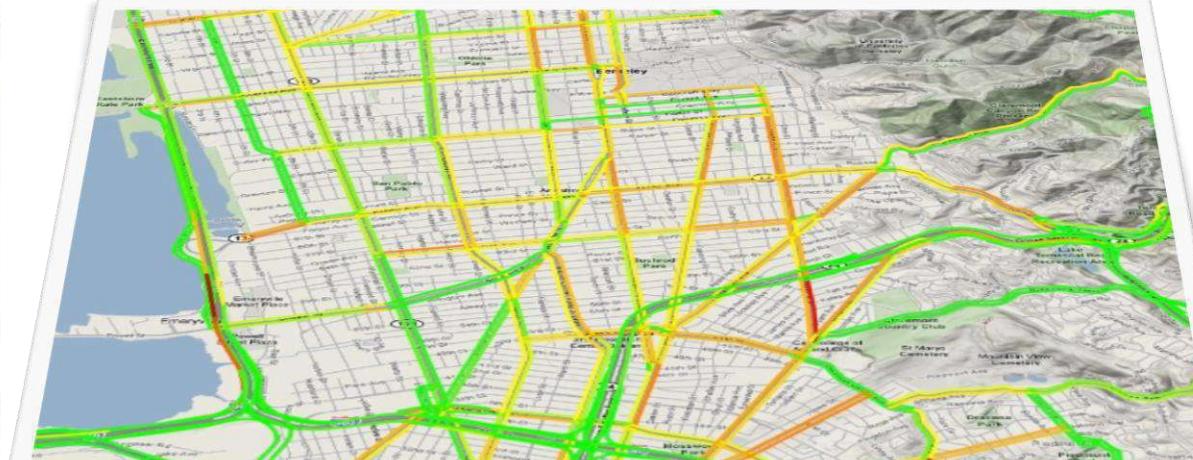


What *can not be done* with the data?

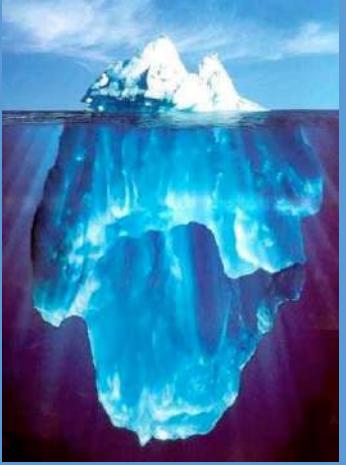


Crowdsourcing + physical modeling + sensing + data assimilation

to produce:



It's All Happening On-line



Every
Click
Ad impression
Billing event
Fast Forward, pause,...
Server request
Transaction
Network message
Fault

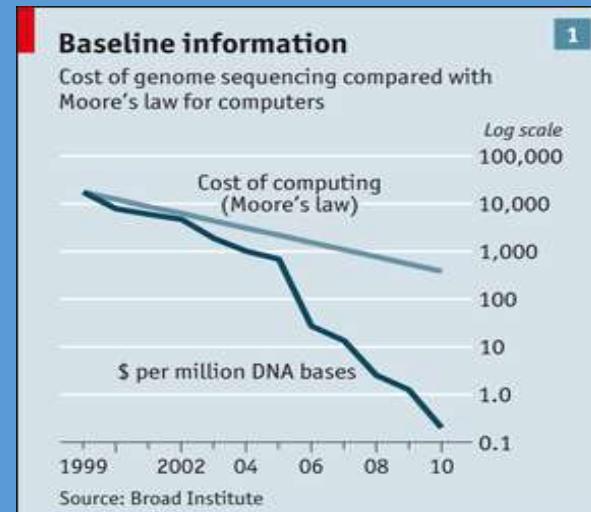
Internet of Things / M2M



User Generated (Web & Mobile)

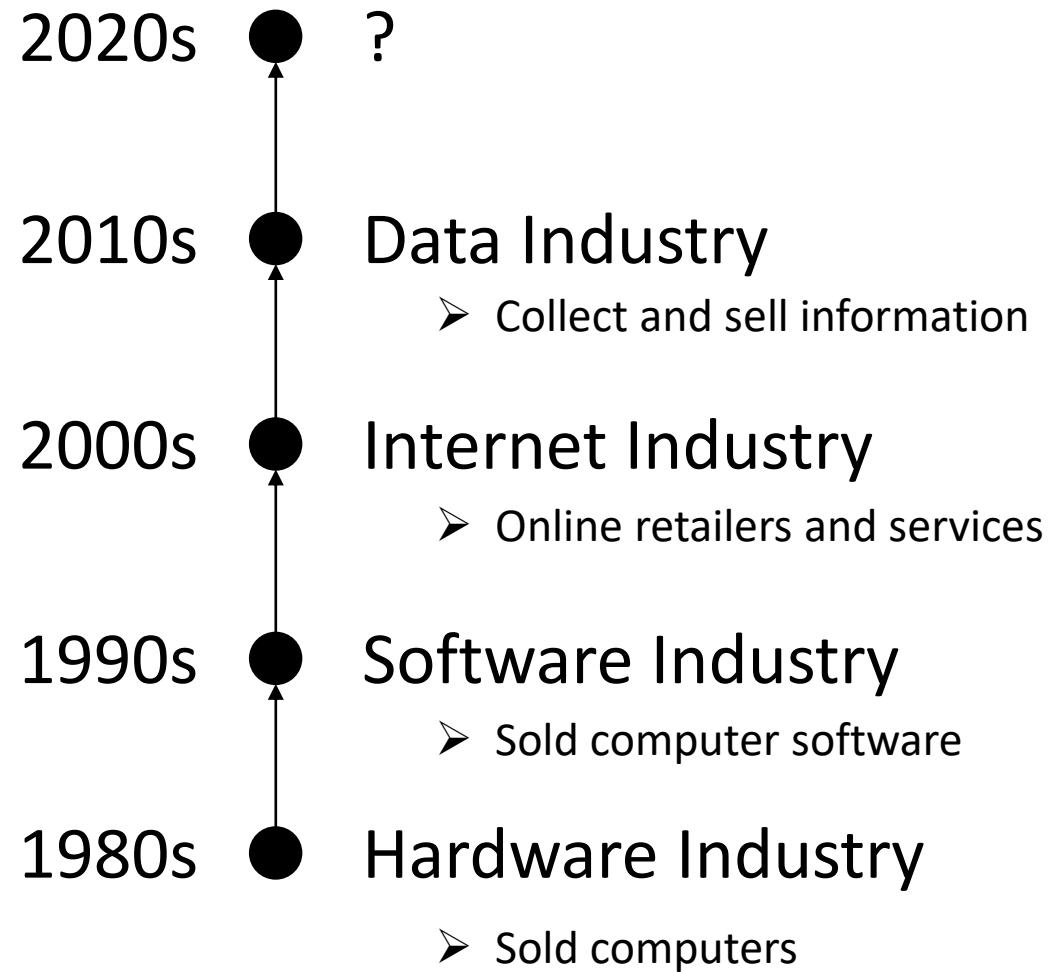


Health/Scientific Computing

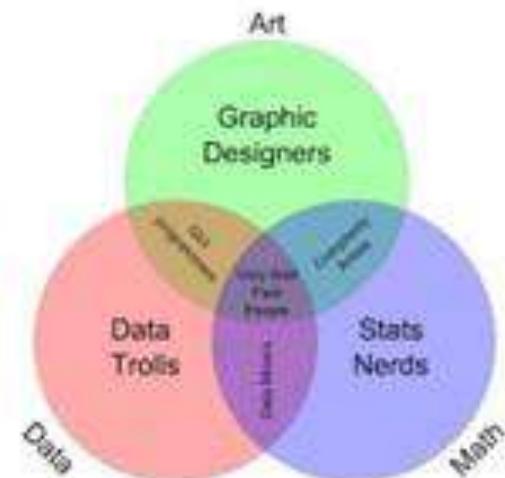
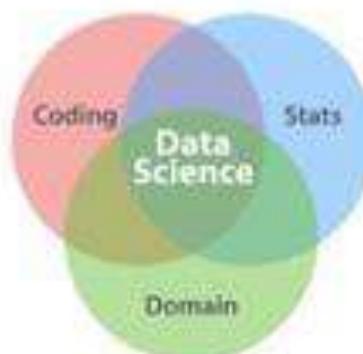
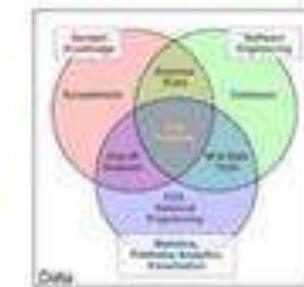
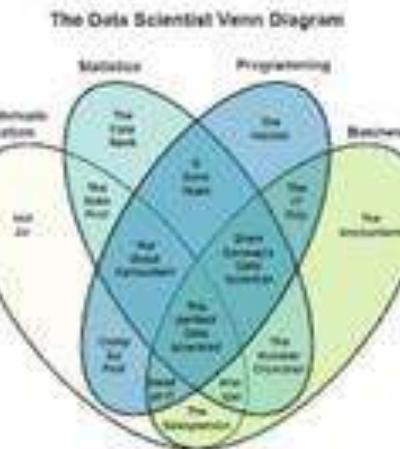


Data
To
“Big
Data”

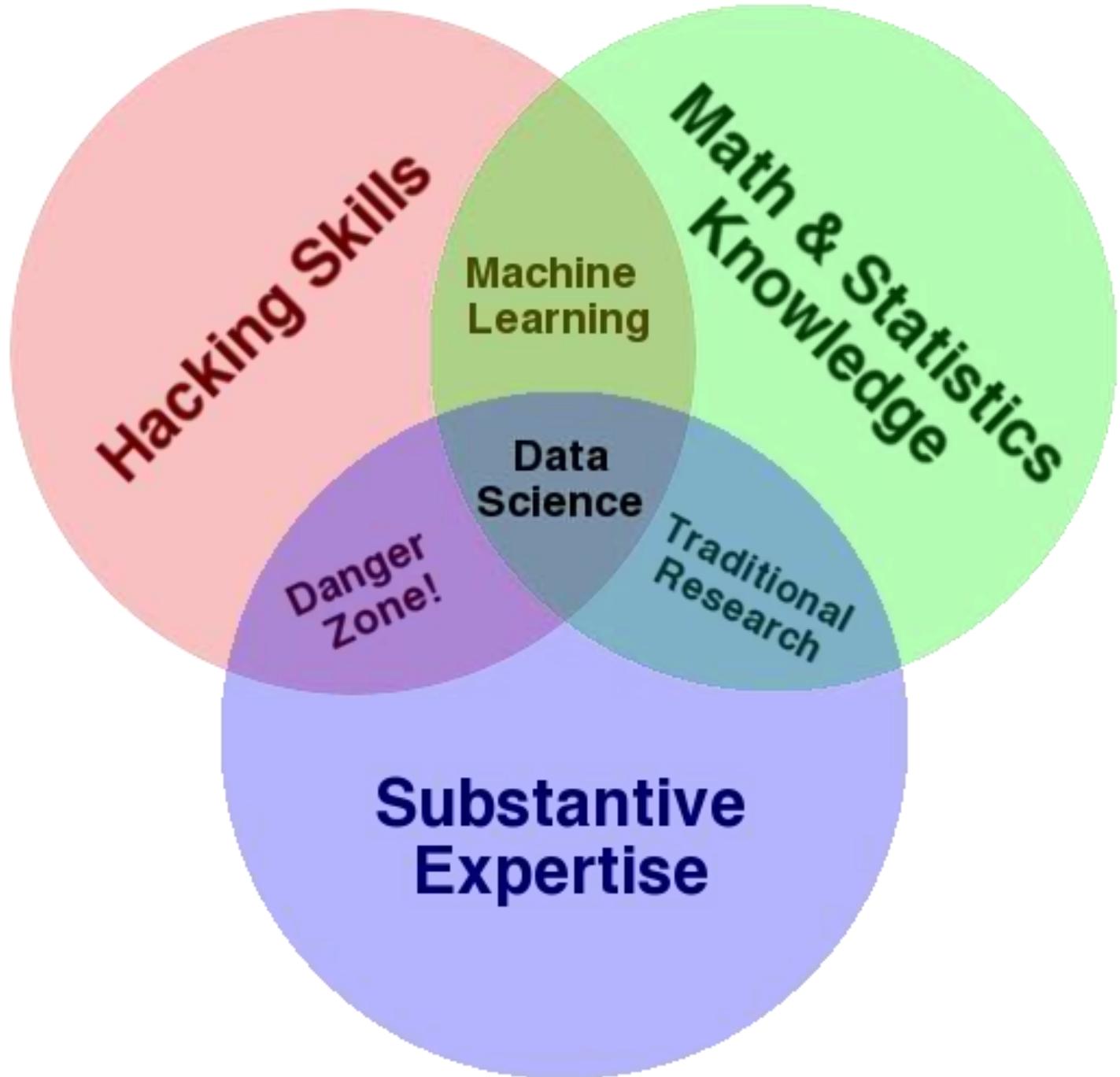
Technology Trends



Data Science: Multiple Domain (2010 SIAM Article)



Data Science : One Definition



Why “Danger Zone?”

Ronny Kohavi* keynote at KDD 2015

- People are incredibly clever at explaining “very surprising results”. Unfortunately most very surprising results are caused by data pipeline errors.
- Beware “HiPPOs” (Highest Paid-Person’s Opinion)

* General Manager for Microsoft’s Analysis & Experimentation Team

Succinct Definition of Data Science

The application of **data centric, computational, and inferential thinking** to

*understand
the world*

*solve
problems*

Science

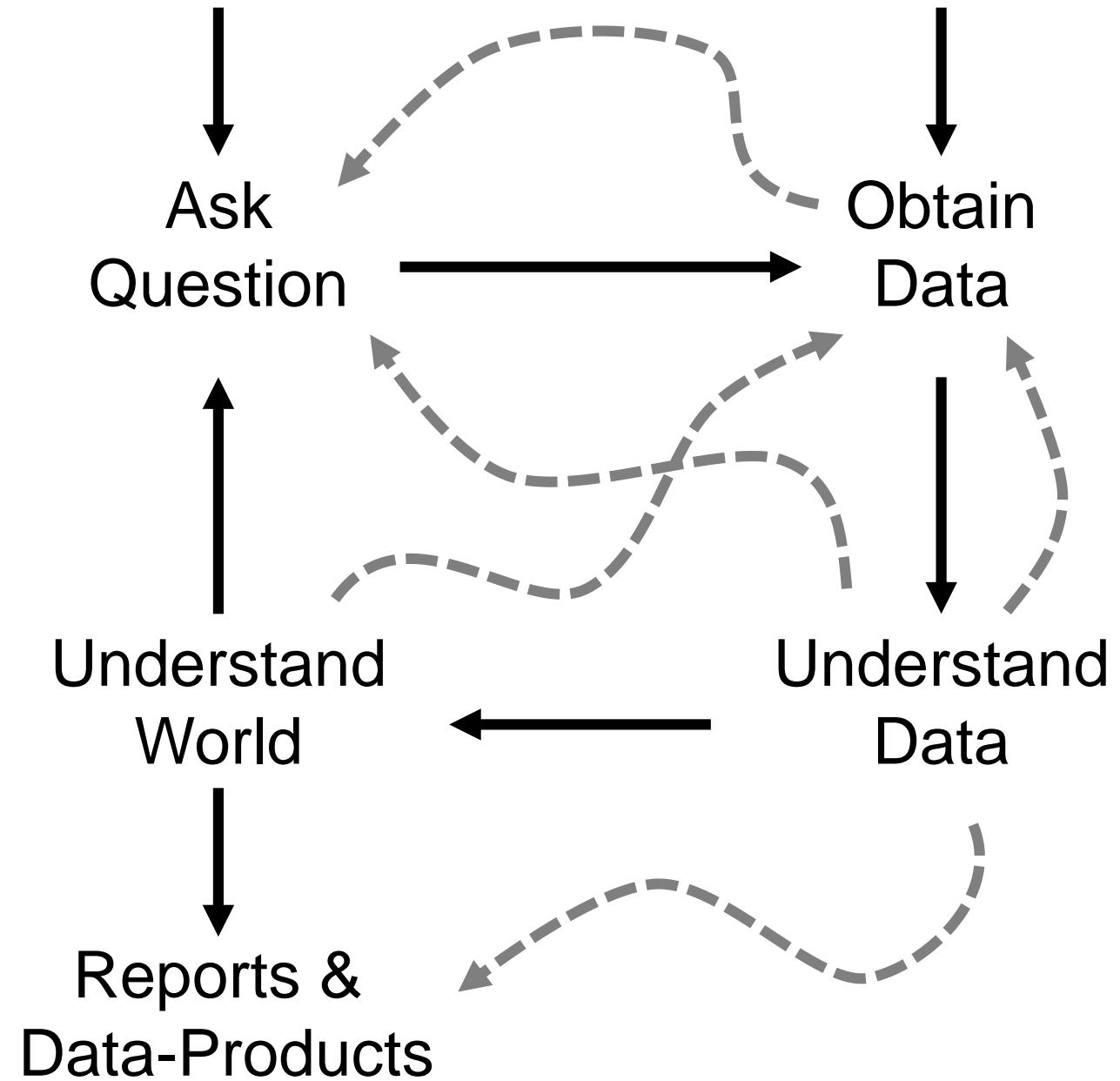
Engineering

➤ *Data science is fundamentally interdisciplinary*

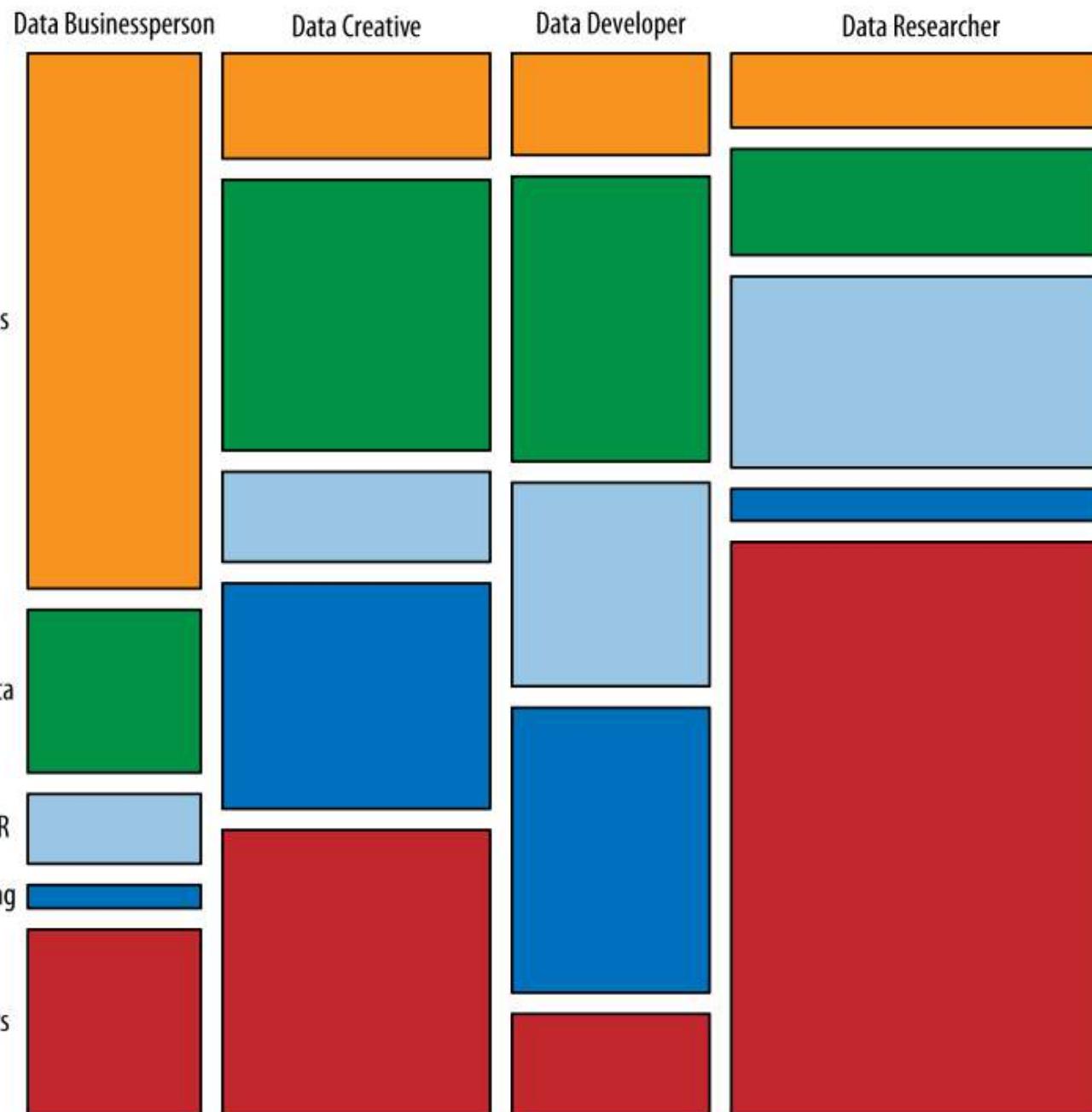
Data Science Lifecycle

High-level description of the data science workflow

- Frame questions & design experiments
- Obtain and clean data
- Summarize and visualize data
- Inference and prediction
- continuous process ...



Skills and Self-ID Top Factors



Skill Patterns

- Different skill profiles
 - Business = Domain Knowledge.
 - Data Creative /Developer

Contrast: Databases

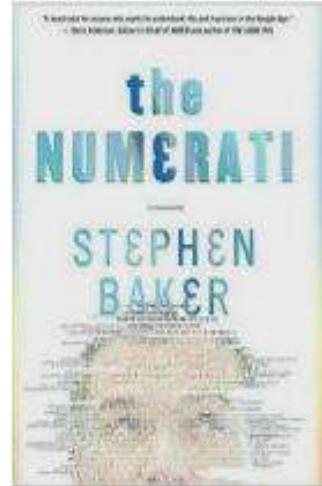
	Databases	Data Science
Data Value	“Precious”	“Cheap”
Data Volume	Modest	Massive
Examples	Bank records, Personnel records, Census, Medical records	Online clicks, GPS logs, Tweets, Building sensor readings
Priorities	Consistency, Error recovery, Auditability	Speed, Availability, Query richness
Structured	Strongly (Schema)	Weakly or none (Text)
Properties	Transactions, ACID*	CAP* theorem (2/3), eventual consistency
Realizations	SQL	NoSQL: Riak, Memcached, Apache River, MongoDB, CouchDB, Hbase, Cassandra,...

ACID = Atomicity, Consistency, Isolation and Durability

CAP = Consistency, Availability, Partition Tolerance

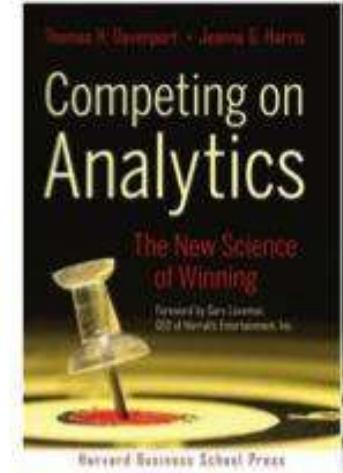
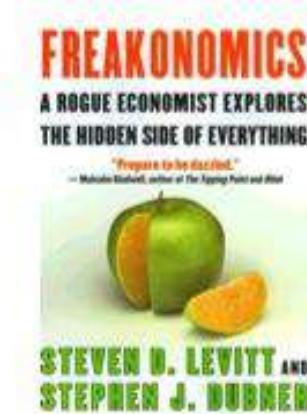
Databases

Querying the past



Data Science

Querying the future



Contrast:
Databases

Business intelligence (BI) is the transformation of raw data into meaningful and useful information for business analysis purposes. BI can handle enormous amounts of unstructured data to help identify, develop and otherwise create new strategic business opportunities - Wikipedia

Contrast: Machine Learning

Machine Learning

Develop new (individual) models

Prove mathematical properties of models

Improve/validate on a few, relatively clean, small datasets

Publish a paper

Data Science

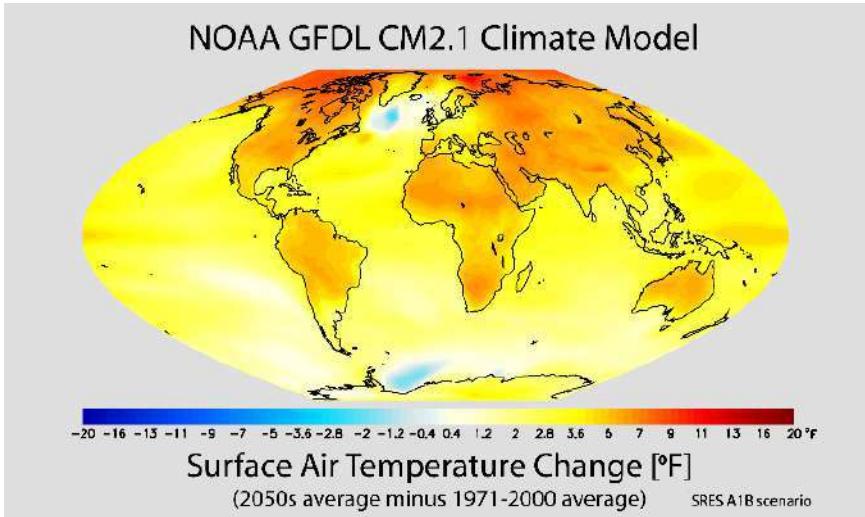
Explore many models, build and tune hybrids

Understand empirical properties of models

Develop/use tools that can handle massive datasets

Take action!

Contrast: Scientific Computing



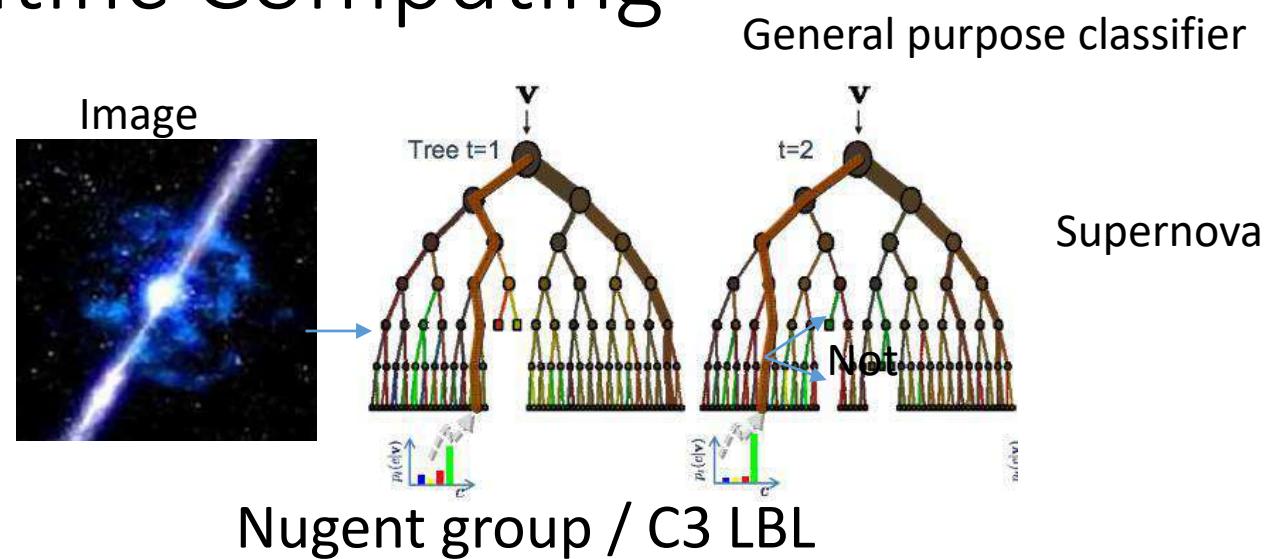
Scientific Modeling

Physics-based models

Problem-Structured

Mostly deterministic, precise

Run on Supercomputer or High-end Computing Cluster



Data-Driven Approach

General inference engine replaces model

Structure not related to problem

Statistical models handle true randomness, and **unmodeled complexity**.

Run on cheaper computer Clusters (EC2)

Hype Cycle

Gartner®

The **five** phases in the Hype Cycle are

- 1. Technology Trigger**
- 2. Peak of Inflated Expectations**
- 3. Trough of Disillusionment**
- 4. Slope of Enlightenment**
- 5. Plateau of Productivity**



What's New in the 2023 Gartner Hype Cycle for Emerging Technologies

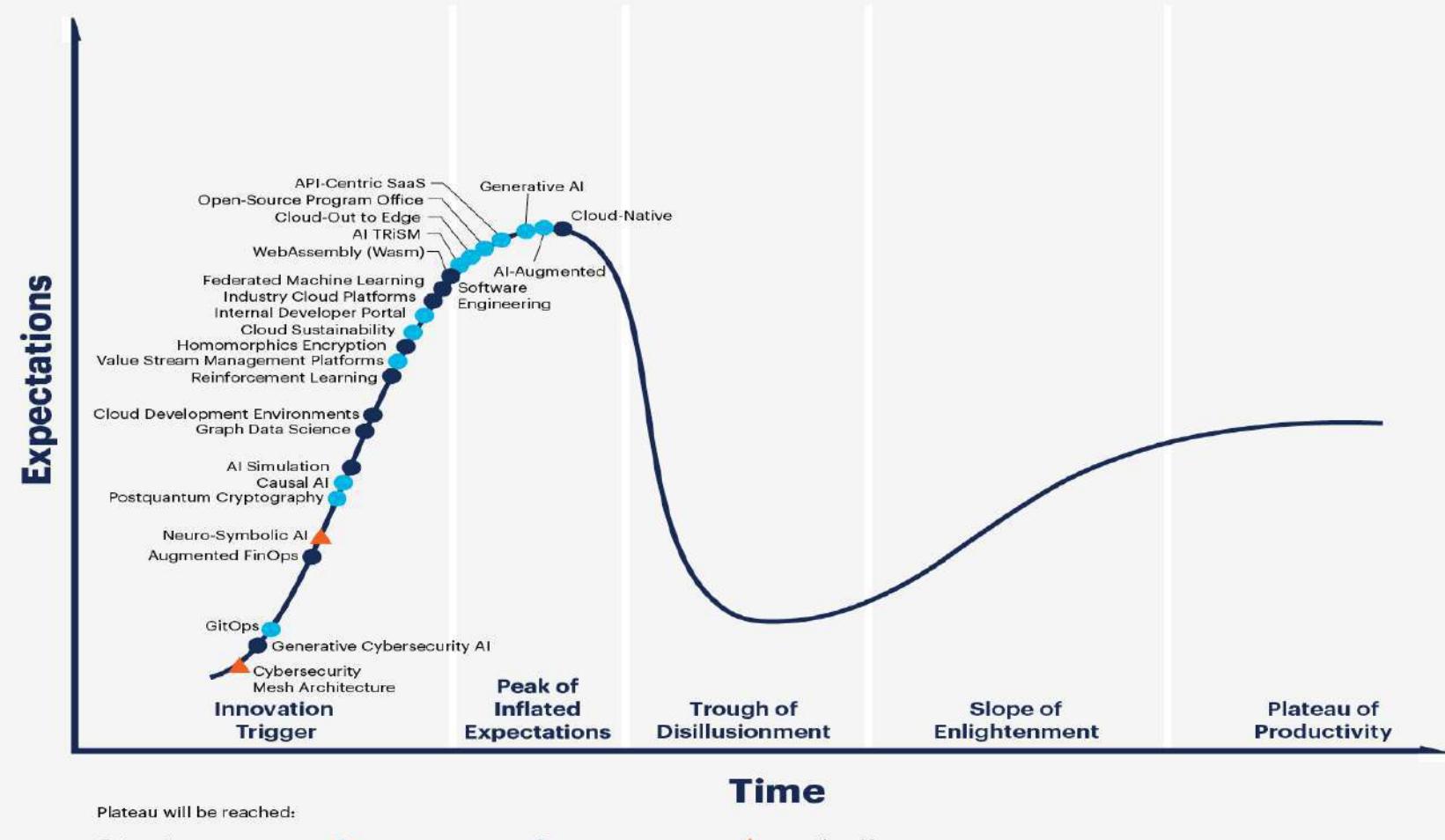
August 23, 2023

Contributor: Lori Perri

They fit into four main themes: emergent AI, developer experience, pervasive cloud, and human-centric security and privacy.

- They fit into 4 themes:
- Emergent AI
 - Developer experience
 - Pervasive clou
 - Human-centric security and privacy

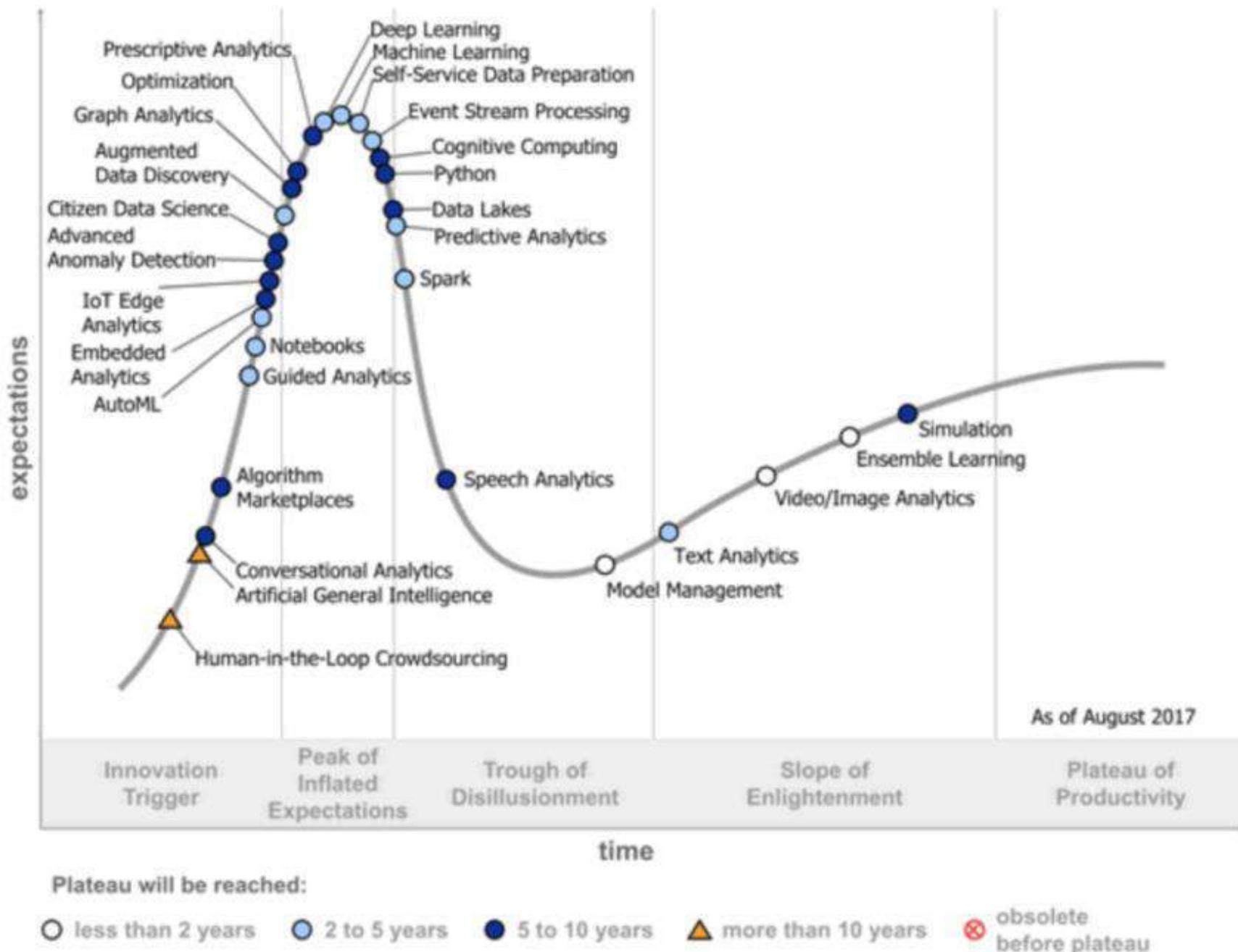
Hype Cycle for Emerging Technologies, 2023



gartner.com

Hype Cycle

Data Science & Machine Learning

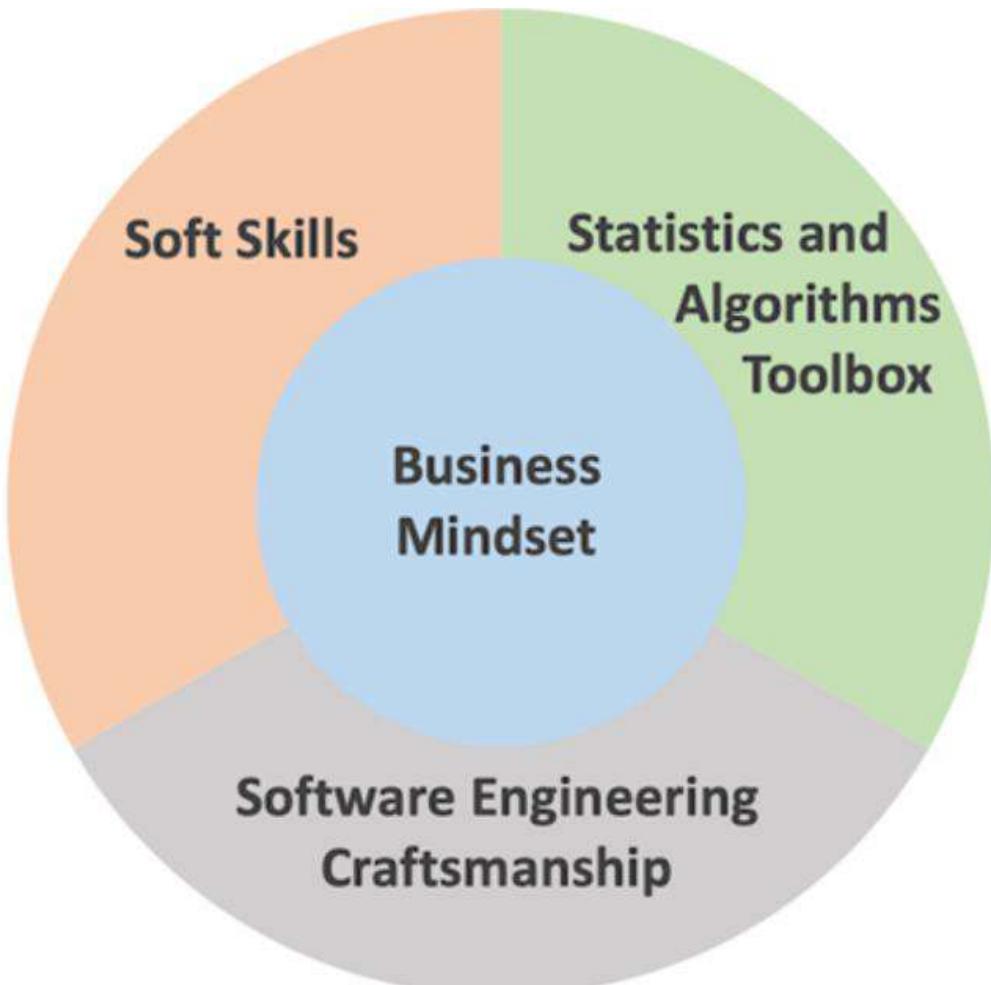


Data Science & Analytics

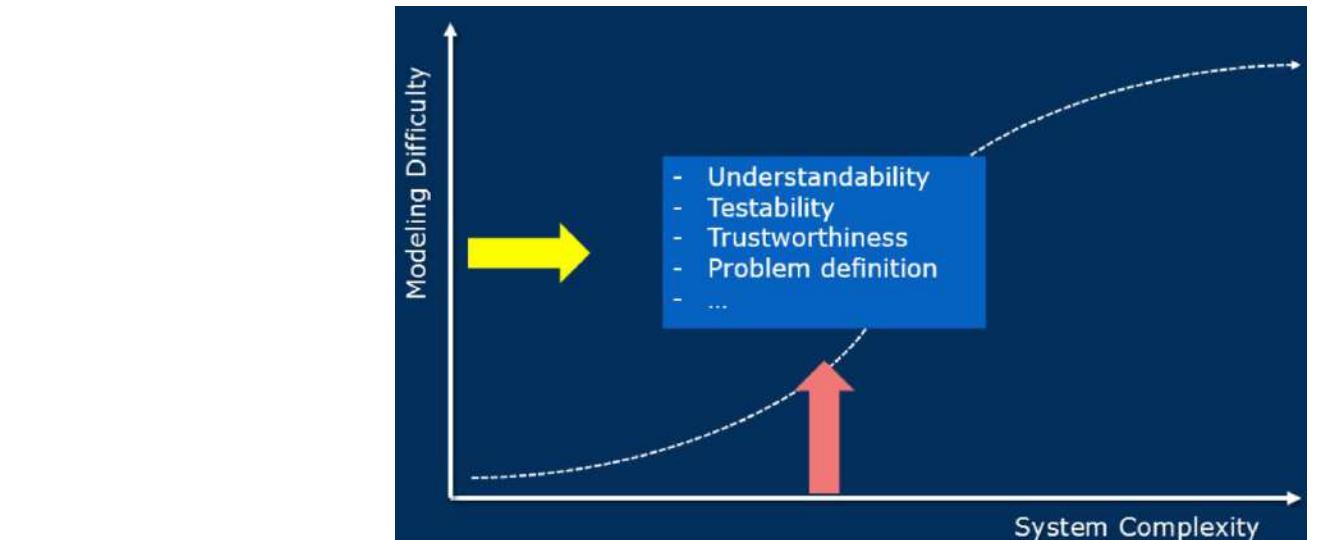
benefit	years to mainstream adoption			
	less than 2 years	2 to 5 years	5 to 10 years	more than 10 years
transformational		Augmented Data Discovery Deep Learning Event Stream Processing Machine Learning	Algorithm Marketplaces Citizen Data Science Cognitive Computing Conversational Analytics	Artificial General Intelligence Human-in-the-Loop Crowdsourcing
high	Ensemble Learning Model Management Video/Image Analytics	AutoML Guided Analytics Predictive Analytics Self-Service Data Preparation	Graph Analytics IoT Edge Analytics Optimization Prescriptive Analytics Speech Analytics	
moderate		Notebooks Spark Text Analytics	Advanced Anomaly Detection Data Lakes Embedded Analytics Python Simulation	
low				

As of August 2017

Path to be a successful Data Scientist /Analyst



"Skill portfolio of the third wave data scientist." Dominik Haitz



The data science landscape with the dimensions system complexity and modeling difficulty (cf. Ramanathan, 2016)

The best data scientists have one thing in common:
unbelievable curiosity

D.J. Patil, First White House Chief Data Scientist

FDSA

About Data

Material Adaptation from Introduction to Data Mining : Tan, Steinbach, Karpatne, Kumar

Outline

- Attributes and Objects
- Types of Data
- Data Quality
- Similarity and Distance
- Data Preprocessing

Data : Distinction

- Data
- Information
- Knowledge

Data : Distinction

- Data
- Information
- Knowledge

12 , 34 , 16 , 32, 18, 35

What are these ?

Data ? Information ?

Data : Distinction

- Data
- Information
- Knowledge

Last Three days
(June 3,4,5)
Min – Max Temp of
Kathmandu

12 , 34 , 16 , 32, 18, 35

What are these ?

*Data ? Information ?
Knowledge ?*

Data : Distinction

- Data
- Information
- Knowledge

Last Three days
Min – Max Temp of
Kathmandu

12 , 34 , 16 , 32, 18, 35

Average Temp ?
23, 24, 26.5

Data : Distinction

- Data
- Information
- Knowledge

Last Three days (June 3,4,5)
Min – Max Temp of Kathmandu

12 , 34 , 16 , 32, 18, 35

A traveler seeks Knowledge
regarding traveling to KTM in
the month of June!

*Whether Jacket is needed or
not if the travel is during the
month of June ?*

What is Data?

- Collection of ***data objects*** and their ***attributes***
- An ***attribute*** is a property or characteristic of an object
 - Examples: eye color of a person, temperature, etc.
 - Attribute is also known as variable, field, characteristic, dimension, or feature
- A collection of attributes describe an ***object***
 - Object is also known as record, point, case, sample, entity, or instance

Attributes

Objects

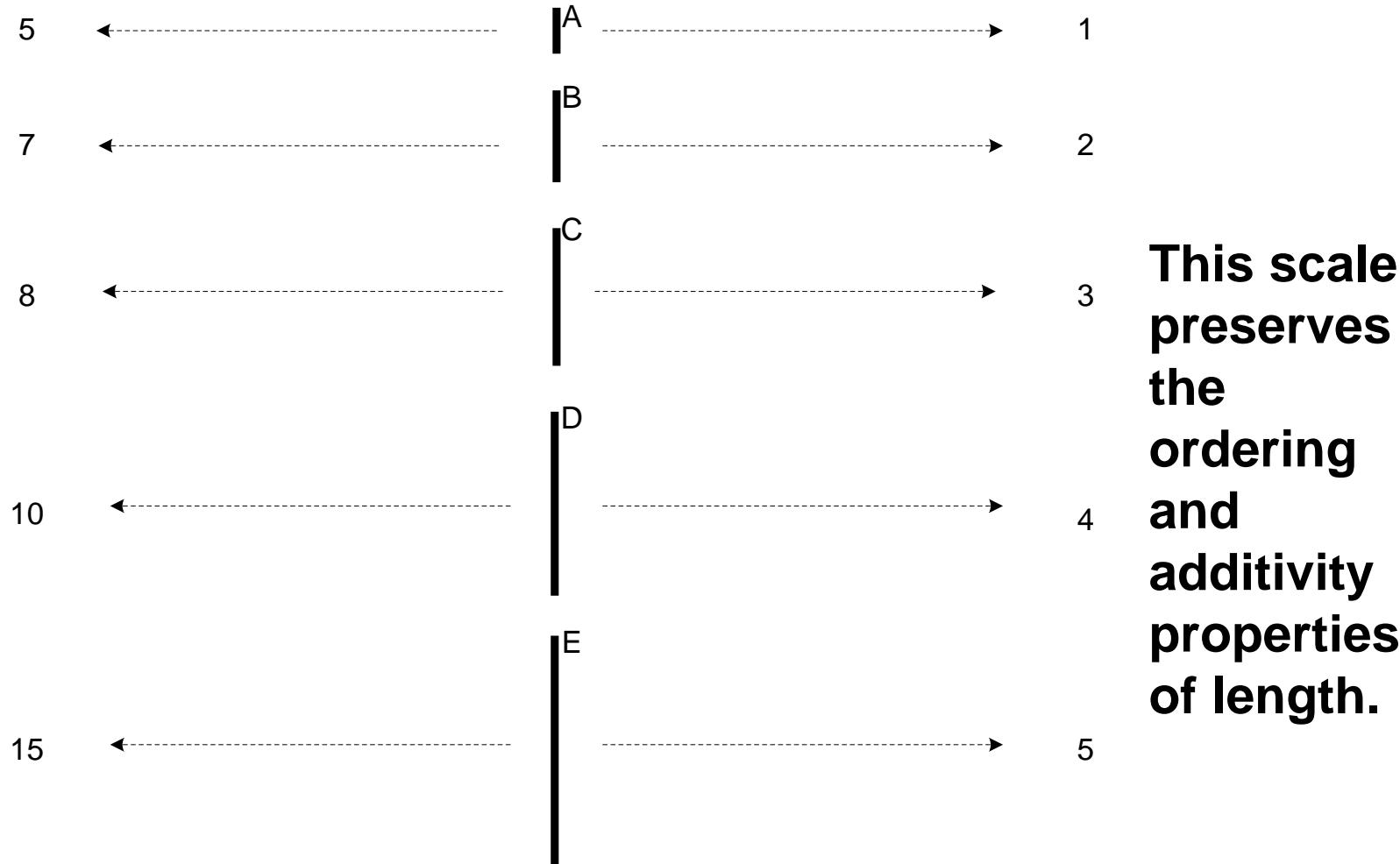
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Attribute Values

- **Attribute values** are numbers or symbols assigned to an attribute for a particular object
- Distinction between attributes and attribute values
 - Same attribute can be mapped to different attribute values
 - ◆ Example: height can be measured in feet or meters
 - Different attributes can be mapped to the same set of values
 - ◆ Example: Attribute values for ID and age are integers
 - ◆ But properties of attribute values can be different

Measurement of Length

- The way you measure an attribute may not match the attribute's properties.



Types of Attributes

- There are different types of attributes
 - Nominal
 - ◆ Examples: ID numbers, eye color, zip codes
 - Ordinal
 - ◆ Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height {tall, medium, short}
 - Interval
 - ◆ Examples: calendar dates, temperatures in Celsius or Fahrenheit.
 - Ratio
 - ◆ Examples: temperature in Kelvin, length, time, counts

Properties of Attribute Values

- The type of an attribute depends on which of the following properties/operations it possesses:
 - Distinctness: $= \neq$
 - Order: $< >$
 - Differences are meaningful : $+ -$
 - Ratios are $* /$
meaningful
 - Nominal attribute: distinctness
 - Ordinal attribute: distinctness & order
 - Interval attribute: distinctness, order & meaningful differences
 - Ratio attribute: all 4 properties/operations

Difference Between Ratio and Interval

- Is it physically meaningful to say that a temperature of 10° is twice that of 5° on
 - the Celsius scale?
 - the Fahrenheit scale?
 - the Kelvin scale?
- The meaning of 0 (Zero) difference
 - Last night temp reached up to 0°C
 - As a full-time student, he is earning 0 Rs per month!

Attribute Type	Description	Examples	Operations
Categorical Qualitative	Nominal Nominal attribute values only distinguish. ($=$, \neq)	zip codes, employee ID numbers, eye color, sex: { <i>male</i> , <i>female</i> }	mode, entropy, contingency correlation, χ^2 test
	Ordinal Ordinal attribute values also order objects. ($<$, $>$)	hardness of minerals, { <i>good</i> , <i>better</i> , <i>best</i> }, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
	Interval For interval attributes, differences between values are meaningful. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, <i>t</i> and <i>F</i> tests
	Ratio For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, current	geometric mean, harmonic mean, percent variation

This categorization of attributes is due to S. S. Stevens

Attribute Type	Transformation	Comments
Nominal	Any permutation of values	If all employee ID numbers were reassigned, would it make any difference?
Ordinal	An order preserving change of values, i.e., $\text{new_value} = f(\text{old_value})$ where f is a monotonic function	An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by { 0.5, 1, 10}.
Interval	$\text{new_value} = a * \text{old_value} + b$ where a and b are constants	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
Ratio	$\text{new_value} = a * \text{old_value}$	Length can be measured in meters or feet.

This categorization of attributes is due to S. S. Stevens

Discrete and Continuous Attributes

- Discrete Attribute

- Has only a finite or countably infinite set of values
- Examples: zip codes, counts, or the set of words in a collection of documents
- Often represented as integer variables.
- Note: **binary attributes** are a special case of discrete attributes

- Continuous Attribute

- Has real numbers as attribute values
- Examples: temperature, height, or weight.
- Practically, real values can only be measured and represented using a finite number of digits.
- Continuous attributes are typically represented as floating-point variables.

Asymmetric Attributes

- Only presence (a non-zero attribute value) is regarded as important
 - ◆ Words present in documents
 - ◆ Items present in customer transactions
- If we met a friend in the grocery store would we ever say the following?

“I see our purchases are very similar since we didn’t buy most of the same things.”
- Asymmetric attributes typically arise from objects that are sets
- We need two asymmetric binary attributes to represent one ordinary binary attribute
 - Association analysis uses asymmetric attributes

Types of data sets (Organization Aspect)

- Record
 - Data Matrix
 - Document Data
 - Transaction Data
- Graph
 - World Wide Web
 - Molecular Structures
- Ordered
 - Spatial Data
 - Temporal Data
 - Sequential Data
 - Genetic Sequence Data

Record Data

- Data that consists of a collection of records, each of which consists of a fixed set of attributes

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Data Matrix

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute
- Such data set can be represented by an m by n matrix, where there are m rows, one for each object, and n columns, one for each attribute

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Document Data

- Each document becomes a ‘term’ vector
 - Each term is a component (attribute) of the vector
 - The value of each component is the number of times the corresponding term occurs in the document.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

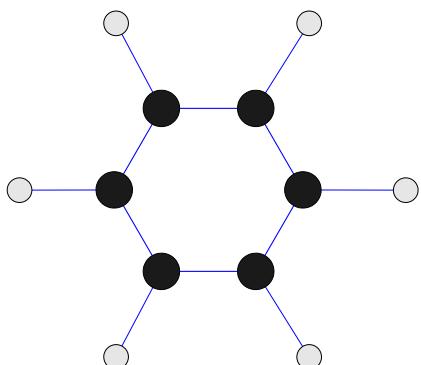
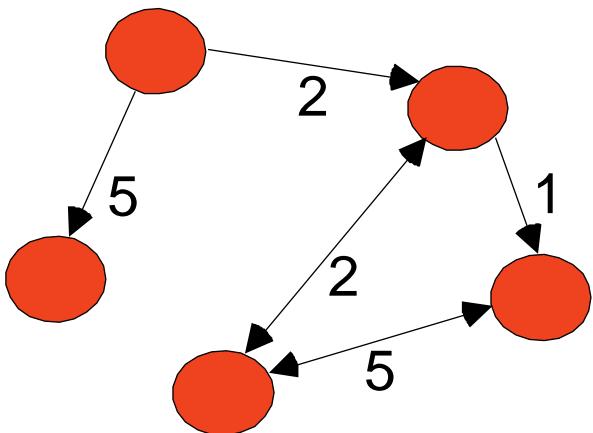
Transaction Data

- A special type of record data, where
 - Each record (transaction) involves a set of items.
 - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Graph Data

- Examples: Generic graph, a molecule, and webpages



Benzene Molecule: C₆H₆

Useful Links:

- [Bibliography](#)
- Other Useful Web sites
 - [ACM SIGKDD](#)
 - [KDnuggets](#)
 - [The Data Mine](#)

Knowledge Discovery and Data Mining Bibliography

(Gets updated frequently, so visit often!)

- [Books](#)
- [General Data Mining](#)

Book References in Data Mining and Knowledge Discovery

Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Iyer, "Advances in Knowledge Discovery and Data Mining", AAAI Press/the MIT Press, 1996.

J. Ross Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993. Michael Berry and Gordon Linoff, "Data Mining Techniques (For Marketing, Sales, and Customer Support)", John Wiley & Sons, 1997.

General Data Mining

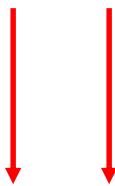
Usama Fayyad, "Mining Databases: Towards Algorithms for Knowledge Discovery", Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, vol. 21, no. 1, March 1998.

Christopher Matheus, Philip Chan, and Gregory Piatetsky-Shapiro, "Systems for Knowledge Discovery in Databases", IEEE Transactions on Knowledge and Data Engineering, 5(6):903-913, December 1993.

Ordered Data

- Sequences of transactions

Items/Events



(A B) (D) (C E)

(B D) (C) (E)

(C D) (B) (A E)



**An element of
the sequence**

Ordered Data

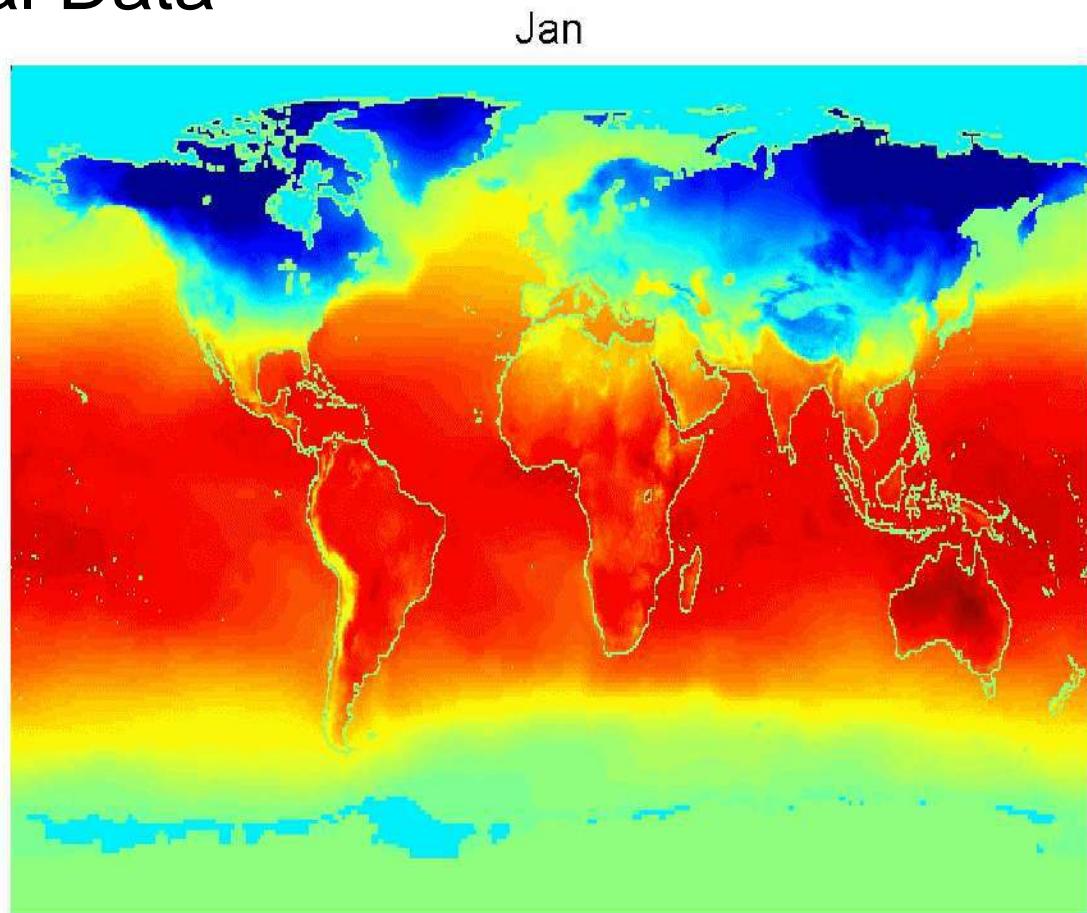
- Genomic sequence data

GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTG
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCAGGGGCCGCCCAGC
CCAACCGAGTCCGACCAAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCAGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG

Ordered Data

- Spatio-Temporal Data

Average Monthly Temperature of land and ocean



Important Characteristics of Data

- Dimensionality (number of attributes)
 - ◆ High dimensional data brings a number of challenges
- Sparsity
 - ◆ Only presence counts
- Resolution
 - ◆ Patterns depend on the scale
- Size
 - ◆ Type of analysis may depend on size of data

Similarity and Dissimilarity Measures

- Similarity measure
 - Numerical measure of how alike two data objects are.
 - Is higher when objects are more alike.
 - Often falls in the range [0,1]
- Dissimilarity measure
 - Numerical measure of how different two data objects are
 - Lower when objects are more alike
 - Minimum dissimilarity is often 0
 - Upper limit varies
- **Proximity** refers to a similarity or dissimilarity

Similarity/Dissimilarity for Simple Attributes

The following table shows the similarity and dissimilarity between two objects, x and y , with respect to a single, simple attribute.

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$
Ordinal	$d = x - y /(n - 1)$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - d$
Interval or Ratio	$d = x - y $	$s = -d, s = \frac{1}{1+d}, s = e^{-d},$ $s = 1 - \frac{d - \min_d}{\max_d - \min_d}$

Euclidean Distance

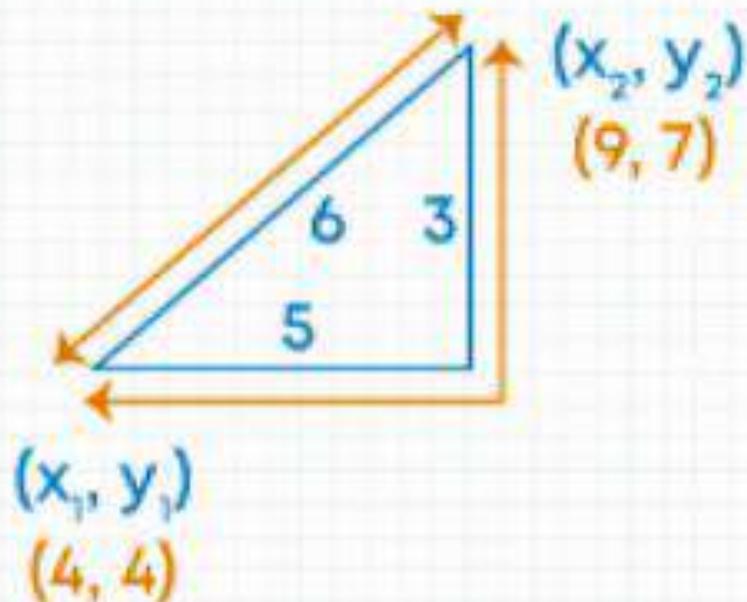
- Euclidean Distance

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^n (x_k - y_k)^2}$$

where n is the number of dimensions (attributes) and x_k and y_k are, respectively, the k^{th} attributes (components) or data objects \mathbf{x} and \mathbf{y} .

- Standardization is necessary, if scales differ.

Example:



Euclidean distance

$$= \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

$$= \sqrt{(9 - 4)^2 + (7 - 4)^2}$$

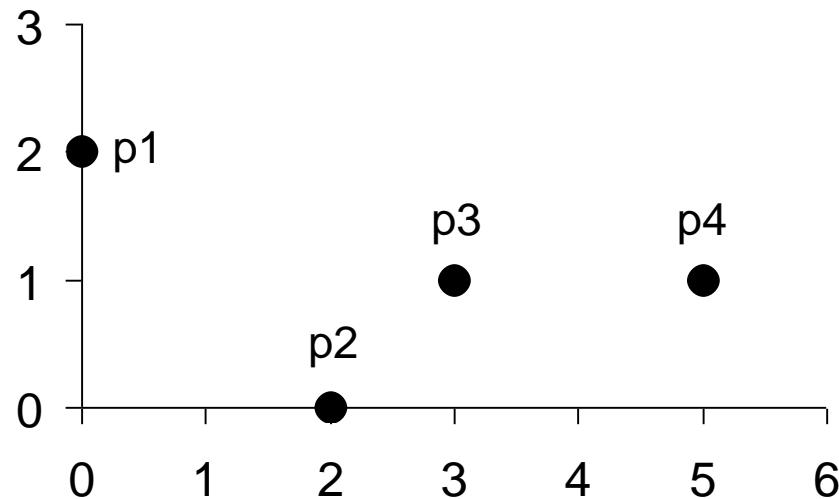
$$= \sqrt{5^2 + 3^2}$$

$$= \sqrt{25 + 9}$$

$$= \sqrt{34}$$

$$= 5.83$$

Euclidean Distance



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix

Minkowski Distance

- Minkowski Distance is a generalization of Euclidean Distance

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^n |x_k - y_k|^r \right)^{1/r}$$

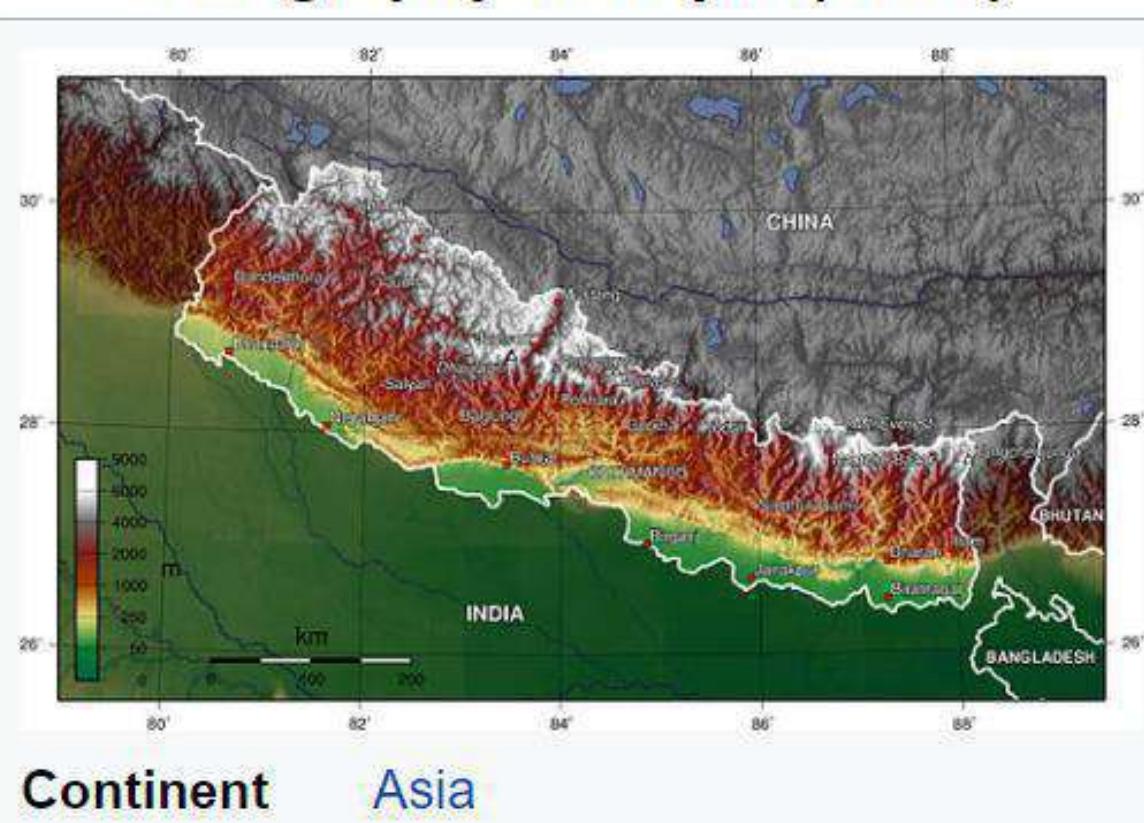
Where r is a parameter, n is the number of dimensions (attributes) and x_k and y_k are, respectively, the k^{th} attributes (components) or data objects x and y .

Minkowski Distance: Examples

- $r = 1$. City block (Manhattan, taxicab, L_1 norm) distance.
 - A common example of this is the Hamming distance, which is just the number of bits that are different between two binary vectors
- $r = 2$. Euclidean distance
- $r \rightarrow \infty$. “supremum” (L_{\max} norm, L_∞ norm) distance.
 - This is the maximum difference between any component of the vectors
- Do not confuse r with n , i.e., all these distances are defined for all numbers of dimensions.

Nepal measures about 880 kilometers (547 mi) along its Himalayan axis by 150 to 250 kilometers (93 to 155 mi) across. It has an area of 147,516 km² (56,956 sq mi).^[1]

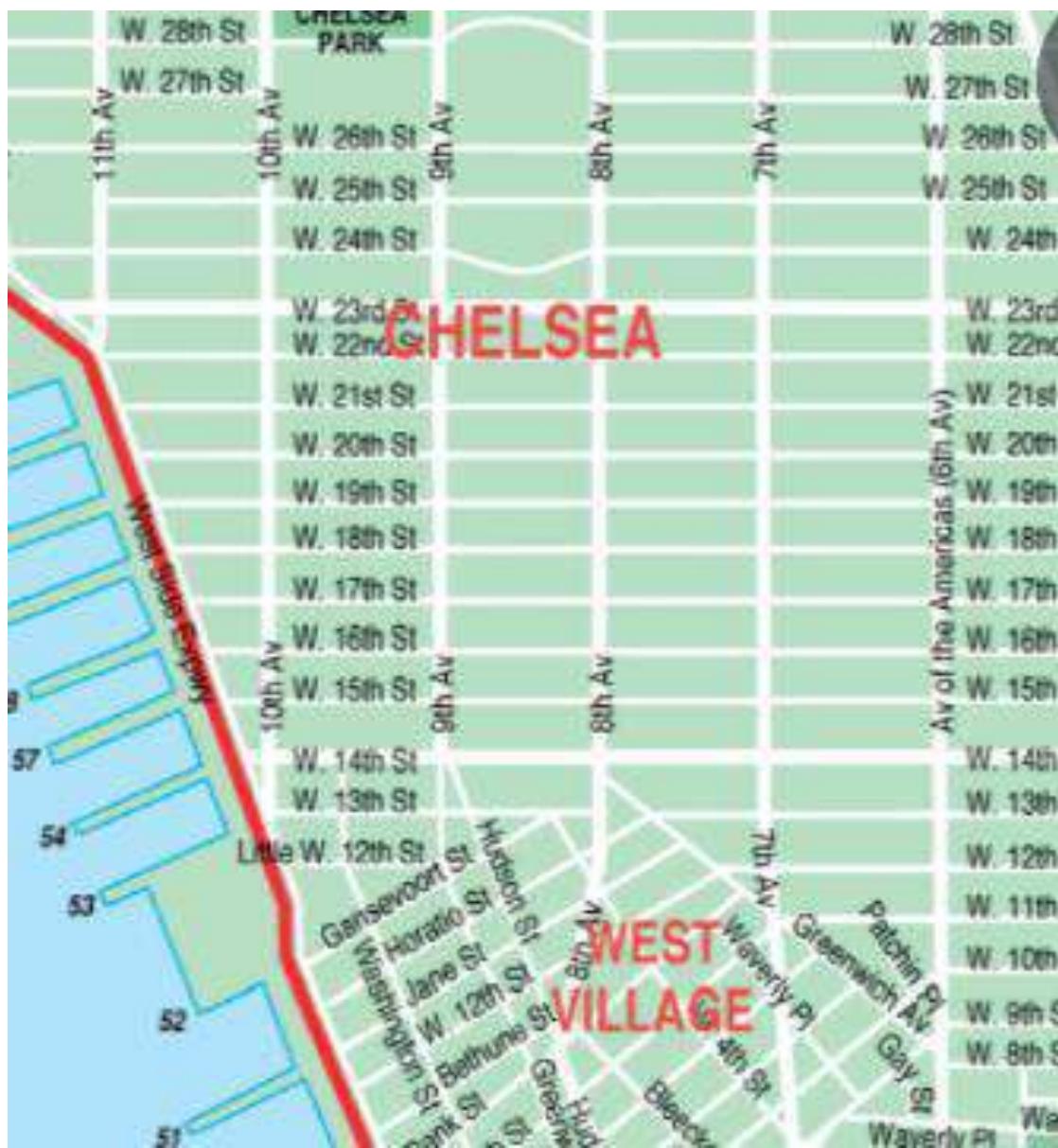
Geography of Nepal (नेपाल)



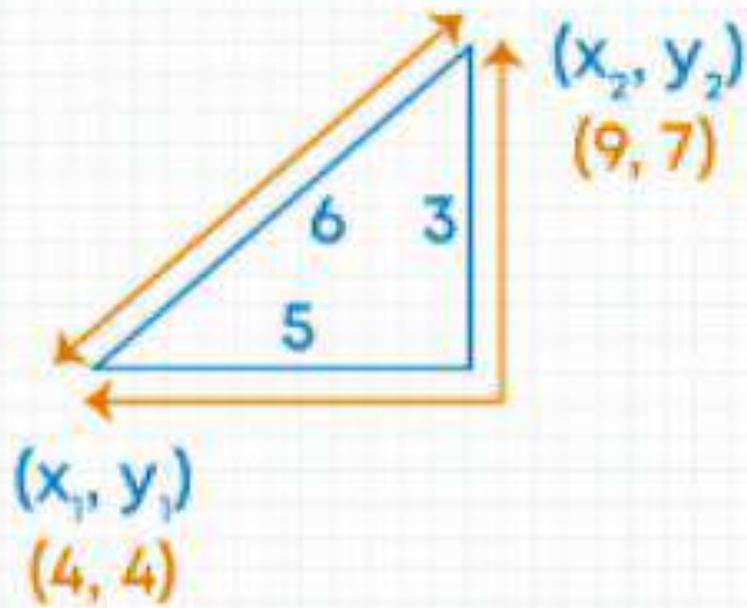
1027.67 km

East west Highway of Nepal is also known as the Mahendra **highway** is the longest roadway Length of **east-west highway** is 1027.67 km. Feb 23, 2016

Manhattan Street Map



Example:



Euclidean distance

$$\begin{aligned} &= \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \\ &= \sqrt{(9 - 4)^2 + (7 - 4)^2} \\ &= \sqrt{5^2 + 3^2} \\ &= \sqrt{25 + 9} \\ &= \sqrt{34} \\ &= 5.83 \end{aligned}$$

Manhattan distance

$$\begin{aligned} &= |x_2 - x_1| + |y_2 - y_1| \\ &= |9 - 4| + |7 - 4| \\ &= 5 + 3 \\ &= 8 \end{aligned}$$

Minkowski Distance

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

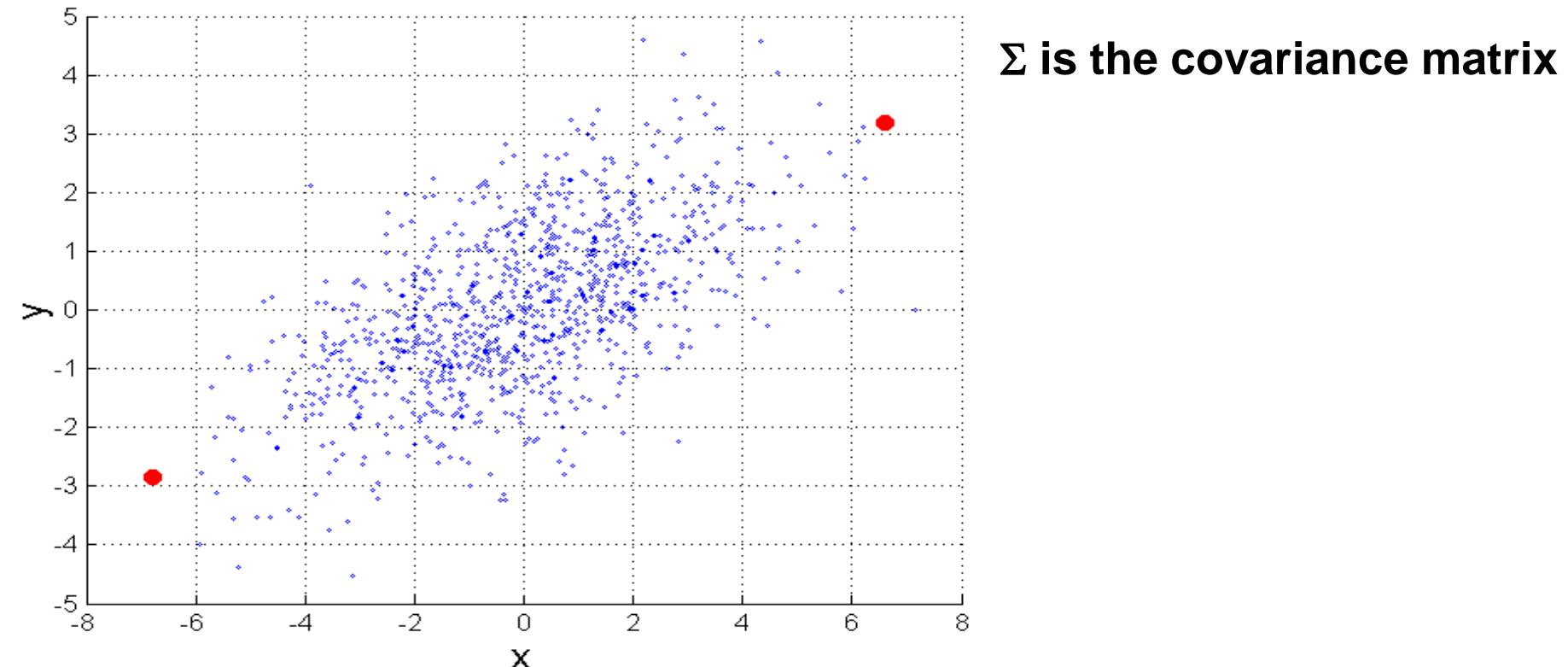
L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

L ∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Distance Matrix

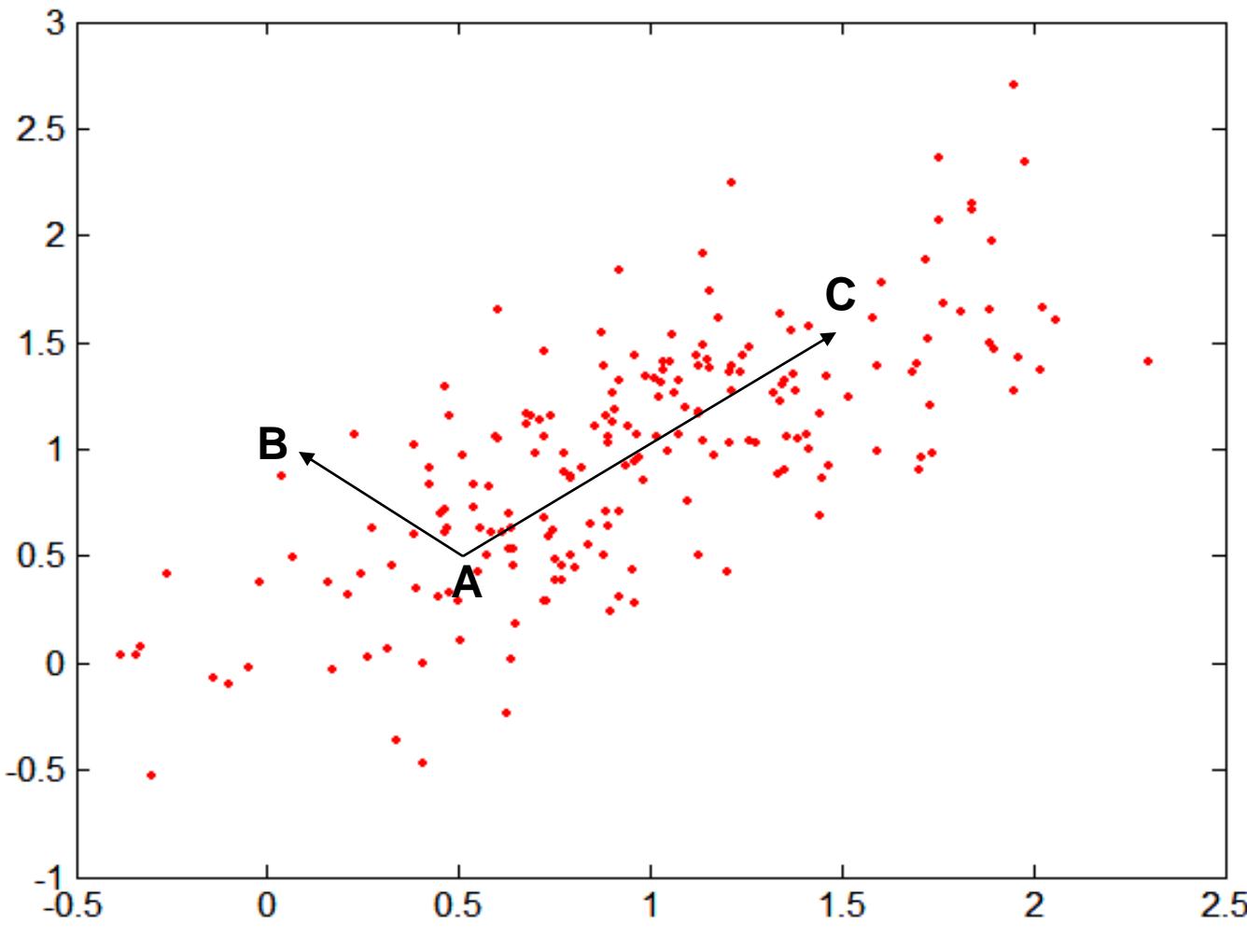
Mahalanobis Distance

$$\text{mahalanobis}(x, y) = (x - y)^T \Sigma^{-1} (x - y)$$



For red points, the Euclidean distance is 14.7, Mahalanobis distance is 6.

Mahalanobis Distance



Covariance Matrix:

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

A: (0.5, 0.5)

B: (0, 1)

C: (1.5, 1.5)

$\text{Mahal}(A,B) = 5$

$\text{Mahal}(A,C) = 4$

Similarity Between Binary Vectors

- Common situation is that objects, p and q , have only binary attributes

- Compute similarities using the following quantities

f_{01} = the number of attributes where p was 0 and q was 1

f_{10} = the number of attributes where p was 1 and q was 0

f_{00} = the number of attributes where p was 0 and q was 0

f_{11} = the number of attributes where p was 1 and q was 1

- Simple Matching and Jaccard Coefficients

SMC = number of matches / number of attributes

$$= (f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00})$$

J = number of 11 matches / number of non-zero attributes

$$= (f_{11}) / (f_{01} + f_{10} + f_{11})$$

SMC versus Jaccard: Example

$\mathbf{x} = 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0$

$\mathbf{y} = 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 1$

$f_{01} = 2$ (the number of attributes where p was 0 and q was 1)

$f_{10} = 1$ (the number of attributes where p was 1 and q was 0)

$f_{00} = 7$ (the number of attributes where p was 0 and q was 0)

$f_{11} = 0$ (the number of attributes where p was 1 and q was 1)

$$\begin{aligned}\text{SMC} &= (f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00}) \\ &= (0+7) / (2+1+0+7) = 0.7\end{aligned}$$

$$J = (f_{11}) / (f_{01} + f_{10} + f_{11}) = 0 / (2 + 1 + 0) = 0$$

Cosine Similarity

- If \mathbf{d}_1 and \mathbf{d}_2 are two document vectors, then

$$\cos(\mathbf{d}_1, \mathbf{d}_2) = \langle \mathbf{d}_1, \mathbf{d}_2 \rangle / \|\mathbf{d}_1\| \|\mathbf{d}_2\|,$$

where $\langle \mathbf{d}_1, \mathbf{d}_2 \rangle$ indicates inner product or vector dot product of vectors, \mathbf{d}_1 and \mathbf{d}_2 , and $\|\mathbf{d}\|$ is the length of vector \mathbf{d} .

- Example:

$$\mathbf{d}_1 = 3 \ 2 \ 0 \ 5 \ 0 \ 0 \ 0 \ 2 \ 0 \ 0$$

$$\mathbf{d}_2 = 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 2$$

$$\langle \mathbf{d}_1, \mathbf{d}_2 \rangle = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$$

$$\|\mathbf{d}_1\| = (3^2 + 2^2 + 0^2 + 5^2 + 0^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2)^{0.5} = (42)^{0.5} = 6.481$$

$$\|\mathbf{d}_2\| = (1^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 1^2 + 0^2 + 2^2)^{0.5} = (6)^{0.5} = 2.449$$

$$\cos(\mathbf{d}_1, \mathbf{d}_2) = 0.3150$$

Extended Jaccard Coefficient (Tanimoto)

- Variation of Jaccard for continuous or count attributes
 - Reduces to Jaccard for binary attributes

$$EJ(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \mathbf{x} \cdot \mathbf{y}}$$

Correlation measures the linear relationship between objects

$$\text{corr}(\mathbf{x}, \mathbf{y}) = \frac{\text{covariance}(\mathbf{x}, \mathbf{y})}{\text{standard_deviation}(\mathbf{x}) * \text{standard_deviation}(\mathbf{y})} = \frac{s_{xy}}{s_x s_y}, \quad (2.11)$$

where we are using the following standard statistical notation and definitions

$$\text{covariance}(\mathbf{x}, \mathbf{y}) = s_{xy} = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y}) \quad (2.12)$$

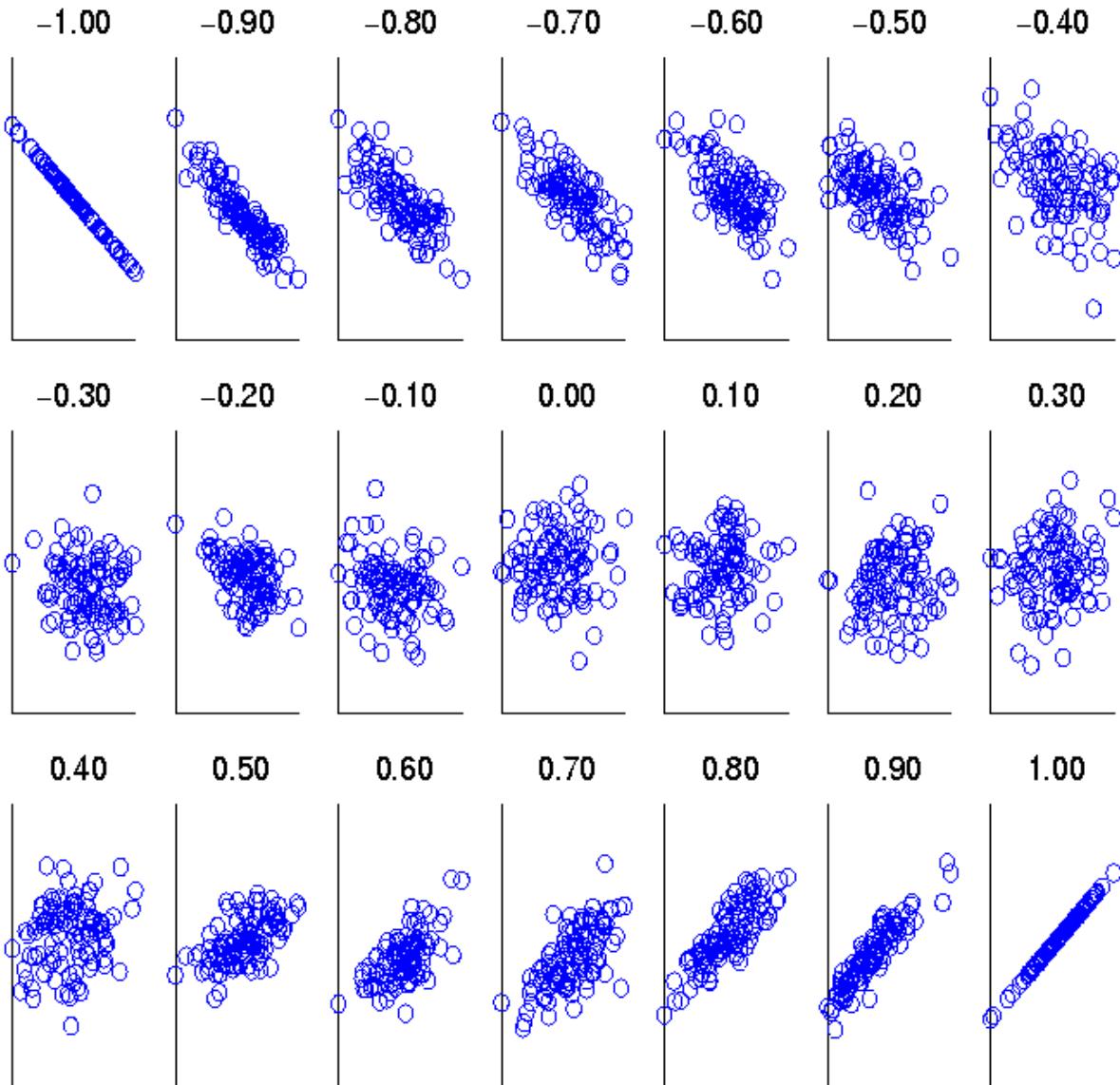
$$\text{standard_deviation}(\mathbf{x}) = s_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2}$$

$$\text{standard_deviation}(\mathbf{y}) = s_y = \sqrt{\frac{1}{n-1} \sum_{k=1}^n (y_k - \bar{y})^2}$$

$$\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k \text{ is the mean of } \mathbf{x}$$

$$\bar{y} = \frac{1}{n} \sum_{k=1}^n y_k \text{ is the mean of } \mathbf{y}$$

Visually Evaluating Correlation



**Scatter plots
showing the
similarity from
-1 to 1.**

Drawback of Correlation

- $\mathbf{x} = (-3, -2, -1, 0, 1, 2, 3)$
- $\mathbf{y} = (9, 4, 1, 0, 1, 4, 9)$

$$y_i = x_i^2$$

- $\text{mean}(\mathbf{x}) = 0, \text{mean}(\mathbf{y}) = 4$
- $\text{std}(\mathbf{x}) = 2.16, \text{std}(\mathbf{y}) = 3.74$
- $\begin{aligned} \text{corr} &= (-3)(5)+(-2)(0)+(-1)(-3)+(0)(-4)+(1)(-3)+(2)(0)+3(5) / (6 * 2.16 * 3.74) \\ &= 0 \end{aligned}$

General Approach for Combining Similarities

- Sometimes attributes are of many different types, but an overall similarity is needed.

1: For the k^{th} attribute, compute a similarity, $s_k(\mathbf{x}, \mathbf{y})$, in the range [0, 1].

2: Define an indicator variable, δ_k , for the k^{th} attribute as follows:

$\delta_k = 0$ if the k^{th} attribute is an asymmetric attribute and both objects have a value of 0, or if one of the objects has a missing value for the k^{th} attribute

$\delta_k = 1$ otherwise

3. Compute $\text{similarity}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^n \delta_k s_k(\mathbf{x}, \mathbf{y})}{\sum_{k=1}^n \delta_k}$

Using Weights to Combine Similarities

- May not want to treat all attributes the same.
 - Use non-negative weights ω_k
 - $similarity(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^n \omega_k \delta_k s_k(\mathbf{x}, \mathbf{y})}{\sum_{k=1}^n \omega_k \delta_k}$
- Can also define a weighted form of distance

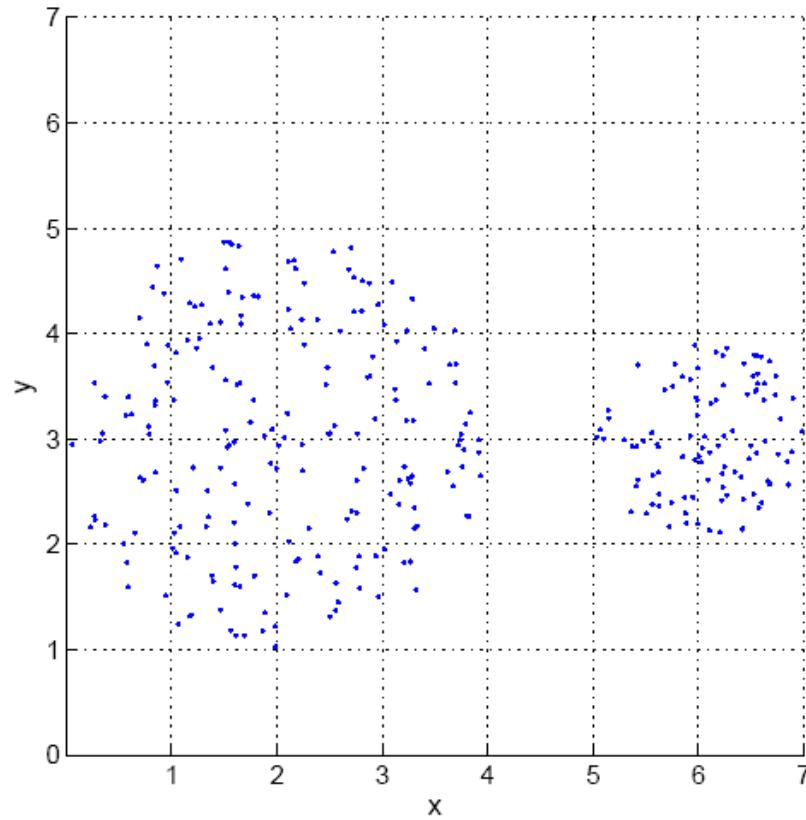
$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^n w_k |x_k - y_k|^r \right)^{1/r}$$

Density

- Measures the degree to which data objects are close to each other in a specified area
- The notion of density is closely related to that of proximity
- Concept of density is typically used for clustering and anomaly detection
- Examples:
 - Euclidean density
 - ◆ Euclidean density = number of points per unit volume
 - Probability density
 - ◆ Estimate what the distribution of the data looks like
 - Graph-based density
 - ◆ Connectivity

Euclidean Density: Grid-based Approach

- Simplest approach is to divide region into a number of rectangular cells of equal volume and define density as # of points the cell contains



0	0	0	0	0	0	0
0	0	0	0	0	0	0
4	17	18	6	0	0	0
14	14	13	13	0	18	27
11	18	10	21	0	24	31
3	20	14	4	0	0	0
0	0	0	0	0	0	0

Euclidean Density: Center-Based

- Euclidean density is the number of points within a specified radius of the point

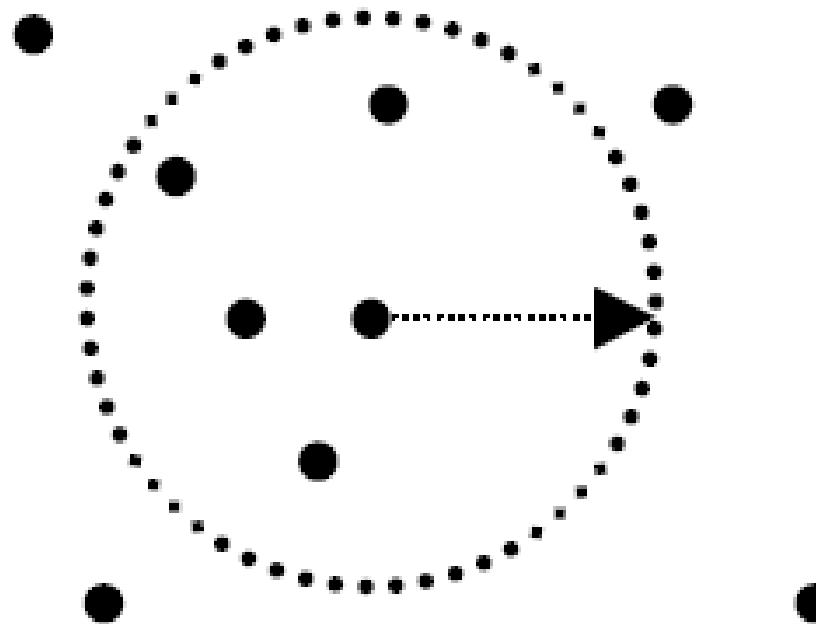


Illustration of center-based density.

Comparison of Proximity Measures

- Domain of application
 - Similarity measures tend to be specific to the type of attribute and data
 - Record data, images, graphs, sequences, 3D-protein structure, etc. tend to have different measures
- However, one can talk about various properties that you would like a proximity measure to have
 - Symmetry is a common one
 - Tolerance to noise and outliers is another
 - Ability to find more types of patterns?
 - Many others possible
- The measure must be applicable to the data and produce results that agree with domain knowledge

Information Based Measures

- Information theory is a well-developed and fundamental discipline with broad applications
- Some similarity measures are based on information theory
 - Mutual information in various versions
 - Maximal Information Coefficient (MIC) and related measures
 - General and can handle non-linear relationships
 - Can be complicated and time intensive to compute

Information and Probability

- Information relates to possible outcomes of an event
 - transmission of a message, flip of a coin, or measurement of a piece of data
- The more certain an outcome, the less information that it contains and vice-versa
 - For example, if a coin has two heads, then an outcome of heads provides no information
 - More quantitatively, the information is related to the probability of an outcome
 - ◆ The smaller the probability of an outcome, the more information it provides and vice-versa
 - Entropy is the commonly used measure



Entropy

- For
 - a variable (event), X ,
 - with n possible values (outcomes), $x_1, x_2 \dots, x_n$
 - each outcome having probability, $p_1, p_2 \dots, p_n$
 - the entropy of X , $H(X)$, is given by

$$H(X) = - \sum_{i=1}^n p_i \log_2 p_i$$

- Entropy is between 0 and $\log_2 n$ and is measured in bits
 - Thus, entropy is a measure of how many bits it takes to represent an observation of X on average

Entropy Examples

- For a coin with probability p of heads and probability $q = 1 - p$ of tails

$$H = -p \log_2 p - q \log_2 q$$

- For $p=0.5, q=0.5$ (fair coin) $H=1$
 - For $p = 1$ or $q = 1$, $H = 0$
-
- What is the entropy of a fair four-sided die?

Entropy for Sample Data: Example

Hair Color	Count	p	$-p \log_2 p$
Black	75	0.75	0.3113
Brown	15	0.15	0.4105
Blond	5	0.05	0.2161
Red	0	0.00	0
Other	5	0.05	0.2161
Total	100	1.0	1.1540

Maximum entropy is $\log_2 5 = 2.3219$

Entropy for Sample Data

- Suppose we have
 - a number of observations (m) of some attribute, X , e.g., the hair color of students in the class,
 - where there are n different possible values
 - And the number of observation in the i^{th} category is m_i
 - Then, for this sample

$$H(X) = - \sum_{i=1}^n \frac{m_i}{m} \log_2 \frac{m_i}{m}$$

- For continuous data, the calculation is harder

Mutual Information

- Information one variable provides about another

Formally, $I(X, Y) = H(X) + H(Y) - H(X, Y)$, where

$H(X, Y)$ is the joint entropy of X and Y ,

$$H(X, Y) = - \sum_i \sum_j p_{ij} \log_2 p_{ij}$$

Where p_{ij} is the probability that the i^{th} value of X and the j^{th} value of Y occur together

- For discrete variables, this is easy to compute
- Maximum mutual information for discrete variables is $\log_2(\min(n_X, n_Y))$, where n_X (n_Y) is the number of values of X (Y)

Mutual Information Example

Student Status	Count	p	$-p \log_2 p$
Undergrad	45	0.45	0.5184
Grad	55	0.55	0.4744
Total	100	1.00	0.9928

Grade	Count	p	$-p \log_2 p$
A	35	0.35	0.5301
B	50	0.50	0.5000
C	15	0.15	0.4105
Total	100	1.00	1.4406

Student Status	Grade	Count	p	$-p \log_2 p$
Undergrad	A	5	0.05	0.2161
Undergrad	B	30	0.30	0.5211
Undergrad	C	10	0.10	0.3322
Grad	A	30	0.30	0.5211
Grad	B	20	0.20	0.4644
Grad	C	5	0.05	0.2161
Total		100	1.00	2.2710

$$\text{Mutual information of Student Status and Grade} = 0.9928 + 1.4406 - 2.2710 = 0.1624$$

Maximal Information Coefficient

- Reshef, David N., Yakir A. Reshef, Hilary K. Finucane, Sharon R. Grossman, Gilean McVean, Peter J. Turnbaugh, Eric S. Lander, Michael Mitzenmacher, and Pardis C. Sabeti. "Detecting novel associations in large data sets." *science* 334, no. 6062 (2011): 1518-1524.
- Applies mutual information to two continuous variables
- Consider the possible binnings of the variables into discrete categories
 - $n_X \times n_Y \leq N^{0.6}$ where
 - ◆ n_X is the number of values of X
 - ◆ n_Y is the number of values of Y
 - ◆ N is the number of samples (observations, data objects)
- Compute the mutual information
 - Normalized by $\log_2(\min(n_X, n_Y))$
- Take the highest value

Data Preprocessing for Datamining

Dr. Arun K. Timalsina

Materials Adaptation :

Jiawei Han, Micheline Kamber, and Jian Pei, **Data Mining: Concepts and Techniques**, 3rd Edition, Morgan Kaufmann, 2011.

Data Preprocessing

- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

Why Data Preprocessing?

- Data in the real world is dirty
 - **incomplete**: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
 - e.g., occupation=" "
 - **noisy**: containing errors or outliers
 - e.g., Salary="-10"
 - **inconsistent**: containing discrepancies in codes or names
 - e.g., Age="42" Birthday="03/07/1997"
 - e.g., Was rating "1,2,3", now rating "A, B, C"
 - e.g., discrepancy between duplicate records

Why Is Data Dirty?

- Incomplete data may come from
 - “Not applicable” data value when collected
 - Different considerations between the time when the data was collected and when it is analyzed.
 - Human/hardware/software problems
- Noisy data (incorrect values) may come from
 - Faulty data collection instruments
 - Human or computer error at data entry
 - Errors in data transmission
- Inconsistent data may come from
 - Different data sources
 - Functional dependency violation (e.g., modify some linked data)
- Duplicate records also need data cleaning

Why Is Data Preprocessing Important?

- No quality data, no quality mining results!
 - Quality decisions must be based on quality data
 - e.g., duplicate or missing data may cause incorrect or even misleading statistics.
 - Data warehouse needs consistent integration of quality data
- Data extraction, cleaning, and transformation comprises the majority of the work of building a data warehouse

Multi-Dimensional Measure of Data Quality

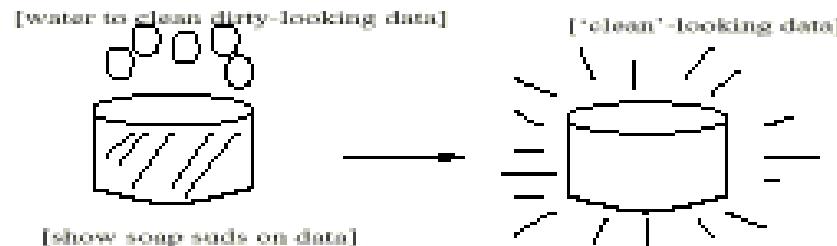
- A well-accepted multidimensional view:
 - Accuracy
 - Completeness
 - Consistency
 - Timeliness
 - Believability
 - Value added
 - Interpretability
 - Accessibility
- Broad categories:
 - Intrinsic, contextual, representational, and accessibility

Major Tasks in Data Preprocessing

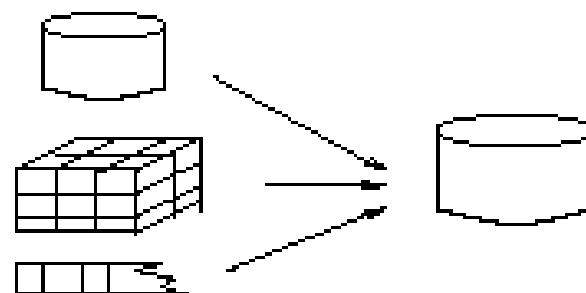
- Data cleaning
 - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration
 - Integration of multiple databases, data cubes, or files
- Data transformation
 - Normalization and aggregation
- Data reduction
 - Obtains reduced representation in volume but produces the same or similar analytical results
- Data discretization
 - Part of data reduction but with particular importance, especially for numerical data

Forms of Data Preprocessing

Data Cleaning



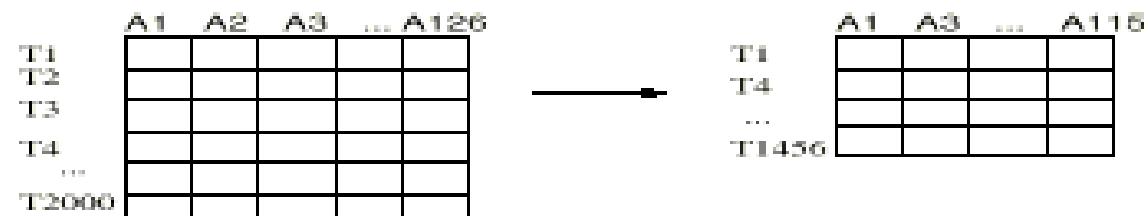
Data Integration



Data Transformation

-2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data Reduction



Data Preprocessing

- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

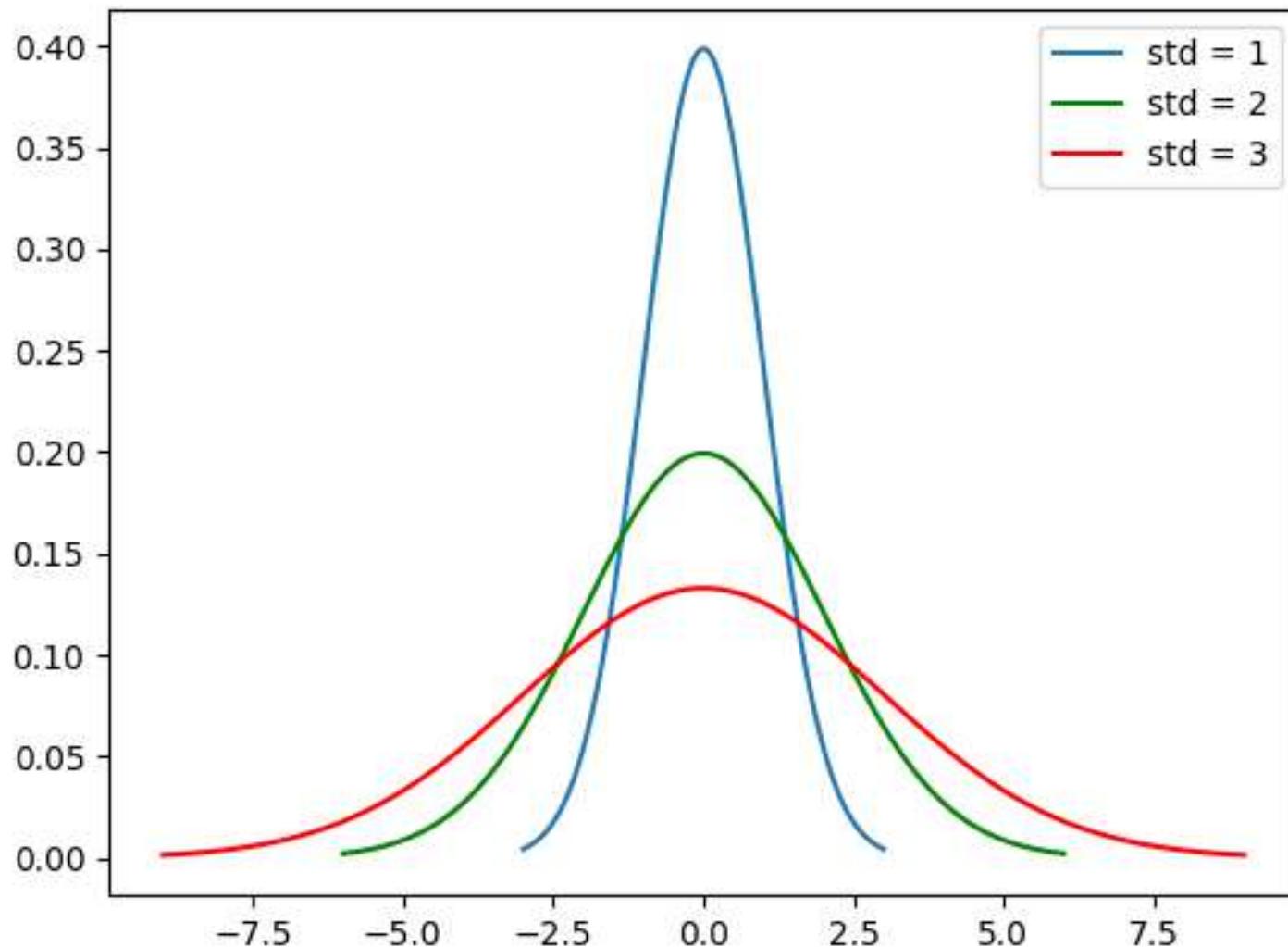
Mining Data Descriptive Characteristics

- Motivation
 - To better understand the data: central tendency, variation and spread
- Data dispersion characteristics
 - median, max, min, quantiles, outliers, variance, etc.
- Numerical dimensions correspond to sorted intervals
 - Data dispersion: analyzed with multiple granularities of precision
 - Boxplot or quantile analysis on sorted intervals
- Dispersion analysis on computed measures
 - Folding measures into numerical dimensions
 - Boxplot or quantile analysis on the transformed cube

Measuring the Central Tendency

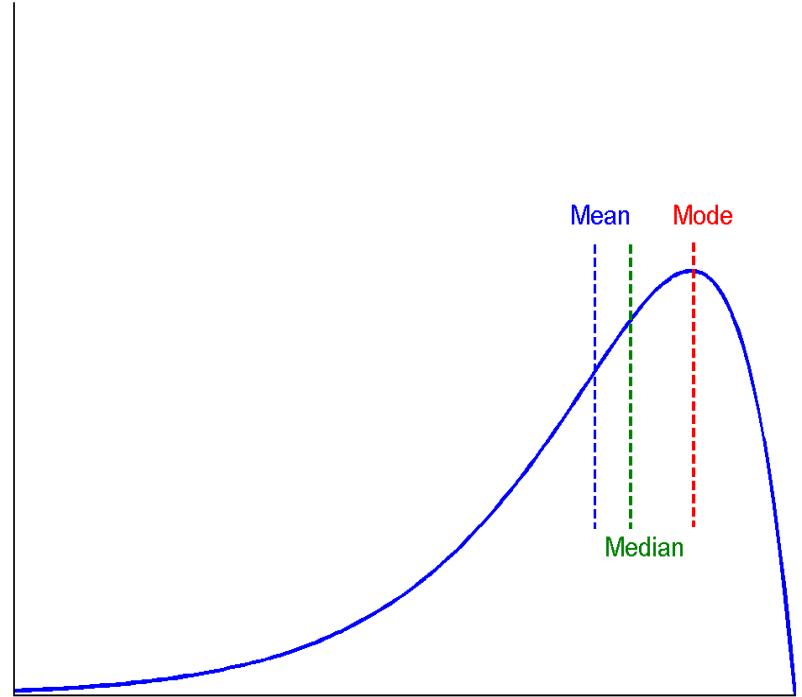
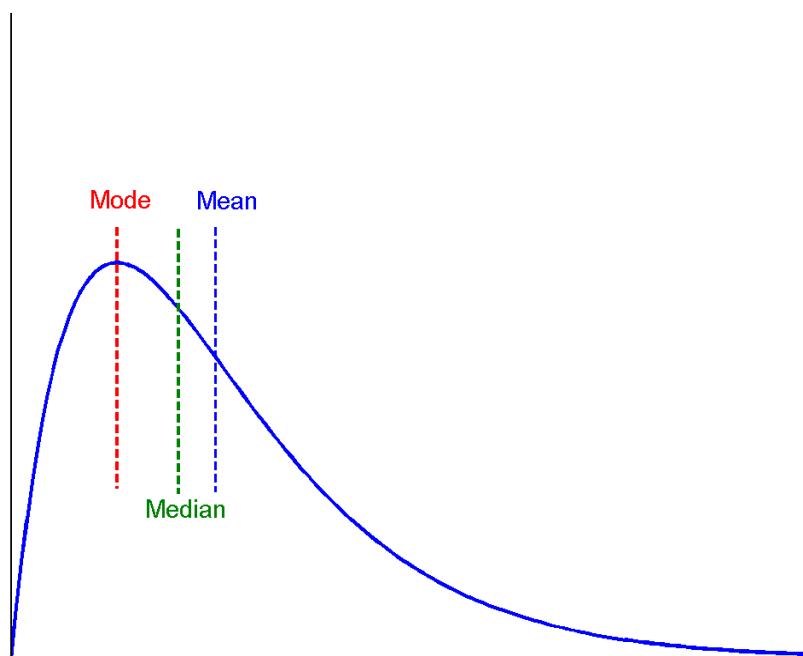
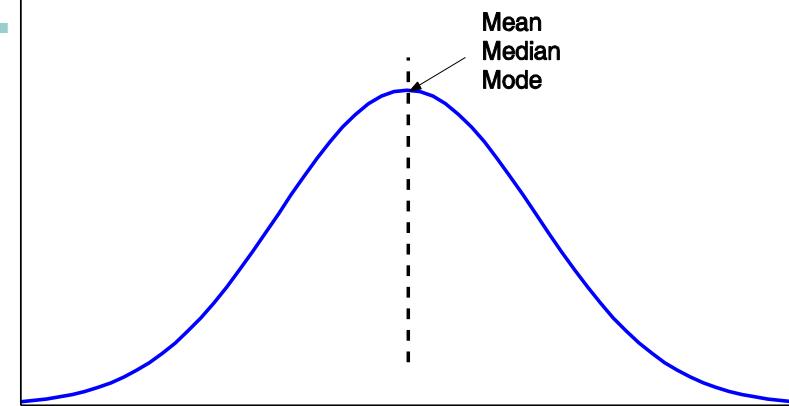
- Mean (algebraic measure) (sample vs. population): $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ $\mu = \frac{\sum x}{N}$
 - Weighted arithmetic mean:
 - Trimmed mean: chopping extreme values
$$\bar{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$
- Median: A holistic measure
 - Middle value if odd number of values, or average of the middle two values otherwise
 - Estimated by interpolation (for *grouped data*):
$$median = L_1 + \left(\frac{n/2 - (\sum f)l}{f_{median}} \right) c$$
- Mode
 - Value that occurs most frequently in the data
 - Unimodal, bimodal, trimodal
 - Empirical formula: $mean - mode = 3 \times (mean - median)$

Effect of standard deviation on distribution

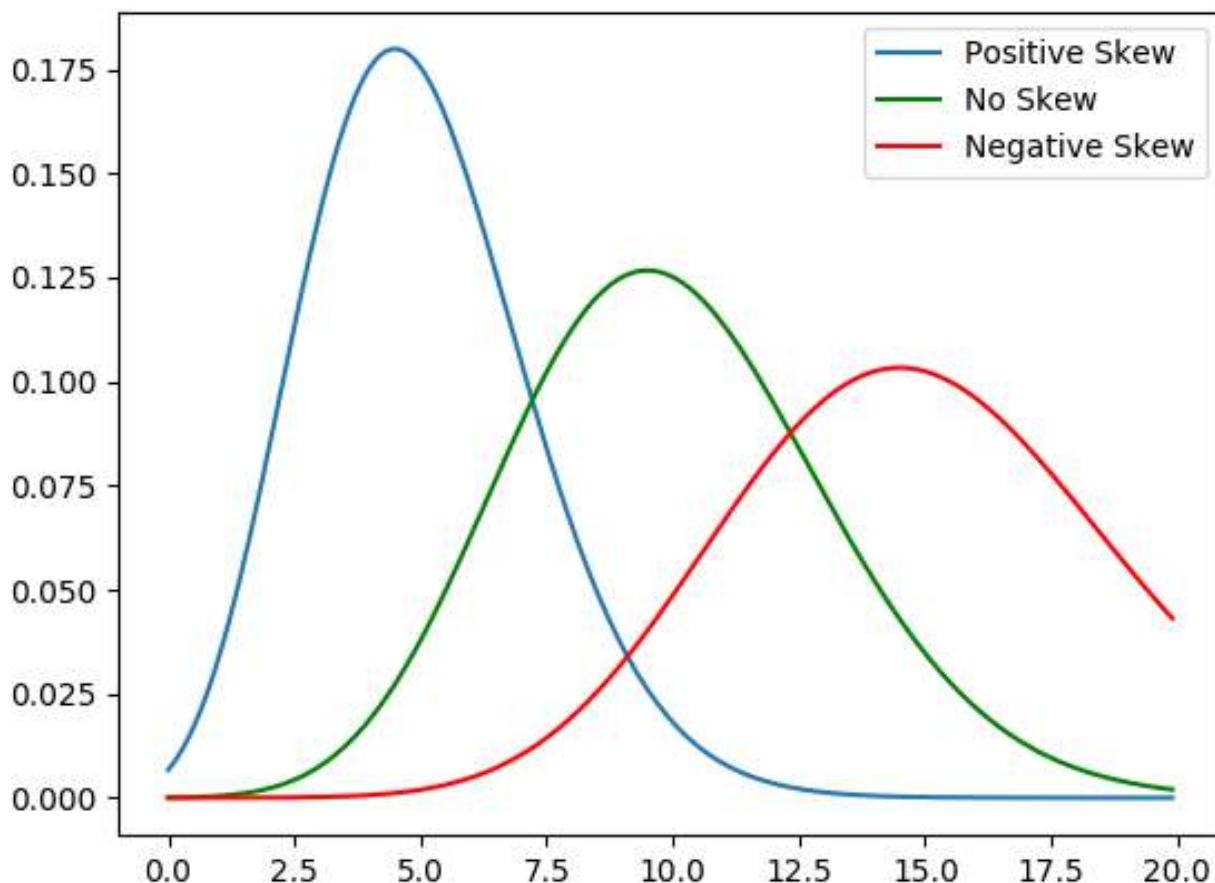


Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data



Skewness Measure



$$skewness = \frac{3 * (Mean - Median)}{S.D.}$$

Measuring the Dispersion of Data

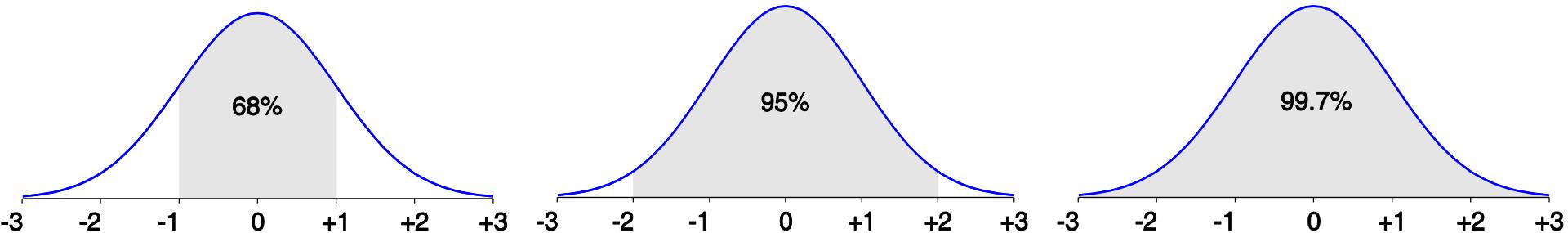
- Quartiles, outliers and boxplots
 - **Quartiles**: Q_1 (25^{th} percentile), Q_3 (75^{th} percentile)
 - **Inter-quartile range**: $\text{IQR} = Q_3 - Q_1$
 - **Five number summary**: min, Q_1 , M, Q_3 , max
 - **Boxplot**: ends of the box are the quartiles, median is marked, whiskers, and plot outlier individually
 - **Outlier**: usually, a value higher/lower than $1.5 \times \text{IQR}$
- Variance and standard deviation (*sample: s, population: σ*)
 - **Variance**: (algebraic, scalable computation)

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - \frac{1}{n} (\sum_{i=1}^n x_i)^2 \right] \quad \sigma^2 = \frac{1}{N} \sum_{i=1}^n (x_i - \mu)^2 = \frac{1}{N} \sum_{i=1}^n x_i^2 - \mu^2$$

- **Standard deviation** s (or σ) is the square root of variance s^2 (or σ^2)

Properties of Normal Distribution Curve

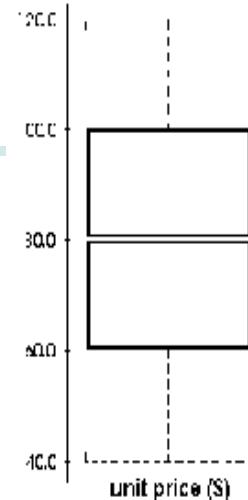
- The normal (distribution) curve
 - From $\mu-\sigma$ to $\mu+\sigma$: contains about 68% of the measurements (μ : mean, σ : standard deviation)
 - From $\mu-2\sigma$ to $\mu+2\sigma$: contains about 95% of it
 - From $\mu-3\sigma$ to $\mu+3\sigma$: contains about 99.7% of it



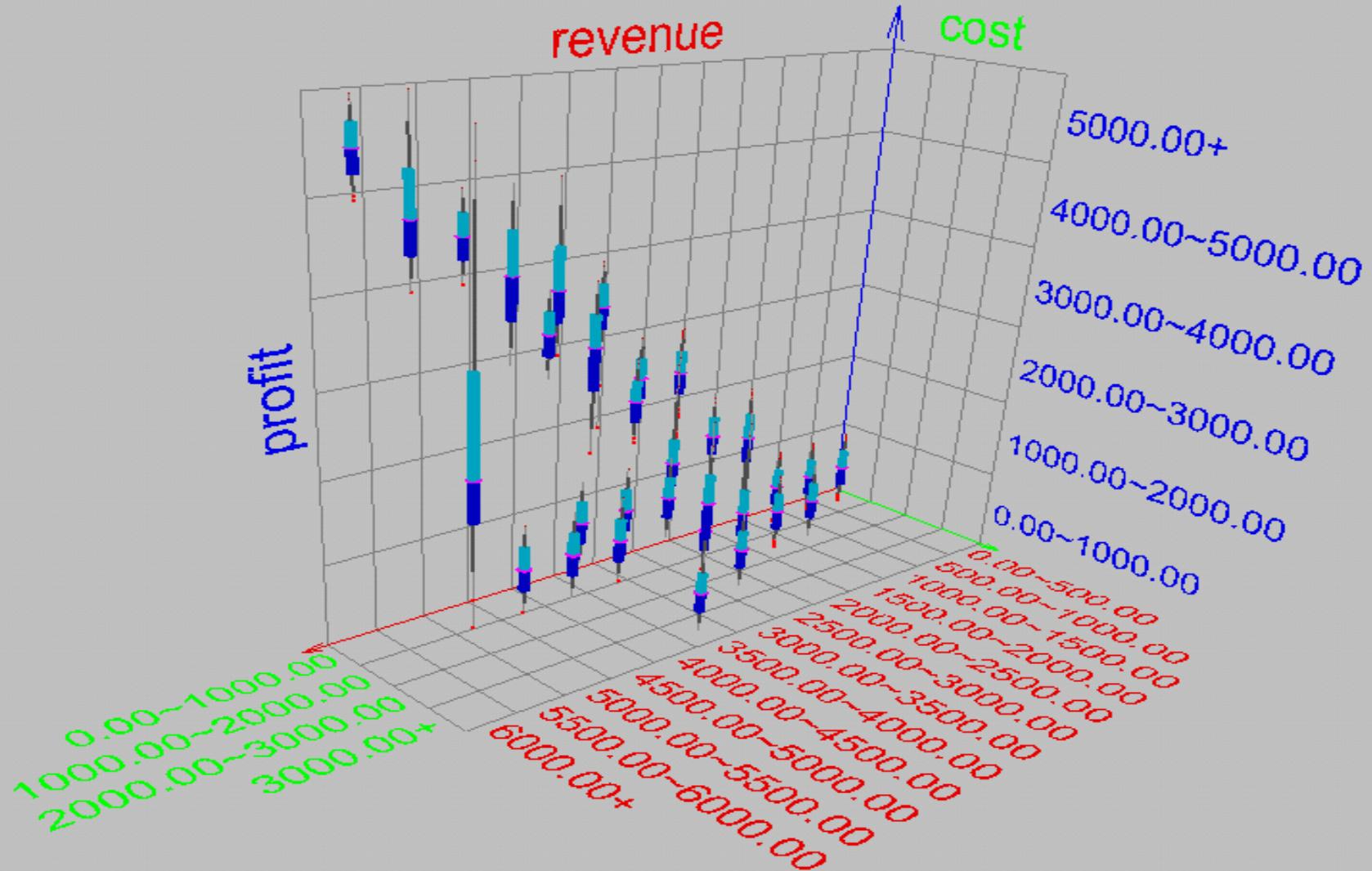
Boxplot Analysis

- Five-number summary of a distribution:
Minimum, Q1, M, Q3, Maximum

- Boxplot
 - Data is represented with a box
 - The ends of the box are at the first and third quartiles, i.e., the height of the box is IRQ
 - The median is marked by a line within the box
 - Whiskers: two lines outside the box extend to Minimum and Maximum

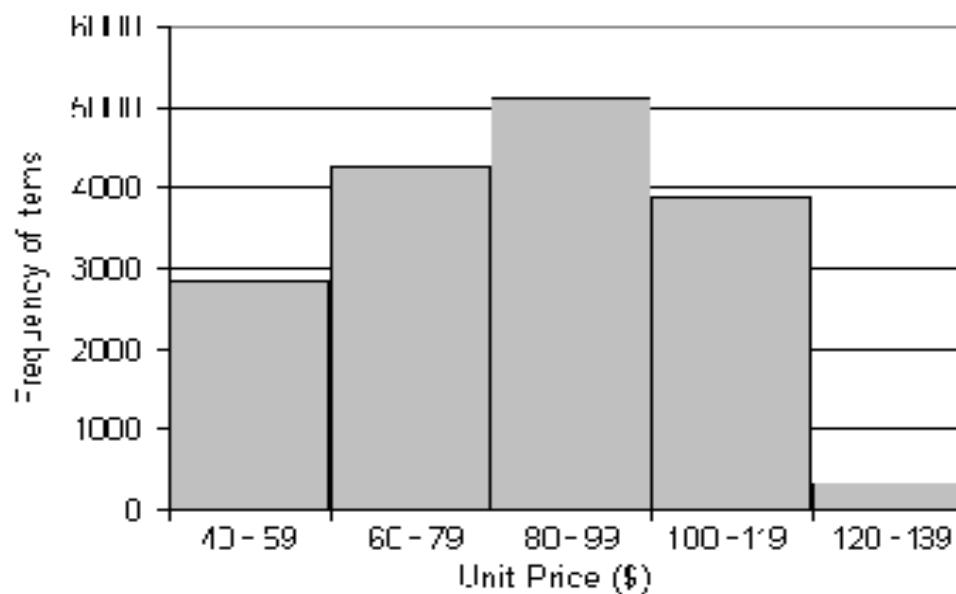


Visualization of Data Dispersion: Boxplot Analysis



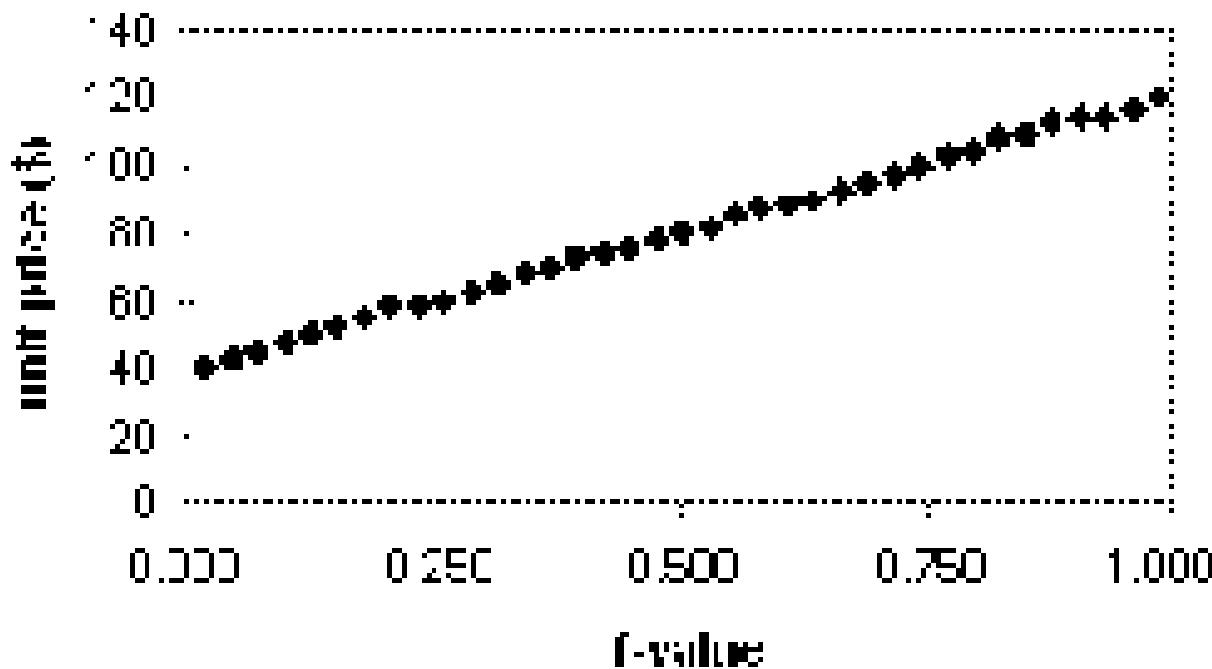
Histogram Analysis

- Graph displays of basic statistical class descriptions
 - Frequency histograms
 - A univariate graphical method
 - Consists of a set of rectangles that reflect the counts or frequencies of the classes present in the given data



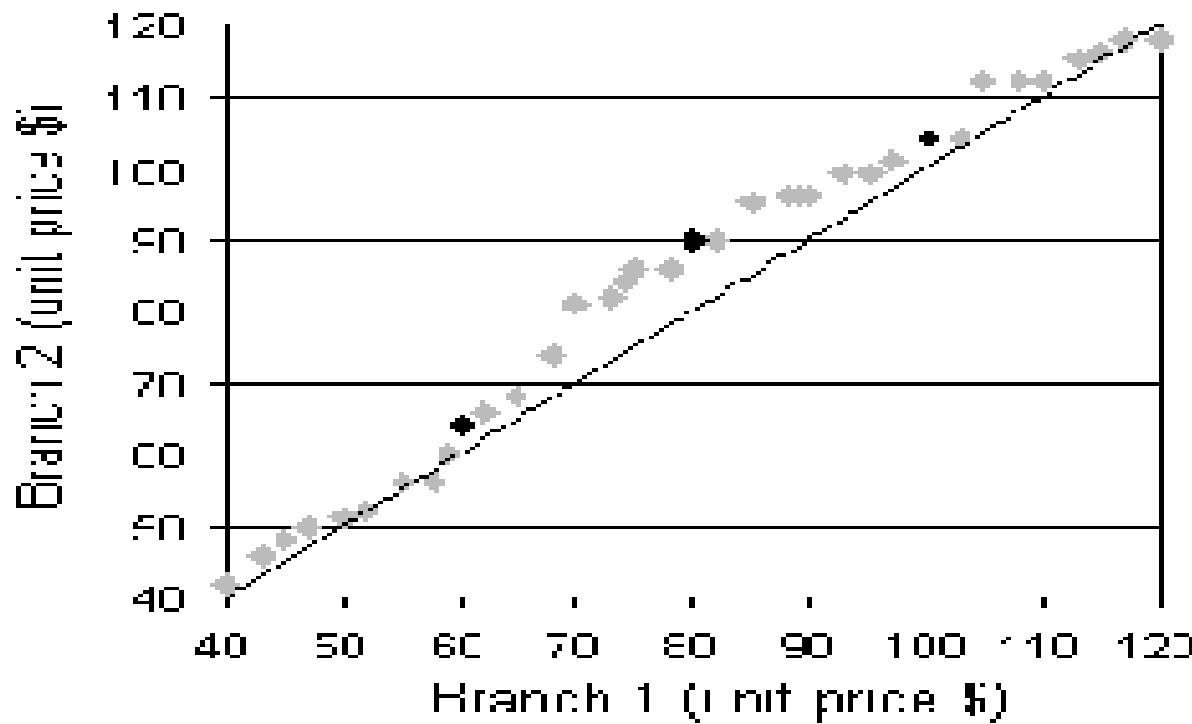
Quantile Plot

- Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)
- Plots **quantile** information
 - For a data x_i , data sorted in increasing order, f_i indicates that approximately $100 f_i\%$ of the data are below or equal to the value x_i .



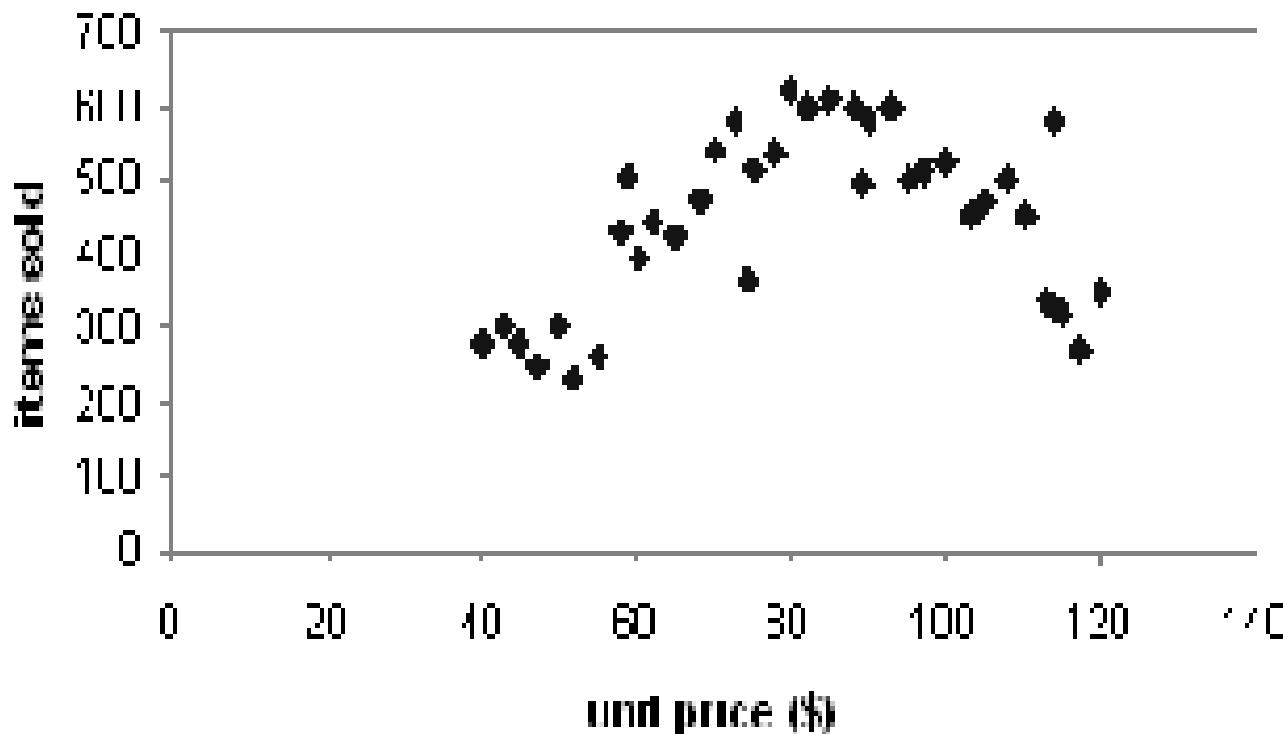
Quantile-Quantile (Q-Q) Plot

- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- Allows the user to view whether there is a shift in going from one distribution to another



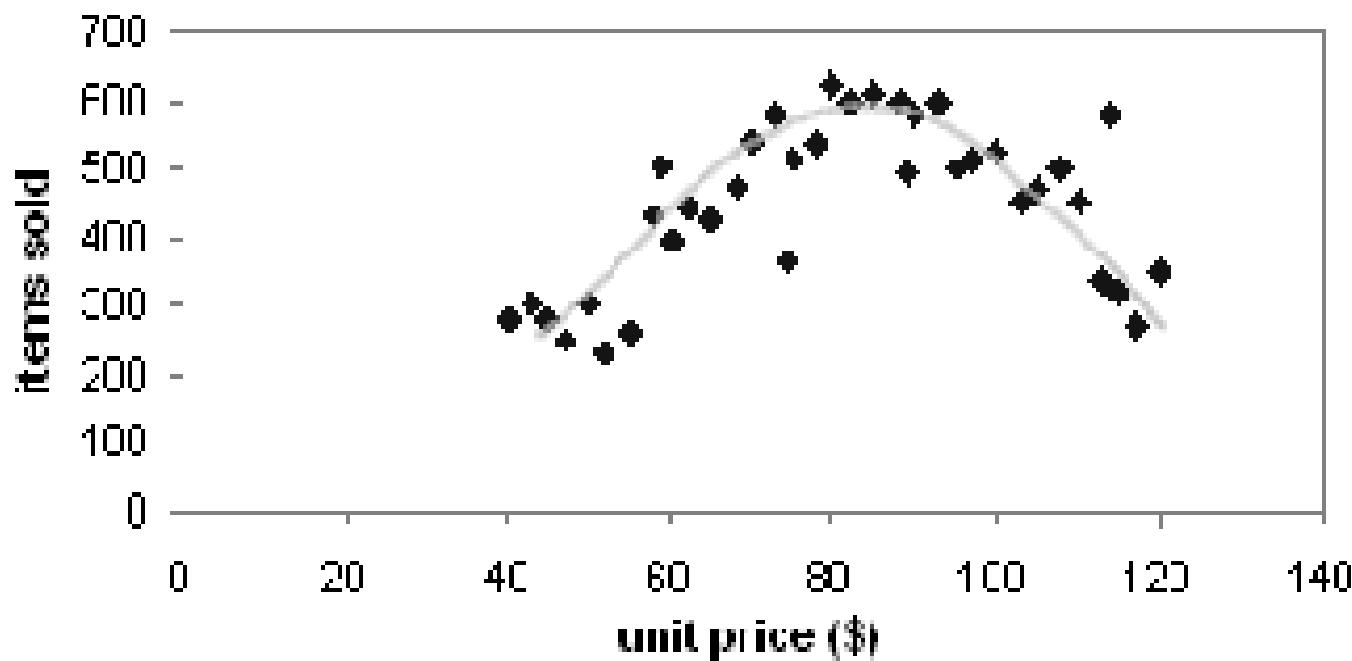
Scatter plot

- 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

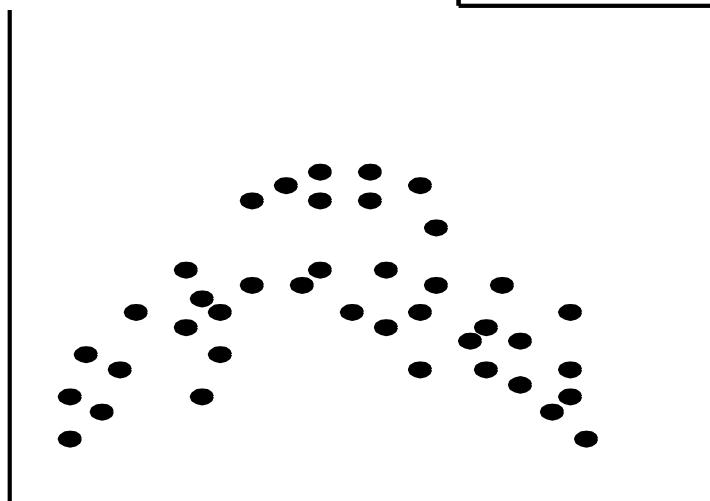
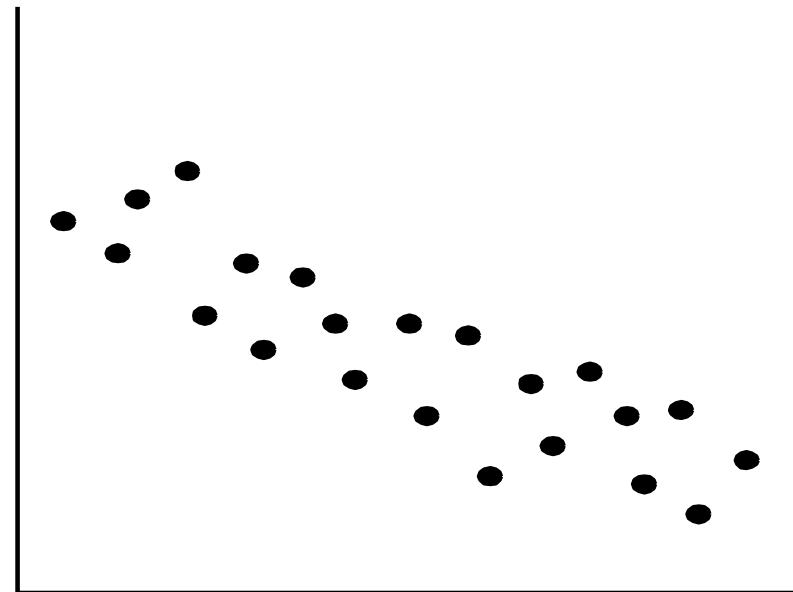
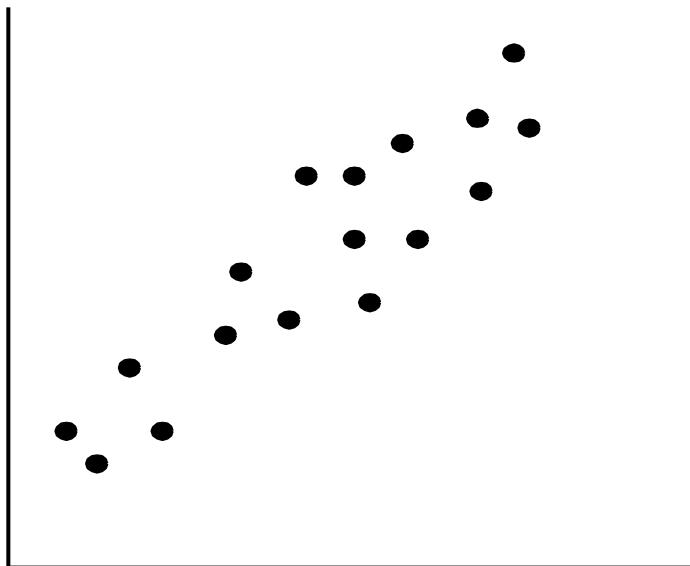


Loess Curve

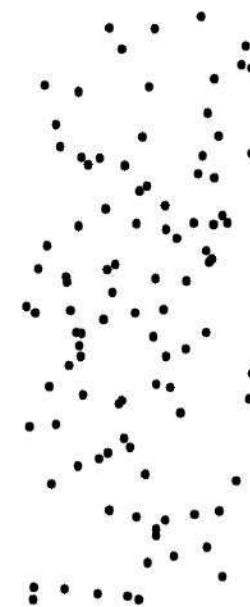
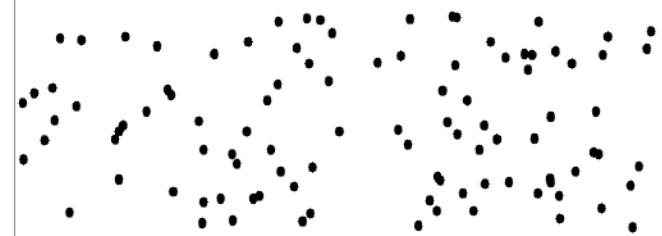
- Adds a smooth curve to a scatter plot in order to provide better perception of the pattern of dependence
- Loess curve is fitted by setting two parameters: a smoothing parameter, and the degree of the polynomials that are fitted by the regression



Positively and Negatively Correlated Data



Not Correlated Data



Graphic Displays of Basic Statistical Descriptions

- Histogram: (shown before)
- Boxplot: (covered before)
- Quantile plot: each value x_i is paired with f_i indicating that approximately $100 f_i\%$ of data are $\leq x_i$
- Quantile-quantile (q-q) plot: graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- Scatter plot: each pair of values is a pair of coordinates and plotted as points in the plane
- Loess (local regression) curve: add a smooth curve to a scatter plot to provide better perception of the pattern of dependence

Data Preprocessing

- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

Data Cleaning

- Importance
 - “Data cleaning is one of the three biggest problems in data warehousing”—Ralph Kimball
 - “Data cleaning is the number one problem in data warehousing”—DCI survey
- Data cleaning tasks
 - Fill in missing values
 - Identify outliers and smooth out noisy data
 - Correct inconsistent data
 - Resolve redundancy caused by data integration

Missing Data

- Data is not always available
 - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
 - equipment malfunction
 - inconsistent with other recorded data and thus deleted
 - data not entered due to misunderstanding
 - certain data may not be considered important at the time of entry
 - not register history or changes of the data
- Missing data may need to be inferred.

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (assuming the tasks in classification—not effective when the percentage of missing values per attribute varies considerably.)
- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically with
 - a global constant : e.g., “unknown”, a new class?!
 - the attribute mean
 - the attribute mean for all samples belonging to the same class: smarter
 - the most probable value: inference-based such as Bayesian formula or decision tree

Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may due to
 - faulty data collection instruments
 - data entry problems
 - data transmission problems
 - technology limitation
 - inconsistency in naming convention
- Other data problems which requires data cleaning
 - duplicate records
 - incomplete data
 - inconsistent data

How to Handle Noisy Data?

- Binning
 - first sort data and partition into (equal-frequency) bins
 - then one can **smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.**
- Regression
 - smooth by fitting the data into regression functions
- Clustering
 - detect and remove outliers
- Combined computer and human inspection
 - detect suspicious values and check by human (e.g., deal with possible outliers)

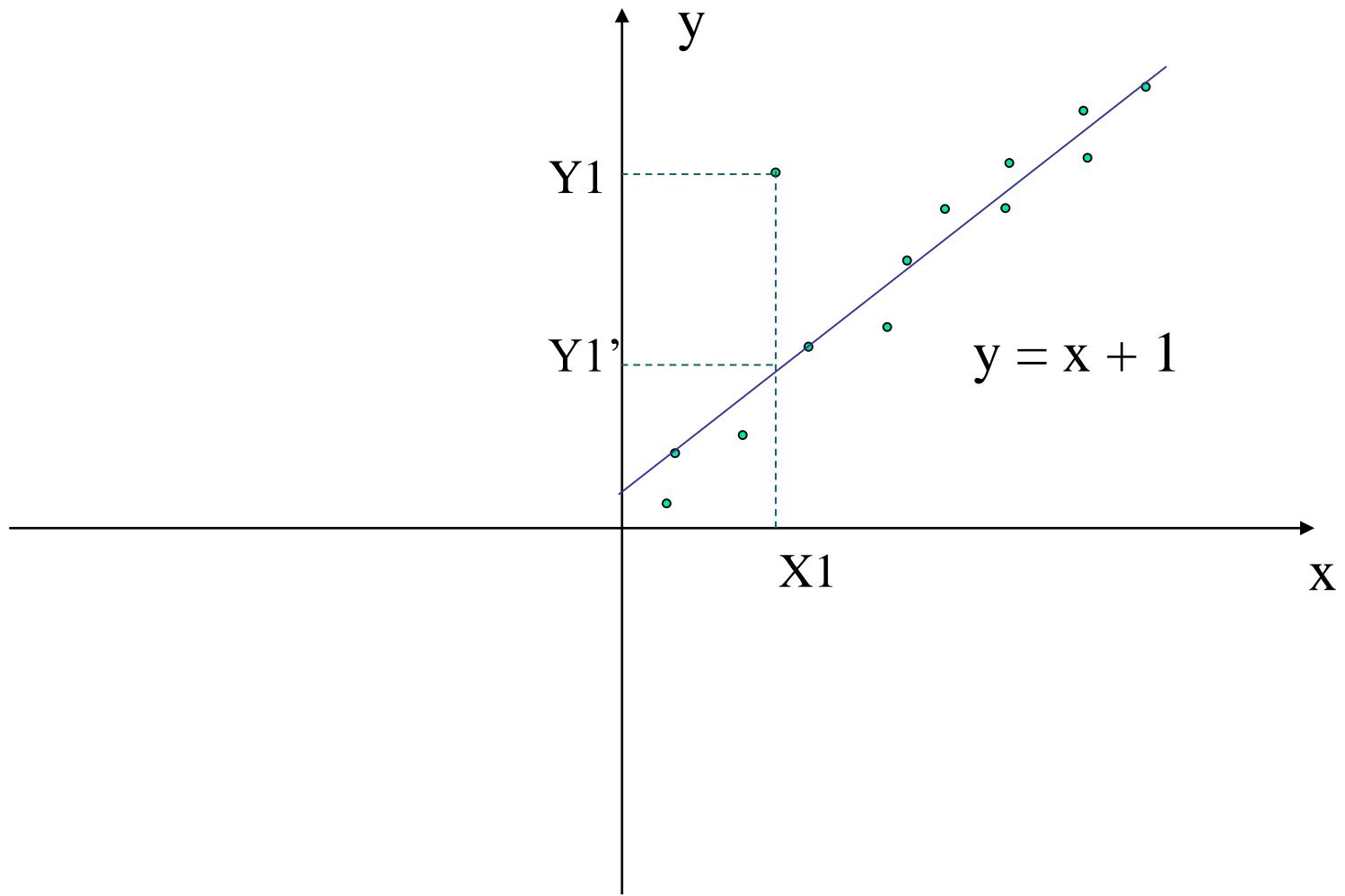
Simple Discretization Methods: Binning

- **Equal-width** (distance) partitioning
 - Divides the range into N intervals of equal size: uniform grid
 - if A and B are the lowest and highest values of the attribute, the width of intervals will be: $W = (B - A)/N$.
 - The most straightforward, but outliers may dominate presentation
 - Skewed data is not handled well
- **Equal-depth** (frequency) partitioning
 - Divides the range into N intervals, each containing approximately same number of samples
 - Good data scaling
 - Managing categorical attributes can be tricky

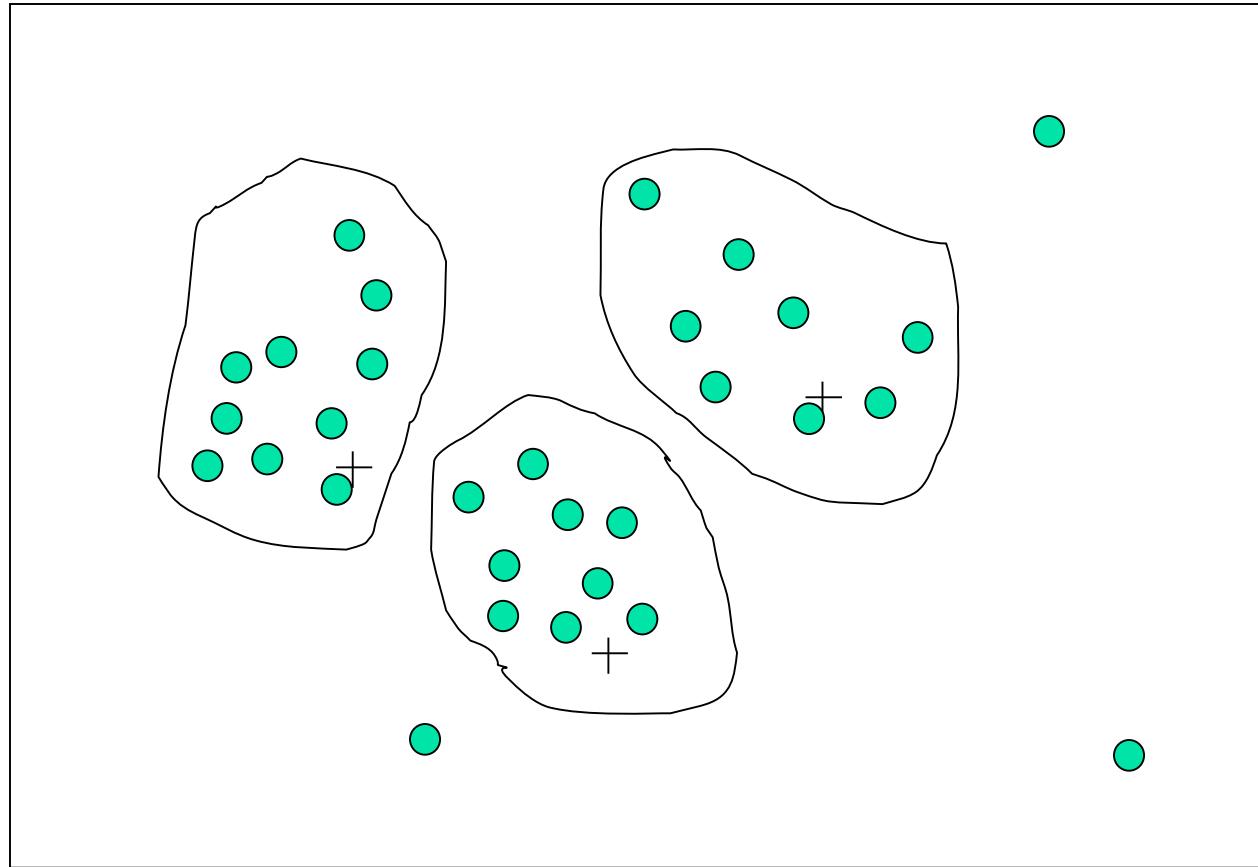
Binning Methods for Data Smoothing

- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
 - * Partition into equal-frequency (equi-depth) bins:
 - Bin 1: 4, 8, 9, 15
 - Bin 2: 21, 21, 24, 25
 - Bin 3: 26, 28, 29, 34
 - * Smoothing by bin means:
 - Bin 1: 9, 9, 9, 9
 - Bin 2: 23, 23, 23, 23
 - Bin 3: 29, 29, 29, 29
 - * Smoothing by bin boundaries:
 - Bin 1: 4, 4, 4, 15
 - Bin 2: 21, 21, 25, 25
 - Bin 3: 26, 26, 26, 34

Regression



Cluster Analysis



Data Cleaning as a Process

- Data discrepancy detection
 - Use metadata (e.g., domain, range, dependency, distribution)
 - Check field overloading
 - Check uniqueness rule, consecutive rule and null rule
 - Use commercial tools
 - Data scrubbing: use simple domain knowledge (e.g., postal code, spell-check) to detect errors and make corrections
 - Data auditing: by analyzing data to discover rules and relationship to detect violators (e.g., correlation and clustering to find outliers)
- Data migration and integration
 - Data migration tools: allow transformations to be specified
 - ETL (Extraction/Transformation>Loading) tools: allow users to specify transformations through a graphical user interface
- Integration of the two processes
 - Iterative and interactive (e.g., Potter's Wheels)

Data Preprocessing

- Why preprocess the data?
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

Data Integration

- Data integration:
 - Combines data from multiple sources into a coherent store
- Schema integration: e.g., A.cust-id \equiv B.cust-#
 - Integrate metadata from different sources
- Entity identification problem:
 - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
- Detecting and resolving data value conflicts
 - For the same real world entity, attribute values from different sources are different
 - Possible reasons: different representations, different scales, e.g., metric vs. British units

Handling Redundancy in Data Integration

- Redundant data occur often when integration of multiple databases
 - *Object identification:* The same attribute or object may have different names in different databases
 - *Derivable data:* One attribute may be a “derived” attribute in another table, e.g., annual revenue
- Redundant attributes may be able to be detected by *correlation analysis*
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Correlation Analysis (Numerical Data)

- Correlation coefficient (also called Pearson's product moment coefficient)

$$r_{A,B} = \frac{\sum (A - \bar{A})(B - \bar{B})}{(n-1)\sigma_A\sigma_B} = \frac{\sum (AB) - n\bar{A}\bar{B}}{(n-1)\sigma_A\sigma_B}$$

where n is the number of tuples, \bar{A} and \bar{B} are the respective means of A and B, σ_A and σ_B are the respective standard deviation of A and B, and $\Sigma(AB)$ is the sum of the AB cross-product.

- If $r_{A,B} > 0$, A and B are positively correlated (A's values increase as B's). The higher, the stronger correlation.
- $r_{A,B} = 0$: independent; $r_{A,B} < 0$: negatively correlated

Correlation Analysis (Categorical Data)

- χ^2 (chi-square) test

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- The larger the χ^2 value, the more likely the variables are related
- The cells that contribute the most to the χ^2 value are those whose actual count is very different from the expected count
- Correlation does not imply causality
 - # of hospitals and # of car-theft in a city are correlated
 - Both are causally linked to the third variable: population

Chi-Square Calculation: An Example

	Play chess	Not play chess	Sum (row)
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

- χ^2 (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- It shows that like_science_fiction and play_chess are correlated in the group

Data Transformation

- Smoothing: remove noise from data
- Aggregation: summarization, data cube construction
- Generalization: concept hierarchy climbing
- Normalization: scaled to fall within a small, specified range
 - min-max normalization
 - z-score normalization
 - normalization by decimal scaling
- Attribute/feature construction
 - New attributes constructed from the given ones

Data Transformation: Normalization

- Min-max normalization: to $[new_min_A, new_max_A]$

$$v' = \frac{v - min_A}{max_A - min_A} (new_max_A - new_min_A) + new_min_A$$

- Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,000 is mapped to $\frac{73,600 - 12,000}{98,000 - 12,000}(1.0 - 0) + 0 = 0.716$
- Z-score normalization (μ : mean, σ : standard deviation):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

- Ex. Let $\mu = 54,000$, $\sigma = 16,000$. Then $\frac{73,600 - 54,000}{16,000} = 1.225$
- Normalization by decimal scaling

$$v' = \frac{v}{10^j} \quad \text{Where } j \text{ is the smallest integer such that } \text{Max}(|v'|) < 1$$

Data Preprocessing

- Why preprocess the data?
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

Data Reduction Strategies

- Why data reduction?
 - A database/data warehouse may store terabytes of data
 - Complex data analysis/mining may take a very long time to run on the complete data set
- Data reduction
 - Obtain a reduced representation of the data set that is much smaller in volume but yet produce the same (or almost the same) analytical results
- Data reduction strategies
 - Dimensionality reduction — e.g., remove unimportant attributes
 - Data Compression
 - Numerosity reduction — e.g., fit data into models
 - Discretization and concept hierarchy generation

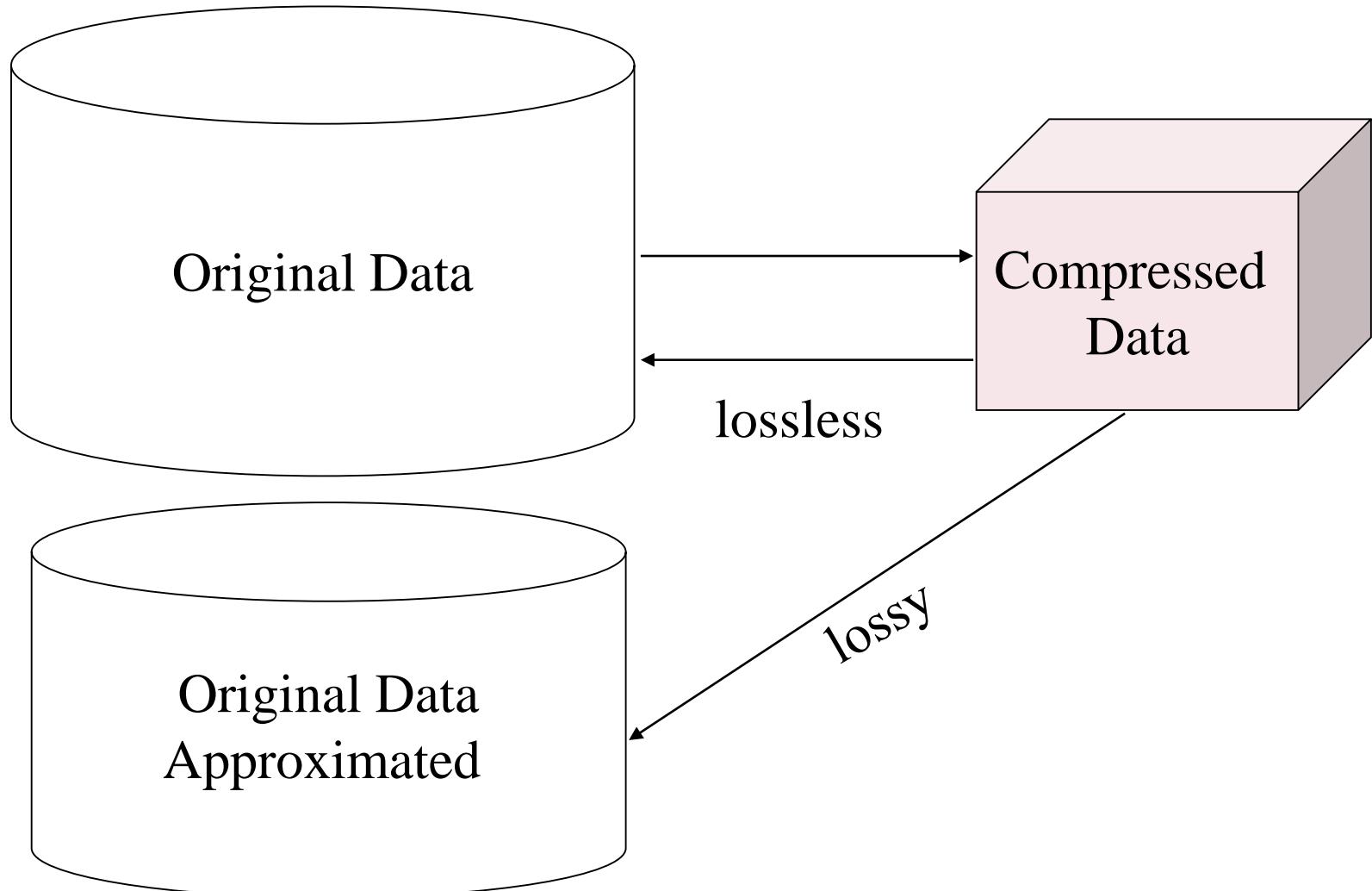
Heuristic Feature Selection Methods

- There are 2^d possible sub-features of d features
- Several heuristic feature selection methods:
 - Best single features under the feature independence assumption: choose by significance tests
 - Best step-wise feature selection:
 - The best single-feature is picked first
 - Then next best feature condition to the first, ...
 - Step-wise feature elimination:
 - Repeatedly eliminate the worst feature
 - Best combined feature selection and elimination
 - Optimal branch and bound:
 - Use feature elimination and backtracking

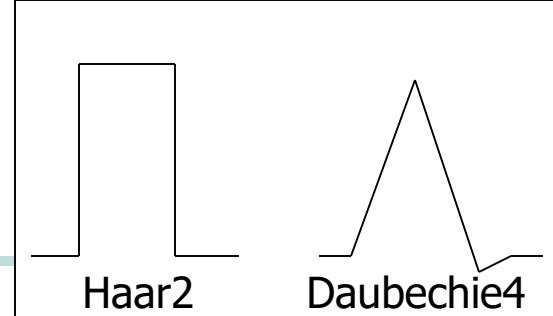
Data Compression

- String compression
 - There are extensive theories and well-tuned algorithms
 - Typically lossless
 - But only limited manipulation is possible without expansion
- Audio/video compression
 - Typically lossy compression, with progressive refinement
 - Sometimes small fragments of signal can be reconstructed without reconstructing the whole
- Time sequence is not audio
 - Typically short and vary slowly with time

Data Compression

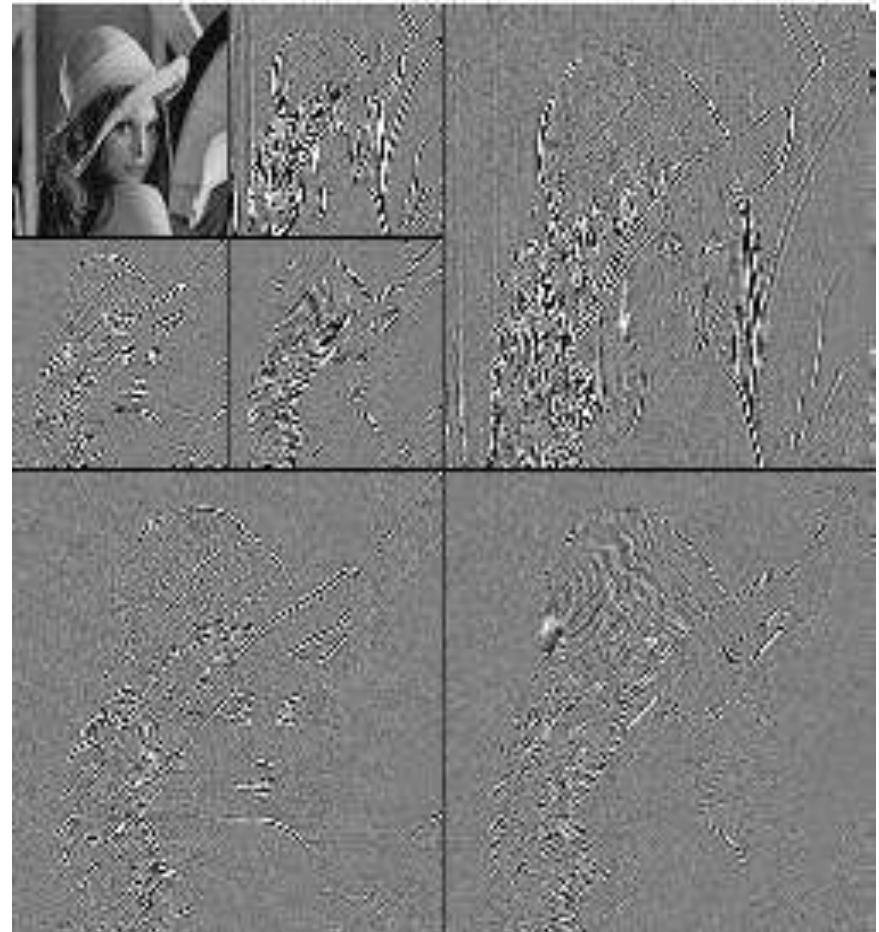
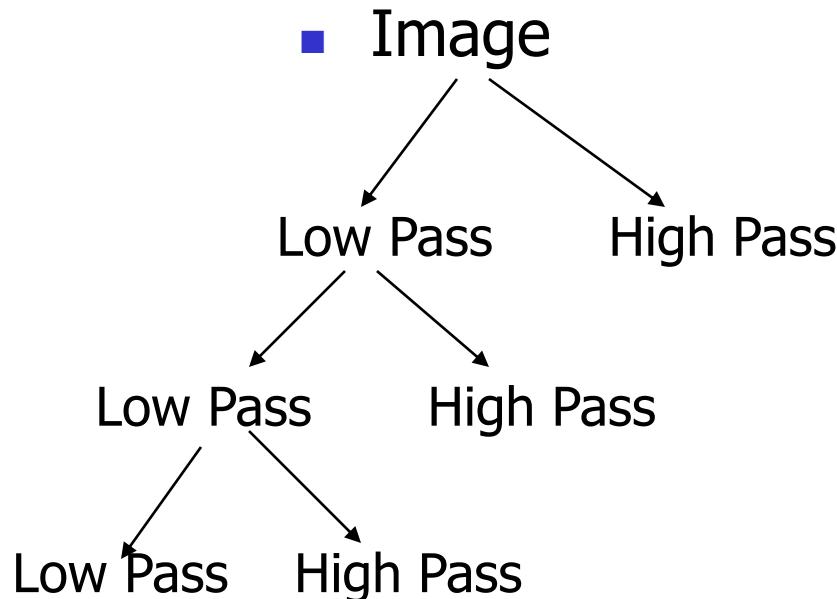


Dimensionality Reduction: Wavelet Transformation



- Discrete wavelet transform (DWT): linear signal processing, multi-resolutonal analysis
- Compressed approximation: store only a small fraction of the strongest of the wavelet coefficients
- Similar to discrete Fourier transform (DFT), but better lossy compression, localized in space
- Method:
 - Length, L , must be an integer power of 2 (padding with 0's, when necessary)
 - Each transform has 2 functions: smoothing, difference
 - Applies to pairs of data, resulting in two set of data of length $L/2$
 - Applies two functions recursively, until reaches the desired length

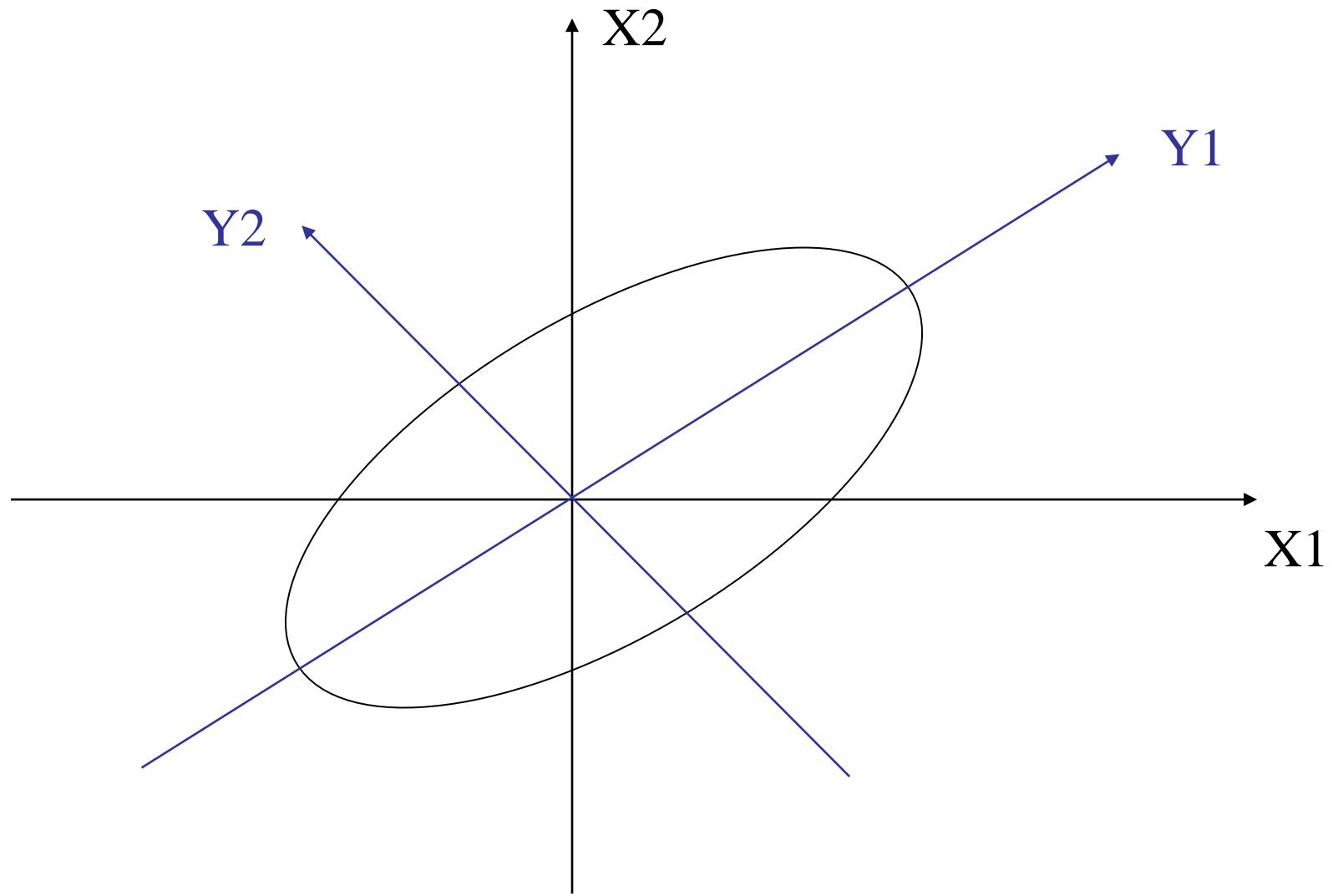
DWT for Image Compression



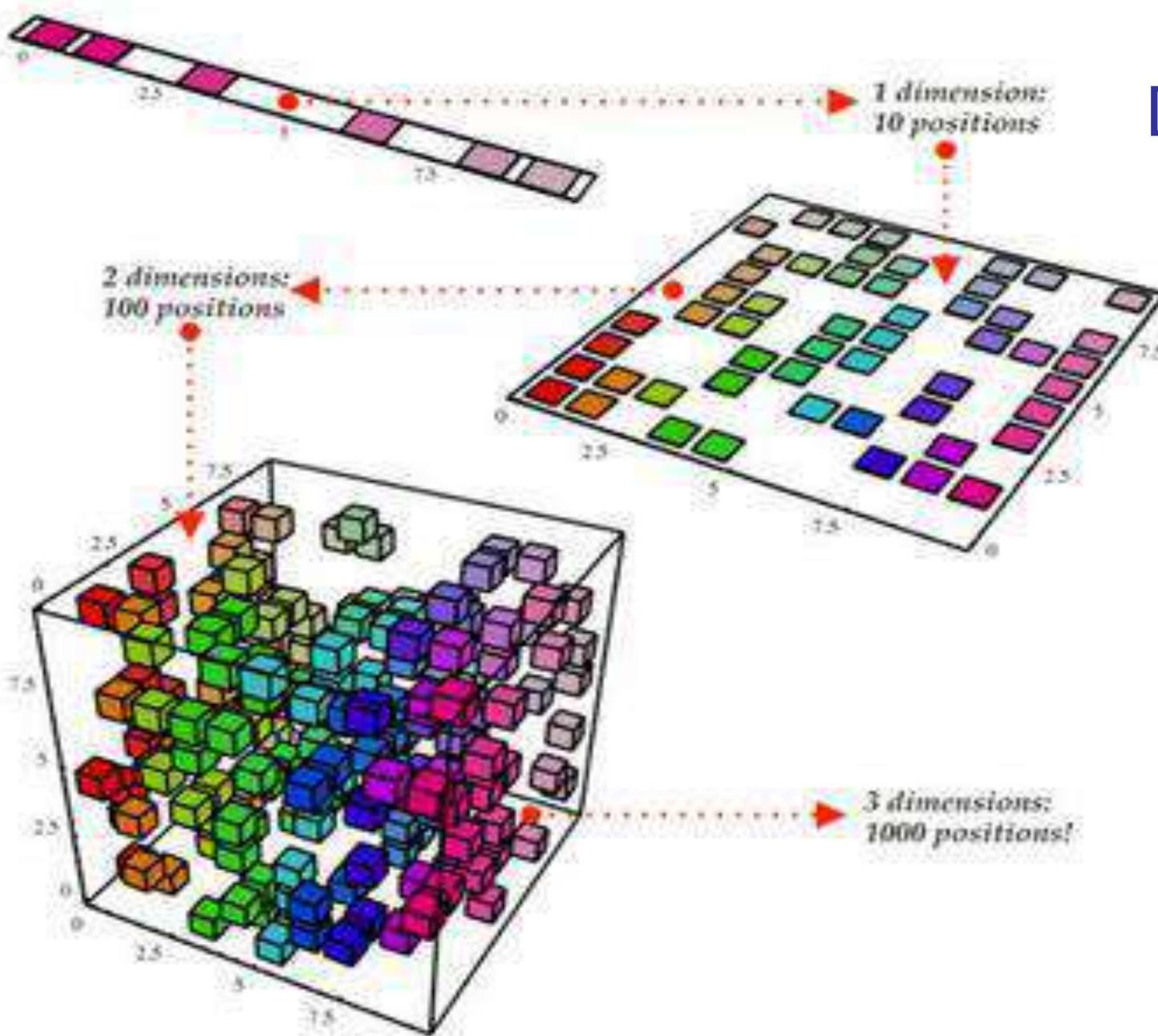
Dimensionality Reduction: Principal Component Analysis (PCA)

- Given N data vectors from n -dimensions, find $k \leq n$ orthogonal vectors (*principal components*) that can be best used to represent data
- Steps
 - Normalize input data: Each attribute falls within the same range
 - Compute k orthonormal (unit) vectors, i.e., *principal components*
 - Each input data (vector) is a linear combination of the k principal component vectors
 - The principal components are sorted in order of decreasing “significance” or strength
 - Since the components are sorted, the size of the data can be reduced by eliminating the weak components, i.e., those with low variance. (i.e., using the strongest principal components, it is possible to reconstruct a good approximation of the original data
- Works for numeric data only
- Used when the number of dimensions is large

Principal Component Analysis



PCA & Dimension Reduction



Numerosity Reduction

- Reduce data volume by choosing alternative, smaller forms of data representation
- Parametric methods
 - Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)
 - Example: Log-linear models—obtain value at a point in m-D space as the product on appropriate marginal subspaces
- Non-parametric methods
 - Do not assume models
 - Major families: histograms, clustering, sampling

Data Reduction Method (1): Regression and Log-Linear Models

- **Linear regression**: Data are modeled to fit a straight line
 - Often uses the least-square method to fit the line
 - $\mathbf{Y} = \mathbf{w} \mathbf{X} + \mathbf{b}$: Two regression coefficients, w and b , specify the line, to be estimated by using the data at hand
 - Using the least squares criterion to the known values of $Y_1, Y_2, \dots, X_1, X_2, \dots$
- **Multiple regression**: allows a response variable Y to be modeled as a linear function of multidimensional feature vector
 - Many nonlinear functions can be transformed into

$$\mathbf{Y} = \mathbf{b}_0 + \mathbf{b}_1 \mathbf{X}_1 + \mathbf{b}_2 \mathbf{X}_2 + \dots$$

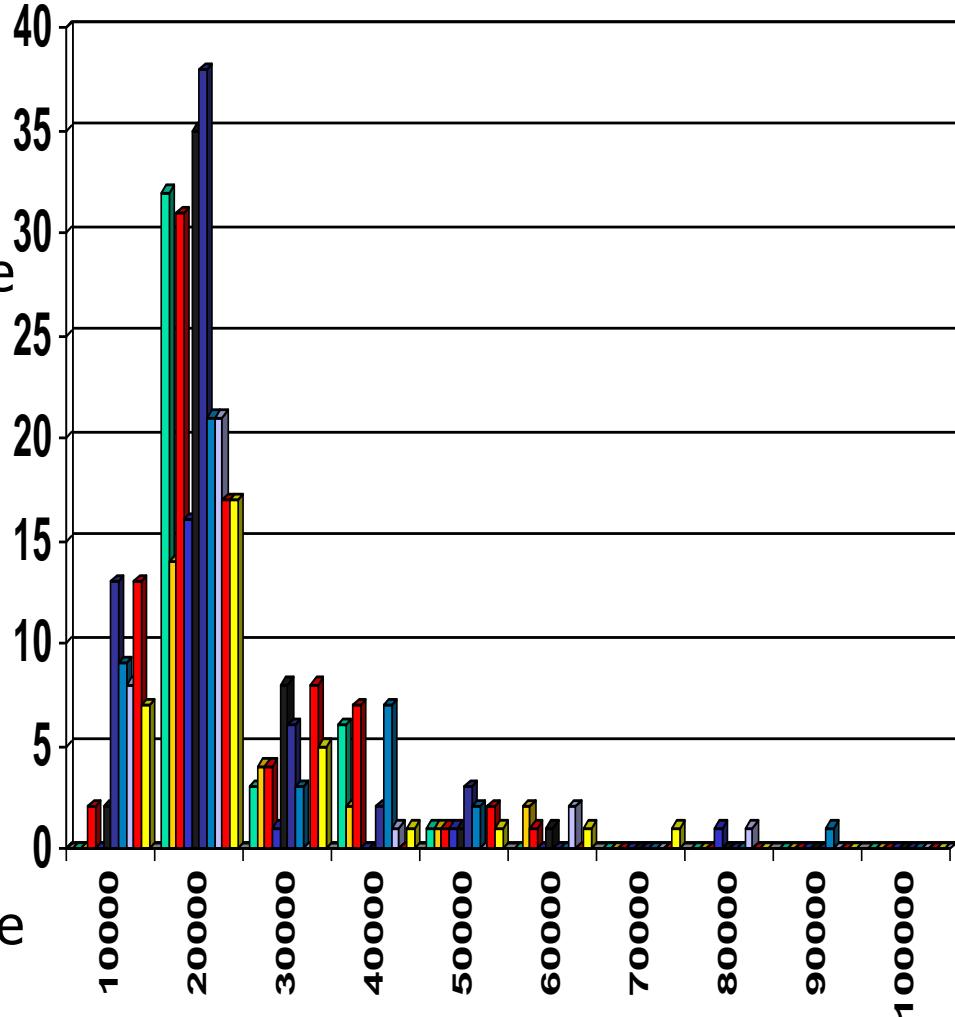
Data Reduction Method (1): Regression and Log-Linear Models

- Log-linear models:

- Log-linear model: approximates discrete multidimensional probability distributions
- The multi-way table of joint probabilities is approximated by a product of lower-order tables
- Probability: $p(a, b, c, d) = \alpha_{ab} \beta_{ac} \chi_{ad} \delta_{bcd}$

Data Reduction Method (2): Histograms

- Divide data into buckets and store average (sum) for each bucket
- Partitioning rules:
 - Equal-width: equal bucket range
 - Equal-frequency (or equal-depth)
 - V-optimal: with the least *histogram variance* (weighted sum of the original values that each bucket represents)
 - MaxDiff: set bucket boundary between each pair for pairs have the $\beta-1$ largest differences



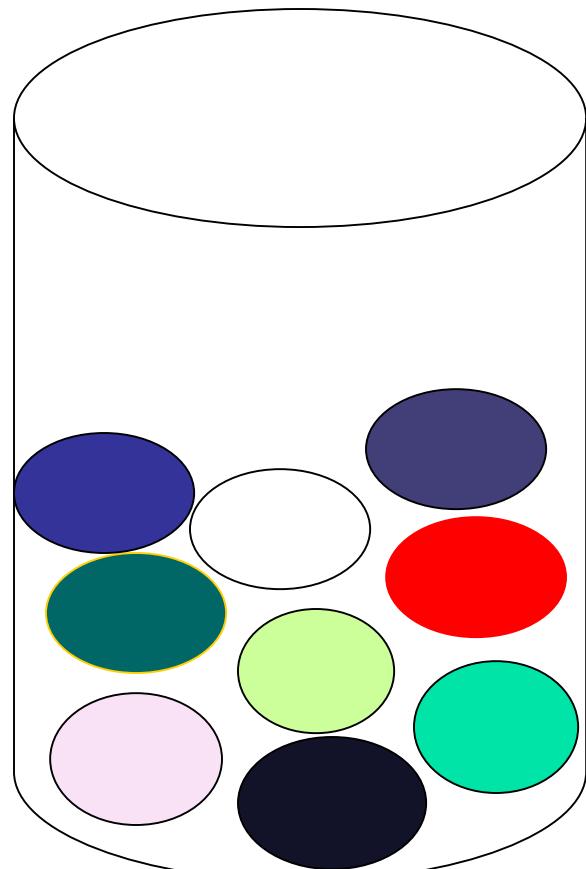
Data Reduction Method (3): Clustering

- Partition data set into clusters based on similarity, and store cluster representation (e.g., centroid and diameter) only
- Can be very effective if data is clustered but not if data is “smeared”
- Can have hierarchical clustering and be stored in multi-dimensional index tree structures
- There are many choices of clustering definitions and clustering algorithms
- Cluster analysis will be studied later in depth in later chapter

Data Reduction Method (4): Sampling

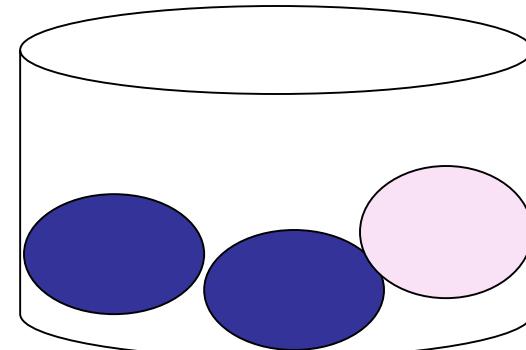
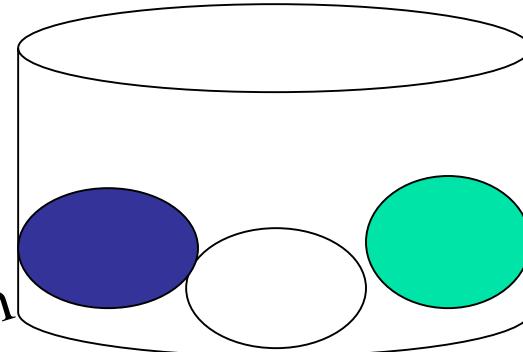
- Sampling: obtaining a small sample s to represent the whole data set N
- Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the data
- Choose a **representative** subset of the data
 - Simple random sampling may have very poor performance in the presence of skew
- Develop adaptive sampling methods
 - Stratified sampling:
 - Approximate the percentage of each class (or subpopulation of interest) in the overall database
 - Used in conjunction with skewed data
- Note: Sampling may not reduce database I/Os (page at a time)

Sampling: with or without Replacement



SRSWOR
(simple random
sample without
replacement)

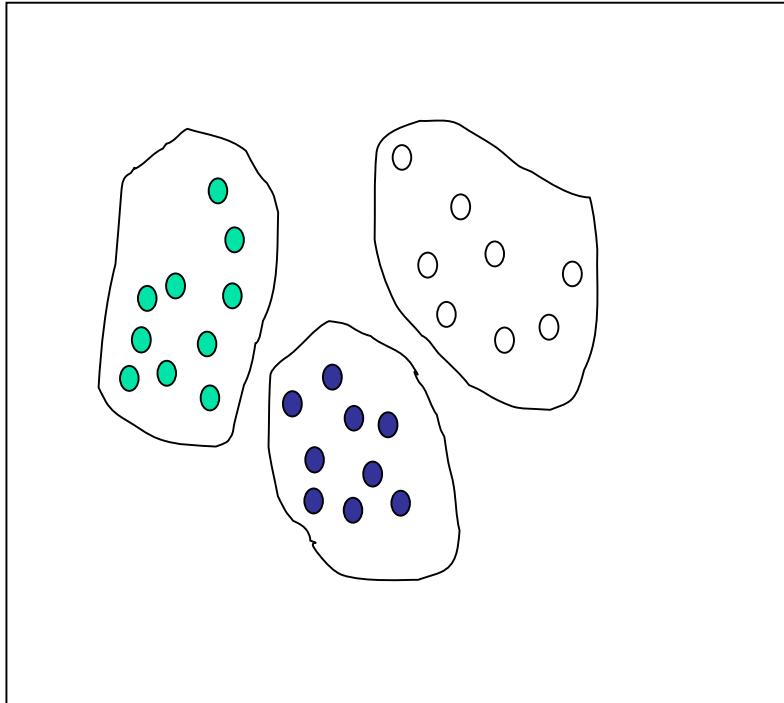
SRSWR



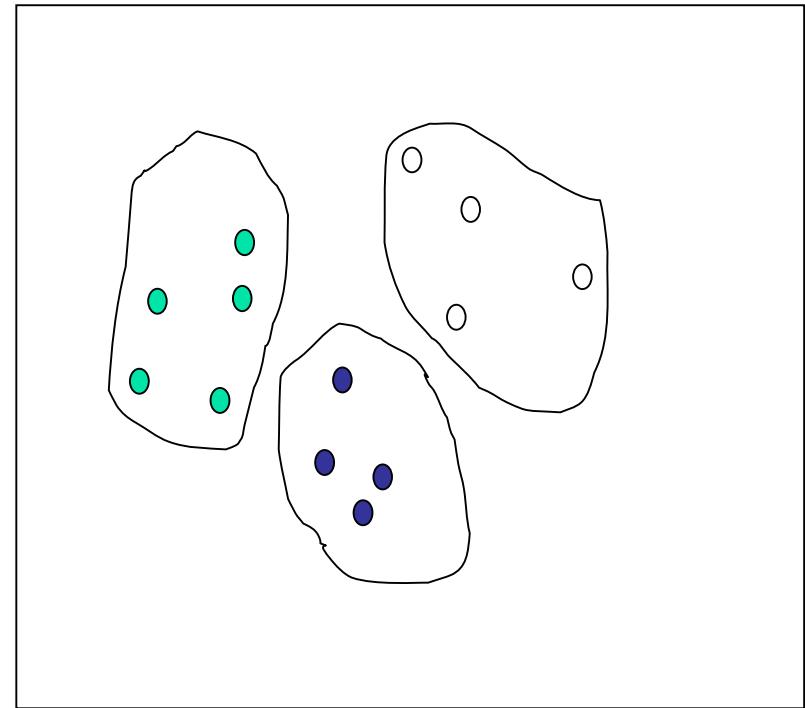
Raw Data

Sampling: Cluster or Stratified Sampling

Raw Data



Cluster/Stratified Sample



Data Preprocessing

- Why preprocess the data?
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

Discretization

- Three types of attributes:
 - Nominal — values from an unordered set, e.g., color, profession
 - Ordinal — values from an ordered set, e.g., military or academic rank
 - Continuous — real numbers, e.g., integer or real numbers
- Discretization:
 - Divide the range of a continuous attribute into intervals
 - Some classification algorithms only accept categorical attributes.
 - Reduce data size by discretization
 - Prepare for further analysis

Discretization and Concept Hierarchy

- Discretization
 - Reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals
 - Interval labels can then be used to replace actual data values
 - Supervised vs. unsupervised
 - Split (top-down) vs. merge (bottom-up)
 - Discretization can be performed recursively on an attribute
- Concept hierarchy formation
 - Recursively reduce the data by collecting and replacing low level concepts (such as numeric values for age) by higher level concepts (such as young, middle-aged, or senior)

Discretization and Concept Hierarchy Generation for Numeric Data

- Typical methods: All the methods can be applied recursively
 - Binning (covered above)
 - Top-down split, unsupervised,
 - Histogram analysis (covered above)
 - Top-down split, unsupervised
 - Clustering analysis (covered above)
 - Either top-down split or bottom-up merge, unsupervised
 - Entropy-based discretization: supervised, top-down split
 - Interval merging by χ^2 Analysis: unsupervised, bottom-up merge
 - Segmentation by natural partitioning: top-down split, unsupervised

Entropy-Based Discretization

- Given a set of samples S , if S is partitioned into two intervals S_1 and S_2 using boundary T , the information gain after partitioning is

$$I(S, T) = \frac{|S_1|}{|S|} \text{Entropy}(S_1) + \frac{|S_2|}{|S|} \text{Entropy}(S_2)$$

- Entropy is calculated based on class distribution of the samples in the set. Given m classes, the entropy of S_1 is

$$\text{Entropy}(S_1) = -\sum_{i=1}^m p_i \log_2(p_i)$$

where p_i is the probability of class i in S_1

- The boundary that minimizes the entropy function over all possible boundaries is selected as a binary discretization
- The process is recursively applied to partitions obtained until some stopping criterion is met
- Such a boundary may reduce data size and improve classification accuracy

Interval Merge by χ^2 Analysis

- Merging-based (bottom-up) vs. splitting-based methods
- Merge: Find the best neighboring intervals and merge them to form larger intervals recursively
- ChiMerge [Kerber AAAI 1992, See also Liu et al. DMKD 2002]
 - Initially, each distinct value of a numerical attr. A is considered to be one interval
 - χ^2 tests are performed for every pair of adjacent intervals
 - Adjacent intervals with the least χ^2 values are merged together, since low χ^2 values for a pair indicate similar class distributions
 - This merge process proceeds recursively until a predefined stopping criterion is met (such as significance level, max-interval, max inconsistency, etc.)

Concept Hierarchy Generation for Categorical Data

- Specification of a partial/total ordering of attributes explicitly at the schema level by users or experts
 - street < city < state < country
- Specification of a hierarchy for a set of values by explicit data grouping
 - {Urbana, Champaign, Chicago} < Illinois
- Specification of only a partial set of attributes
 - E.g., only street < city, not others
- Automatic generation of hierarchies (or attribute levels) by the analysis of the number of distinct values
 - E.g., for a set of attributes: {street, city, state, country}

Automatic Concept Hierarchy Generation

- Some hierarchies can be automatically generated based on the analysis of the number of distinct values per attribute in the data set
 - The attribute with the most distinct values is placed at the lowest level of the hierarchy
 - Exceptions, e.g., weekday, month, quarter, year



Data Preprocessing

- Why preprocess the data?
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

Summary

- Data preparation or preprocessing is a big issue for both data warehousing and data mining
- Descriptive data summarization is need for quality data preprocessing
- Data preparation includes
 - Data cleaning and data integration
 - Data reduction and feature selection
 - Discretization
- A lot a methods have been developed but data preprocessing still an active area of research

References

- D. P. Ballou and G. K. Tayi. Enhancing data quality in data warehouse environments. Communications of ACM, 42:73-78, 1999
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley & Sons, 2003
- T. Dasu, T. Johnson, S. Muthukrishnan, V. Shkapenyuk. Mining Database Structure; Or, How to Build a Data Quality Browser. SIGMOD'02.
- H.V. Jagadish et al., Special Issue on Data Reduction Techniques. Bulletin of the Technical Committee on Data Engineering, 20(4), December 1997
- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999
- E. Rahm and H. H. Do. Data Cleaning: Problems and Current Approaches. *IEEE Bulletin of the Technical Committee on Data Engineering*. Vol.23, No.4
- V. Raman and J. Hellerstein. Potters Wheel: An Interactive Framework for Data Cleaning and Transformation, VLDB'2001
- T. Redman. Data Quality: Management and Technology. Bantam Books, 1992
- Y. Wand and R. Wang. Anchoring data quality dimensions ontological foundations. Communications of ACM, 39:86-95, 1996
- R. Wang, V. Storey, and C. Firth. A framework for analysis of data quality research. IEEE Trans. Knowledge and Data Engineering, 7:623-640, 1995

Foundation of Data Science and Analytics

Probability Distribution

Arun K. Timalsina

Clarification on Probability

What does it mean :

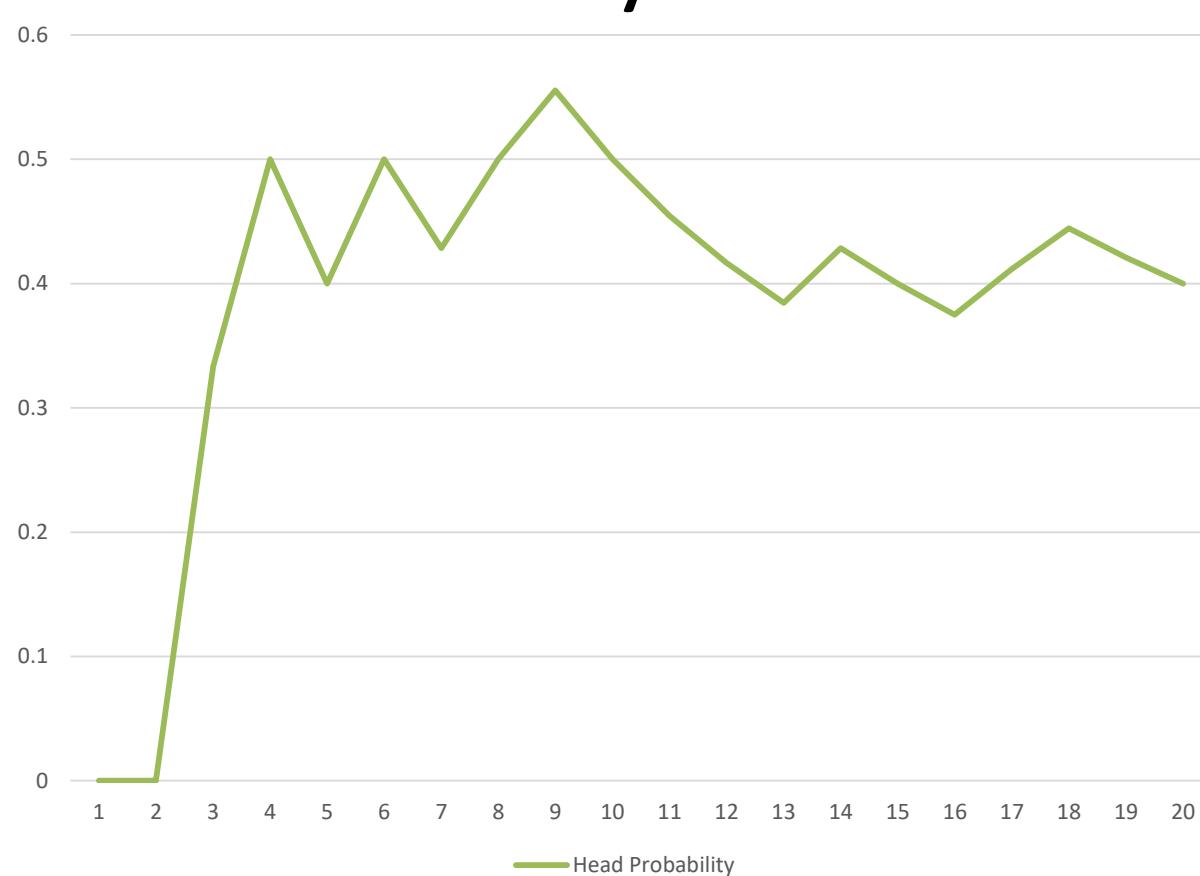
Probability of getting
“*Head*” on coin toss
experiment is *0.5* ?

Head / Tail Experiment

Event Count	Outcome	He
1	Head	
2	Tail	
3	Head	
4	Tail	
5	Tail	
6	Tail	
7	Tail	
8	Tail	
9	Head	
10	Tail	
11	Head	
12	Tail	
13	Head	
14	Tail	
15	Head	
16	Head	
17	Head	
18	Tail	
19	Tail	
20	Head	

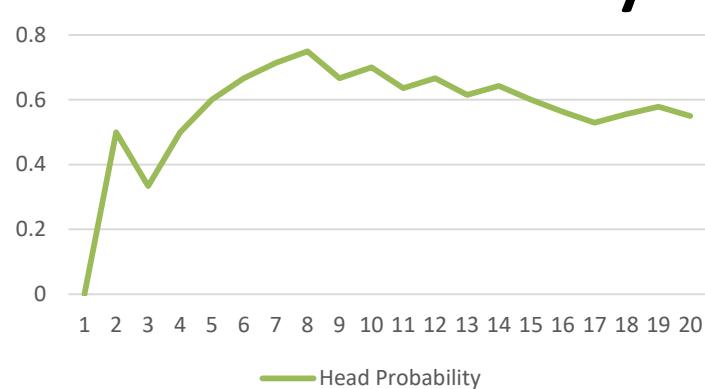
Probability

Event Count	Outcome	He
1	Head	
2	Tail	
3	Head	
4	Tail	
5	Tail	
6	Tail	
7	Tail	
8	Tail	
9	Head	
10	Tail	
11	Head	
12	Tail	
13	Head	
14	Tail	
15	Head	
16	Head	
17	Head	
18	Tail	
19	Tail	
20	Head	

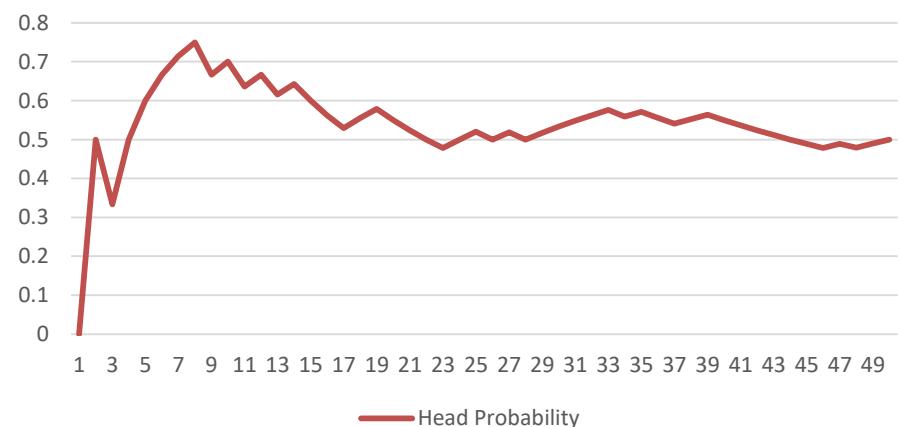


Probability

Event Count	Outcome	He
1	Head	
2	Tail	
3	Head	
4	Tail	
5	Tail	
6	Tail	
7	Tail	
8	Tail	
9	Head	
10	Tail	
11	Head	
12	Tail	
13	Head	
14	Tail	
15	Head	
16	Head	
17	Head	
18	Tail	
19	Tail	
20	Head	

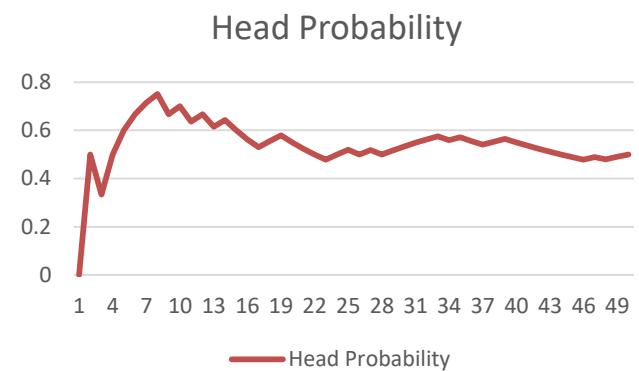
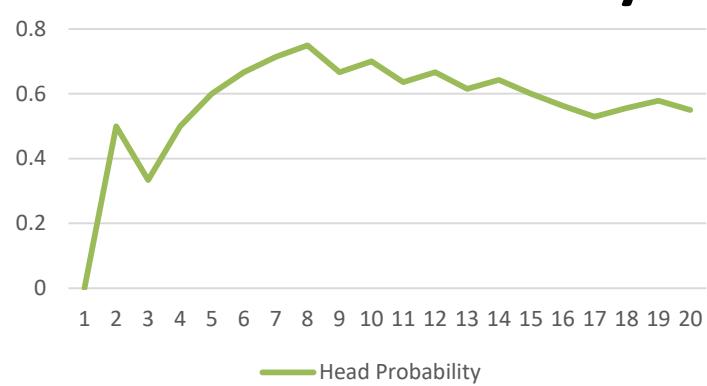


Head Probability



Probability

Event Count	Outcome	He
1	Head	
2	Tail	
3	Head	
4	Tail	
5	Tail	
6	Tail	
7	Tail	
8	Tail	
9	Head	
10	Tail	
11	Head	
12	Tail	
13	Head	
14	Tail	
15	Head	
16	Head	
17	Head	
18	Tail	
19	Tail	
20	Head	



Probability

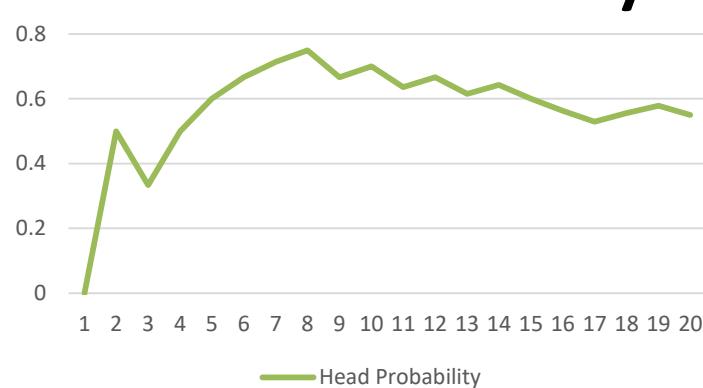
Head Probability

Event Count	Outcome	He
1	Head	
2	Tail	
3	Head	
4	Tail	
5	Tail	
6	Tail	
7	Tail	
8	Tail	
9	Head	
10	Tail	
11	Head	
12	Tail	
13	Head	
14	Tail	
15	Head	
16	Head	
17	Head	
18	Tail	
19	Tail	
20	Head	



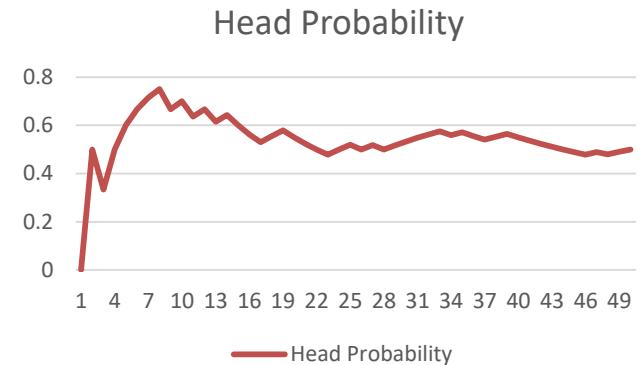
Probability

Event Count	Outcome	He
1	Head	
2	Tail	
3	Head	
4	Tail	
5	Tail	
6	Tail	
7	Tail	
8	Tail	
9	Head	
10	Tail	
11	Head	
12	Tail	
13	Head	
14	Tail	
15	Head	
16	Head	
17	Head	
18	Tail	
19	Tail	
20	Head	



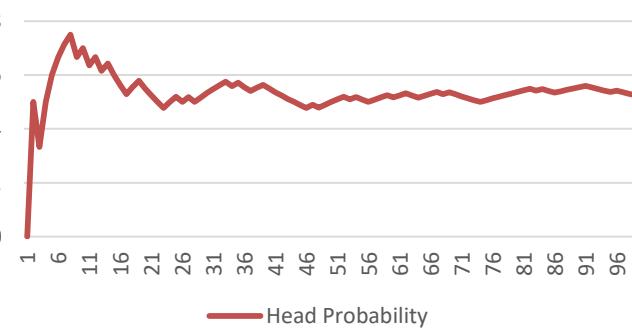
Head Probability

— Head Probability



Head Probability

— Head Probability



Head Probability

— Head Probability

Random Experiments

The goal is to understand, quantify and model the variation affecting a physical system's behavior. The model is used to analyze and predict the physical system's behavior as system inputs affect system outputs. The predictions are verified through experimentation with the physical system.

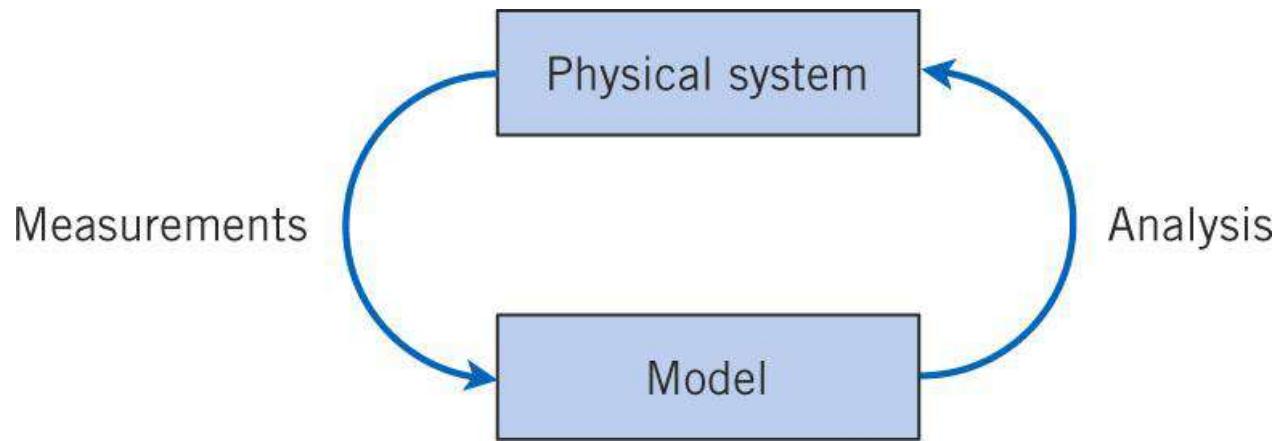


Figure 2-1 Continuous iteration between model and physical system.

Noise Produces Output Variation

Random values of the noise variables cannot be controlled and cause the random variation in the output variables. Holding the controlled inputs constant does not keep the output values constant.

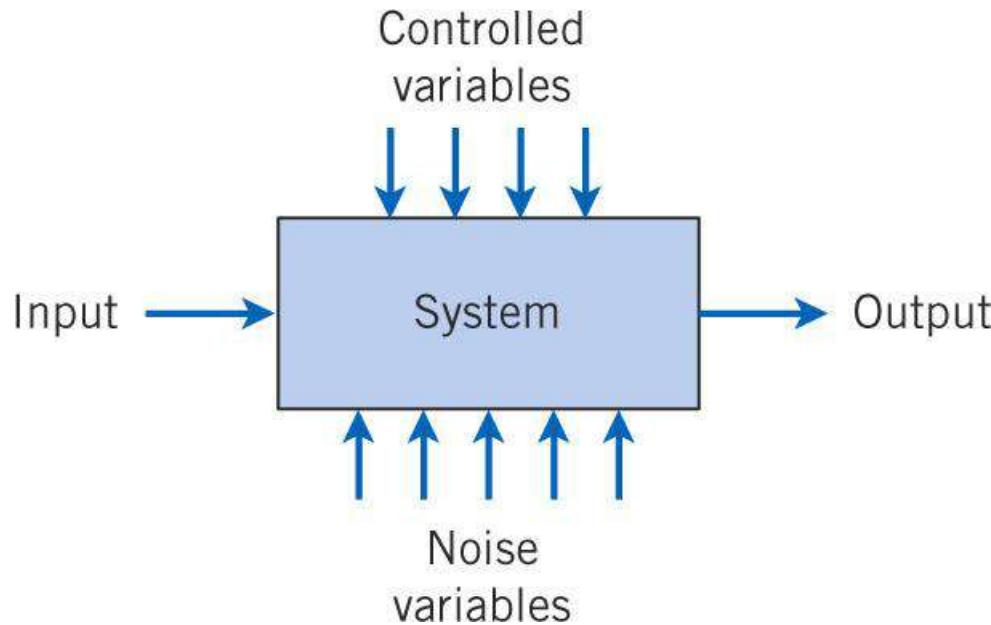


Figure 2-2 Noise variables affect the transformation of inputs to outputs.

Random Experiment

- An experiment is an operation or procedure, carried out under controlled conditions, executed to discover an unknown result or to illustrate a known law.
- An experiment that can result in different outcomes, even if repeated in the same manner every time, is called a **random experiment**.

Randomness Affects Natural Law

Ohm's Law current is a linear function of voltage. However, current will vary due to noise variables, even under constant voltage.

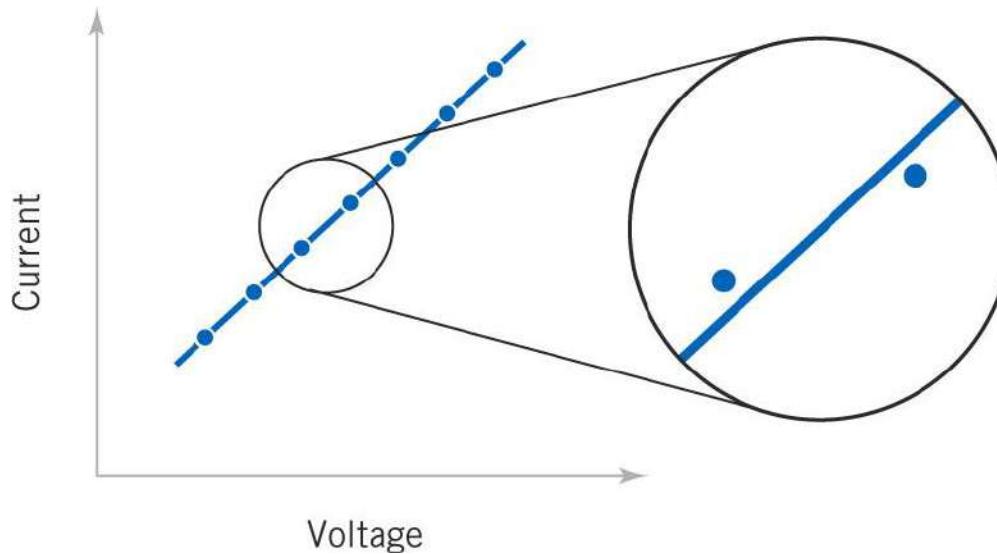


Figure 2-3 A closer examination of the system identifies deviations from the model.

Randomness Can Disrupt a System

- Telephone systems must have sufficient capacity (lines) to handle a random number of callers at a random point in time whose calls are of a random duration.
- If calls arrive exactly every 5 minutes and last for exactly 5 minutes, only 1 line is needed – a deterministic system.
- Practically, times between calls are random and the call durations are random. Calls can come into conflict as shown in following slide.
- Conclusion: Telephone system design must include provision for input variation.

Deterministic & Random Call Behavior

Calls arrive every 5 minutes. In top system, call durations are all of 5 minutes exactly. In bottom system, calls are of random duration, averaging 5 minutes, which can cause blocked calls, a “busy” signal.

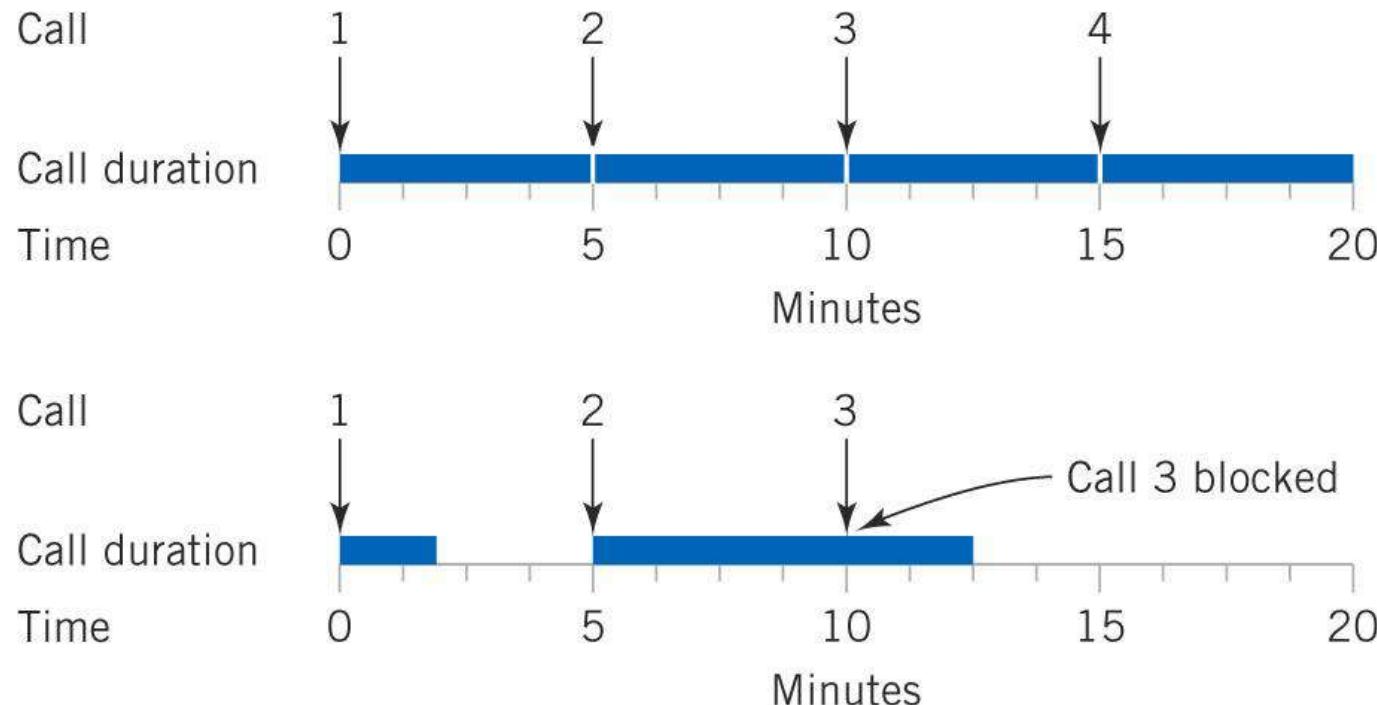


Figure 2-4 Variation causes disruption in the system.

Sample Spaces

- Random experiments have unique **outcomes**.
- The set of all possible outcome of a random experiment is called the **sample space**, S .
- S is **discrete** if it consists of a finite or countable infinite set of outcomes.
- S is **continuous** if it contains an interval (either a finite or infinite width) of real numbers.

Example 2-1: Defining Sample Spaces

- Randomly select and measure the thickness of a part. $S = R^+ = \{x | x > 0\}$, the positive real line. Negative or zero thickness is not possible.
 S is continuous.
- It is known that the thickness is between 10 and 11 mm. $S = \{x | 10 < x < 11\}$, continuous.
- It is known that the thickness has only three values. $S = \{\text{low, medium, high}\}$, discrete.
- Does the part thickness meet specifications?
 $S = \{\text{yes, no}\}$, discrete.

Example 2-2: Defining Sample Spaces, n=2

- Two parts are randomly selected & measured.
 $S = R^+ * R^+$, S is continuous.
- Do the 2 parts conform to specifications?
 $S = \{yy, yn, ny, nn\}$, S is discrete.
- Number of conforming parts?
 $S = \{1, 1, 2\}$, S is discrete.
- Parts are randomly selected until a non-conforming part is found.
 $S = \{n, yn, yyn, yyyn, \dots\}$,
 S is countably infinite.

Sample Space Is Defined By A Tree Diagram

Example 2-3: Messages are classified as on-time or late. 3 messages are classified. There are $2^3 = 8$ outcomes in the sample space.

$$S = \{ooo, ool, olo, oll, loo, lol, llo, lll\}$$

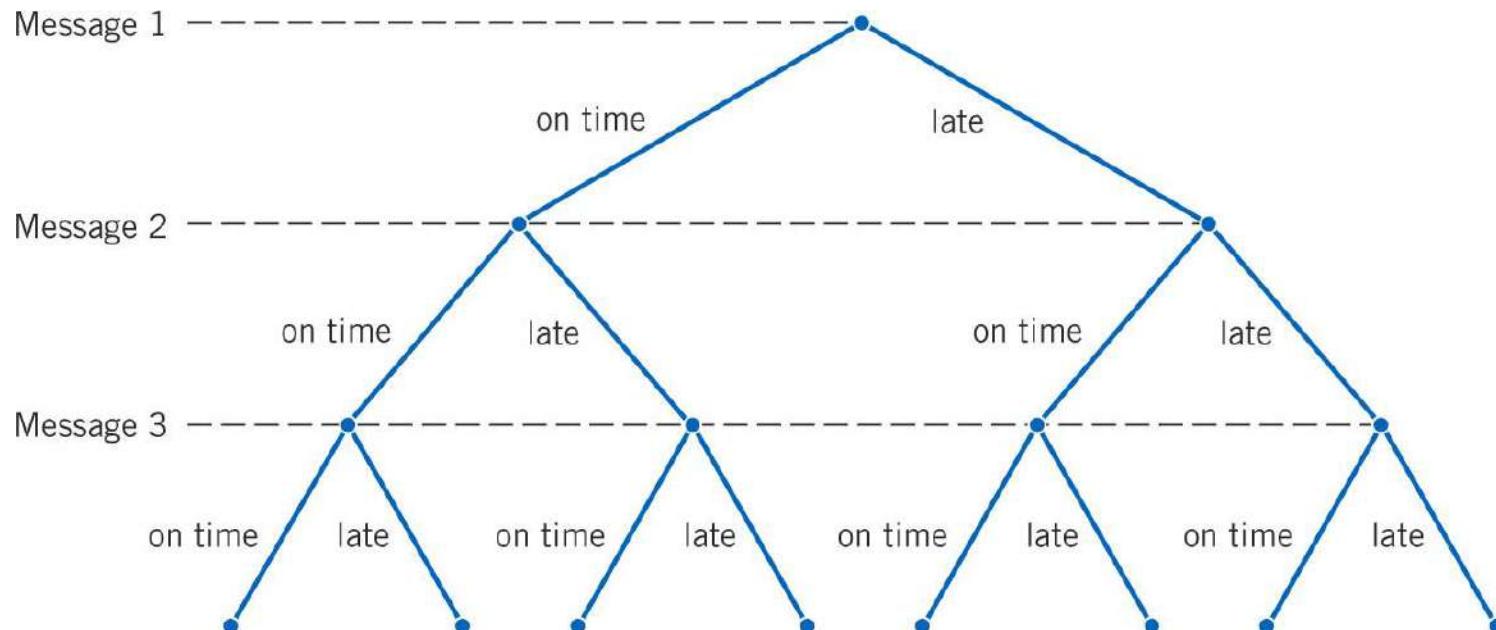


Figure 2-5 Tree diagram for three messages.

Tree Diagrams Can Fit The Situation

Example 2-4: New cars can be equipped with selected options as follows:

1. Manual or automatic transmission
2. With or without air conditioning
3. Three choices of stereo sound systems
4. Four exterior color choices

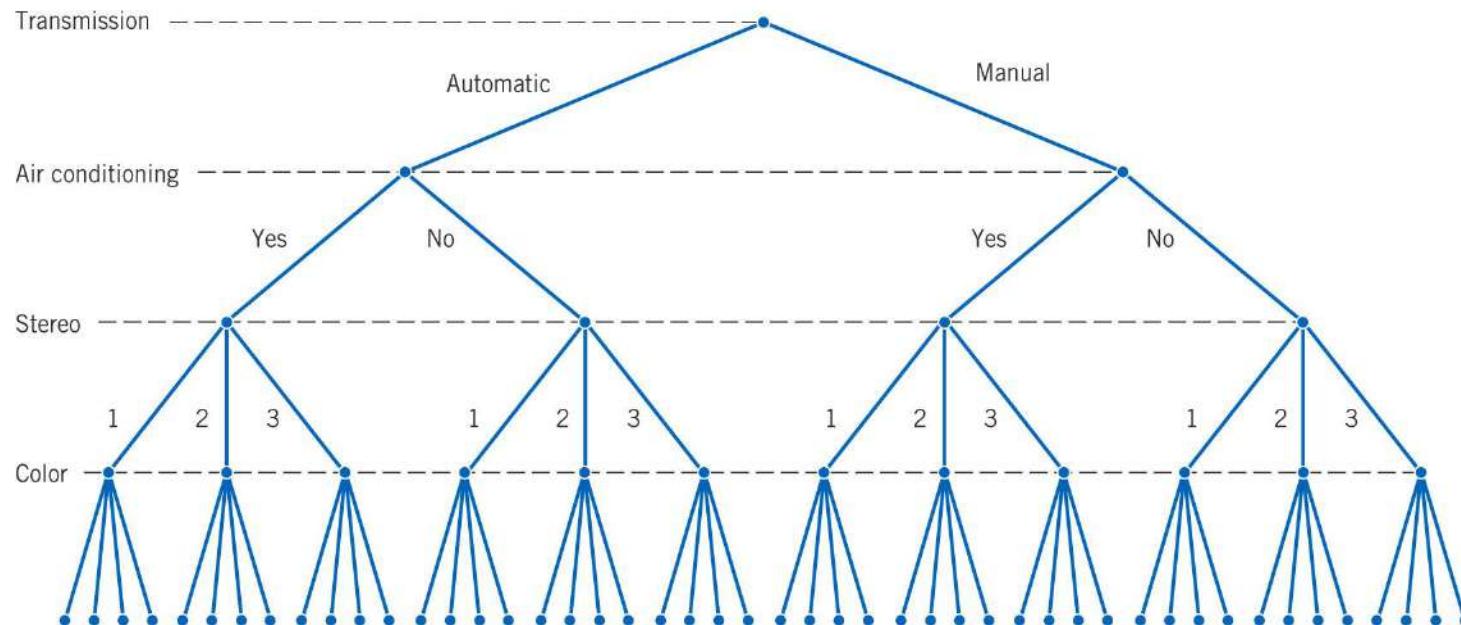


Figure 2-6 Tree diagram for different configurations of vehicles. Note that S has $2 \times 2 \times 3 \times 4 = 48$ outcomes.

Tree Diagrams Help Count Outcomes

Example 2-5: The interior car color can depend on the exterior color as shown in the tree diagrams below. There are 12 possibilities without considering color combinations.

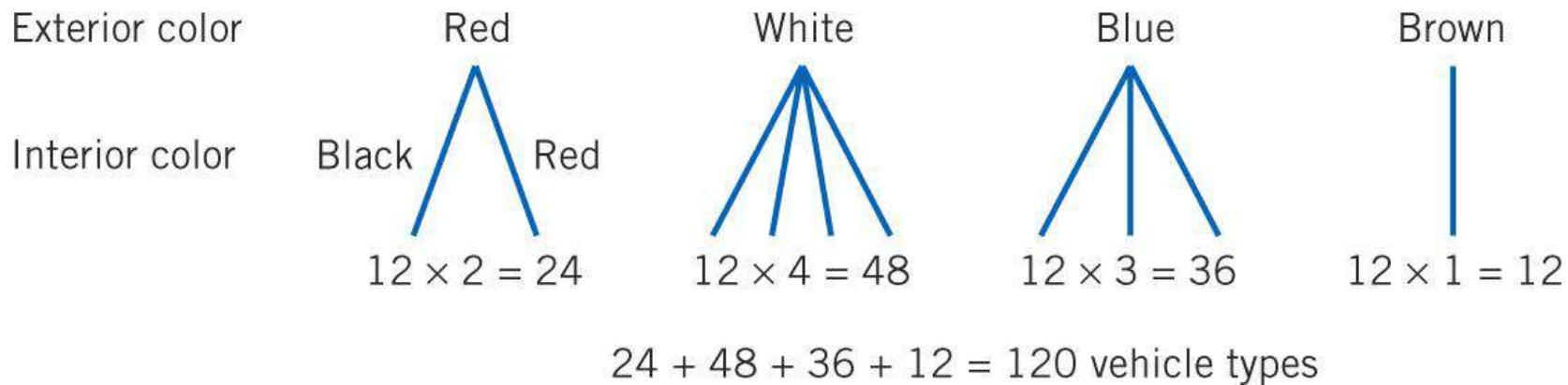


Figure 2-7 Tree diagram for different vehicle configurations with interior colors.

Events Are Sets of Outcomes

- An event (E) is a subset of the sample space of a random experiment, i.e., one or more outcomes of the sample space.
- Event combinations are:
 - **Union** of two events is the event consisting of all outcomes that are contained in either of two events, $E_1 \cup E_2$. Called E_1 or E_2 .
 - **Intersection** of two events is the event consisting of all outcomes that contained in both of two events, $E_1 \cap E_2$. Called E_1 and E_2 .
 - **Complement** of an event is the set of outcomes that are not contained in the event, E' or not E .

Example 2-6, Discrete Event Algebra

- Recall the sample space from Example 2-2, $S = \{yy, yn, ny, nn\}$ concerning conformance to specifications.
 - Let E_1 denote the event that at least one part does conform to specifications, $E_1 = \{yy, yn, ny\}$
 - Let E_2 denote the event that no part conforms to specifications, $E_2 = \{nn\}$
 - Let $E_3 = \emptyset$, the null or empty set.
 - Let $E_4 = S$, the universal set.
 - Let $E_5 = \{yn, ny, nn\}$, at least one part does not conform.
 - Then $E_1 \cup E_5 = S$
 - Then $E_1 \cap E_5 = \{yn, ny\}$
 - Then $E_1' = \{nn\}$

Example 2-7, Continuous Event Algebra

Measurements of the thickness of a part are modeled with the sample space: $S = \mathbb{R}^+$.

- Let $E_1 = \{x | 10 \leq x < 12\}$, show on the real line below.
- Let $E_2 = \{x | 11 < x < 15\}$
- Then $E_1 \cup E_2 = \{x | 10 \leq x < 15\}$
- Then $E_1 \cap E_2 = \{x | 11 < x < 12\}$
- Then $E_1' = \{x | x < 10 \text{ or } x \geq 12\}$
- Then $E_1' \cap E_2 = \{x | 12 \geq x < 15\}$



Example 2-8, Hospital Emergency Visits

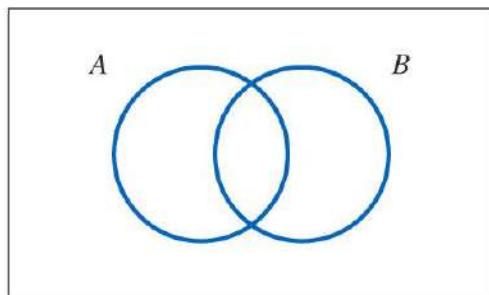
- This table summarizes the ER visits at 4 hospitals. People may leave without being seen by a physician(LWBS). The remaining people are seen, and may or may not be admitted.

	Hospital					
	1	2	3	4	Total	Answers
Total	5,292	6,991	5,640	4,329	22,252	$A \rightarrow B = 195$
LWBS	195	270	246	242	953	$A' = 16,960$
Admitted	1,277	1,558	666	984	4,485	$A \leftarrow B = 6,050$
Not admitted	3,820	5,163	4,728	3,103	16,814	

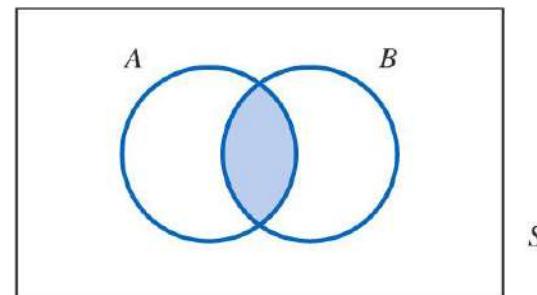
- Let A be the event of a visit to Hospital 1.
- Let B be the event that the visit is LWBS.
- Find number of outcomes in:
 - $A \cap B$
 - A'
 - $A \cup B$

Venn Diagrams Show Event Relations

Events A & B contain their respective outcomes. The shaded regions indicate the event relation of each diagram.

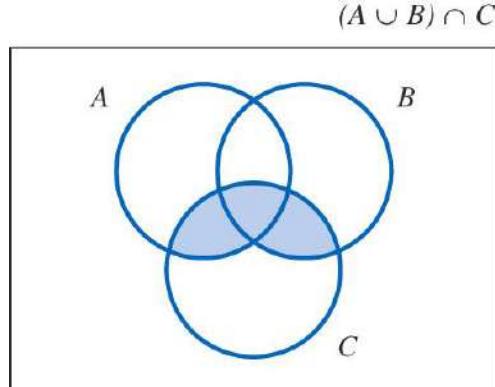


(a)

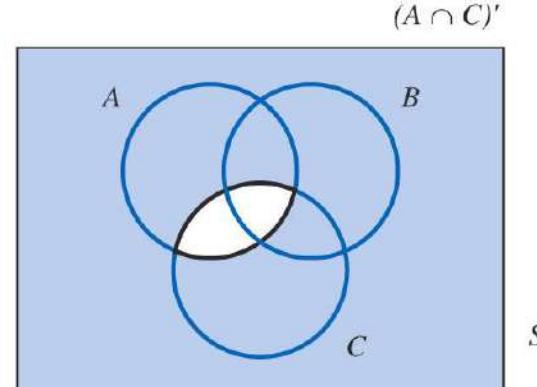


(b)

Sample space S with events A and B



(c)



(d)

Material Adaptation :
Applied Statistics and
Probability for Engineers, by
Montgomery and Runger.
John Wiley & Sons,

Figure 2-8 Venn diagrams

Venn Diagram of Mutually Exclusive Events

- Events A & B are mutually exclusive because they share no common outcomes.
- The occurrence of one event precludes the occurrence of the other.
- Symbolically, $A \cap B = \emptyset$

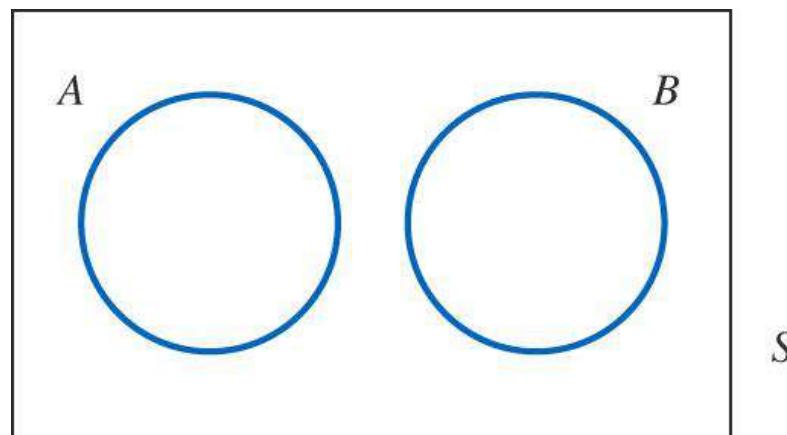


Figure 2-9 Mutually exclusive events

Event Relation Laws

- Transitive law (event order is unimportant):
 - $A \cap B = B \cap A$ and $A \cup B = B \cup A$
- Distributive law (like in algebra):
 - $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$
 - $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$
- DeMorgan's laws:
 - $(A \cup B)' = A' \cap B'$ The complement of the union is the intersection of the complements.
 - $(A \cap B)' = A' \cup B'$ The complement of the intersection is the union of the complements.

Counting Techniques

- These are three special rules, or counting techniques, used to determine the number of outcomes in the events and the sample space.
- They are the:
 1. Multiplication rule
 2. Permutation rule
 3. Combination rule
- Each has its special purpose that must be applied properly – the right tool for the right job.

Counting – Multiplication Rule

- Multiplication rule:
 - Let an operation consist of k steps and
 - n_1 ways of completing step 1,
 - n_2 ways of completing step 2, ... and
 - n_k ways of completing step k .
 - Then, the total number of ways or outcomes are:
 - $n_1 * n_2 * \dots * n_k$

Example 2-9: Multiplication Rule

- In the design for a gear housing, we can choose to use among:
 - 4 different fasteners,
 - 3 different bolt lengths and
 - 2 different bolt locations.
- How many designs are possible?
- Answer: $4 * 3 * 2 = 24$

Counting – Permutation Rule

- A permutation is a unique sequence of distinct items.
- If $S = \{a, b, c\}$, then there are 6 permutations
 - Namely: abc, acb, bac, bca, cab, cba (**order matters**)
 - The # of ways 3 people can be arranged.
- # of permutations for a set of n items is $n!$
- $n! \text{ (factorial function)} = n*(n-1)*(n-2)*...*2*1$
- $7! = 7*6*5*4*3*2*1 = 5,040 = \text{FACT}(7)$ in Excel
- By definition: $0! = 1$

Counting – Sub-set Permutations

- To sequence only r items from a set of n items:

$$P_r^n = n(n-1)(n-2)\dots(n-r+1) = \frac{n!}{(n-r)!}$$

$$P_3^7 = \frac{7!}{(7-3)!} = \frac{7!}{4!} = \frac{7*6*5*4!}{4!} = 7*6*5 = 210$$

In Excel: permut(7,3) = 210

Example 2-10: Circuit Board Designs

- A printed circuit board has eight different locations in which a component can be placed. If four different components are to be placed on the board , how many designs are possible?
- Answer: order is important, so use the permutation formula with n = 8, r = 4.

$$P_4^8 = \frac{8!}{(8-4)!} = \frac{8*7*6*5*4!}{4!} = 8*7*6*5 = 1,680$$

Counting - Similar Item Permutations

- Used for counting the sequences when not all the items are different.
- The number of permutations of:
 - $n = n_1 + n_2 + \dots + n_r$ items of which
 - n_1 are identical,
 - n_2 are identical, ... , and
 - n_r are identical.
- Is calculated as:

$$\frac{n!}{n_1! n_2! \dots n_r!}$$

Example 2-11: Machine Shop Schedule

- In a machining operation, a piece of sheet metal needs two identical-diameter holes drilled and two identical-size notched cut. The drilling operation is denoted as d and the notching as n.
 - How many sequences are there?

$$\frac{4!}{2!2!} = \frac{4 * 3 * 2!}{2 * 1 * 2!} = 6$$

- What is the set of sequences?

{ddnn, dndn, dnnd, nddn, ndnd, nndd}

Example 2-12: Bar Codes

- A part is labeled with 4 thick lines, 3 medium lines, and two thin lines. Each sequence is a different label.
 - How many unique labels can be created?

$$\frac{9!}{4!3!2!} = \frac{9*8*7*6*5*4!}{2*1*3*2*1*4!} = 1,260$$

- In Excel:

1,260	=FACT(9) / (FACT(4)*FACT(3)*FACT(2))
-------	--------------------------------------

Counting – Combination Rule

- A combination is a selection of r items from a set of n where **order does not matter**.
- If $S = \{a, b, c\}$, $n = 3$, then there is 1 combination.
 - If $r = 3$, there is 1 combination, namely: abc
 - If $r = 2$, there are 3 combinations, namely ab, ac, bc
- # of permutations \geq # of combinations

$$C_r^n = \frac{n!}{r!(n-r)!} \quad (2-4)$$

Example 2-13: Applying the Combination Rule

- A circuit board has eight locations in which a component can be placed. If 5 identical components are to be placed on a board, how many different designs are possible?
- The order of the components is not important, so the combination rule is appropriate.

$$C_5^8 = \frac{8!}{5!(8-5)!} = \frac{8 * 7 * 6 * 5!}{3 * 2 * 1 * 5!} = 56$$

- Excel: 56 = COMBIN(8,5)

Example 2-14: Sampling w/o Replacement-1

- A bin of 50 parts contains 3 defectives & 47 good parts. A sample of 6 parts is selected from the 50 **without** replacement.
- How many different samples of size 6 are there that contain exactly 2 defective parts?

$$C_2^3 = \frac{3!}{2!1!} = 3 \text{ different ways}$$

- In Excel: 3 = COMBIN(3,2)

Example 2-14: Sampling w/o Replacement-2

- Now, how many ways are there of selecting 4 parts from the 47 acceptable parts?

$$C_4^{47} = \frac{47!}{4!43!} = \frac{47*46*45*44*43!}{4*3*2*1*43!} = 178,365 \text{ different ways}$$

- In Excel: 178,365 = COMBIN(47,4)

Example 2-14: Sampling w/o Replacement-3

- Now, how many ways are there to obtain:
 - 2 from the 3 defectives, and
 - 4 from the 47 non-defectives?

$$C_2^3 C_4^{47} = 3 * 178,365 = 535,095 \text{ different ways}$$

– In Excel: 535,095 = COMBIN(3,2)*COMBIN(47,4)

Example 2-14: Sampling w/o Replacement-4

- Furthermore, how many ways are there to obtain 6 parts (0-6 defectives) from the set of 50?

$$C_6^{50} = \frac{50!}{6!*44!} = 15,890,700$$

- So the ratio of obtaining 2 defectives out 6 to any number (0-6) defectives out of 6 is:

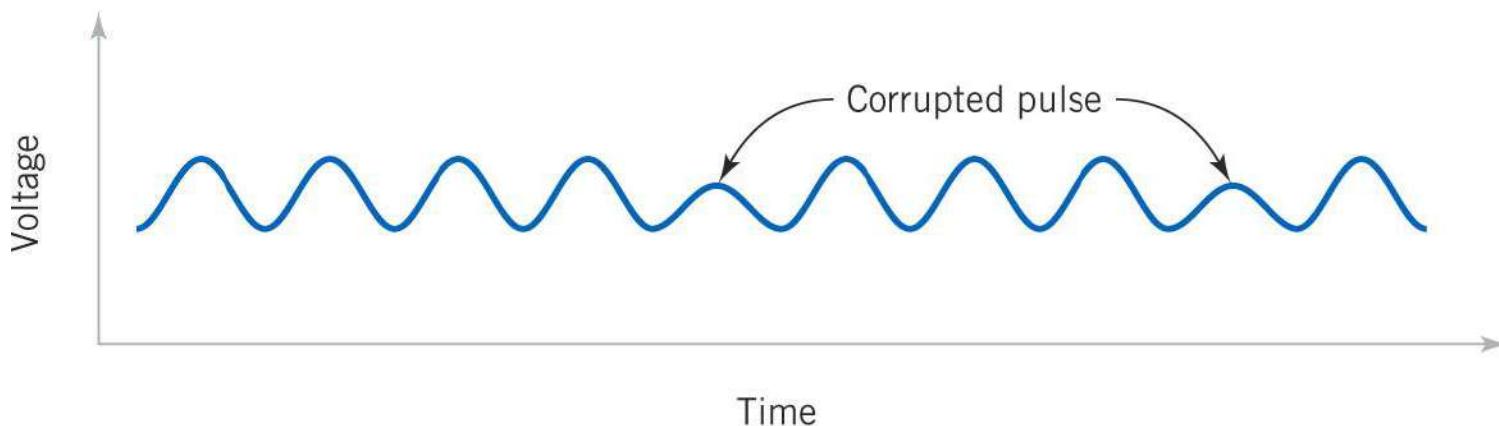
$$\frac{C_2^3 C_4^{47}}{C_6^{50}} = \frac{3 * 178,365}{15,890,700} = 0.034$$

What Is Probability?

- Probability is the likelihood or chance that a particular outcome or event from a random experiment will occur.
- Here, only finite sample spaces ideas apply.
- Probability is a number in the $[0,1]$ interval.
- May be expressed as a:
 - proportion (0.15)
 - percent (15%)
 - fraction ($3/20$)
- A probability of:
 - 1 means certainty
 - 0 means impossibility

Types of Probability

- **Subjective probability** is a “degree of belief.”
 - “There is a 50% chance that I’ll study tonight.”
- **Relative frequency** probability is based how often an event occurs over a very large sample space.



$$\text{Relative frequency of corrupted pulse} = \frac{2}{10}$$

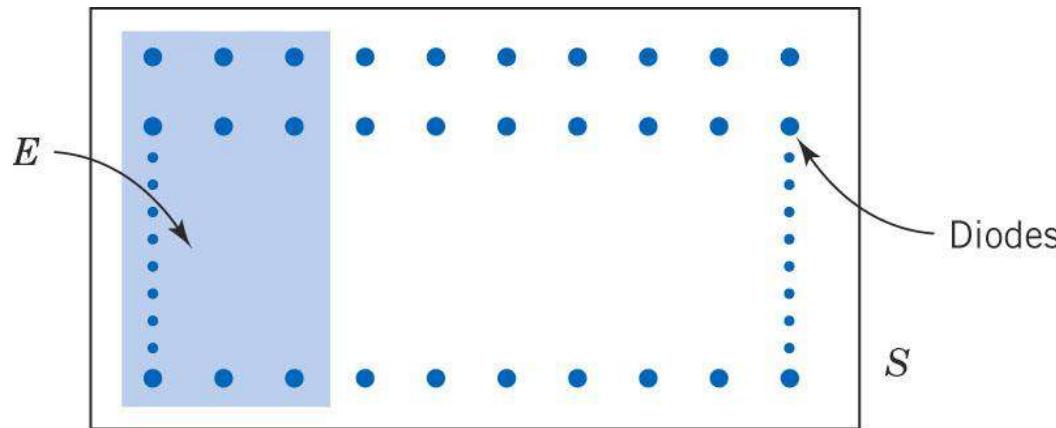
Figure 2-10 Relative frequency of corrupted pulses over a communications channel

Probability Based on Equally-Likely Outcomes

- Whenever a sample space consists of N possible outcomes that are equally likely, the probability of each outcome is $1/N$.
- Example: In a batch of 100 diodes, 1 is colored red. A diode is randomly selected from the batch. Random means each diode has an equal chance of being selected. The probability of choosing the red diode is $1/100$ or 0.01, because each outcome in the sample space is equally likely.

Example 2-15: Laser Diodes

- Assume that 30% of the laser diodes in a batch of 100 meet a customer requirements.
- A diode is selected randomly. Each diode has an equal chance of being selected. The probability of selecting an acceptable diode is 0.30.



$$P(E) = 30(0.01) = 0.30$$

Figure 2-11 Probability of the event E is the sum of the probabilities of the outcomes in E .

Probability of an Event

- For a discrete sample space, the *probability of an event E*, denoted by $P(E)$, equals the sum of the probabilities of the outcomes in E .
- The discrete sample space may be:
 - A finite set of outcomes
 - A countably infinite set of outcomes.
- Further explanation is necessary to describe probability with respect to continuous sample spaces.

Example 2-16: Probabilities of Events

- A random experiment has a sample space $\{w,x,y,z\}$. These outcomes are not equally-likely; their probabilities are: 0.1, 0.3, 0.5, 0.1.
- Event A = $\{w,x\}$, event B = $\{x,y,z\}$, event C = $\{z\}$
 - $P(A) = 0.1 + 0.3 = 0.4$
 - $P(B) = 0.3 + 0.5 + 0.1 = 0.9$
 - $P(C) = 0.1$
 - $P(A') = 0.6$ and $P(B') = 0.1$ and $P(C') = 0.9$
 - Since event $A \cap B = \{x\}$, then $P(A \cap B) = 0.3$
 - Since event $A \cup B = \{w,x,y,z\}$, then $P(A \cup B) = 1.0$
 - Since event $A \cap C = \{\text{null}\}$, then $P(A \cap C) = 0.0$

Example 2-17: Contamination Particles

- An inspection of a large number of semiconductor wafers revealed the data for this table. A wafer is selected randomly.
- Let E be the event of selecting a 0 particle wafer. $P(E) = 0.40$
- Let E be the event of selecting a wafer with 3 or more particles. $P(E) = 0.10+0.05+0.10 = 0.25$

Number of Contamination Particles	Proportion of Wafers
0	0.40
1	0.20
2	0.15
3	0.10
4	0.05
5 or more	0.10
Total	1.00

Example 2-18: Sampling w/o Replacement

- A batch of parts contains 6 parts $\{a,b,c,d,e,f\}$. Two are selected at random. Suppose part f is defective. What is the probability that part f appears in the sample?
- How many possible samples can be drawn?
 - Excel: $15 = \text{COMBIN}(6,2)$
- How many samples contain part f ?
 - 5 by enumeration: $\{af, bf, cf, df, ef\}$
- $P(\text{defective part}) = 5/15 = 1/3.$

Axioms of Probability

- Probability is a number that is assigned to each member of a collection of events from a random experiment that satisfies the following properties:
 1. $P(S) = 1$
 2. $0 \leq P(E) \leq 1$
 3. For each two events E_1 and E_2 with $E_1 \cap E_2 = \emptyset$,
$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$
- These imply that:
 - $P(\emptyset) = 0$ and $P(E') = 1 - P(E)$
 - If E_1 is contained in E_2 , then $P(E_1) \leq P(E_2)$.

Addition Rules

- Joint events are generated by applying basic set operations to individual events, specifically:
 - Unions of events, $A \cup B$
 - Intersections of events, $A \cap B$
 - Complements of events, A'
- Probabilities of joint events can often be determined from the probabilities of the individual events that comprise it. And conversely.

Example 2-19: Semiconductor Wafers

A wafer is randomly selected from a batch as shown in the table.

- Let H be the event of high concentrations of contaminants. Then $P(H) = 358/940$.
- Let C be the event of the wafer being located at the center of a sputtering tool used in manufacture. Then $P(C) = 626/940$.
- $P(H \cap C) = 112/940$

Table 2-1		Location of Tool		
Contamination		Center	Edge	Total
Low		514	68	582
High		112	246	358
Total		626	314	940

- $P(H \cup C) = P(H) + P(C) - P(H \cap C) = (358+626-112)/940$ This is the **addition rule**.

-
- The probability of a union:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (2-5)$$

and, as rearranged:

$$P(A \cap B) = P(A) + P(B) - P(A \cup B)$$

- If events A and B are mutually exclusive:

$$P(A \cap B) = \emptyset$$

therefore:

$$P(A \cup B) = P(A) + P(B) \quad (2-6)$$

Example 2-20: Contaminants & Location

Wafers in last example are now classified by degree of contamination per table of proportions.

- E_1 is the event that a wafer has 4 or more particles.

$$P(E_1) = 0.15$$

- E_2 is the event that a wafer was on edge. $P(E_2) = 0.28$

$$\bullet P(E_1 \cap E_2) = 0.04$$

$$\begin{aligned}\bullet P(E_1 \cup E_2) \\ = 0.15 + 0.28 - 0.04 \\ = 0.39\end{aligned}$$

Number of Contamination Particles	Table 2-2		
	Location of Tool	Center	Edge
0	0.30	0.10	0.40
1	0.15	0.05	0.20
2	0.10	0.05	0.15
3	0.06	0.04	0.10
4	0.04	0.01	0.05
5 or more	0.07	0.03	0.10
Totals	0.72	0.28	1.00

Addition Rule: 3 or More Events

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) \\ &\quad - P(A \cap B) - P(A \cap C) - P(B \cap C) \\ &\quad + P(A \cap B \cap C) \end{aligned} \tag{2-7}$$

Note the alternating signs.

If a collection of events E_i is mutually exclusive,
thus for all pairs: $E_i \cap E_j = \emptyset$

$$\text{Then: } P(E_1 \cup E_2 \cup \dots \cup E_k) = \sum_{i=1}^k P(E_i) \tag{2-8}$$

Venn Diagram of Mutually Exclusive Events

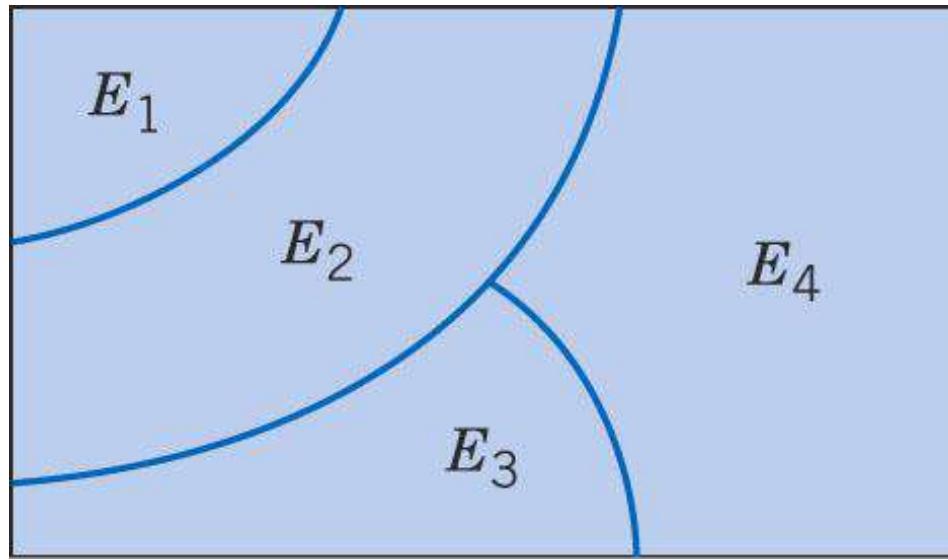


Figure 2-12 Venn diagram of four mutually exclusive events. Note that no outcomes are common to more than one event, i.e. all intersections are null.

Example 2-21: pH

- Let X denote the pH of a sample. Consider the event that $P(6.5 < X \leq 7.5) =$
$$P(6.5 < X \leq 7.0) + P(7.0 < X \leq 7.5) + P(7.5 < X \leq 7.8)$$
- The partition of an event into mutually exclusive subsets is widely used to allocate probabilities.

Conditional Probability

- Probabilities should be reevaluated as additional information becomes available.
- $P(B|A)$ is called the probability of event B occurring, given that event A has already occurred.
- A communications channel has an error rate of 1 per 1000 bits transmitted. Errors are rare, but do tend to occur in bursts. If a bit is in error, the probability that the next bit is also an error ought to be greater than 1/1000.

An Example of Conditional Probability

- In a thin film manufacturing process, the proportion of parts that are not acceptable is 2%. However the process is sensitive to contamination that can increase the rate of parts rejection.
- If we know that the plant is having filtration problems that increase film contamination, we would presume that the rejection rate has increased.

Another Example of Conditional Probability

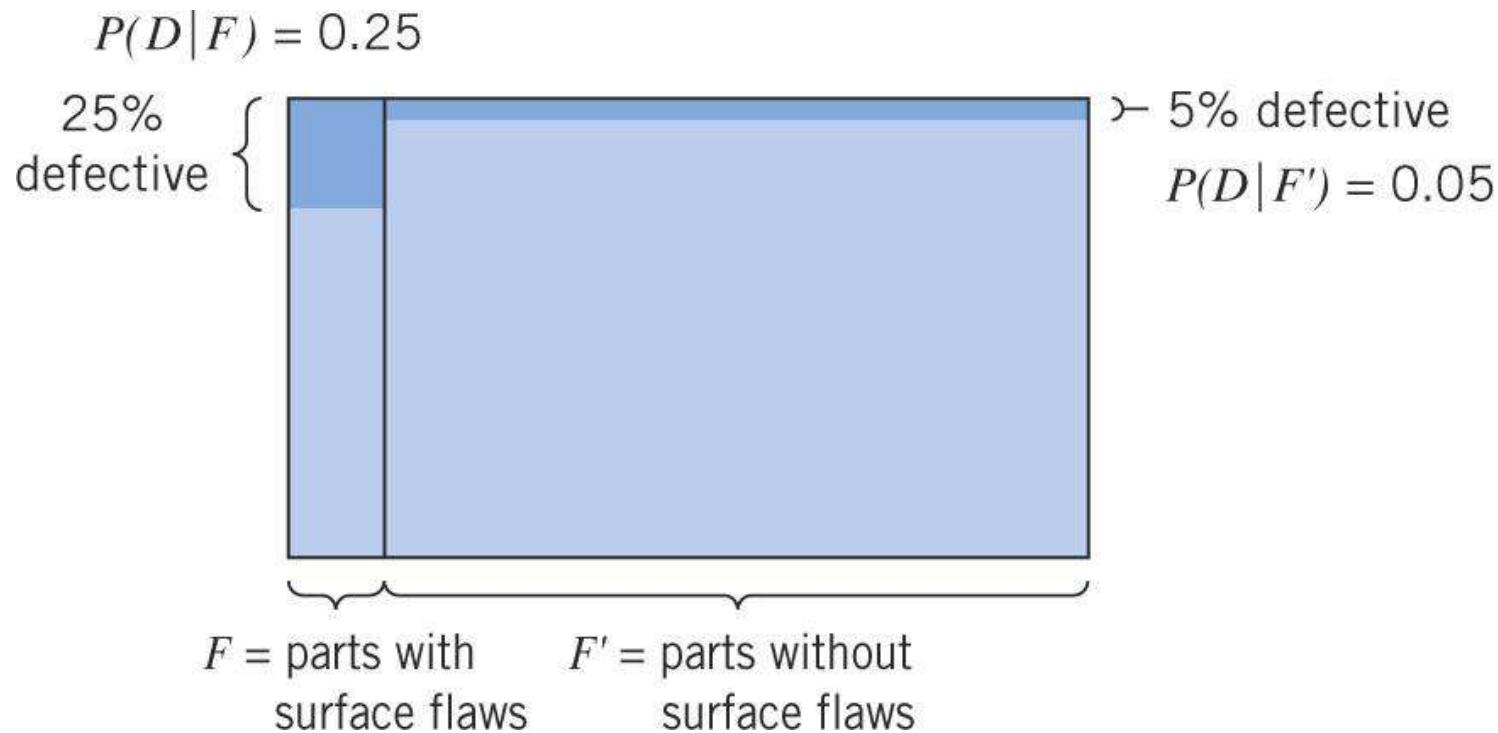


Figure 2-13 Conditional probability of rejection for parts with surface flaws and for parts without surface flaws. The probability of a defective part is not evenly distributed. Flawed parts are five times more likely to be defective than non-flawed parts, i.e., $P(D|F) / P(D|F')$.

Example 2-22: A Sample From Prior Graphic

- Table 2-3 shows that 400 parts are classified by surface flaws and as functionally defective. Observe that:
 - $P(D|F) = 10/40 = 0.25$
 - $P(D|F') = 18/360 = 0.05$

Table 2-3 Parts Classified

		Surface Flaws		
Defective	Yes (F)	No (F')	Total	
Yes (D)	10	18	28	
No (D')	30	342	372	
Total	40	360	400	

Conditional Probability Rule

- The **conditional probability** of event B given event A , denoted as $P(B|A)$, is:

$$P(B|A) = P(A \cap B) / P(A) \quad (2-9)$$

for $P(A) > 0$.

- From a relative frequency perspective of n equally likely outcomes:
 - $P(A) = (\text{number of outcomes in } A) / n$
 - $P(A \cap B) = (\text{number of outcomes in } A \cap B) / n$

Example 2-23: More Surface Flaws

Refer to Table 2-3 again. There are 4 probabilities conditioned on flaws.

$$P(F) = 40/400 \text{ and } P(D) = 28/400$$

Table 2-3 Parts Classified			
	Surface Flaws		
Defective	Yes (F)	No (F')	Total
Yes (D)	10	18	28
No (D')	30	342	372
Total	40	360	400

$$P(D|F) = P(D \cap F)/P(F) = \frac{10}{400}/\frac{40}{400} = \frac{10}{40}$$

$$P(D'|F) = P(D' \cap F)/P(F) = \frac{30}{400}/\frac{40}{400} = \frac{30}{40}$$

$$P(D|F') = P(D \cap F')/P(F') = \frac{18}{400}/\frac{360}{400} = \frac{18}{360}$$

$$P(D'|F') = P(D' \cap F')/P(F') = \frac{342}{400}/\frac{360}{400} = \frac{342}{360}$$

Example 2-23: Tree Diagram

Tree illustrates sampling two parts without replacement:

- At the 1st stage (flaw), every original part of the 400 is equally likely.
- At the 2nd stage (defect), the probability is conditional upon the part drawn in the prior stage.

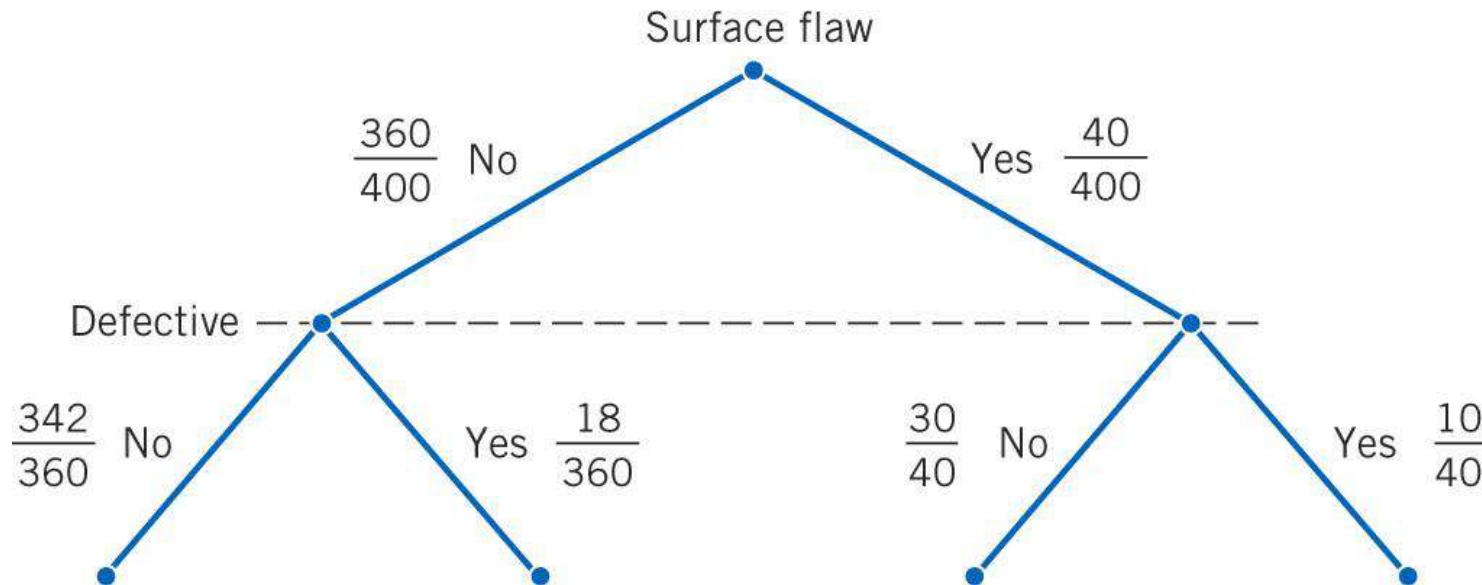


Figure 2-14 Tree diagram for parts classification

Random Samples & Conditional Probabilities

- Random means each item is equally likely to be chosen. If more than one item is sampled, random means that every sampling outcome is equally likely.
- 2 items are taken from $S = \{a,b,c\}$ without replacement.
- Ordered sample space: $S = \{ab,ac,bc,ba,bc,cb\}$
- Unordered sample space: $S = \{ab,ac,bc\}$
- This is done by enumeration – too hard 😞

Sampling Without Enumeration

- Use conditional probability to avoid enumeration. To illustrate: A batch of 50 parts contains 10 made by Tool 1 and 40 made by Tool 2. We take a sample of n=2.
- What is the probability that the 2nd part came from Tool 2, given that the 1st part came from Tool 1?
 - $P(1^{\text{st}} \text{ part came from Tool 1}) = 10/50$
 - $P(2^{\text{nd}} \text{ part came from Tool 2}) = 40/49$
 - $P(\text{Tool 1, then Tool 2 part sequence}) = (10/50)*(40/49)$
- To select randomly implies that, at each step of the sample, the items remaining in the batch are equally likely to be selected.

Example 2-24: Sampling Without Replacement

- A production lot of 850 parts contains 50 defectives. Two parts are selected at random.
- What is the probability that the 2nd is defective, given that the first part is defective?
- Let A denote the event that the 1st part selected is defective.
- Let B denote the event that the 2nd part selected is defective.
- Probability desired is $P(B|A) = 49/849$.

Example 2-25: Continuing Prior Example

- Now, 3 parts are sampled randomly.
- What is the probability that the first two are defective, while the third is not?

$$P(ddn) = \frac{50}{850} * \frac{49}{849} * \frac{800}{848} = 0.0032$$

- In Excel: $0.0032 = (50*49*800)/(850*849*848)$

Multiplication Rule

- The conditional probability definition of Equation 2-9 can be rewritten to generalize it as the **multiplication rule**.
- $P(A \cap B) = P(B|A)*P(A) = P(A|B)*P(B)$ (2-10)
- The last expression is obtained by exchanging the roles of A and B .

Example 2-26: Machining Stages

- The probability that, a part made in the 1st stage of a machining operation passes inspection, is 0.90. The probability that, it passes inspection after the 2nd stage, is 0.95.
- What is the probability that the part meets specifications?
- Let A & B denote the events that the 1st & 2nd stages meet specs.
- $P(A \cap B) = P(B|A) * P(A) = 0.95 * 0.90 = 0.955$

Two Mutually Exclusive Subsets

- A & A' are mutually exclusive.
- $A \cap B$ and $A' \cap B$ are mutually exclusive
- $B = (A \cap B) \cup (A' \cap B)$

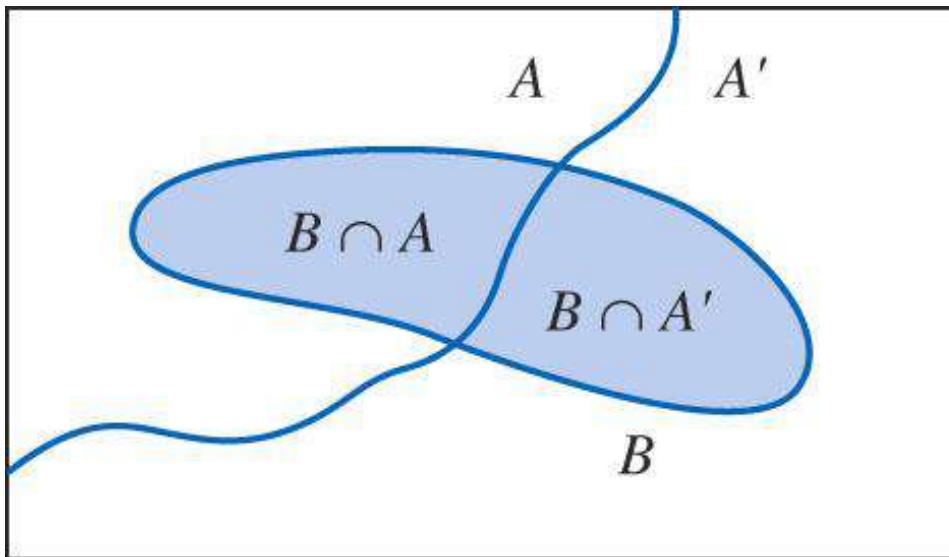


Figure 2-15 Partitioning an event into two mutually exclusive subsets.

Total Probability Rule

For any two events A and B :

$$\begin{aligned} P(B) &= P(B \cap A) + P(B \cap A') \\ &= P(B | A) * P(A) + P(B | A') * P(A') \end{aligned} \quad (2-11)$$

Example 2-27: Semiconductor Contamination

- Information about product failure based on chip manufacturing process contamination.

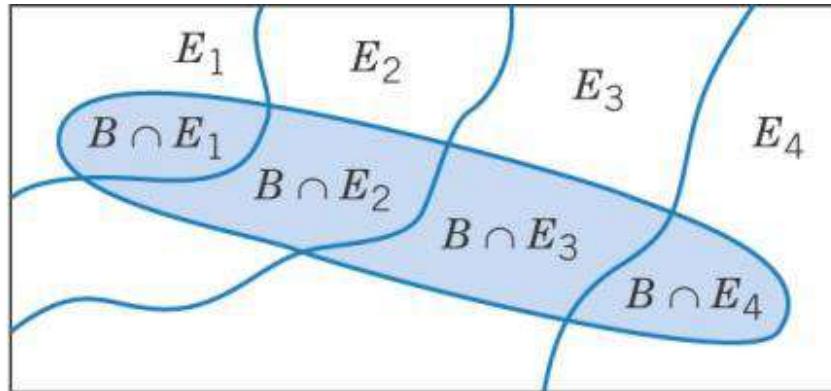
Probability of Failure	Level of Contamination	Probability of Level
0.100	High	0.2
0.005	Not High	0.8

- F denotes the event that the product fails.
- H denotes the event that the chip is exposed to high contamination during manufacture.
- $P(F|H) = 0.100$ & $P(H) = 0.2$, so $P(F \cap H) = 0.02$
- $P(F|H') = 0.005$ and $P(H') = 0.8$, so $P(F \cap H') = 0.004$
- $P(F) = P(F \cap H) + P(F \cap H') = 0.020 + 0.004 = 0.024$

Total Probability Rule (multiple events)

- Assume E_1, E_2, \dots, E_k are k mutually exclusive & exhaustive subsets. Then:

$$\begin{aligned} P(B) &= P(B \cap E_1) + P(B \cap E_2) + \dots + P(B \cap E_k) \\ &= P(B|E_1)*P(E_1) + P(B|E_2)*P(E_2) + \dots + P(B|E_k)*P(E_k) \quad (2-11) \end{aligned}$$



$$B = (B \cap E_1) \cup (B \cap E_2) \cup (B \cap E_3) \cup (B \cap E_4)$$

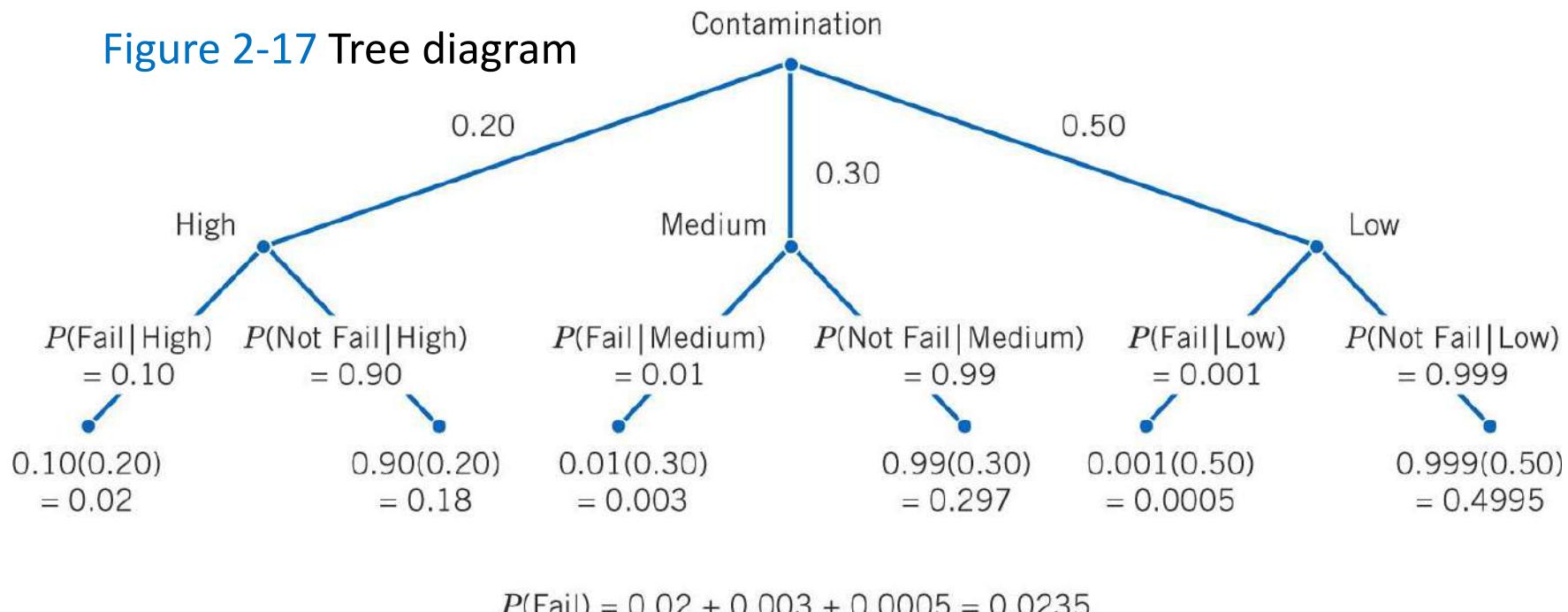
Figure 2-16 Partitioning an event into several mutually exclusive subsets.

Example 2-28: Refined Contamination Data

Continuing the discussion of contamination during chip manufacture:

Probability of Failure	Level of Contamination	Probability of Level
0.100	High	0.2
0.010	Medium	0.3
0.001	Low	0.5

Figure 2-17 Tree diagram



Event Independence

- Two events are independent if any one of the following equivalent statements are true:
 1. $P(B|A) = P(A)$
 2. $P(A|B) = P(B)$
 3. $P(A \cap B) = P(A)*P(B)$
- This means that occurrence of one event has no impact on the occurrence of the other event.

Example 2-29: Sampling With Replacement

- A production lot of 850 parts contains 50 defectives. Two parts are selected at random, but the first **is replaced** before selecting the 2nd.
- Let A denote the event that the 1st part selected is defective. $P(A) = 50/850$
- Let B denote the event that the 2nd part selected is defective. $P(B) = 50/850$
- What is the probability that the 2nd is defective, given that the first part is defective? The same.
- Probability that both are defective is:
$$P(A)*P(B) = 50/850 *50/850 = 0.0035.$$

Example 2-30: Flaw & Functions

The data shows whether the events are independent.

Table 2-3 Parts Classified				Table 2-4 Parts Classified (data chg'd)			
Defective	Surface Flaws			Defective	Surface Flaws		
	Yes (F)	No (F')	Total		Yes (F)	No (F')	Total
Yes (D)	10	18	28	Yes (D)	2	18	20
No (D')	30	342	372	No (D')	38	342	380
Total	40	360	400	Total	40	360	400
$P(D F) = 10/40 = 0.25$				$P(D F) = 2/40 = 0.05$			
$P(F) = 40/400 = 0.10$				$P(F) = 20/400 = 0.05$			
not same				same			
Events D & F are dependent				Events D & F are independent			

Example 2.31: Conditioned vs. Unconditioned

- A production lot of 850 parts contains 50 defectives. Two parts are selected at random, without replacement.
- Let A denote the event that the 1st part selected is defective. $P(A) = 50/850$
- Let B denote the event that the 2nd part selected is defective. $P(B|A) = 49/849$
- Probability that the 2nd is defective is:
$$P(B) = P(B|A)*P(A) + P(B|A')*P(A')$$
$$P(B) = (49/849) * (50/850) + (50/849)*(800/850)$$
$$P(B) = (49*50+50*800) / (849*850)$$
$$P(B) = 50*(49+800) / (849*850)$$
$$P(B) = 50/850$$
 is unconditional, same as $P(A)$
- Since $P(B|A) \neq P(A)$, then A and B are dependent.

Independence with Multiple Events

The events E_1, E_2, \dots, E_k are independent if and only if, for any subset of these events:

$$P(E_1 \cap E_2 \cap \dots \cap E_k) = P(E_1) * P(E_2) * \dots * P(E_k) \quad (2-14)$$

Be aware that, if E_1 & E_2 are independent,
 E_2 & E_3 may or may not be independent.

Example 2-32: Series Circuit

This circuit operates only if there is a path of functional devices from left to right. The probability that each device functions is shown on the graph. Assume that the devices fail independently. What is the probability that the circuit operates?



Let L & R denote the events that the left and right devices operate. The probability that the circuit operates is:

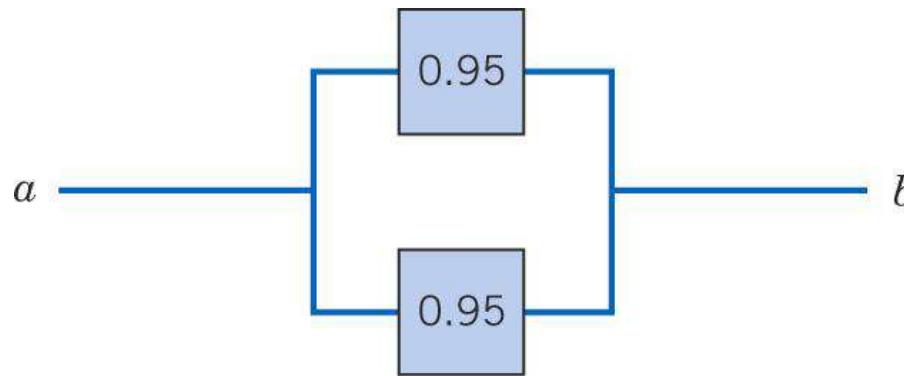
$$P(L \cap R) = P(L) * P(R) = 0.8 * 0.9 = 0.72.$$

Example 2-33: Another Series Circuit

- The probability that a wafer contains a large particle of contamination is 0.01. The wafer events are independent.
- $P(E_i)$ denotes the event that the i^{th} wafer contain no particles and $P(E_i) = 0.99$.
- If 15 wafers are analyzed, what is the probability that no large particles are found?
- $$P(E_1 \cap E_2 \cap \dots \cap E_k) = P(E_1) * P(E_2) * \dots * P(E_k)$$
$$= (0.99)^{15} = 0.86.$$

Example 2-34: Parallel Circuit

This circuit operates only if there is a path of functional devices from left to right. The probability that each device functions is shown. Each device fails independently.



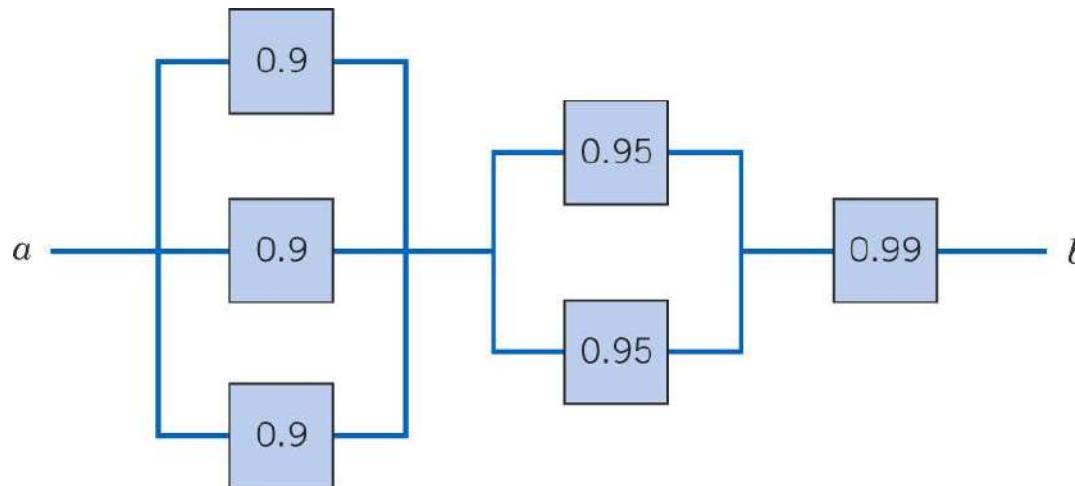
Let T & B denote the events that the top and bottom devices operate. The probability that the circuit operates is:

$$P(T \cup B) = 1 - P(T' \cap B') = 1 - P(T') * P(B') = 1 - 0.05^2 = 1 - 0.0025 = 0.9975.$$

(this is 1 minus the probability that they both don't fail)

Example 2-35: Advanced Circuit

This circuit operates only if there is a path of functional devices from left to right. The probability that each device functions is shown. Each device fails independently.



Partition the graph into 3 columns with L & M denoting the left & middle columns.

$P(L) = 1 - 0.1^3$, and $P(M) = 1 - 0.5^2$, so the probability that the circuit operates is: $(1 - 0.1^3)(1 - 0.05^2)(0.99) = 0.9875$ (this is a series of parallel circuits). In Excel: $0.98752 = (1 - 0.01^3) * (1 - 0.05^2) * 0.99$

Bayes Theorem

- Thomas Bayes (1702-1761) was an English mathematician and Presbyterian minister.
- His idea is that we observe conditional probabilities through prior information.
- The short formal statement is:

$$P(A|B) = \frac{P(B|A)*P(A)}{P(B)} \quad \text{for } P(B) > 0 \quad (2-15)$$

- Note the reversal of the condition!

Example 2-36:

- From Example 2-27, find $P(F)$ which is not given:

Probability of Failure	Level of Contamination	Probability of Level
0.100	High	0.2
0.005	Not High	0.8

$$P(H | F) = \frac{P(B | A) * P(A)}{P(F)} = \frac{0.10 * 0.20}{0.024} = 0.83$$

$$\begin{aligned}P(F) &= P(F | H) * P(H) + P(F | H') * P(H') \\&= 0.1 * .2 + 0.005 * 0.8 = 0.024\end{aligned}$$

Bayes Theorem with Total Probability

- If E_1, E_2, \dots, E_k are k mutually exclusive and exhaustive events and B is any event,

- Note that the:
 - Total probability expression of the denominator
 - Numerator is always one term of the denominator.

Example 2-37: Medical Diagnostic-1

Because a new medical procedure has been shown to be effective in the early detection of a disease, a medical screening of the population is proposed. The probability that the test correctly identifies someone with the disease as positive is 0.99, and probability that the test correctly identifies someone without the disease as negative is 0.95. The incidence of the illness in the general population is 0.0001. You take the test and the result is positive. What is the probability that you have the illness?

Let D denote the event that you have the disease and let S denote the event that the test signals positive. Given info is:

- $P(S'|D') = 0.95$, so $P(S|D') = 0.05$, and $P(D) = 0.0001$,
- $P(S|D) = 0.99$. We desire $P(D|S)$.

Example 2-37: Medical Diagnostic-2

$$\begin{aligned} P(D|S) &= \frac{P(S|D)*P(D)}{P(S|D)*P(D)+P(S|D')*P(D')} \\ &= \frac{0.99 * 0.0001}{0.99 * 0.0001 + 0.05 * (1 - 0.0001)} \\ &= 1/506 = 0.002 \end{aligned}$$

Excel:

$$0.00198 = (0.99*0.0001) / (0.99*0.0001 + 0.05*(1-0.0001))$$

Example 2-37: Medical Diagnostic-2

$$\begin{aligned} P(D|S) &= \frac{P(S|D)*P(D)}{P(S|D)*P(D)+P(S|D')*P(D')} \\ &= \frac{0.99*0.0001}{0.99*0.0001+0.05*(1-0.0001)} \\ &= 1/506 = 0.002 \end{aligned}$$

Excel: $0.00198 = (0.99*0.0001) / (0.99*0.0001 + 0.05*(1-0.0001))$

Before the test, your chance was 0.0001. After the positive result, your chance is 0.00198. So your risk of having the disease has increased 20 times = $0.00198/0.00010$, but is still tiny.

Example 2-38: Bayesian Network-1

- Bayesian networks are used on Web sites of high-tech manufacturers to allow customers to quickly diagnose problems with products. A printer manufacturer obtained the following probabilities from its database. Printer failures are of 3 types: hardware $P(H) = 0.3$, software $P(S)=0.6$, and other $P(O)=0.1$. Also:
 - $P(F|H) = 0.9$, $P(F|S) = 0.2$, $P(F|O) = 0.5$.
- Find the max of $P(H|F)$, $P(S|F)$, $P(O|F)$ to direct the diagnostic effort.

Example 2-38: Bayesian Network-2

$$\begin{aligned}P(F) &= P(F|H)P(H) + P(F|S)P(S) + P(F|O)P(O) \\&= 0.9(0.1) + 0.2(0.6) + 0.5(0.3) = 0.36\end{aligned}$$

$$P(H|F) = \frac{P(F|H)*P(H)}{P(F)} = \frac{0.9*0.1}{0.36} = 0.250$$

$$P(S|F) = \frac{P(F|S)*P(S)}{P(F)} = \frac{0.2*0.6}{0.36} = 0.333$$

$$P(O|F) = \frac{P(F|O)*P(O)}{P(F)} = \frac{0.5*0.3}{0.36} = 0.417$$

Note that the conditionals based on Failure add to 1. Since the Other category is the most likely cause of the failure, diagnostic effort should be so initially directed.

Random Variables

- A variable that associates a number with the outcome of a random experiment is called a **random variable**.
- A **random variable** is a function that assigns a real number to each outcome in the sample space of a random experiment.
- Particular notation is used to distinguish the random variable (rv) from the real number. The rv is denoted by an uppercase letter, such as X . After the experiment is conducted, the measured value is denoted by a lowercase letter, such as $x = 70$. X and x are shown in italics, e.g., $P(X=x)$.

Continuous & Discrete Random Variables

- A **discrete** random variable is a rv with a finite (or countably infinite) range. They are usually integer counts, e.g., number of errors or number of bit errors per 100,000 transmitted (rate). The ends of the range of rv values may be finite ($0 \leq x \leq 5$) or infinite ($x \geq 0$).
- A **continuous** random variable is a rv with an interval (either finite or infinite) of real numbers for its range. Its precision depends on the measuring instrument.

Examples of Discrete & Continuous RVs

- Discrete rv's:
 - Number of scratches on a surface.
 - Proportion of defective parts among 100 tested.
 - Number of transmitted bits received in error.
 - Number of common stock shares traded per day.
- Continuous rv's:
 - Electrical current and voltage.
 - Physical measurements, e.g., length, weight, time, temperature, pressure.

Foundation of Data Science and Analytics

Probability Distribution - 2

Arun K. Timalsina

Example 3-1: Voice Lines

- A voice communication system for a business contains 48 external lines. At a particular time, the system is observed, and some of the lines are being used.
- Let X denote the number of lines in use. Then, X can assume any of the integer values 0 through 48.
- The system is observed at a random point in time. If 10 lines are in use, then $x = 10$.

Example 3-2: Wafers

In a semiconductor manufacturing process, 2 wafers from a lot are sampled. Each wafer is classified as *pass* or *fail*.

Assume that the probability that a wafer passes is 0.8, and that wafers are independent.

The sample space for the experiment and associated probabilities are shown in Table 3-1. The probability that the 1st wafer passes and the 2nd fails, denoted as *pf* is $P(pf) = 0.8 * 0.2 = 0.16$.

The random variable X is defined as the number of wafers that pass.

Table 3-1 Wafer Tests

Outcome			
Wafer #			
1	2	Probability	x
Pass	Pass	0.64	2
Fail	Pass	0.16	1
Pass	Fail	0.16	1
Fail	Fail	0.04	0
		1.00	

Example 3-3: Particles on Wafers

- Define the random variable X to be the number of contamination particles on a wafer. Although wafers possess a number of characteristics, the random variable X summarizes the wafer only in terms of the number of particles. The possible values of X are the integers 0 through a very large number, so we write $x \geq 0$.
- We can also describe the random variable Y as the number of chips made from a wafer that fail the final test. If there can be 12 chips made from a wafer, then we write $0 \leq y \leq 12$. **(changed)**

Probability Distributions

- A random variable X associates the outcomes of a random experiment to a number on the number line.
- The probability distribution of the random variable X is a description of the probabilities with the possible numerical values of X .
- A probability distribution of a discrete random variable can be:
 1. A list of the possible values along with their probabilities.
 2. A formula that is used to calculate the probability in response to an input of the random variable's value.

Example 3-4: Digital Channel

- There is a chance that a bit transmitted through a digital transmission channel is received in error.
- Let X equal the number of bits received in error of the next 4 transmitted.
- The associated probability distribution of X is shown as a graph and as a table.

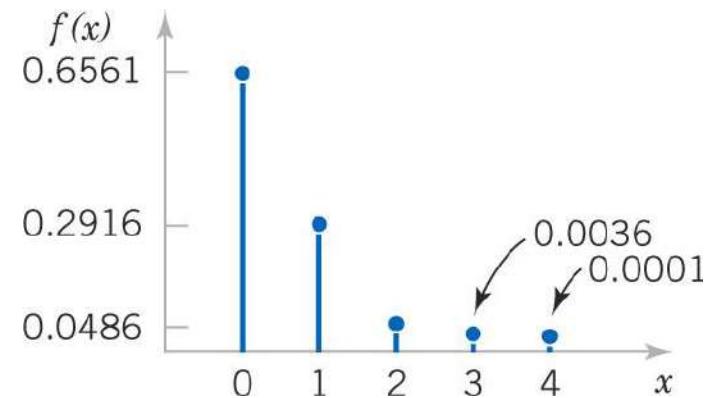


Figure 3-1 Probability distribution for bits in error.

$P(X=0) =$	0.6561
$P(X=1) =$	0.2916
$P(X=2) =$	0.0486
$P(X=3) =$	0.0036
$P(X=4) =$	0.0001
	1.0000

Probability Mass Function

Suppose a loading on a long, thin beam places mass only at discrete points. This represents a probability distribution where the beam is the number line over the range of x and the probabilities represent the mass. That's why it is called a probability **mass** function.

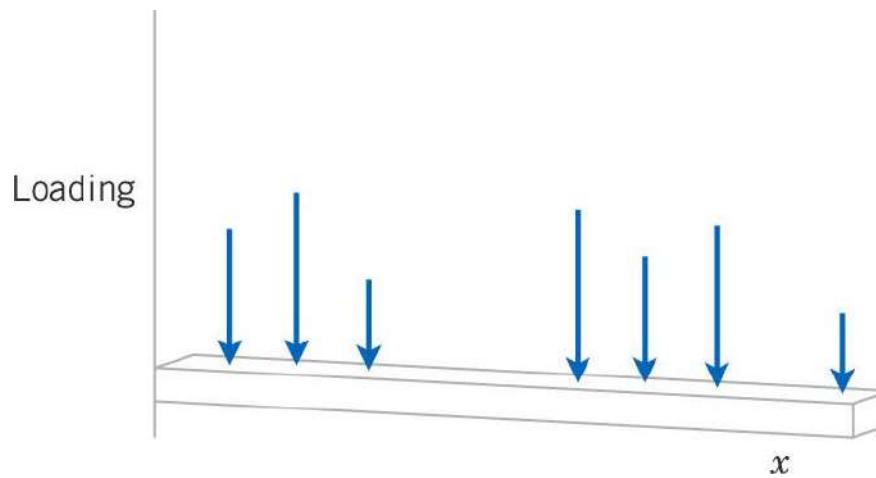


Figure 3-2 Loading at discrete points on a long, thin beam.

Probability Mass Function Properties

For a discrete random variable X with possible values x_1, x_2, \dots, x_n ,
a **probability mass function** is a function such that:

$$(1) \quad f(x_i) \geq 0$$

$$(2) \quad \sum_{i=1}^n f(x_i) = 1$$

$$(3) \quad f(x_i) = P(X = x_i)$$

Example 3-5: Wafer Contamination

- Let the random variable X denote the number of wafers that need to be analyzed to detect a large particle. Assume that the probability that a wafer contains a large particle is 0.1, and that the wafers are independent. Determine the probability distribution of X .
- Let p denote a wafer for which a large particle is **present** & let a denote a wafer in which it is **absent**.
- The sample space is: $S = \{p, ap, aap, aaap, \dots\}$
- The range of the values of X is: $x = 1, 2, 3, 4, \dots$

$P(X=1) =$	0.1	0.1
$P(X=2) =$	$(0.9)*0.1$	0.09
$P(X=3) =$	$(0.9)^2*0.1$	0.081
$P(X=4) =$	$(0.9)^3*0.1$	0.0729
		0.3439

Cumulative Distribution Functions

- Example 3-6: From Example 3.4, we can express the probability of three or fewer bits being in error, denoted as $P(X \leq 3)$.
- The event $(X \leq 3)$ is the union of the **mutually exclusive** events: $(X=0), (X=1), (X=2), (X=3)$.
- From the table:

x	$P(X=x)$	$P(X \leq x)$
0	0.6561	0.6561
1	0.2916	0.9477
2	0.0486	0.9963
3	0.0036	0.9999
4	0.0001	1.0000
	1.0000	

$$P(X \leq 3) = P(X=0) + P(X=1) + P(X=2) + P(X=3) = 0.9999$$

$$P(X = 3) = P(X \leq 3) - P(X \leq 2) = 0.0036$$

Cumulative Distribution Function Properties

The cumulative distribution function is built from the probability mass function and vice versa.

The cumulative distribution function of a discrete random variable X , denoted as $F(x)$, is:

$$F(x) = F(X = x) = \sum_{x_i \leq x} x_i$$

For a discrete random variable X , $F(x)$ satisfies the following properties:

$$(1) \quad F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i)$$

$$(2) \quad 0 \leq F(x) \leq 1$$

$$(3) \quad \text{If } x \leq y, \text{ then } F(x) \leq F(y)$$

Example 3-7: Cumulative Distribution Function

- Determine the probability mass function of X from this cumulative distribution function:

$F(x) =$	
0.0	$x < -2$
0.2	$-2 \leq x < 0$
0.7	$0 \leq x < 2$
1.0	$2 \leq x$

PMF
$f(2) = 0.2$
$f(0) = 0.5$
$f(-2) = 0.3$

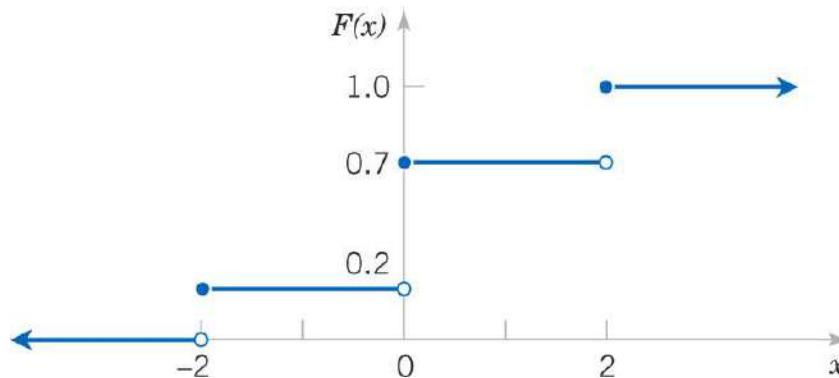


Figure 3-3 Graph of the CDF

Example 3-8: Sampling without Replacement

A day's production of 850 parts contains 50 defective parts.

Two parts are selected at random without replacement.

Let the random variable X equal the number of defective parts in the sample. Create the CDF of X .

$$P(X = 0) = \frac{800}{850} \cdot \frac{799}{849} = 0.886$$

$$P(X = 1) = 2 \cdot \frac{800}{850} \cdot \frac{50}{849} = 0.111$$

$$P(X = 2) = \frac{50}{850} \cdot \frac{49}{849} = 0.003$$

Therefore,

$$F(0) = P(X \leq 0) = 0.886$$

$$F(1) = P(X \leq 1) = 0.997$$

$$F(2) = P(X \leq 2) = 1.000$$

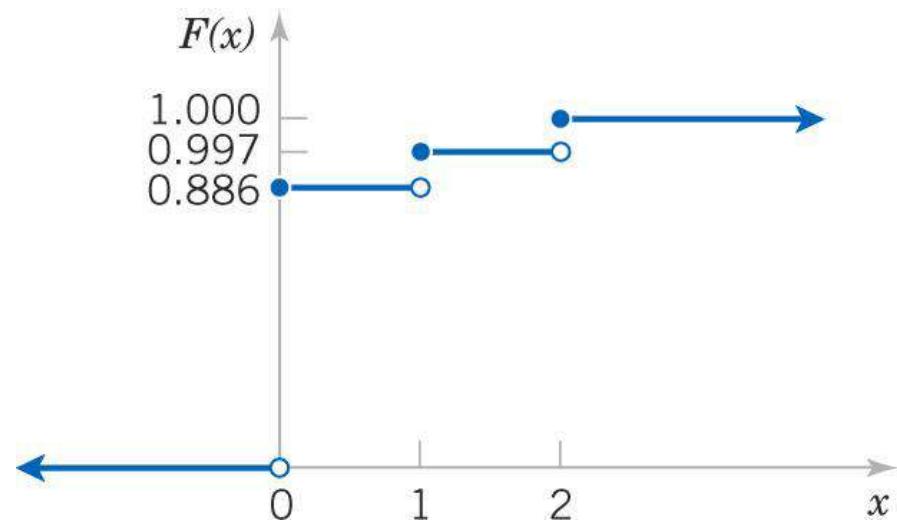


Figure 3-4 CDF. Note that $F(x)$ is defined for all x , $-\infty < x < \infty$, not just 0, 1 and 2.

Summary Numbers of a Probability Distribution

- The **mean** is a measure of the center of a probability distribution.
- The **variance** is a measure of the dispersion or variability of a probability distribution.
- The **standard deviation** is another measure of the dispersion. It is the square root of the variance.

Mean Defined

The **mean** or **expected value** of the discrete random variable X , denoted as μ or $E(X)$, is

$$\mu = E(X) = \sum_x x \cdot f(x)$$

- The mean is the weighted average of the possible values of X , the weights being the probabilities where the beam balances. It represents the center of the distribution. It is also called the arithmetic mean.
- If $f(x)$ is the probability mass function representing the loading on a long, thin beam, then $E(X)$ is the fulcrum or point of balance for the beam.
- The mean value may, or may not, be a given value of x .

Variance Defined

The **variance** of X , denoted as σ^2 or $V(X)$, is

$$\sigma^2 = V(X) = E(X - \mu)^2 = \sum_x (x - \mu)^2 \cdot f(x) = \sum_x x^2 \cdot f(x) - \mu^2$$

- The variance is the measure of dispersion or scatter in the possible values for X .
- It is the average of the squared deviations from the distribution mean.

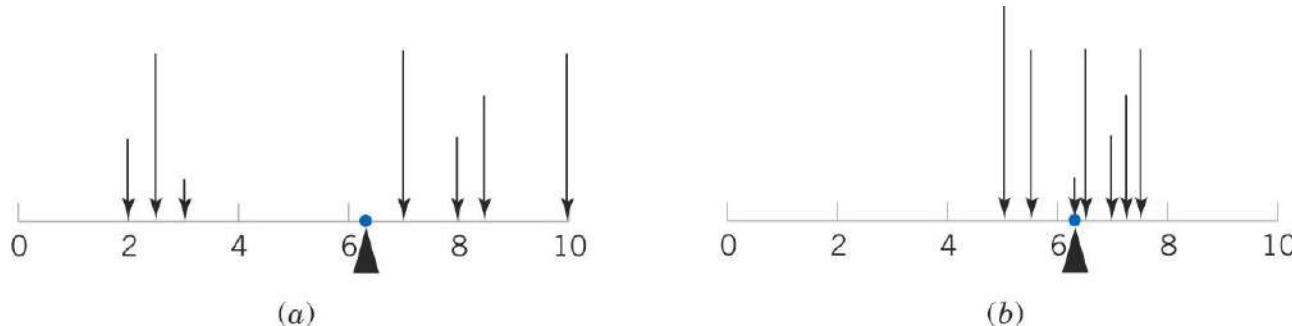


Figure 3-5 The mean is the balance point. Distributions (a) & (b) have equal mean, but (a) has a larger variance.

Variance Formula Derivations

$$\begin{aligned} V(X) &= \sum_x (x - \mu)^2 f(x) \text{ is the } \text{definitional} \text{ formula} \\ &= \sum_x (x^2 - 2\mu x + \mu^2) f(x) \\ &= \sum_x x^2 f(x) - 2\mu \sum_x x f(x) + \mu^2 \sum_x f(x) \\ &= \sum_x x^2 f(x) - 2\mu^2 + \mu^2 \\ &= \sum_x x^2 f(x) - \mu^2 \text{ is the } \text{computational} \text{ formula} \end{aligned}$$

The computational formula is easier to calculate manually.

Different Distributions Have Same Measures

These measures do not uniquely identify a probability distribution – different distributions could have the same mean & variance.

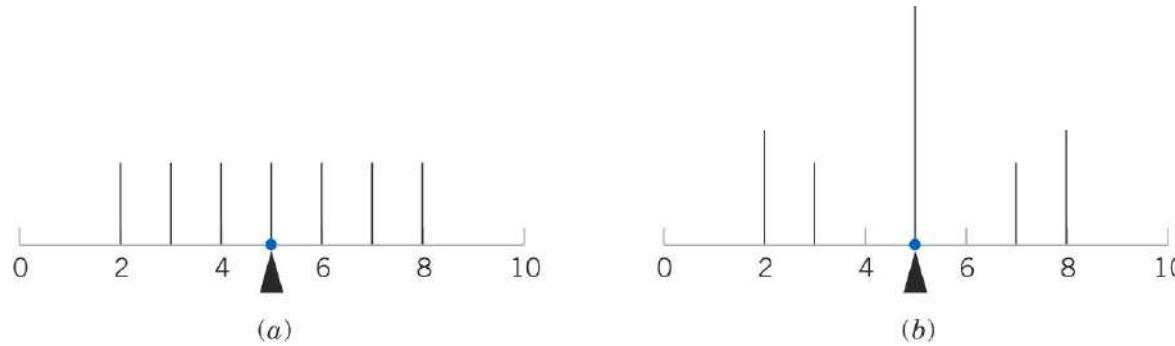


Figure 3-6 These probability distributions have the same mean and variance measures, but are very different in shape.

Exercise 3-9: Digital Channel

In Exercise 3-4, there is a chance that a bit transmitted through a digital transmission channel is an error. X is the number of bits received in error of the next 4 transmitted. Use table to calculate the mean & variance.

Definitional formula					
x	$f(x)$	$x * f(x)$	$(x-0.4)^2$	$(x-0.4)^2 * f(x)$	$x^2 * f(x)$
0	0.6561	0.0000	0.160	0.1050	0.0000
1	0.2916	0.2916	0.360	0.1050	0.2916
2	0.0486	0.0972	2.560	0.1244	0.1944
3	0.0036	0.0108	6.760	0.0243	0.0324
4	0.0001	0.0004	12.960	0.0013	0.0016
Totals =		0.4000		0.3600	0.5200
= Mean			= Variance (σ^2)		$= E(x^2)$
$= \mu$			$\sigma^2 = E(x^2) - \mu^2 =$		0.3600
Computational formula					

Exercise 3-10 Marketing

- Two new product designs are to be compared on the basis of revenue potential. Revenue from Design A is predicted to be \$3 million. But for Design B, the revenue could be \$7 million with probability 0.3 or only \$2 million with probability 0.7. Which design is preferable?
- Answer:
 - Let X & Y represent the revenues for products A & B.
 - $E(X) = \$3$ million. $V(X) = 0$ because x is certain.
 - $E(Y) = \$3.5$ million = $7*0.3 + 2*0.7 = 2.1 + 1.4$
 - $V(X) = 5.25$ million dollars² or $(7-3.5)^2*.3 + (2-3.5)^2*.7 = 3.675 + 1.575$
 - $SD(X) = 2.29$ million dollars , the square root of the variance.
 - Standard deviation has the same units as the mean, not the squared units of the variance.

Exercise 3-11: Messages

The number of messages sent per hour over a computer network has the following distribution. Find the mean & standard deviation of the number of messages sent per hour.

x	$f(x)$	$x * f(x)$	$x^2 * f(x)$
10	0.08	0.80	8
11	0.15	1.65	18.15
12	0.30	3.60	43.2
13	0.20	2.60	33.8
14	0.20	2.80	39.2
15	0.07	1.05	15.75
	1.00	12.50	158.10
		$= E(X)$	$= E(X^2)$

$$\text{Mean} = 12.5$$

$$\text{Variance} = 158.10^2 - 12.5^2 = 1.85$$

$$\text{Standard deviation} = 1.36$$

$$\text{Note that: } E(X^2) \neq [E(X)]^2$$

A Function of a Random Variable

If X is a discrete random variable with probability mass function $f(x)$,

$$E[h(X)] = \sum_x h(x)f(x) \quad (3-4)$$

If $h(x) = (X - \mu)^2$, then its expectation is the variance of X .

Example 3-12: Digital Channel

In Example 3-9, X is the number of bits in error in the next four bits transmitted. What is the expected value of the square of the number of bits in error?

x	$f(x)$	$x^2 * f(x)$
0	0.6561	0.0000
1	0.2916	0.2916
2	0.0486	0.1944
3	0.0036	0.0324
4	0.0001	0.0016
	1.0000	0.5200
		$= E(x^2)$

Discrete Uniform Distribution

- Simplest discrete distribution.
- The random variable X assumes only a finite number of values, each with equal probability.
- A random variable X has a discrete uniform distribution if each of the n values in its range, say x_1, x_2, \dots, x_n , has equal probability.

$$f(x_i) = 1/n \quad (3-5)$$

Example 3-13: Discrete Uniform Random Variable

The first digit of a part's serial number is equally likely to be the digits 0 through 9. If one part is selected from a large batch & X is the 1st digit of the serial number, then X has a discrete uniform distribution as shown.

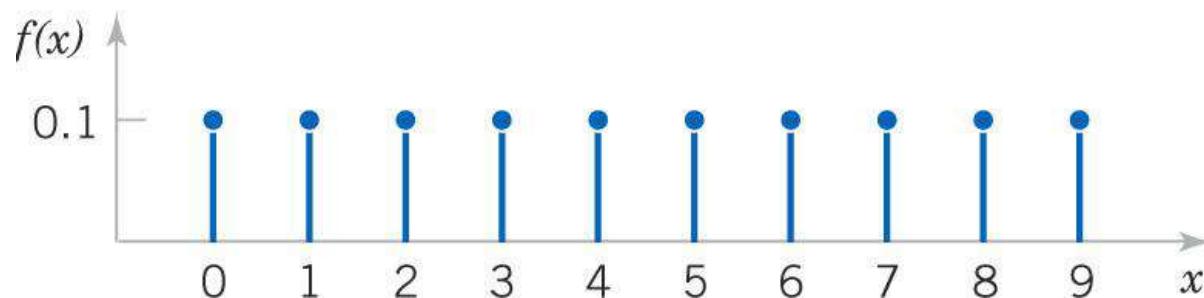


Figure 3-7 Probability mass function, $f(x) = 1/10$ for $x = 0, 1, 2, \dots, 9$

General Discrete Uniform Distribution

- Let X be a discrete uniform random variable from a to b for $a < b$. There are $b - (a-1)$ values in the inclusive interval. Therefore:

$$f(x) = 1/(b-a+1)$$

- Its measures are:

$$\mu = E(x) = 1/(b-a)$$

$$\sigma^2 = V(x) = [(b-a+1)^2-1]/12 \quad (3-6)$$

Note that the mean is the midpoint of a & b .

Example 3-14: Number of Voice Lines

Per Example 3-1, let the random variable X denote the number of the 48 voice lines that are in use at a particular time. Assume that X is a discrete uniform random variable with a range of 0 to 48. Find $E(X)$ & $SD(X)$.

Answer:

$$\mu = \frac{48+0}{2} = 24$$

$$\sigma_x = \sqrt{\frac{(48-0+1)^2 - 1}{12}} = \sqrt{\frac{2400}{12}} = 14.142$$

Example 3-15 Proportion of Voice Lines

Let the random variable Y denote the proportion of the 48 voice line that are in use at a particular time & X as defined in the prior example. Then $Y = X/48$ is a proportion. Find $E(Y)$ & $V(Y)$.

Answer:

$$E(Y) = \frac{E(X)}{48} = \frac{24}{48} = 0.5$$

$$V(Y) = \frac{V(X)}{48^2} = \frac{14.142^2}{2304} = 0.0868$$

Examples of Binomial Random Variables

1. Flip a coin 10 times. $X = \#$ heads obtained.
2. A worn tool produces 1% defective parts. $X = \#$ defective parts in the next 25 parts produced.
3. A multiple-choice test contains 10 questions, each with 4 choices, and you guess. $X = \#$ of correct answers.
4. Of the next 20 births, let $X = \#$ females.

These are binomial experiments having the following characteristics:

1. Fixed number of trials (n).
2. Each trial is termed a success or failure. X is the # of successes.
3. The probability of success in each trial is constant (p).
4. The outcomes of successive trials are independent.

Example 3-16: Digital Channel

The chance that a bit transmitted through a digital transmission channel is received in error is 0.1. Assume that the transmission trials are independent. Let X = the number of bits in error in the next 4 bits transmitted. Find $P(X=2)$.

Answer:

Let E denote a bit in error

Let O denote an OK bit.

Sample space & x listed in table.

6 outcomes where $x = 2$.

Prob of each is $0.1^2 * 0.9^2 = 0.0081$

$Prob(X=2) = 6 * 0.0081 = 0.0486$

$$P(X=2) = C_2^4 (0.1)^2 (0.9)^2$$

Outcome	x	Outcome	x
0000	0	E000	1
000E	1	EO00	2
00OE	1	EO0E	2
0OEE	2	EOEE	3
OEO0	1	EEO0	2
OEOE	2	EEOE	3
OEE0	2	EEE0	3
OEEE	3	EEEE	4

Binomial Distribution Definition

- The random variable X that equals the number of trials that result in a success is a binomial random variable with parameters $0 < p < 1$ and $n = 0, 1, \dots$
- The probability mass function is:

$$f(x) = C_x^n p^x (1-p)^{n-x} \quad \text{for } x = 0, 1, \dots, n \quad (3-7)$$

- Based on the binomial expansion:

$$(a+b)^n = \sum_{k=0}^n C_k^n a^k b^{n-k}$$

Binomial Distribution Shapes

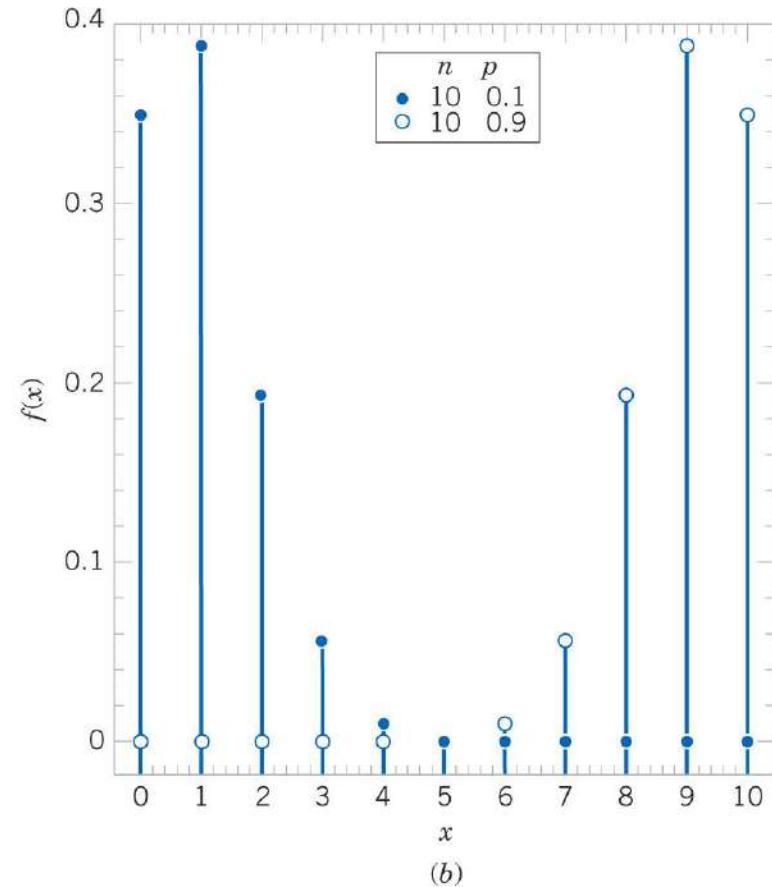
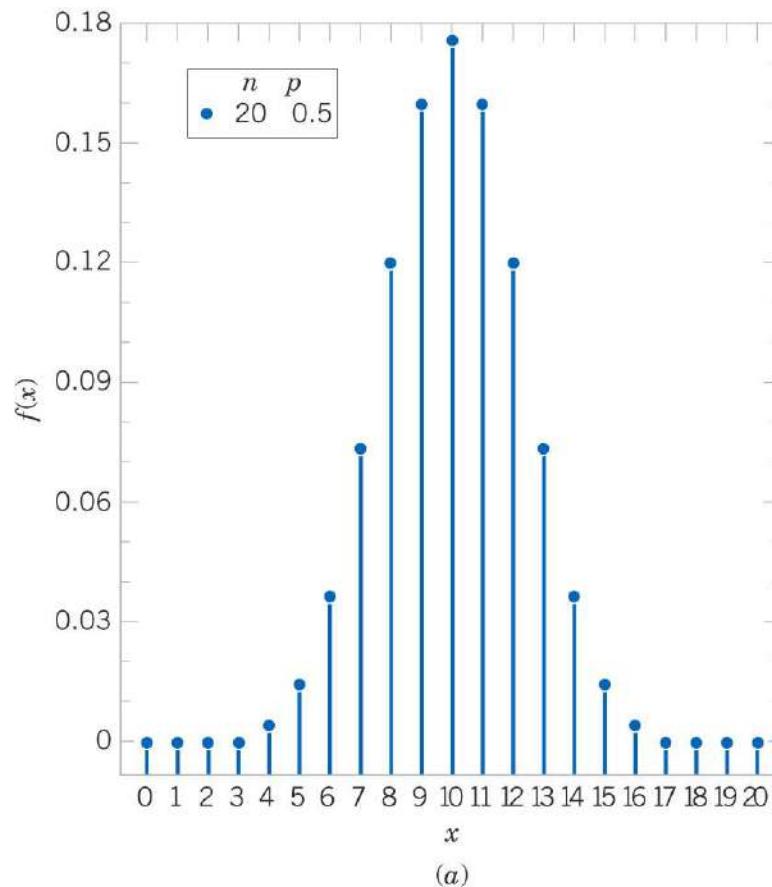


Figure 3-8 Binomial Distributions for selected values of n and p . Distribution (a) is symmetrical, while distributions (b) are skewed. The skew is right if p is small.

Example 3-17: Binomial Coefficients

Exercises in binomial coefficient calculation:

$$C_3^{10} = \frac{10!}{3!7!} = \frac{10 \cdot 9 \cdot 8 \cdot 7!}{3 \cdot 2 \cdot 1 \cdot 7!} = 120$$

$$C_{10}^{15} = \frac{15!}{10!5!} = \frac{15 \cdot 14 \cdot 13 \cdot 12 \cdot 11}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 3,003$$

$$C_4^{100} = \frac{100!}{4!96!} = \frac{100 \cdot 99 \cdot 98 \cdot 97}{4 \cdot 3 \cdot 2 \cdot 1} 3,921,225$$

Exercise 3-18: Organic Pollution-1

Each sample of water has a 10% chance of containing a particular organic pollutant. Assume that the samples are independent with regard to the presence of the pollutant. Find the probability that, in the next 18 samples, exactly 2 contain the pollutant.

Answer: Let X denote the number of samples that contain the pollutant in the next 18 samples analyzed. Then X is a binomial random variable with $p = 0.1$ and $n = 18$

$$P(X = 2) = C_2^{18} (0.1)^2 (0.9)^{16} = 153(0.1)^2 (0.9)^{16} = 0.2835$$

0.2835	= BINOMDIST(2,18,0.1,FALSE)
--------	-----------------------------

Exercise 3-18: Organic Pollution-2

Determine the probability that at least 4 samples contain the pollutant.

Answer:

$$\begin{aligned} P(X \geq 4) &= \sum_{x=4}^{18} C_x^{18} (0.1)^x (0.9)^{18-x} \\ &= 1 - P(X < 4) \\ &= 1 - \sum_{x=0}^3 C_x^{18} (0.1)^x (0.9)^{18-x} \\ &= 0.098 \end{aligned}$$

0.0982	= 1 - BINOMDIST(3,18,0.1,TRUE)
--------	--------------------------------

Exercise 3-18: Organic Pollution-3

Now determine the probability that $3 \leq X \leq 7$.

Answer:

$$P(3 \leq X \leq 7) = \sum_{x=3}^7 C_x^{18} (0.1)^x (0.9)^{18-x} = 0.265$$

$$P(X \leq 7) - P(X \leq 2)$$

0.2660	= BINOMDIST(7,18,0.1,TRUE) - BINOMDIST(2,18,0.1,TRUE)
--------	---

Appendix A, Table II (pg. 705) is a cumulative binomial table for selected values of p and n .

Binomial Mean and Variance

If X is a binomial random variable with parameters p and n ,

$$\mu = E(X) = np \quad \text{and} \quad \sigma^2 = V(X) = np(1-p) \quad (3-8)$$

Example 3-19:

For the number of transmitted bit received in error in Example 3-16, $n = 4$ and $p = 0.1$. Find the mean and variance of the binomial random variable.

Answer:

$$\mu = E(X) = np = 4 * 0.1 = 0,4$$

$$\sigma^2 = V(X) = np(1-p) = 4 * 0.1 * 0.9 = 3.6$$

$$\sigma = SD(X) = 1.9$$

Example 3-20: New Idea

The probability that a bit, sent through a digital transmission channel, is received in error is 0.1. Assume that the transmissions are independent. Let X denote the number of bits transmitted until the 1st error.

$P(X=5)$ is the probability that the 1st four bits are transmitted correctly and the 5th bit is in error.

$$P(X=5) = P(OOOOE) = 0.9^4 0.1 = 0.0656.$$

x is the total number of bits sent.

This illustrates the geometric distribution.

Geometric Distribution

- Similar to the binomial distribution – a series of Bernoulli trials with fixed parameter p .
- Binomial distribution has:
 - Fixed number of trials.
 - Random number of successes.
- Geometric distribution has reversed roles:
 - Random number of trials.
 - Fixed number of successes, in this case 1.
- $f(x) = p(1-p)^{x-1}$ where: (3-9)
 - $x = 1, 2, \dots \infty$, the number of failures until the 1st success.
 - $0 < p < 1$, the probability of success.

Geometric Graphs

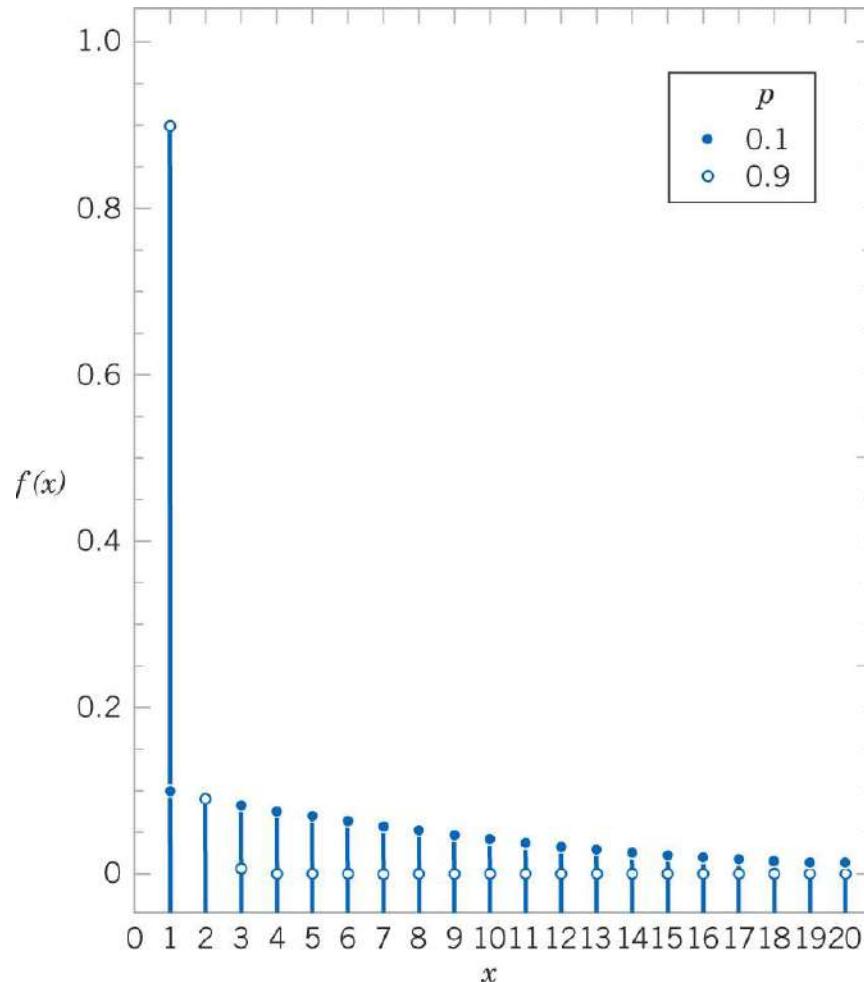


Figure 3-9 Geometric distributions for parameter p values of 0.1 and 0.9. The graphs coincide at $x = 2$.

Example 3.21: Geometric Problem

The probability that a wafer contains a large particle of contamination is 0.01. Assume that the wafers are independent. What is the probability that exactly 125 wafers need to be analyzed before a particle is detected?

Answer:

Let X denote the number of samples analyzed until a large particle is detected. Then X is a geometric random variable with parameter $p = 0.01$.

$$P(X=125) = (0.99)^{124}(0.01) = 0.00288.$$

Geometric Mean & Variance

- If X is a geometric random variable with parameter p ,

$$\mu = E(X) = \frac{1}{p} \quad \text{and} \quad \sigma^2 = V(X) = \frac{(1-p)}{p^2} \quad (3-10)$$

Exercise 3-22: Geometric Problem

Consider the transmission of bits in Exercise 3-20.

Here, $p = 0.1$. Find the mean and standard deviation.

Answer:

$$\text{Mean} = \mu = E(X) = 1 / p = 1 / 0.1 = 10$$

$$\text{Variance} = \sigma^2 = V(X) = (1-p) / p^2 = 0.9 / 0.01 = 90$$

$$\text{Standard deviation} = \sqrt{90} = 9.487$$

Lack of Memory Property

- For a geometric random variable, the trials are independent. Thus the count of the number of trials until the next success can be started at any trial without changing the probability.
- The probability that the next bit error will occur on bit 106, given that 100 bits have been transmitted, is the same as it was for bit 006.
- Implies that the system does not wear out!

Example 3-23: Lack of Memory

In Example 3-20, the probability that a bit is transmitted in error is 0.1. Suppose 50 bits have been transmitted. What is the mean number of bits transmitted until the next error?

Answer:

The mean number of bits transmitted until the next error, after 50 bits have already been transmitted, is $1 / 0.1 = 10$.

Example 3-24: New Idea

The probability that a bit, sent through a digital transmission channel, is received in error is 0.1.

Assume that the transmissions are independent. Let X denote the number of bits transmitted until the 4th error.

$P(X=10)$ is the probability that 3 errors occur over the first 9 trials, then the 4th success occurs on the 10th trial.

$$3 \text{ errors occur over the first 9 trials} = C_3^9 p^3 (1-p)^6$$

$$4\text{th error occurs on the 10th trial} = C_3^9 p^4 (1-p)^6$$

In general, probabilities for X can be determined as follows. Here $P(X = x)$ implies that $r - 1$ successes occur in the first $x - 1$ trials and the r th success occurs on trial x . The probability that $r - 1$ successes occur in the first $x - 1$ trials is obtained from the binomial distribution to be

$$\binom{x-1}{r-1} p^{r-1} (1-p)^{x-r}$$

for $r \leq x$. The probability that trial x is a success is p . Because the trials are independent, these probabilities are multiplied so that

$$P(X = x) = \binom{x-1}{r-1} p^{r-1} (1-p)^{x-r} p$$

This leads to the following result.

In a series of Bernoulli trials (independent trials with constant probability p of a success), the random variable X that equals the number of trials until r successes occur is a **negative binomial random variable** with parameters $0 < p < 1$ and $r = 1, 2, 3, \dots$, and

$$f(x) = \binom{x-1}{r-1} (1-p)^{x-r} p^r \quad x = r, r+1, r+2, \dots \quad (3-11)$$

Negative Binomial Definition

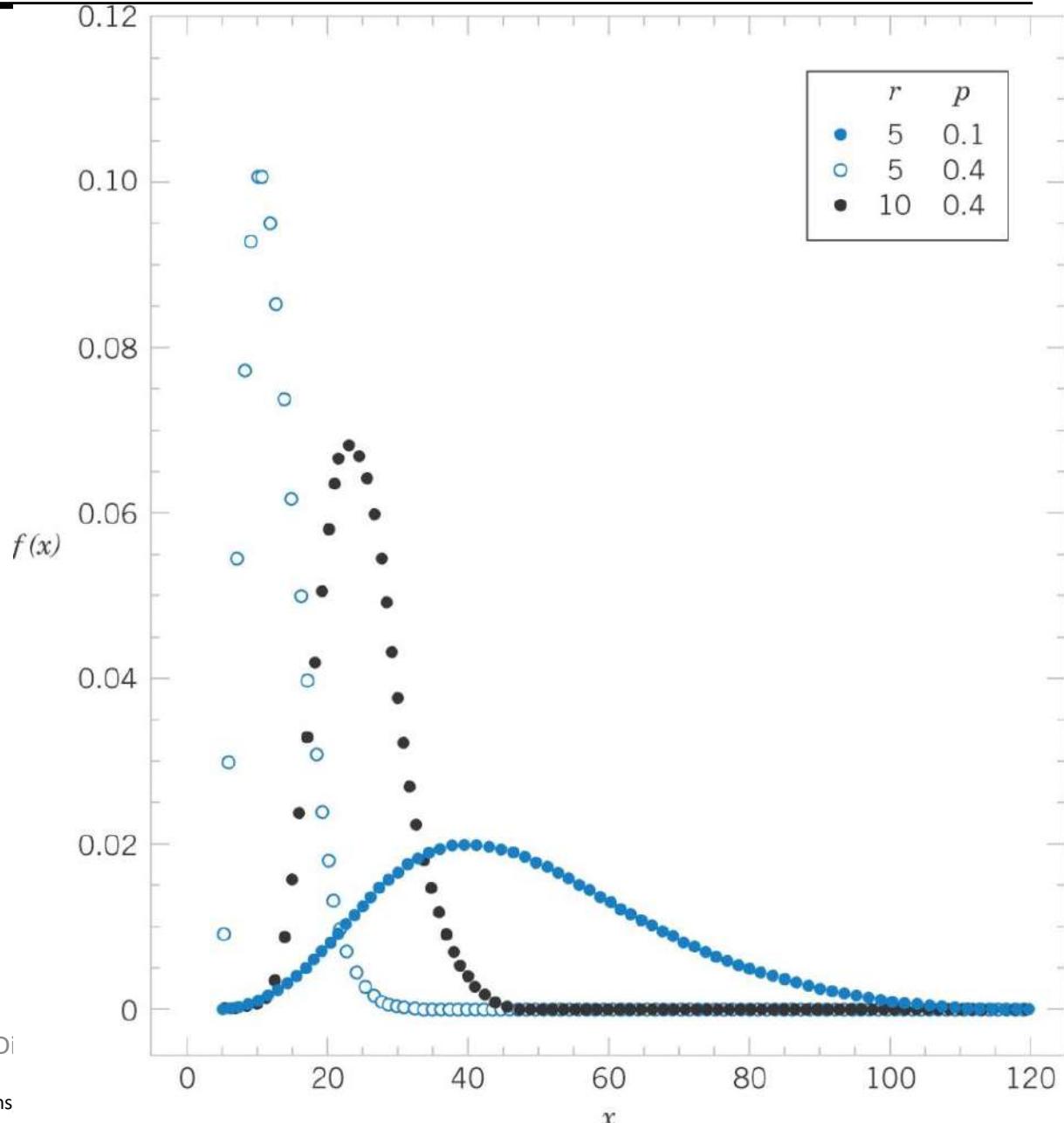
- In a series of independent trials with constant probability of success, let the random variable X denote the number of trials until r successes occur. Then X is a **negative binomial** random variable with parameters $0 < p < 1$ and $r = 1, 2, 3, \dots$
- The probability mass function is:

$$f(x) = C_{r-1}^{x-1} p^r (1-p)^{x-r} \quad \text{for } x = r, r+1, r+2, \dots \quad (3-11)$$

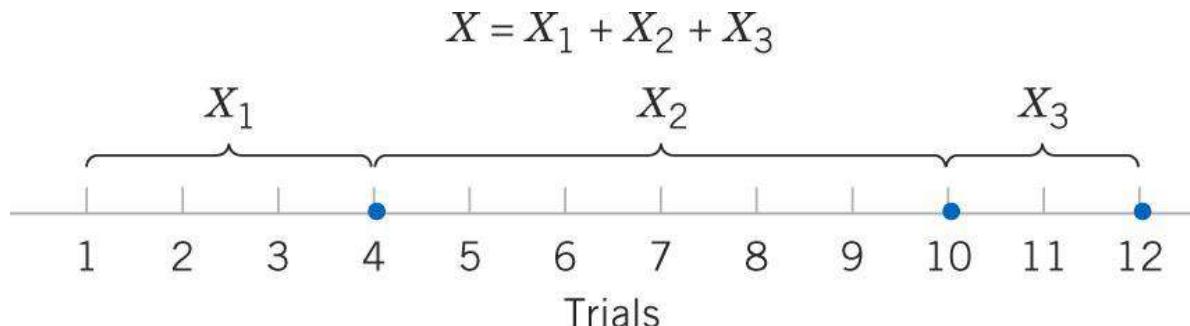
- From the prior example for $f(X=10 | r=4)$:
 - $x-1 = 9$
 - $r-1 = 3$

Negative Binomial Graphs

Figure 3-10
Negative binomial distributions for 3 different parameter combinations.



Lack of Memory Property



- indicates a trial that results in a "success."

- Let X_1 denote the number of trials to the 1st success.
- Let X_2 denote the number of trials to the 2nd success, since the 1st success.
- Let X_3 denote the number of trials to the 3rd success, since the 2nd success.
- Let the X_i be geometric random variables – independent, so without memory.
- Then $X = X_1 + X_2 + X_3$
- Therefore, X is a negative binomial random variable, a sum of three geometric rv's.

Negative Binomial Mean & Variance

- If X is a negative binomial random variable with parameters p and r ,

$$\mu = E(X) = \frac{r}{p} \quad \text{and} \quad \sigma^2 = V(X) = \frac{r(1-p)}{p^2} \quad (3-12)$$

What's In A Name?

- Binomial distribution:
 - Fixed number of trials (n).
 - Random number of successes (x).
- Negative binomial distribution:
 - Random number of trials (x).
 - Fixed number of successes (r).
- Because of the reversed roles, a negative binomial can be considered the opposite or negative of the binomial.

Example 3-25: Web Servers-1

A Web site contains 3 identical computer servers. Only one is used to operate the site, and the other 2 are spares that can be activated in case the primary system fails. The probability of a failure in the primary computer (or any activated spare) from a request for service is 0.0005. Assume that each request represents an independent trial. What is the mean number of requests until failure of all 3 servers?

Answer:

- Let X denote the number of requests until all three servers fail.
- Let $r = 3$ and $p=0.0005 = 1/2000$
- Then $\mu = 3 / 0.0005 = 6,000$ requests

Example 3-25: Web Servers-2

What is the probability that all 3 servers fail within 5 requests? ($X = 5$)

Answer:

$$\begin{aligned}P(X \leq 5) &= P(X = 3) + P(X = 4) + P(X = 5) \\&= 0.005^3 + C_2^3 0.0005^3 0.9995 + C_2^4 0.0005^3 0.9995^2\end{aligned}$$

In Excel	
1.250E-10	= 0.0005^3
3.748E-10	= NEGBINOMDIST(1, 3, 0.0005)
7.493E-10	= NEGBINOMDIST(2, 3, 0.0005)
1.249E-09	

Note that Excel uses a different definition of X ; # of failures before the r^{th} success, not # of trials.

Hypergeometric Distribution

- Applies to sampling without replacement.
- Trials are not independent & a tree diagram used.
- A set of N objects contains:
 - K objects classified as success
 - $N - K$ objects classified as failures
- A sample of size n objects is selected without replacement from the N objects, where:
 - $K \leq N$ and $n \leq N$
- Let the random variable X denote the number of successes in the sample. Then X is a hypergeometric random variable.

$$f(x) = \frac{\binom{K}{x} \binom{N-K}{n-x}}{\binom{N}{n}} \text{ where } x = \max(0, n+K-N) \text{ to } \min(K, n) \quad (3-13)$$

Hypergeometric Graphs

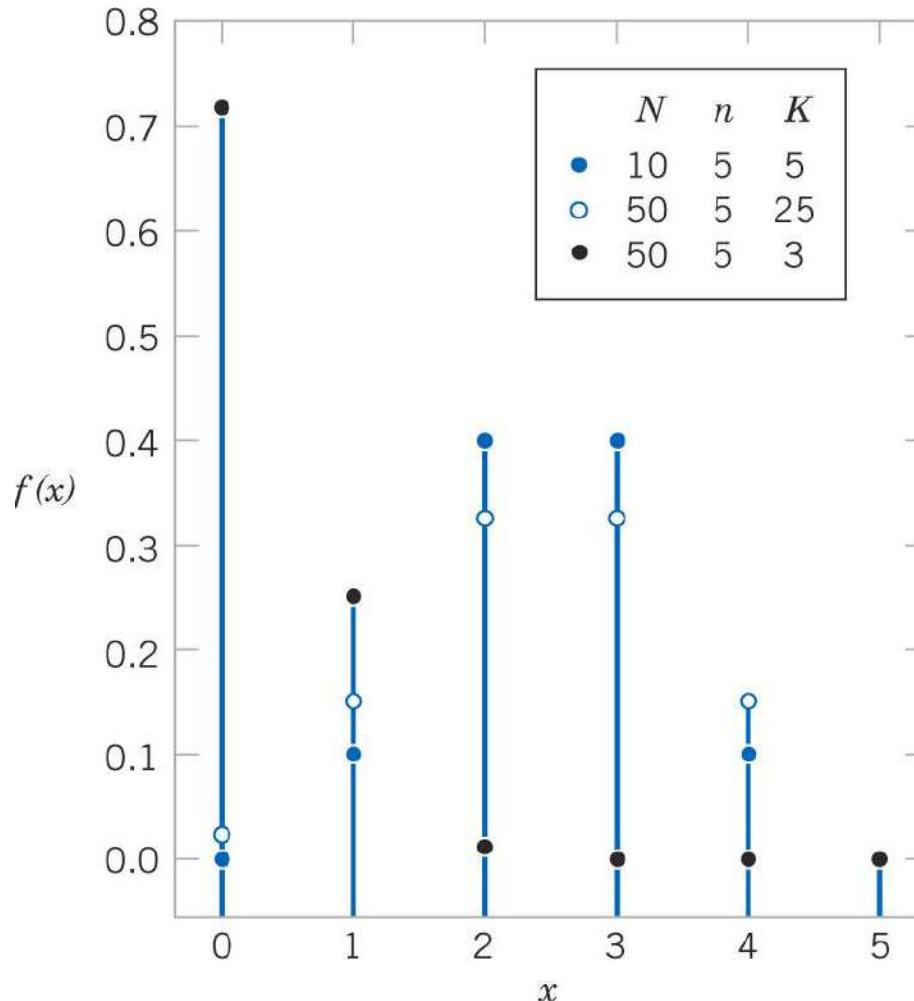


Figure 3-12 Hypergeometric distributions for 3 parameter sets of N , K , and n .

Example 3-26: Sampling without Replacement

From an earlier example, 50 parts are defective on a lot of 850. Two are sampled. Let X denote the number of defectives in the sample. Use the hypergeometric distribution to find the probability distribution.

Answer:

In Excel	
0.8857	= HYPGEOMDIST(0,2,50,850)
0.1109	= HYPGEOMDIST(1,2,50,850)
0.0034	= HYPGEOMDIST(2,2,50,850)

$$P(X = 0) = \frac{\binom{50}{0} \binom{800}{2}}{\binom{850}{2}} = \frac{319,660}{360,825} = 0.886$$

$$P(X = 1) = \frac{\binom{50}{1} \binom{800}{1}}{\binom{850}{2}} = \frac{40,000}{360,825} = 0.111$$

$$P(X = 2) = \frac{\binom{50}{2} \binom{800}{0}}{\binom{850}{2}} = \frac{1,225}{360,825} = 0.003$$

Example 3-27: Parts from Suppliers-1

A batch of parts contains 100 parts from supplier A and 200 parts from Supplier B. If 4 parts are selected randomly, without replacement, what is the probability that they are all from Supplier A?

Answer:

Let X equal the number
of parts in the sample
from Supplier A.

$$P(X = 4) = \frac{\binom{100}{4} \binom{200}{0}}{\binom{300}{4}} = 0.0119$$

In Excel

0.01185 = HYPGEOMDIST(4,100,4,300)

Example 3-27: Parts from Suppliers-2

What is the probability that two or more parts are from Supplier A?

Answer:

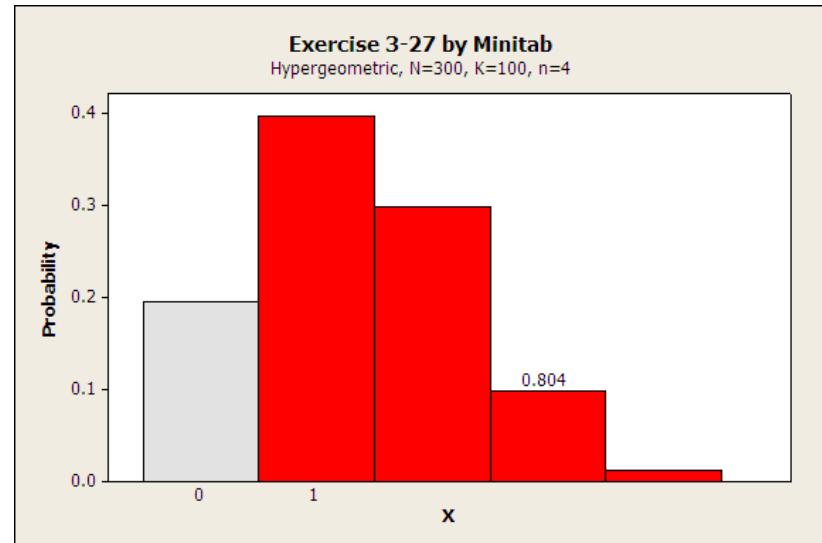
$$\begin{aligned} P(X \geq 2) &= P(X = 2) + P(X = 3) + P(X = 4) \\ &= \frac{\binom{100}{2}\binom{200}{2}}{\binom{300}{4}} + \frac{\binom{100}{3}\binom{200}{1}}{\binom{300}{4}} + \frac{\binom{100}{4}\binom{200}{1}}{\binom{300}{4}} \\ &= 0.298 + 0.098 + 0.0119 = 0.408 \end{aligned}$$

In Excel	
0.40741	= HYPGEOMDIST(2,100,4,300) + HYPGEOMDIST(3,100,4,300) + HYPGEOMDIST(4,100,4,300)

Example 3-27: Parts from Suppliers-3

What is the probability that at least one part is from Supplier A?

Answer:



$$P(X \geq 1) = 1 - P(X = 0) = 1 - \frac{\binom{100}{0} \binom{200}{4}}{\binom{300}{4}} = 0.804$$

In Excel

`0.80445 = 1 - HYPGEOMDIST(0,100,4,300)`

Hypergeometric Mean & Variance

- If X is a hypergeometric random variable with parameters N , K , and n , then

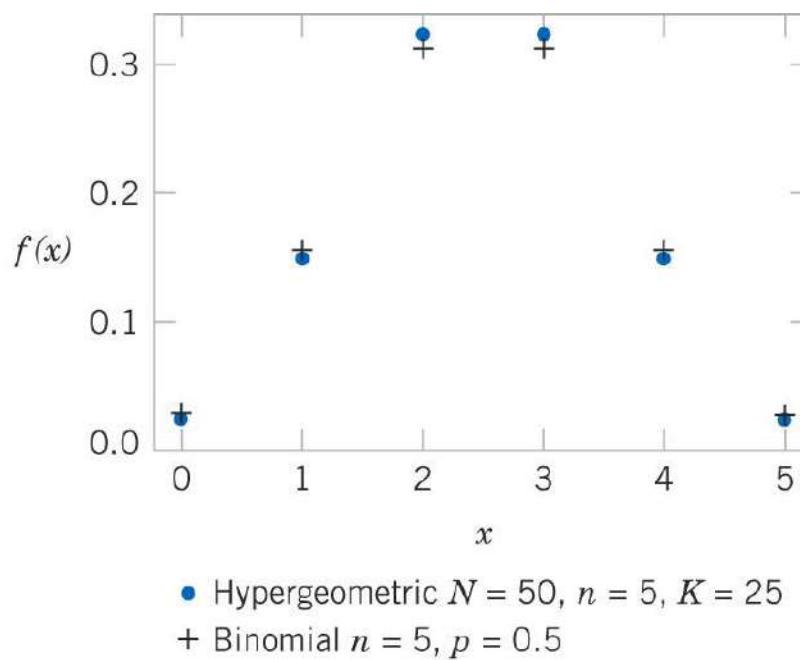
$$\mu = E(X) = np \quad \text{and} \quad \sigma^2 = V(X) = np(1-p)\left(\frac{N-n}{N-1}\right) \quad (3-14)$$

where $p = K/N$

and $\left(\frac{N-n}{N-1}\right)$ is the finite population correction factor.

σ^2 approaches the binomial variance as n/N becomes small.

Hypergeometric & Binomial Graphs



	0	1	2	3	4	5
Hypergeometric probability	0.025	0.149	0.326	0.326	0.149	0.025
Binomial probability	0.031	0.156	0.312	0.312	0.156	0.031

Figure 3-13 Comparison of hypergeometric and binomial distributions.

Example 3-29: Customer Sample-1

A listing of customer accounts at a large corporation contains 1,000 accounts. Of these, 700 have purchased at least one of the company's products in the last 3 months. To evaluate a new product, 50 customers are sampled at random from the listing. What is the probability that more than 45 of the sampled customers have purchased in the last 3 months?

Let X denote the number of customers in the sample who have purchased from the company in the last 3 months. Then X is a hypergeometric random variable with $N = 1,000$, $K = 700$, $n = 50$. This a lengthy problem! ☹

$$P(X > 45) = \sum_{x=46}^{50} \frac{\binom{700}{x} \binom{300}{50-x}}{\binom{1,000}{50}}$$

Example 3-29: Customer Sample-2

Since n/N is small, the binomial will be used to approximate the hypergeometric. Let $p = K/N = 0.7$

$$P(X > 45) = \sum_{x=46}^{50} \binom{50}{x} 0.7^x (1-0.7)^{50-x} = 0.00017$$

In Excel

0.000172 = 1 - BINOMDIST(45, 50, 0.7, TRUE)

The hypergeometric value is 0.00013. The absolute error is 0.00004, but the percent error in using the approximation is $(17-13)/13 = 31\%$.

Poisson Distribution

As the number of trials (n) in a binomial experiment increases to infinity while the binomial mean (np) remains constant, the binomial distribution becomes the Poisson distribution.

Example 3-30:

$$\text{Let } \lambda = np = E(x), \text{ so } p = \lambda/n$$

$$\begin{aligned} P(X = x) &= \binom{n}{x} p^x (1-p)^{n-x} \\ &= \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \rightarrow \\ &= \frac{e^{-\lambda} \lambda^x}{x!} \end{aligned}$$

Example 3-31: Wire Flaws

Flaws occur at random along the length of a thin copper wire.

Let X denote the random variable that counts the number of flaws in a length of L mm of wire. Suppose the average number of flaws in L is λ .

Partition L into n subintervals ($1 \mu\text{m}$) each. If the subinterval is small enough, the probability that more than one flaw occurs is negligible.

Assume that the:

- Flaws occur at random, implying that each subinterval has the same probability of containing a flaw.
- Probability that a subinterval contains a flaw is independent of other subintervals.

X is now binomial. $E(X) = np = \lambda$ and $p = \lambda/n$

As n becomes large, p becomes small and a Poisson process is created.

Examples of Poisson Processes

In general, the Poisson random variable X is the number of events (counts) per interval.

1. Particles of contamination per wafer.
2. Flaws per roll of textile.
3. Calls at a telephone exchange per hour.
4. Power outages per year.
5. Atomic particles emitted from a specimen per second.
6. Flaws per unit length of copper wire.

Poisson Distribution Definition

- The random variable X that equals the number of events in a Poisson process is a Poisson random variable with parameter $\lambda > 0$, and the probability mass function is:

$$f(x) = \frac{e^{-\lambda} \lambda^x}{x!} \quad \text{for } x = 0, 1, 2, 3, \dots \infty \quad (3-16)$$

Poisson Graphs

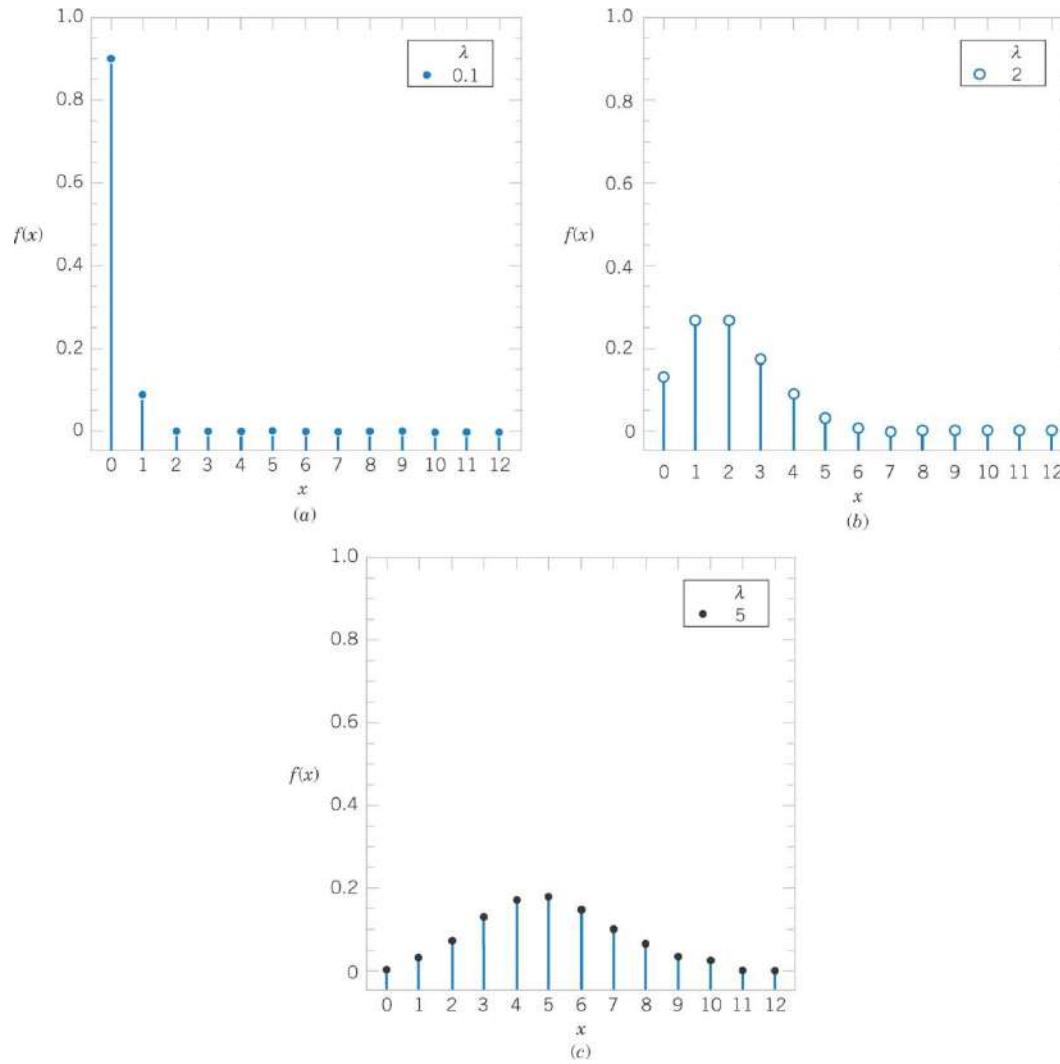


Figure 3-14 Poisson distributions for $\lambda = 0.1, 2, 5$.

Poisson Requires Consistent Units

It is important to use consistent units in the calculation of Poisson:

- Probabilities
- Means
- Variances
- Example of unit conversions:
 - Average # of flaws per mm of wire is 3.4.
 - Average # of flaws per 10 mm of wire is 34.
 - Average # of flaws per 20 mm of wire is 68.

Example 3-32: Calculations for Wire Flaws-1

For the case of the thin copper wire, suppose that the number of flaws follows a Poisson distribution of 2.3 flaws per mm. Let X denote the number of flaws in 1 mm of wire. Find the probability of exactly 2 flaws in 1 mm of wire.

Answer:

$$P(X = 2) = \frac{e^{-2.3} 2.3^2}{2!} = 0.265$$

In Excel

0.26518 = POISSON(2, 2.3, FALSE)

Example 3-32: Calculations for Wire Flaws-2

Determine the probability of 10 flaws in 5 mm of wire.

Now let X denote the number of flaws in 5 mm of wire.

Answer:

$$E(X) = \lambda = 5 \text{ mm} \cdot 2.3 \text{ flaws/mm} = 11.5 \text{ flaws}$$

$$P(X = 10) = e^{-11.5} \frac{11.5^{10}}{10!} = 0.113$$

In Excel

0.1129	=POISSON(10, 11.5, FALSE)
--------	---------------------------

Example 3-32: Calculations for Wire Flaws-3

Determine the probability of at least 1 flaw in 2 mm of wire. Now let X denote the number of flaws in 2 mm of wire. Note that $P(X \geq 1)$ requires ∞ terms. ☹

Answer:

$$E(X) = \lambda = 2 \text{ mm} \cdot 2.3 \text{ flaws/mm} = 4.6 \text{ flaws}$$

$$P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-4.6} \frac{4.6^0}{0!} = 0.9899$$

In Excel

0.989948 = 1 - POISSON(0, 4.6, FALSE)

Example 3-33: CDs-1

Contamination is a problem in the manufacture of optical storage disks (CDs). The number of particles of contamination that occur on a CD has a Poisson distribution. The average number of particles per square cm of media is 0.1. The area of a disk under study is 100 cm^2 . Let X denote the number of particles of a disk. Find $P(X = 12)$.

Answer:

$$E(X) = \lambda = 100 \text{ cm}^2 \cdot 0.1 \text{ particles/cm}^2 = 10 \text{ particles}$$

$$P(X = 12) = e^{-10} \frac{10^{12}}{12!} = 0.095$$

In Excel

0.0948 = POISSON(12, 10, FALSE)

Example 3-33: CDs-2

Find the probability that zero particles occur on the disk. Recall that $\lambda = 10$ particles.

Answer:

$$P(X = 0) = e^{-10} \frac{10^0}{0!} = 4.54 \cdot 10^{-5}$$

In Excel

4.540E-05 = POISSON(0, 10, FALSE)

Example 3-33: CDs-3

Determine the probability that 12 or fewer particles occur on the disk. That will require 13 terms in the sum of probabilities. ☹ Recall that $\lambda = 10$ particles.

Answer:

$$\begin{aligned} P(X \leq 12) &= P(X = 0) + P(X = 1) + \dots + P(X = 12) \\ &= \sum_{x=0}^{12} e^{-10} \frac{10^x}{x!} = 0.792 \end{aligned}$$

In Excel

0.7916 = POISSON(12, 10, TRUE)

Poisson Mean & Variance

If X is a Poisson random variable with parameter λ , then:

$$\mu = E(X) = \lambda \quad \text{and} \quad \sigma^2 = V(X) = \lambda \quad (3-17)$$

The mean and variance of the Poisson model are the same. If the mean and variance of a data set are not about the same, then the Poisson model would not be a good representation of that set.

The derivation of the mean and variance is shown in the text.

Discrete Random Variables

Many physical systems can be modeled by the same or similar random experiments and random variables. The distribution of the random variable involved in each of these common systems can be analyzed, and the results can be used in different applications and examples.

In this chapter, we present the analysis of several random experiments and **discrete random variables** that frequently arise in applications.

We often omit a discussion of the underlying sample space of the random experiment and directly describe the distribution of a particular random variable.

Important Terms & Concepts of Chapter 3

Bernoulli trial	Mean – discrete random variable
Binomial distribution	Mean – function of a discrete random variable
Cumulative probability distribution – discrete random variable	Negative binomial distribution
Discrete uniform distribution	Poisson distribution
Expected value of a function of a random variable	Poisson process
Finite population correction factor	Probability distribution – discrete random variable
Geometric distribution	Probability mass function
Hypergeometric distribution	Standard deviation – discrete random variable
Lack of memory property – discrete random variable	Variance – discrete random variable

3

Discrete Random Variables and Probability Distributions

CHAPTER OUTLINE

- 3-1 Discrete Random Variables
- 3-2 Probability Distributions and Probability Mass Functions
- 3-3 Cumulative Distribution Functions
- 3-4 Mean and Variance of a Discrete Random Variable
- 3-5 Discrete Uniform Distribution
- 3-6 Binomial Distribution
- 3-7 Geometric and Negative Binomial Distributions
 - 3-7.1 Geometric Distribution
 - 3-7.2 Negative Binomial Distribution
- 3-8 Hypergeometric Distribution
- 3-9 Poisson Distribution

Foundation of Data Science and Analytics

Continuous RV and Probability Distributions

Arun K. Timalsina

Continuous Random Variables

The dimensional length of a manufactured part is subject to small variations in measurement due to vibrations, temperature fluctuations, operator differences, calibration, cutting tool wear, bearing wear, and raw material changes.

This length X would be a **continuous random variable** that would occur in an interval (finite or infinite) of real numbers.

The number of possible values of X , in that interval, is uncountably infinite and limited only by the precision of the measurement instrument.

Continuous Density Functions

Density functions, in contrast to mass functions, distribute probability continuously along an interval.

The loading on the beam between points a & b is the integral of the function between points a & b.

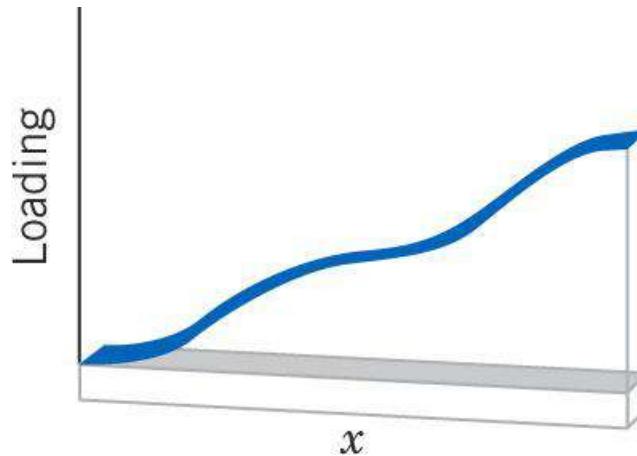


Figure 4-1 Density function as a loading on a long, thin beam. Most of the load occurs at the larger values of x .

A probability density function $f(x)$ describes the probability distribution of a continuous random variable. It is analogous to the beam loading.

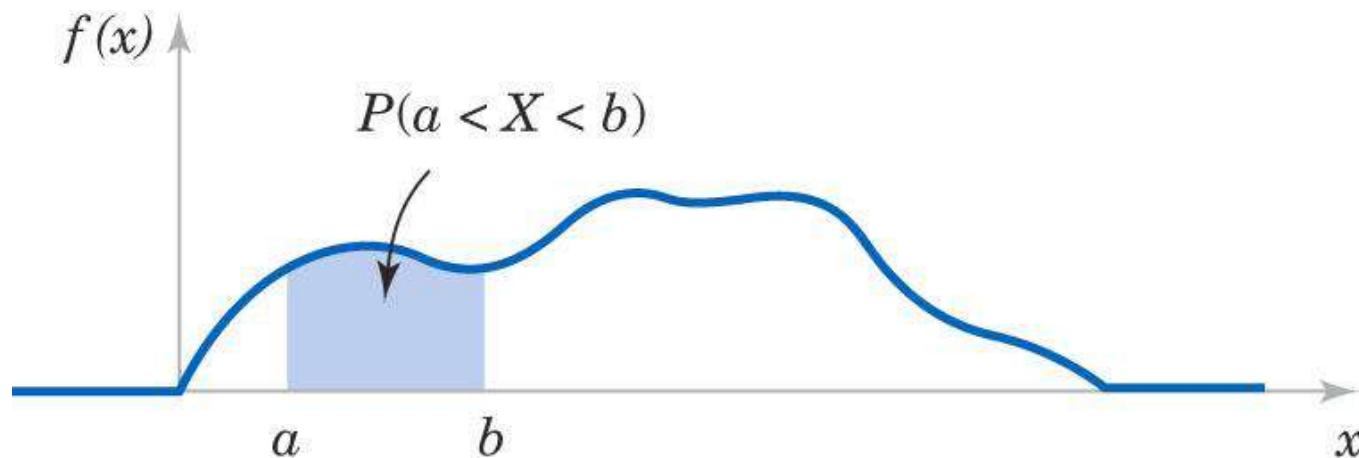


Figure 4-2 Probability is determined from the area under $f(x)$ from a to b .

Probability Density Function

For a continuous random variable X ,

a **probability density function** is a function such that

(1) $f(x) \geq 0$ means that the function is always non-negative.

$$(2) \int_{-\infty}^{\infty} f(x)dx = 1$$

(3) $P(a \leq X \leq b) = \int_a^b f(x)dx =$ area under $f(x)dx$ from a to b

(4) $f(x) = 0$ means there is no area exactly at x .

Histograms

A **histogram** is graphical display of data showing a series of adjacent rectangles. Each rectangle has a base which represents an interval of data values. The height of the rectangle creates an **area** which represents the relative frequency associated with the values included in the base.

A continuous probability distribution $f(x)$ is a model approximating a histogram. A bar has the same area of the integral of those limits.



Figure 4-3 Histogram approximates a probability density function.

Area of a Point

If X is a continuous random variable, for any x_1 and x_2 ,

$$P(x_1 \leq X \leq x_2) = P(x_1 < X \leq x_2) = P(x_1 \leq X < x_2) = P(x_1 < X < x_2) \quad (4-2)$$

which implies that $P(X = x) = 0$.

From another perspective:

As x_1 approaches x_2 , the area or probability becomes smaller and smaller.

As x_1 becomes x_2 , the area or probability becomes zero.

Example 4-1: Electric Current

Let the continuous random variable X denote the current measured in a thin copper wire in milliamperes (mA). Assume that the range of X is $0 \leq x \leq 20$ and $f(x) = 0.05$. What is the probability that a current is less than 10mA?

Answer:

$$P(X < 10) = \int_0^{10} 0.5 dx = 0.5$$

Another example,

$$P(5 < X < 20) = \int_5^{20} 0.5 dx = 0.75$$

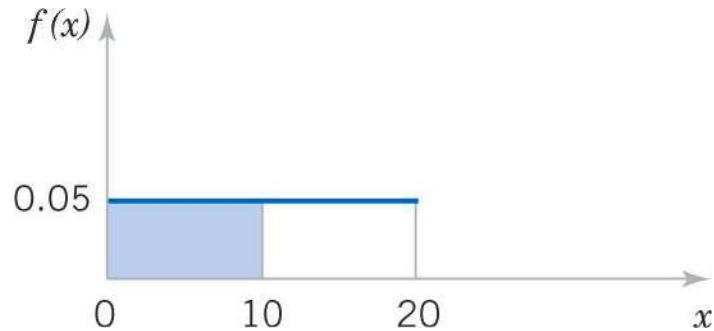
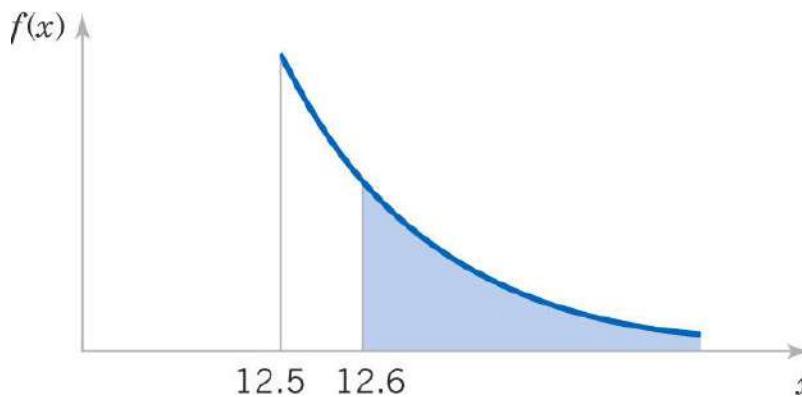


Figure 4-4 $P(X < 10)$ illustrated.

Example 4-2: Hole Diameter

Let the continuous random variable X denote the diameter of a hole drilled in a sheet metal component. The target diameter is 12.5 mm. Random disturbances to the process result in larger diameters. Historical data shows that the distribution of X can be modeled by $f(x) = 20e^{-20(x-12.5)}$, $x \geq 12.5$ mm. If a part with a diameter larger than 12.60 mm is scrapped, what proportion of parts is scrapped?

Answer:



$$\text{Figure 4-5 } P(X > 12.60) = \int_{12.6}^{\infty} 20e^{-20(x-12.5)} dx = 0.135$$

Cumulative Distribution Functions

The **cumulative distribution function** of a continuous random variable X is,

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(u) du \quad \text{for } -\infty < x < \infty \quad (4-3)$$

Example 4-3: Electric Current

For the copper wire current measurement in Exercise 4-1, the cumulative distribution function (CDF) consists of three expressions to cover the entire real number line.

	0	$x < 0$
$F(x) =$	$0.05x$	$0 \leq x \leq 20$
	1	$20 < x$

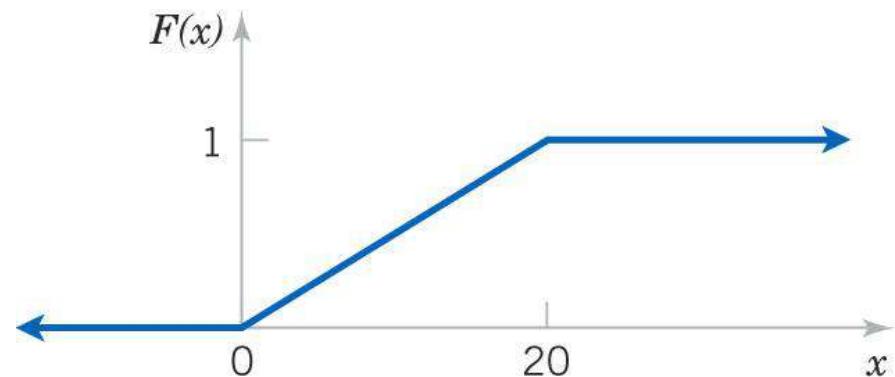


Figure 4-6 This graph shows the CDF as a continuous function.

Example 4-4: Hole Diameter

For the drilling operation in Example 4-2, $F(x)$ consists of two expressions. This shows the proper notation.

$$F(x) = 0 \quad \text{for } x < 12.5$$

$$\begin{aligned} F(x) &= \int_{12.5}^x 20e^{-20(u-12.5)} du \\ &= 1 - e^{-20(x-12.5)} \quad \text{for } x \geq 12.5 \end{aligned}$$

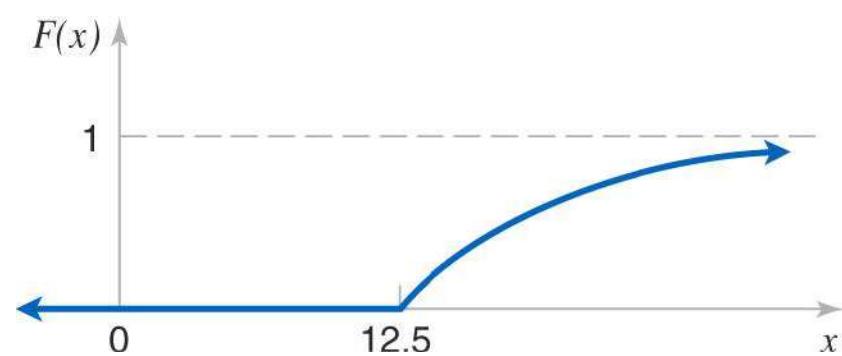


Figure 4-7 This graph shows $F(x)$ as a continuous function.

Density vs. Cumulative Functions

- The probability density function (PDF) is the derivative of the cumulative distribution function (CDF).
- The cumulative distribution function (CDF) is the integral of the probability density function (PDF).

Given $F(x)$, $f(x) = \frac{dF(x)}{dx}$ as long as the derivative exists.

Exercise 4-5: Reaction Time

- The time until a chemical reaction is complete (in milliseconds, ms) is approximated by this CDF:

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 - e^{-0.01x} & \text{for } 0 \leq x \end{cases}$$

- What is the PDF?

$$f(x) = \frac{dF(x)}{dx} = \frac{d}{dx} \left\{ \begin{cases} 0 & \text{for } x < 0 \\ 1 - e^{-0.01x} & \text{for } 0 \leq x \end{cases} \right\} = \begin{cases} 0 & \text{for } x < 0 \\ 0.01e^{-0.01x} & \text{for } 0 \leq x \end{cases}$$

- What proportion of reactions is complete within 200 ms?

$$P(X < 200) = F(200) = 1 - e^{-2} = 0.8647$$

Mean & Variance

Suppose X is a continuous random variable with probability density function $f(x)$. The **mean** or **expected value** of X , denoted as μ or $E(X)$, is

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx \quad (4-4)$$

The **variance** of X , denoted as $V(X)$ or σ^2 , is

$$\sigma^2 = V(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx = \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2$$

The **standard deviation** of X is $\sigma = \sqrt{\sigma^2}$.

Example 4-6: Electric Current

For the copper wire current measurement in Exercise 4-1, the PDF is $f(x) = 0.05$ for $0 \leq x \leq 20$. Find the mean and variance.

$$E(X) = \int_0^{20} x \cdot f(x) dx = \frac{0.05x^2}{2} \Big|_0^{20} = 10$$

$$V(X) = \int_0^{20} (x - 10)^2 f(x) dx = \frac{0.05(x-10)^3}{3} \Big|_0^{20} = 33.33$$

Mean of a Function of a Random Variable

If X is a continuous random variable
with a probability density function $f(x)$,

$$E[h(x)] = \int_{-\infty}^{\infty} h(x)f(x)dx \quad (4-5)$$

Example 4-7: In Example 4-1, X is the current measured in mA. What is the expected value of the squared current?

$$\begin{aligned} E[h(x)] &= E[X^2] = \int_0^{20} x^2 f(x)dx \\ &= \int_0^{20} 0.05x^2 dx = \frac{0.05x^3}{3} \Big|_0^{20} = 133.33 \text{ mA}^2 \end{aligned}$$

Example 4-8: Hole Diameter

For the drilling operation in Example 4-2, find the mean and variance of X using integration by parts. Recall that $f(x) = 20e^{-20(x-12.5)}dx$ for $x \geq 12.5$.

$$\begin{aligned} E(X) &= \int_{12.5}^{\infty} xf(x)dx = \int_{12.5}^{\infty} x20e^{-20(x-12.5)}dx \\ &= -xe^{-20(x-12.5)} - \frac{e^{-20(x-12.5)}}{20} \Big|_{12.5}^{\infty} = 12.5 + 0.05 = 12.55 \text{ mm} \end{aligned}$$

$$V(X) = \int_{12.5}^{\infty} (x-12.55)^2 f(x)dx = 0.0025 \text{ mm}^2 \text{ and } \sigma = 0.05 \text{ mm}$$

Continuous Uniform Distribution

- This is the simplest continuous distribution and analogous to its discrete counterpart.
- A continuous random variable X with probability density function

$$f(x) = 1 / (b-a) \text{ for } a \leq x \leq b \quad (4-6)$$

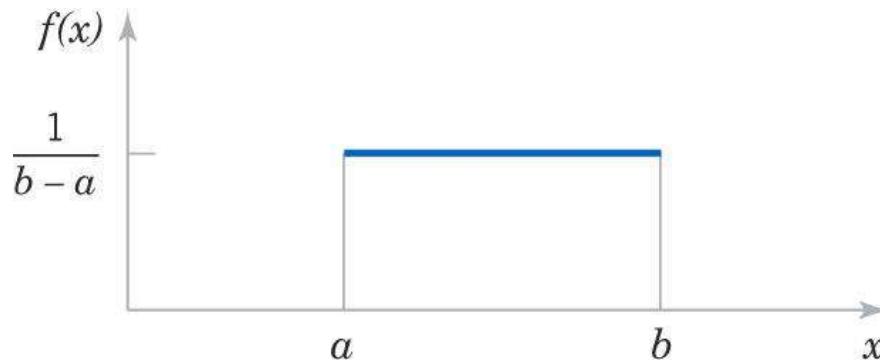


Figure 4-8 Continuous uniform PDF

Mean & Variance

- Mean & variance are:

$$\mu = E(X) = \frac{(a+b)}{2} \quad \text{and} \quad \sigma^2 = V(X) = \frac{(b-a)^2}{12} \quad (4-7)$$

- Derivations are shown in the text. Be reminded that $b^2 - a^2 = (b + a)(b - a)$

Example 4-9: Uniform Current

Let the continuous random variable X denote the current measured in a thin copper wire in mA. Recall that the PDF is $F(x) = 0.05$ for $0 \leq x \leq 20$.

What is the probability that the current measurement is between 5 & 10 mA?

$$P(5 < x < 10) = \int_{5}^{10} 0.05 dx = 5(0.05) = 0.25$$

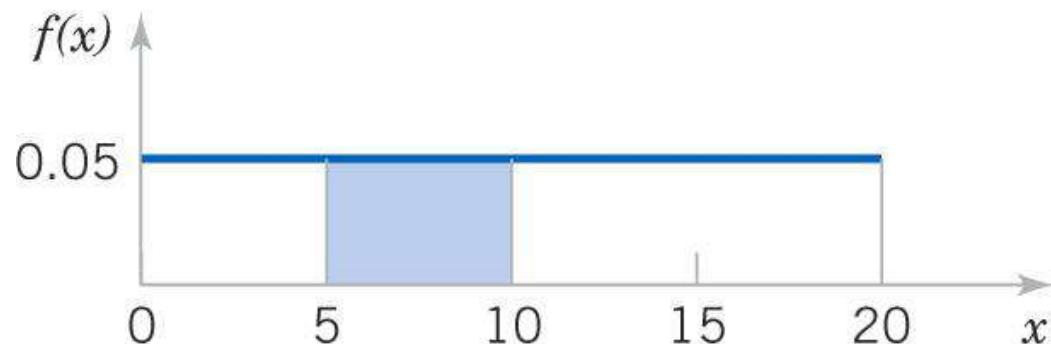


Figure 4-9

Continuous Uniform CDF

$$F(x) = \int_a^x \frac{1}{(b-a)} du = \frac{x-a}{b-a}$$

The CDF is completely described as

$$F(x) = \begin{cases} 0 & x < a \\ (x-a)/(b-a) & a \leq x < b \\ 1 & b \leq x \end{cases}$$

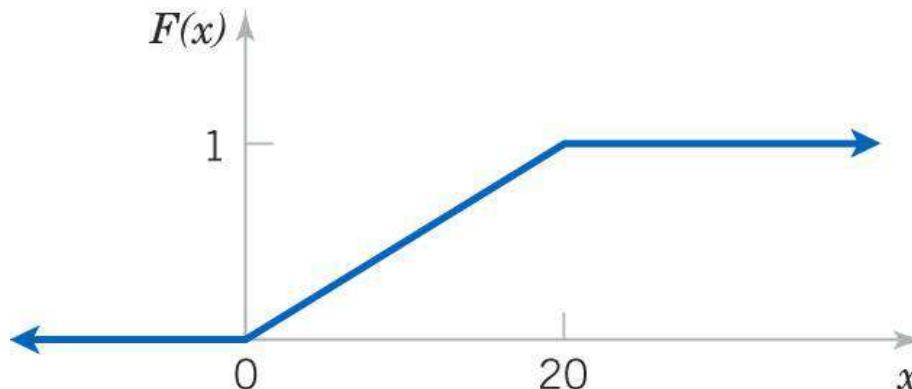


Figure 4-6 (again) Graph of the Cumulative Uniform CDF

Normal Distribution

- The most widely used distribution is the **normal distribution**, also known as the Gaussian distribution.
- Random variation of many physical measurements are normally distributed.
- The location and spread of the normal are independently determined by mean (μ) and standard deviation (σ).

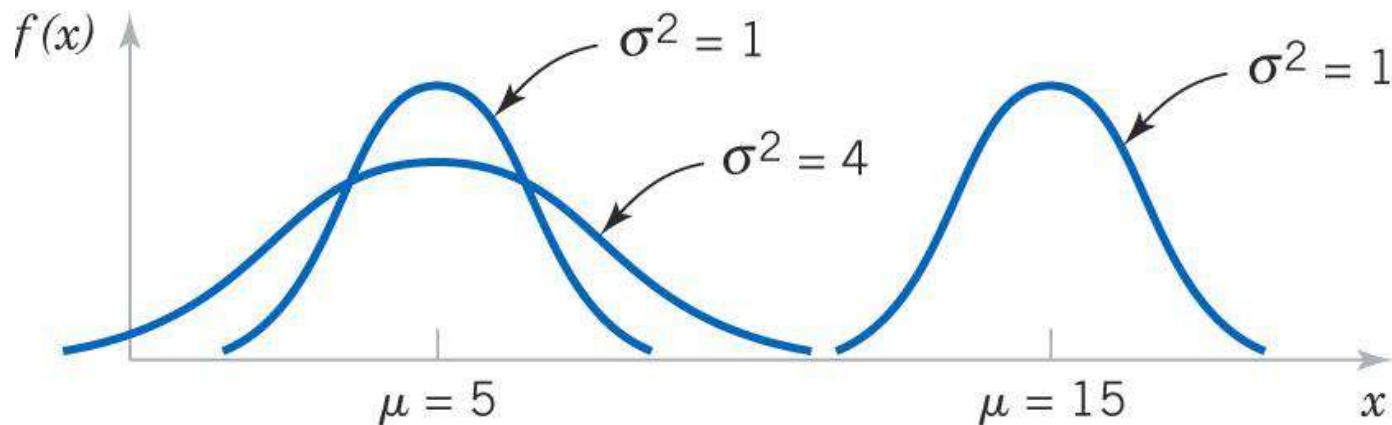


Figure 4-10 Normal probability density functions

Normal Probability Density Function

A random variable X with probability density function

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad -\infty < x < \infty \quad (4-8)$$

is a **normal random variable** with parameters μ ,

where $-\infty < \mu < \infty$, and $\sigma > 0$. Also,

$$E(X) = \mu \quad \text{and} \quad V(X) = \sigma^2 \quad (4-9)$$

and the notation $N(\mu, \sigma^2)$ is used to denote the distribution.

Note that $f(X)$ cannot be intergrated analytically, so $F(X)$ is expressed through numerical integration with Excel or Minitab, and written as Appendix A, Table III.

Example 4-10: Normal Application

Assume that the current measurements in a strip of wire follows a normal distribution with a mean of 10 mA & a variance of 4 mA². Let X denote the current in mA.

What is the probability that a measurement exceeds 13 mA?

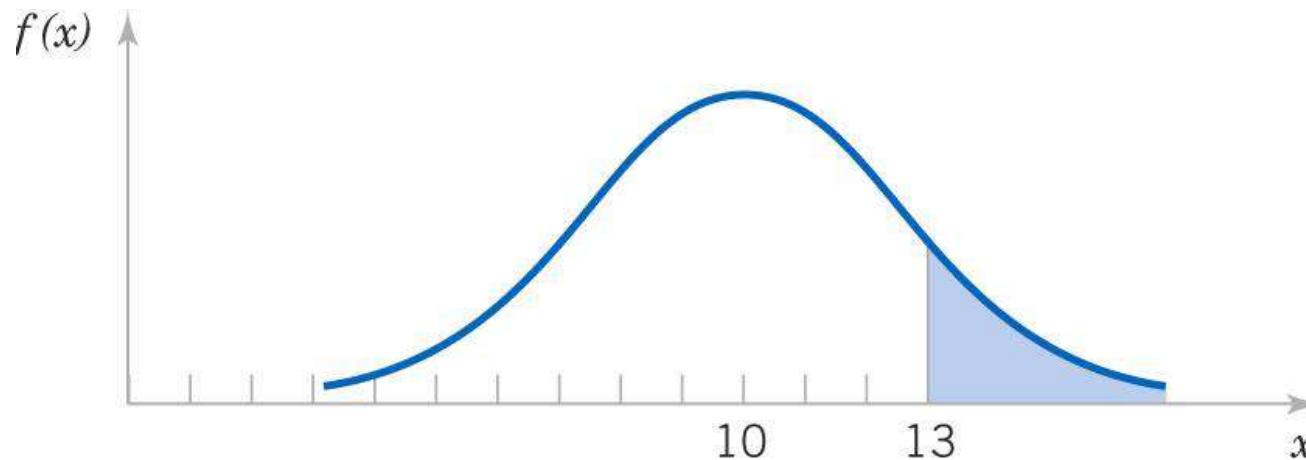


Figure 4-11 Graphical probability that $X > 13$ for a normal random variable with $\mu = 10$ and $\sigma^2 = 4$.

Empirical Rule

$$P(\mu - \sigma < X < \mu + \sigma) = 0.6827$$

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.9545$$

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.9973$$

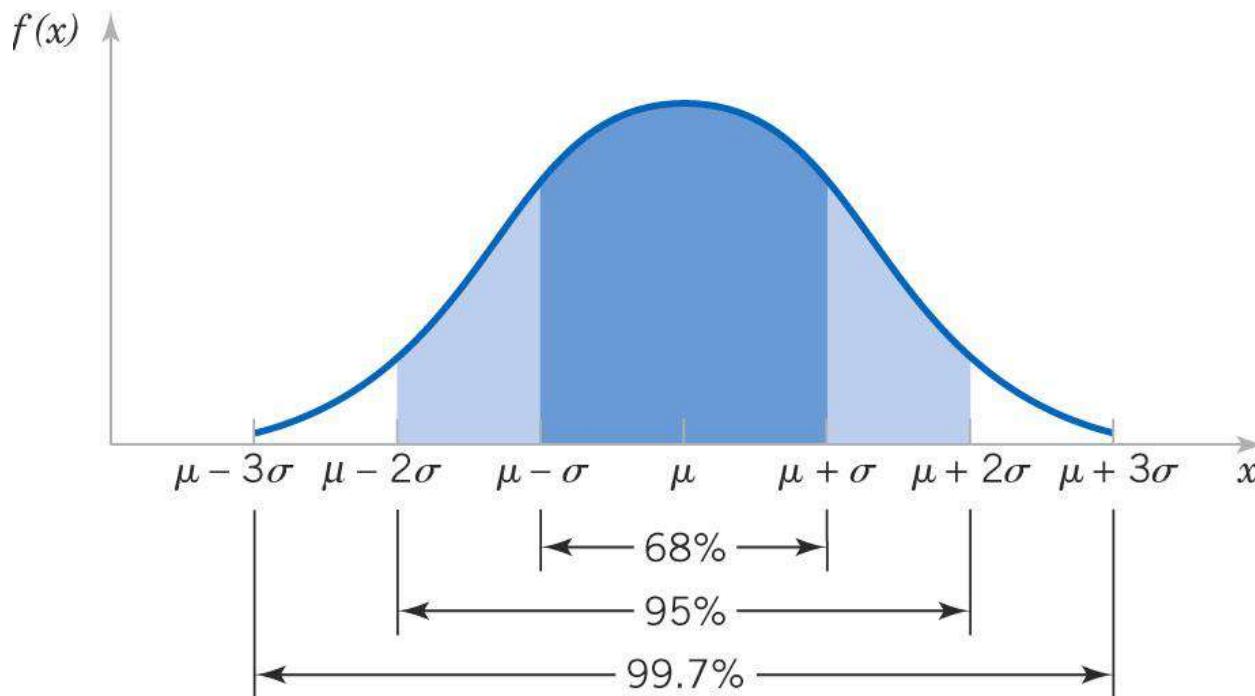


Figure 4-12 Probabilities associated with a normal distribution – well worth remembering to quickly estimate probabilities.

Standard Normal Distribution

A normal random variable with

$$\mu = 0 \text{ and } \sigma^2 = 1$$

Is called a **standard normal random variable** and is denoted as Z . The cumulative distribution function of a standard normal random variable is denoted as:

$$\Phi(z) = P(Z \leq z) = F(z)$$

Values are found in Appendix Table III and by using Excel and Minitab.

Example 4-11: Standard Normal Distribution

Assume Z is a standard normal random variable.

Find $P(Z \leq 1.50)$. Answer: 0.93319

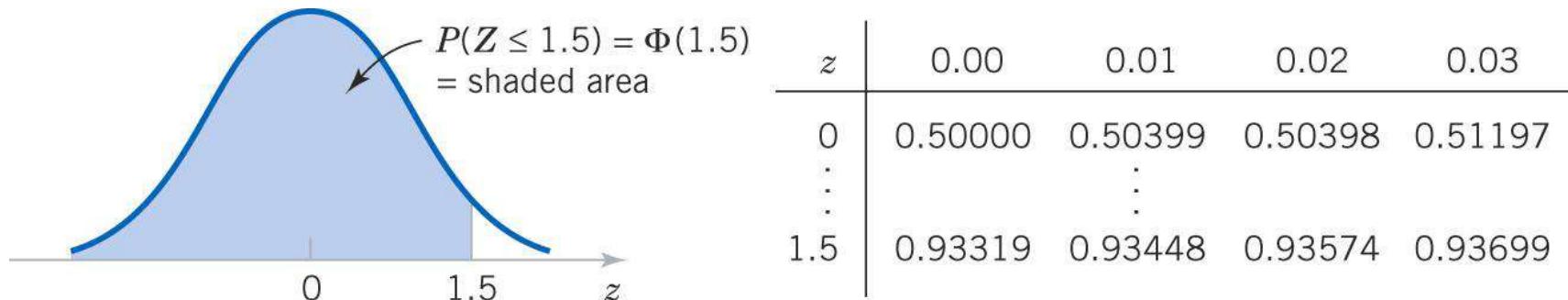


Figure 4-13 Standard normal PDF

Find $P(Z \leq 1.53)$. Answer: 0.93699

Find $P(Z \leq 0.02)$. Answer: 0.50398

Example 4-12: Standard Normal Exercises

1. $P(Z > 1.26) = 0.1038$

2. $P(Z < -0.86) = 0.195$

3. $P(Z > -1.37) = 0.915$

4. $P(-1.25 < Z < 0.37) = 0.5387$

5. $P(Z \leq -4.6) \approx 0$

6. Find z for $P(Z \leq z) = 0.05$, $z = -1.65$

7. Find z for $(-z < Z < z) = 0.99$, $z = 2.58$

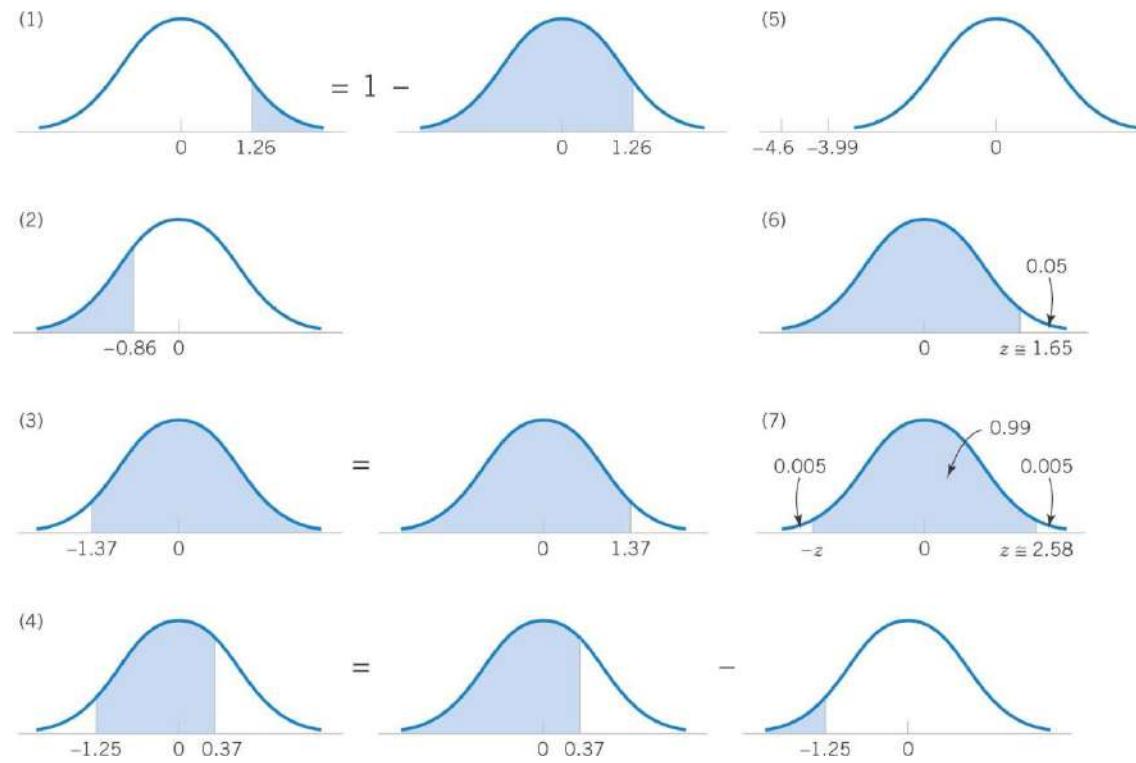


Figure 4-14 Graphical displays for standard normal distributions.

Standardizing

Suppose X is a normal random variable with mean μ and variance σ^2 .

$$\text{Then, } P(X \leq x) = P\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) = P(Z \leq z) \quad (4-11)$$

where Z is a **standard normal random variable**, and

$z = \frac{(x - \mu)}{\sigma}$ is the z-value obtained by **standardizing** X .

The probability is obtained by using Appendix Table III

with $z = \frac{(x - \mu)}{\sigma}$.

Example 4-14: Normally Distributed Current-1

From a previous example
with $\mu = 10$ and $\sigma = 2$ mA,
what is the probability
that the current
measurement is between
9 and 11 mA?

$$\begin{aligned} P(9 < X < 11) &= P\left(\frac{9-10}{2} < \frac{x-10}{2} < \frac{11-10}{2}\right) \\ &= P(-0.5 < z < 0.5) \\ &= P(z < 0.5) - P(z < -0.5) \\ &= 0.69146 - 0.30854 = 0.38292 \end{aligned}$$

Answer:

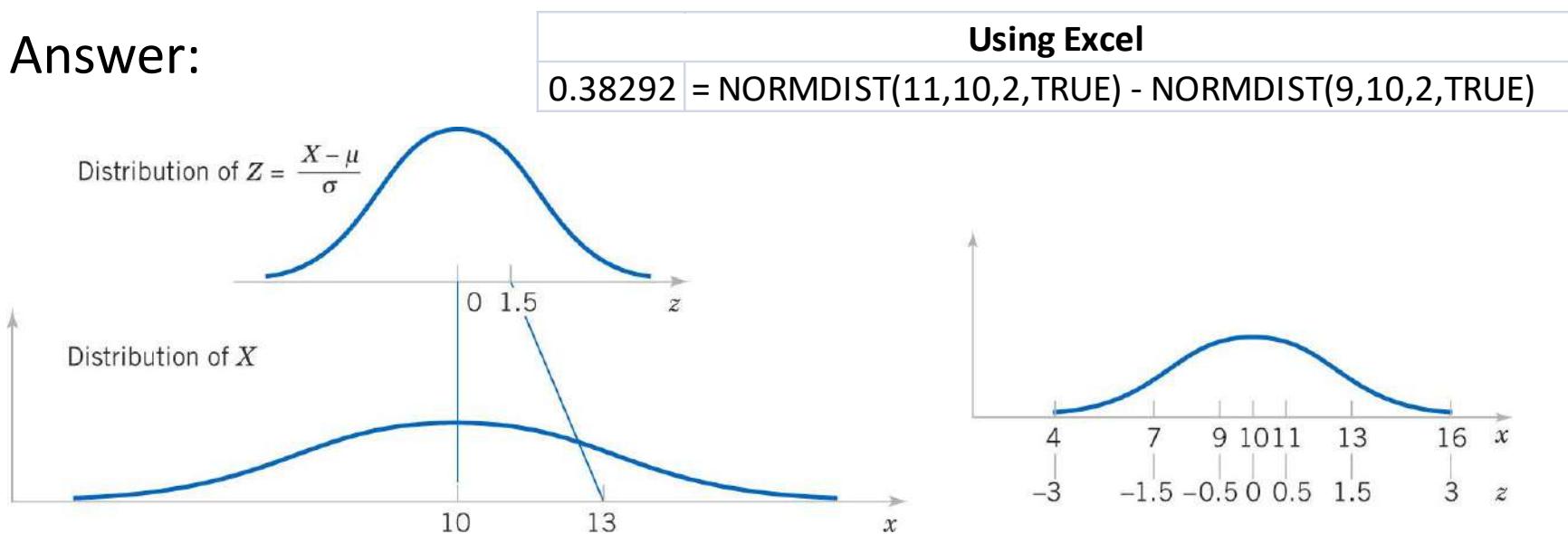


Figure 4-15 Standardizing a normal random variable.

Example 4-14: Normally Distributed Current-2

Determine the value for which the probability that a current measurement is below this value is 0.98.

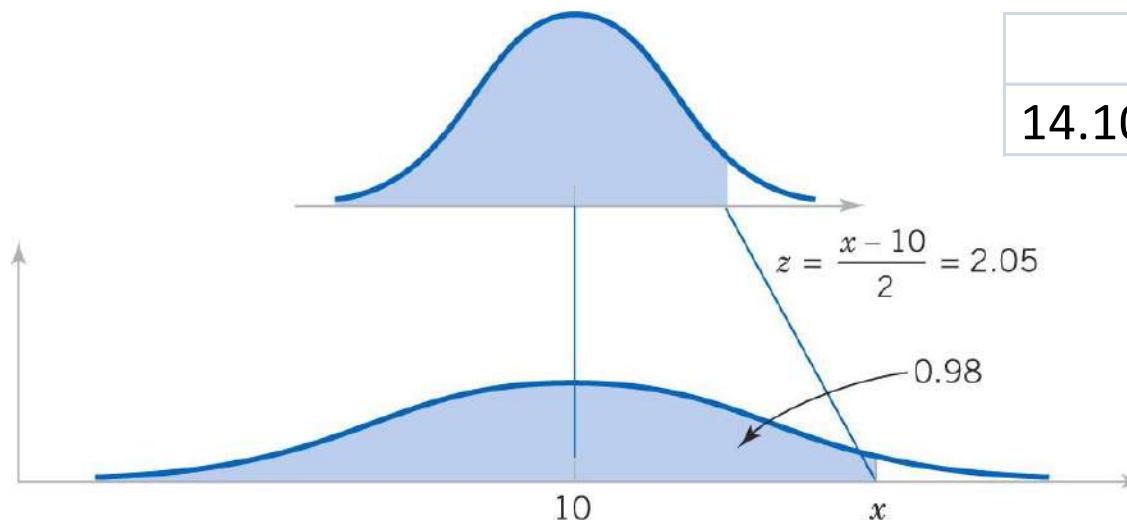
Answer:

$$P(X < x) = P\left(\frac{X - 10}{2} < \frac{x - 10}{2}\right)$$

$$= P\left(Z < \frac{x - 10}{2}\right) = 0.98$$

$z = 2.05$ is the closest value.

$$z = 2(2.05) + 10 = 14.1 \text{ mA.}$$



Using Excel

14.107 = NORMINV(0.98,10,2)

Figure 4-16 Determining the value of x to meet a specified probability.

Example 4-15: Signal Detection-1

Assume that in the detection of a digital signal, the background noise follows a normal distribution with $\mu = 0$ volt and $\sigma = 0.45$ volt. The system assumes a signal 1 has been transmitted when the voltage exceeds 0.9. What is the probability of detecting a digital 1 when none was sent? Let the random variable N denote the voltage of noise.

$$\begin{aligned} P(N > 0.9) &= P\left(\frac{N - 0}{0.45} > \frac{0.9}{0.45}\right) = P(Z > 2) \\ &= 1 - 0.97725 = 0.02275 \end{aligned}$$

Using Excel

`0.02275 = 1 - NORMDIST(0.9,0,0.45,TRUE)`

This probability can be described as the probability of a false detection.

Example 4-15: Signal Detection-2

Determine the symmetric bounds about 0 that include 99% of all noise readings. We need to find x such that $P(-x < N < x) = 0.99$.

$$\begin{aligned} P(-x < N < x) &= P\left(\frac{-x}{0.45} < \frac{N}{0.45} < \frac{x}{0.45}\right) \\ &= P\left(\frac{-x}{0.45} < Z < \frac{x}{0.45}\right) = P(-2.58 < Z < 2.58) \\ x &= 2.58(0.45) + 0 = 1.16 \end{aligned}$$

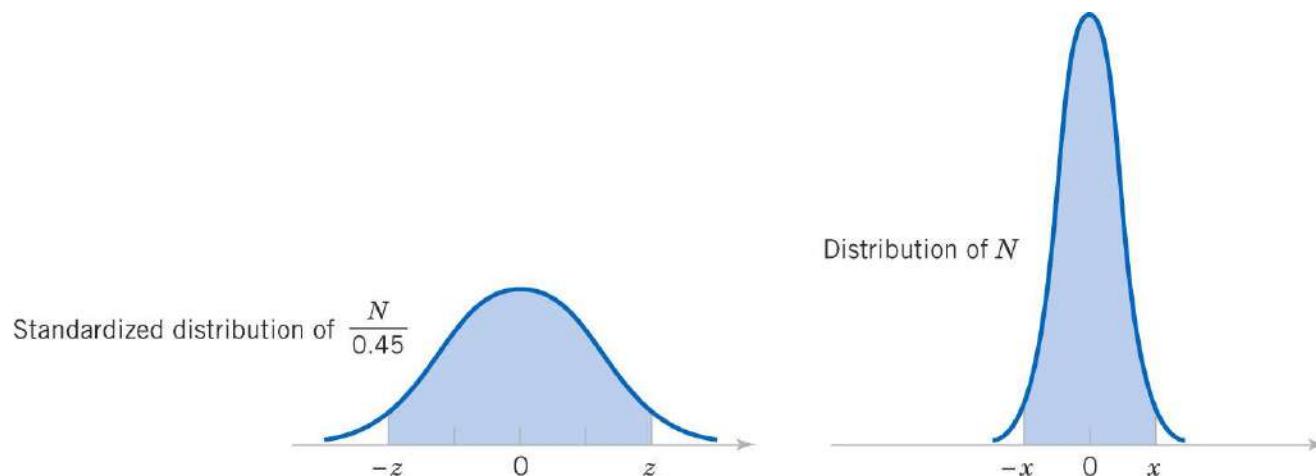


Figure 4-17 Determining the value of x to meet a specified probability.

Example 4-15: Signal Detection-3

Suppose that when a digital 1 signal is transmitted, the mean of the noise distribution shifts to 1.8 volts. What is the probability that a digital 1 is not detected? Let S denote the voltage when a digital 1 is transmitted.

$$\begin{aligned} P(S < 0.9) &= P\left(\frac{S - 1.8}{0.45} < \frac{0.9 - 1.8}{0.45}\right) \\ &= P(Z < -2) = 0.02275 \end{aligned}$$

Using Excel

0.02275 = NORMDIST(0.9, 1.8, 0.45, TRUE)

This probability can be interpreted as the probability of a missed signal.

Example 4-16: Shaft Diameter-1

The diameter of the shaft is normally distributed with $\mu = 0.2508$ inch and $\sigma = 0.0005$ inch. The specifications on the shaft are 0.2500 ± 0.0015 inch. What proportion of shafts conform to the specifications? Let X denote the shaft diameter in inches.

Answer: $P(0.2485 < X < 0.2515)$

$$= P\left(\frac{0.2485 - 0.2508}{0.0005} < Z < \frac{0.2515 - 0.2508}{0.0005}\right)$$

$$= P(-4.6 < Z < 1.4)$$

$$= P(Z < 1.4) - P(Z < -4.6)$$

$$= 0.91924 - 0.0000 = 0.91924$$

Using Excel

`0.91924 = NORMDIST(0.2515, 0.2508, 0.0005, TRUE) - NORMDIST(0.2485, 0.2508, 0.0005, TRUE)`

Example 4-16: Shaft Diameter-2

Most of the nonconforming shafts are too large, because the process mean is near the upper specification limit. If the process is centered so that the process mean is equal to the target value, what proportion of the shafts will now conform?

Answer:

$$\begin{aligned} & P(0.2485 < X < 0.2515) \\ &= P\left(\frac{0.2485 - 0.2500}{0.0005} < Z < \frac{0.2515 - 0.2500}{0.0005}\right) \\ &= P(-3 < Z < 3) \\ &= P(Z < 3) - P(Z < -3) \\ &= 0.99865 - 0.00135 = 0.99730 \end{aligned}$$

Using Excel

`0.99730 = NORMDIST(0.2515, 0.25, 0.0005, TRUE) - NORMDIST(0.2485, 0.25, 0.0005, TRUE)`

By centering the process, the yield increased from 91.924% to 99.730%, an increase of 7.806%

Normal Approximations

- The binomial and Poisson distributions become more bell-shaped and symmetric as their means increase.
- For manual calculations, the normal approximation is practical – exact probabilities of the binomial and Poisson, with large means, require technology (Minitab, Excel).
- The normal is a good approximation for the:
 - Binomial if $np > 5$ and $n(1-p) > 5$.
 - Poisson if $\lambda > 5$.

Normal Approximation to the Binomial

Suppose we have a binomial distribution with $n = 10$ and $p = 0.5$. Its mean and standard deviation are 5.0 and 1.58 respectively.

Draw the normal distribution over the binomial distribution.

The areas of the normal approximate the areas of the bars of the binomial with a **continuity correction**.

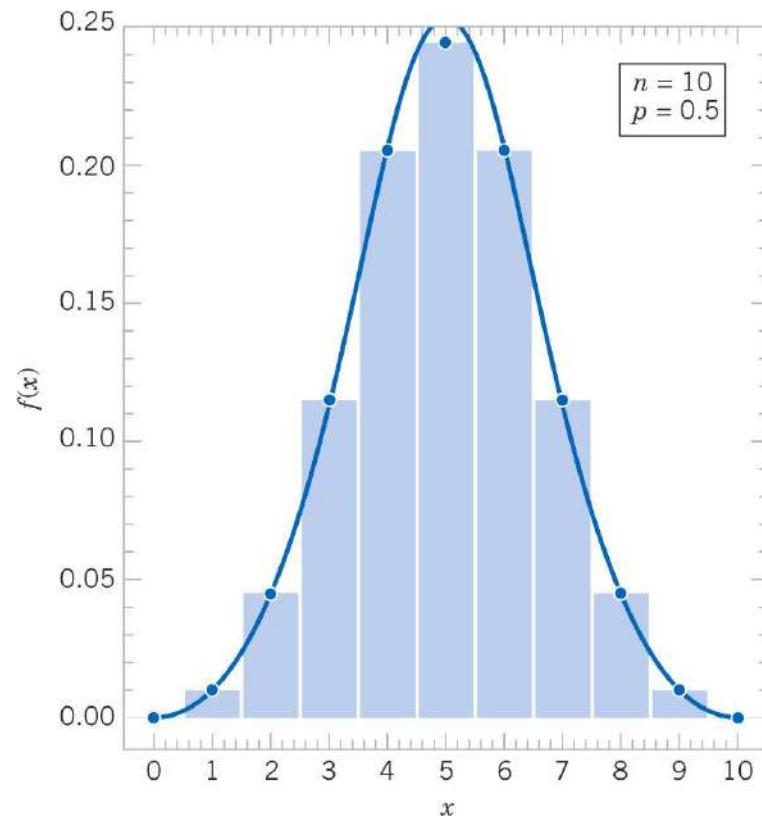


Figure 4-19 Overlaying the normal distribution upon a binomial with matched parameters.

Example 4-17:

In a digital comm channel, assume that the number of bits received in error can be modeled by a binomial random variable. The probability that a bit is received in error is 10^{-5} . If 16 million bits are transmitted, what is the probability that 150 or fewer errors occur? Let X denote the number of errors.

Answer:

$$P(X \leq 150) = \sum_{x=0}^{150} C_x^{16000000} (10^{-5})^x (1-10^{-5})^{16000000-x}$$

Using Excel

0.2280 = BINOMDIST(150,16000000,0.00001,TRUE)

Can only be evaluated with technology. Manually, we must use the normal approximation to the binomial.

Normal Approximation Method

If X is a binomial random variable with parameters n and p ,

$$Z = \frac{X - np}{\sqrt{np(1-p)}} \quad (4-12)$$

is approximately a standard normal random variable. To approximate a binomial probability with a normal distribution, a **continuity correction** is applied as follows:

$$P(X \leq x) = P(X \leq x + 0.5) = P\left(Z \leq \frac{x + 0.5 - np}{\sqrt{np(1-p)}}\right)$$

and

$$P(X \geq x) = P(X \leq x - 0.5) = P\left(Z \leq \frac{x - 0.5 - np}{\sqrt{np(1-p)}}\right)$$

The approximation is good for $np > 5$ and $n(1-p) > 5$. Refer to Figure 4-19 to see the rationale for adding and subtracting the 0.5 continuity correction.

Example 4-18: Applying the Approximation

The digital comm problem in the previous example is solved using the normal approximation to the binomial as follows:

$$\begin{aligned} P(X \leq 150) &= P(X \leq 150.5) \\ &= P\left(\frac{X - 160}{\sqrt{160(1-10^{-5})}} \leq \frac{150.5 - 160}{\sqrt{160(1-10^{-5})}}\right) \\ &= P\left(Z \leq \frac{-9.5}{12.6491}\right) = P(-0.75104) = 0.2263 \end{aligned}$$

Using Excel

0.2263 = NORMDIST(150.5, 160, SQRT(160*(1-0.00001)), TRUE)

-0.7% = (0.2263-0.228)/0.228 = percent error in the approximation

Example 4-19: Normal Approximation-1

Again consider the transmission of bits. To judge how well the normal approximation works, assume $n = 50$ bits are transmitted and the probability of an error is $p = 0.1$. The exact and approximated probabilities are:

$$P(X \leq 2) = C_0^{50} 0.9^{50} + C_1^{50} 0.1(0.9^{49}) + C_2^{50} 0.1^2 (0.9^{48}) = 0.112$$

$$\begin{aligned} P(X \leq 2) &= P\left(\frac{X - 5}{\sqrt{50(0.1)(0.9)}} < \frac{2.5 - 5}{\sqrt{50(0.1)(0.9)}}\right) \\ &= P(Z < -1.18) = 0.119 \end{aligned}$$

Using Excel	
0.1117	= BINOMDIST(2,50,0.1,TRUE)
0.1193	= NORMDIST(2.5, 5, SQRT(5*0.9), TRUE)
6.8%	= (0.1193 - 0.1117) / 0.1117 = percent error

Example 4-19: Normal Approximation-2

$$\begin{aligned} P(X > 8) &= P(X \geq 9) \approx P(X > 8.5) \\ &= P\left(Z > \frac{8.5 - 5}{2.12}\right) = P(Z > 1.65) = 0.05 \end{aligned}$$

$$\begin{aligned} P(X = 5) &\approx P(4.5 < X < 5.5) \\ &= P\left(\frac{4.5 - 5}{2.12} < Z < \frac{5.5 - 5}{2.12}\right) \\ &= P(-0.24 < Z < 0.24) \\ &= P(Z < 0.24) - P(Z < -0.24) = 0.19 \end{aligned}$$

Using Excel

0.1849 = BINOMDIST(5,50,0.1,FALSE)

0.1863 = NORMDIST(5.5, 5, SQRT(5*0.9), TRUE) - NORMDIST(4.5, 5, SQRT(5*0.9), TRUE)

0.8% = (0.1863 - 0.1849) / 0.1849 = percent error

Reason for the Approximation Limits

The $np > 5$ and $n(1-p) > 5$ approximation rule is needed to keep the tails of the normal distribution from getting out-of-bounds.

As the binomial mean approaches the endpoints of the range of x , the standard deviation must be small enough to prevent overrun.

Figure 4-20 shows the asymmetric shape of the binomial when the approximation rule is not met.

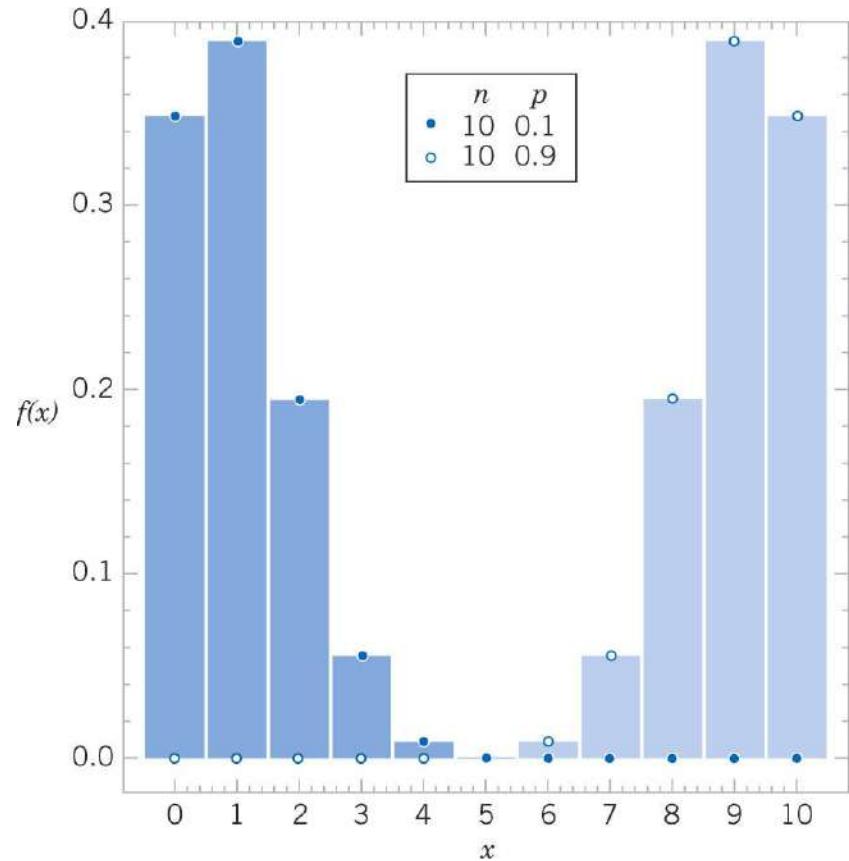


Figure 4-20 Binomial distribution is not symmetric as p gets near 0 or 1.

Normal Approximation to Hypergeometric

Recall that the hypergeometric distribution is similar to the binomial such that $p = K / N$ and when sample sizes are small relative to population size.

Thus the normal can be used to approximate the hypergeometric distribution also.

hypergeometric distribution	\approx	binomial distribution	\approx	normal distribution
	$n / N < 0.1$		$np < 5$	
			$n(1-p) < 5$	

Figure 4-21 Conditions for approximating hypergeometric and binomial with normal probabilities

Normal Approximation to the Poisson

If X is a Poisson random variable with $E(X) = \lambda$ and $V(X) = \lambda$,

$$Z = \frac{X - \lambda}{\sqrt{\lambda}} \quad (4-13)$$

is approximately a standard normal random variable. The same continuity correction used for the binomial distribution can also be applied. The approximation is good for

$$\lambda \geq 5$$

Example 4-20: Normal Approximation to Poisson

Assume that the number of asbestos particles in a square meter of dust on a surface follows a Poisson distribution with a mean of 100. If a square meter of dust is analyzed, what is the probability that 950 or fewer particles are found?

$$\begin{aligned} P(X \leq 950) &= \sum_{x=0}^{950} \frac{e^{-1000} 1000^x}{x!} \quad \dots \text{too hard manually!} \\ &\approx P(X < 950.5) = P\left(Z < \frac{950.5 - 1000}{\sqrt{1000}}\right) \\ &= P(Z < -1.57) = 0.058 \end{aligned}$$

Using Excel	
0.0578	= POISSON(950,1000,TRUE)
0.0588	= NORMDIST(950.5, 1000, SQRT(1000), TRUE)
1.6%	= (0.0588 - 0.0578) / 0.0578 = percent error

Exponential Distribution

- The Poisson distribution defined a random variable as the number of flaws along a length of wire (flaws per mm).
- The exponential distribution defines a random variable as the interval between flaws (mm's between flaws – the inverse).

Let X denote the number of flaws in x mm of wire.

If the mean number of flaws is λ per mm,

N has a Poisson distribution with mean λx .

$$P(X > x) = P(N = 0) = \frac{e^{-\lambda x} (\lambda x)^0}{0!} = e^{-\lambda x}$$

$$F(x) = P(X \leq x) = 1 - e^{-\lambda x}, \quad x \geq 0, \text{ the CDF.}$$

Now differentiating:

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0, \text{ the PDF.}$$

Exponential Distribution Definition

The random variable X that equals the distance between successive events of a Poisson process with mean number of events $\lambda > 0$ per unit interval is an exponential random variable with parameter λ . The probability density function of X is:

$$f(x) = \lambda e^{-\lambda x} \quad \text{for } 0 \leq x < \infty \quad (4-14)$$

Exponential Distribution Graphs

The y-intercept of the exponential probability density function is λ .

The random variable is non-negative and extends to infinity.

$F(x) = 1 - e^{-\lambda x}$ is well-worth committing to memory – it is used often.

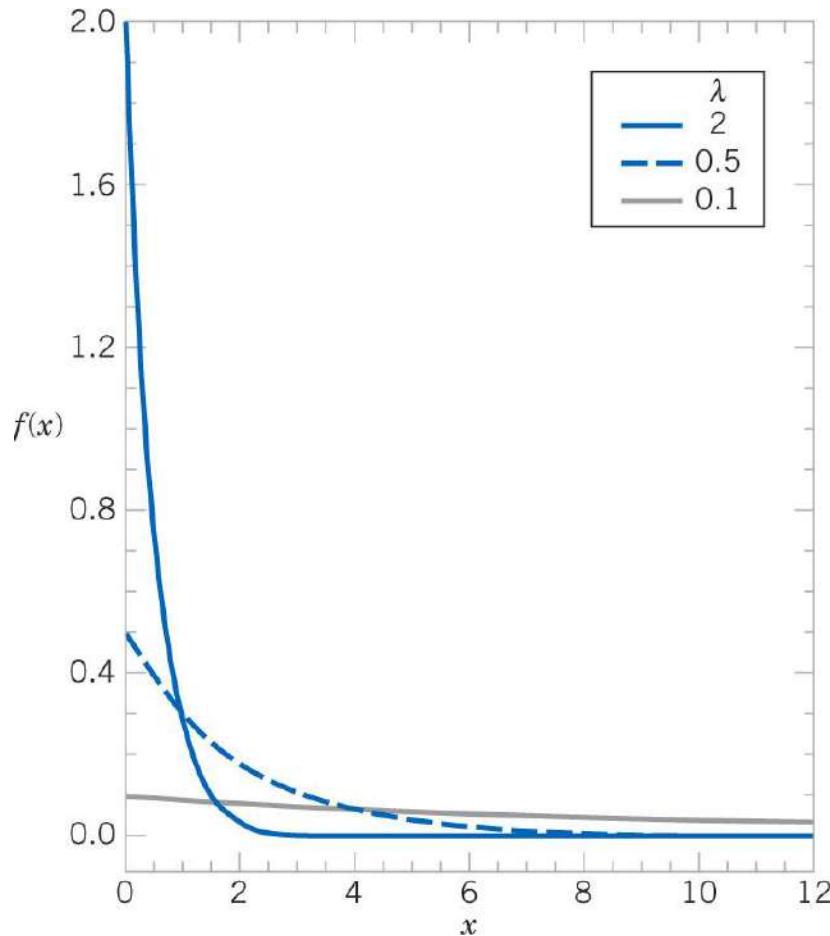


Figure 4-22 PDF of exponential random variables of selected values of λ .

Exponential Mean & Variance

If the random variable X has an exponential distribution with parameter λ ,

$$\mu = E(X) = \frac{1}{\lambda} \quad \text{and} \quad \sigma^2 = V(X) = \frac{1}{\lambda^2} \quad (4-15)$$

Note that, for the:

- Poisson distribution, the mean and **variance** are the same.
- Exponential distribution, the mean and **standard deviation** are the same.

Example 4-21: Computer Usage-1

In a large corporate computer network, user log-ons to the system can be modeled as a Poisson process with a mean of 25 log-ons per hour. What is the probability that there are no log-ons in the next 6 minutes (0.1 hour)? Let X denote the time in hours from the start of the interval until the first log-on.

$$\begin{aligned} P(X > 0.1) &= \int_{0.1}^{\infty} 25e^{25x} dx = e^{-25(0.1)} \\ &= 1 - F(0.1) = 0.082 \end{aligned}$$

Using Excel	
0.0821	= 1 - EXPONDIST(0.1, 25, TRUE)

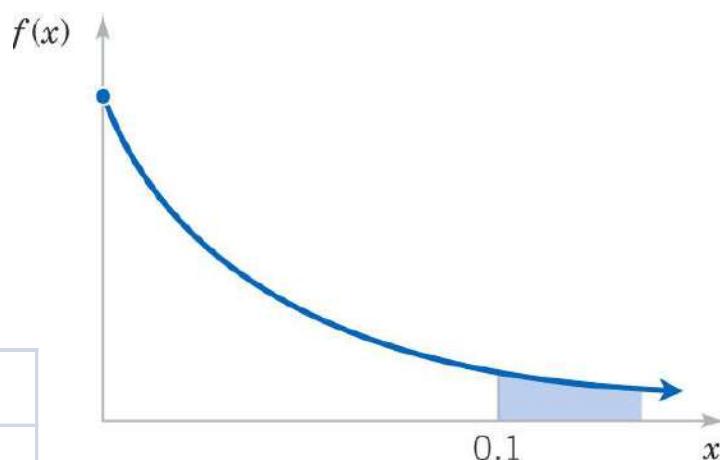


Figure 4-23 Desired probability.

Example 4-21: Computer Usage-2

Continuing, what is the probability that the time until the next log-on is between 2 and 3 minutes (0.033 & 0.05 hours)?

$$\begin{aligned} P(0.033 < X < 0.05) &= \int_{0.033}^{0.05} 25e^{-25x} \\ &= -e^{-25x} \Big|_{0.033}^{0.05} = 0.152 \\ &= F(0.05) - F(0.033) = 0.152 \end{aligned}$$

Using Excel	
0.148	= EXPONDIST(3/60, 25, TRUE) - EXPONDIST(2/60, 25, TRUE) (difference due to round-off error)

Example 4-21: Computer Usage-3

- Continuing, what is the interval of time such that the probability that no log-on occurs during the interval is 0.90?

$$P(X > x) = e^{-25x} = 0.90, \quad -25x = \ln(0.90)$$

$$x = \frac{-0.10536}{-25} = 0.00421 \text{ hour} = 0.253 \text{ minute}$$

- What is the mean and standard deviation of the time until the next log-in?

$$\mu = \frac{1}{\lambda} = \frac{1}{25} = 0.04 \text{ hour} = 2.4 \text{ minutes}$$

$$\sigma = \frac{1}{\lambda} = \frac{1}{25} = 0.04 \text{ hour} = 2.4 \text{ minutes}$$

Characteristic of a Poisson Process

- The starting point for observing the system does not matter.
- The probability of no log-in in the next 6 minutes [$P(X > 0.1 \text{ hour}) = 0.082$], regardless of whether:
 - A log-in has just occurred or
 - A log-in has not occurred for the last hour.
- A system may have different means:
 - High usage period , e.g., $\lambda = 250$ per hour
 - Low usage period, e.g., $\lambda = 25$ per hour

Example 4-22: Lack of Memory Property

- Let X denote the time between detections of a particle with a Geiger counter. Assume X has an exponential distribution with $E(X) = 1.4$ minutes. What is the probability that a particle is detected in the next 30 seconds?

$$P(X < 0.5) = F(0.5) = 1 - e^{-0.5/1.4} = 0.30$$

Using Excel
0.300 = EXPONDIST(0.5, 1/1.4, TRUE)

- No particle has been detected in the last 3 minutes. Will the probability increase since it is “due”?

$$P(X < 3.5 | X > 3) = \frac{P(3 < X < 3.5)}{P(X > 3)} = \frac{F(3.5) - F(3)}{1 - F(3)} = \frac{0.035}{0.117} = 0.30$$

- No, the probability that a particle will be detected depends only on the interval of time, not its detection history.

Lack of Memory Property

- Areas $A+B+C+D=1$
- $A = P(X < t_2)$
- $A+B+C = P(X < t_1 + t_2)$
- $C = P(X < t_1 + t_2 \cap X > t_1)$
- $C+D = P(X > t_1)$
- $C/(C+D) = P(X < t_1 + t_2 | X > t_1)$
- $A = C/(C+D)$

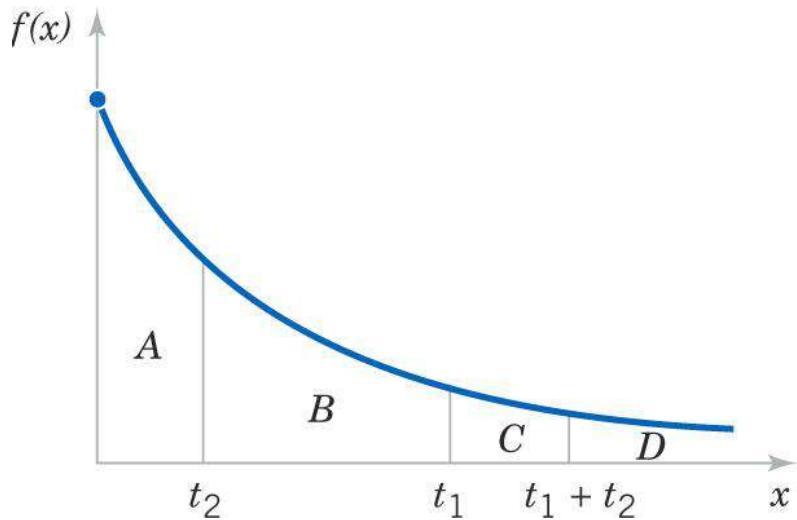


Figure 4-24 Lack of memory property of an exponential distribution.

Exponential Application in Reliability

- The reliability of electronic components is often modeled by the exponential distribution. A chip might have mean time to failure of 40,000 operating hours.
- The memoryless property implies that the component does not wear out – the probability of failure in the next hour is constant, regardless of the component age.
- The reliability of mechanical components **do** have a memory – the probability of failure in the next hour increases as the component ages. The Weibull distribution is used to model this situation.

Erlang & Gamma Distributions

- The Erlang distribution is a generalization of the exponential distribution.
- The exponential models the interval to the 1st event, while the Erlang models the interval to the r^{th} event, i.e., a sum of exponentials.
- If r is not required to be an integer, then the distribution is called gamma.
- The exponential, as well as its Erlang and gamma generalizations, is based on the Poisson process.

Example 4-23: Processor Failure

The failures of CPUs of large computer systems are often modeled as a Poisson process. Assume that units that fail are repaired immediately and the mean number of failures per hour is 0.0001. Let X denote the time until 4 failures occur. What is the probability that X exceed 40,000 hours?

Let the random variable N denote the number of failures in 40,000 hours. The time until 4 failures occur exceeds 40,000 hours *iff* the number of failures in 40,000 hours is ≤ 3 .

$$P(X > 40,000) = P(N \leq 3)$$

$$E(N) = 40,000(0.0001) = 4 \text{ failure in 40,000 hours}$$

$$P(N \leq 3) = \sum_{k=0}^3 \frac{e^{-4} 4^k}{k!} = 0.433$$

Using Excel	
0.433	= POISSON(3, 4, TRUE)

Erlang Distribution

Generalizing from the prior exercise:

$$P(X > x) = \sum_{k=0}^{r-1} \frac{e^{\lambda x} (\lambda x)^k}{k!} = 1 - F(x)$$

Now differentiating $F(x)$:

$$f(x) = \frac{\lambda^r x^{r-1} e^{-\lambda x}}{(r-1)!} \quad \text{for } x > 0 \text{ and } r = 1, 2, \dots$$

Gamma Function

The gamma function is the generalization of the factorial function for $r > 0$, not just non-negative integers.

$$\Gamma(r) = \int_0^{\infty} x^{r-1} e^{-x} dx, \quad \text{for } r > 0 \quad (4-17)$$

Properties of the gamma function

$$\Gamma(r) = (r - 1)\Gamma(r - 1) \quad \text{recursive property}$$

$$\Gamma(r) = (r - 1)! \quad \text{factorial function}$$

$$\Gamma(1) = 0! = 1$$

$$\Gamma(1/2) = \pi^{1/2} = 1.77 \quad \text{useful if manual}$$

Gamma Distribution

The random variable X with a probability density function:

$$f(x) = \frac{\lambda^r x^{r-1} e^{-\lambda x}}{\Gamma(r)}, \text{ for } x > 0 \quad (4-18)$$

has a gamma random distribution with parameters $\lambda > 0$ and $r > 0$. If r is an positive integer, then X has an Erlang distribution.

Mean & Variance of the Gamma

- If X is a gamma random variable with parameters λ and r ,

$$\mu = E(X) = r / \lambda \quad \text{and} \quad \sigma^2 = V(X) = r / \lambda^2 \quad (4-19)$$

- r and λ work together to describe the shape of the gamma distribution.

Gamma Distribution Graphs

The λ and r parameters are often called the “shape” and “scale”, but may take on different meanings.

Different parameter combinations change the distribution.

The distribution becomes symmetric as r (and μ) increases.

Name	Text	Excel	Minitab
Scale	λ	$\beta = 1 / \lambda$	$1 / \lambda$
Shape	r	α	r

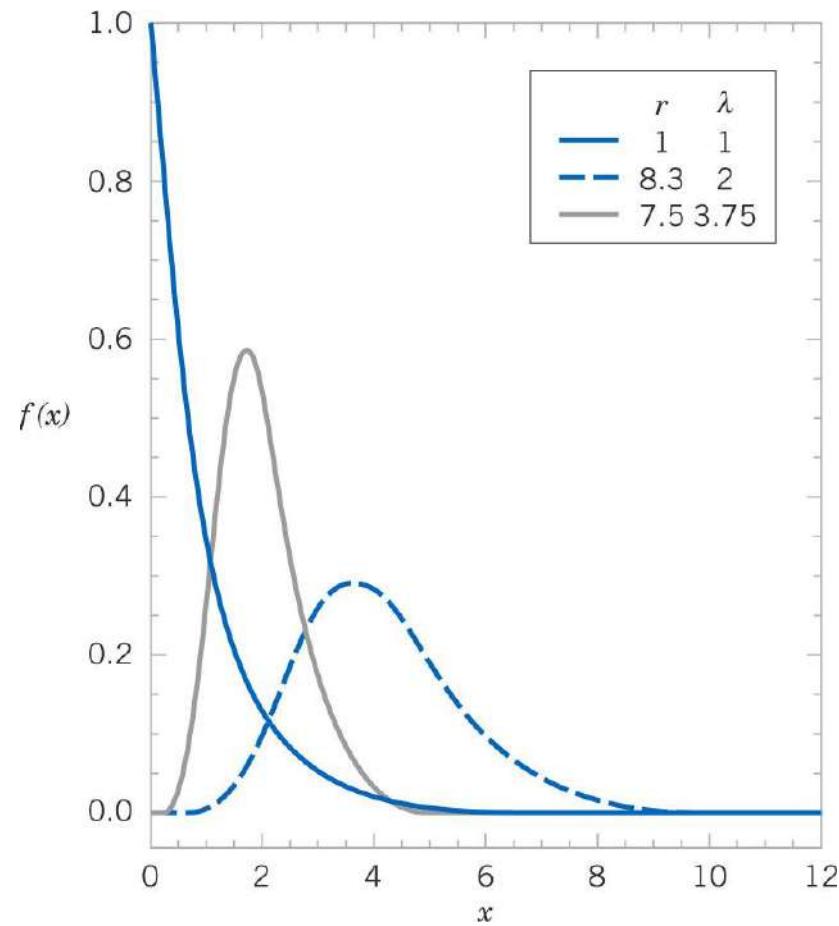


Figure 4-25 Gamma probability density functions for selected values of λ and r .

Example 4-24: Gamma Application-1

The time to prepare a micro-array slide for high-output genomics is a Poisson process with a mean of 2 hours per slide. What is the probability that 10 slides require more than 25 hours?

Let X denote the time to prepare 10 slides. Because of the assumption of a Poisson process, X has a gamma distribution with $\lambda = \frac{1}{2}$, $r = 10$, and the requested probability is $P(X > 25)$.

Using the Poisson distribution, let the random variable N denote the number of slides made in 10 hours. The time until 10 slides are made exceeds 25 hours *iff* the number of slides made in 25 hours is ≤ 9 .

$$P(X > 25) = P(N \leq 9)$$

$$E(N) = 25(1/2) = 12.5 \text{ slides in 25 hours}$$

$$P(N \leq 9) = \sum_{k=0}^9 \frac{e^{-12.5} (12.5)^k}{k!} = 0.2014$$

Using Excel

0.2014	= POISSON(9, 12.5, TRUE)
--------	--------------------------

Using the gamma distribution, the same result is obtained.

$$P(X > 25) = 1 - \int_0^{25} \frac{0.5^{10} x^9 e^{-0.5x}}{\Gamma(10)} dx$$

Using Excel

0.2014	= 1 - GAMMADIST(25,10,2,TRUE)
--------	-------------------------------

Example 4-24: Gamma Application-2

What is the mean and standard deviation of the time to prepare 10 slides?

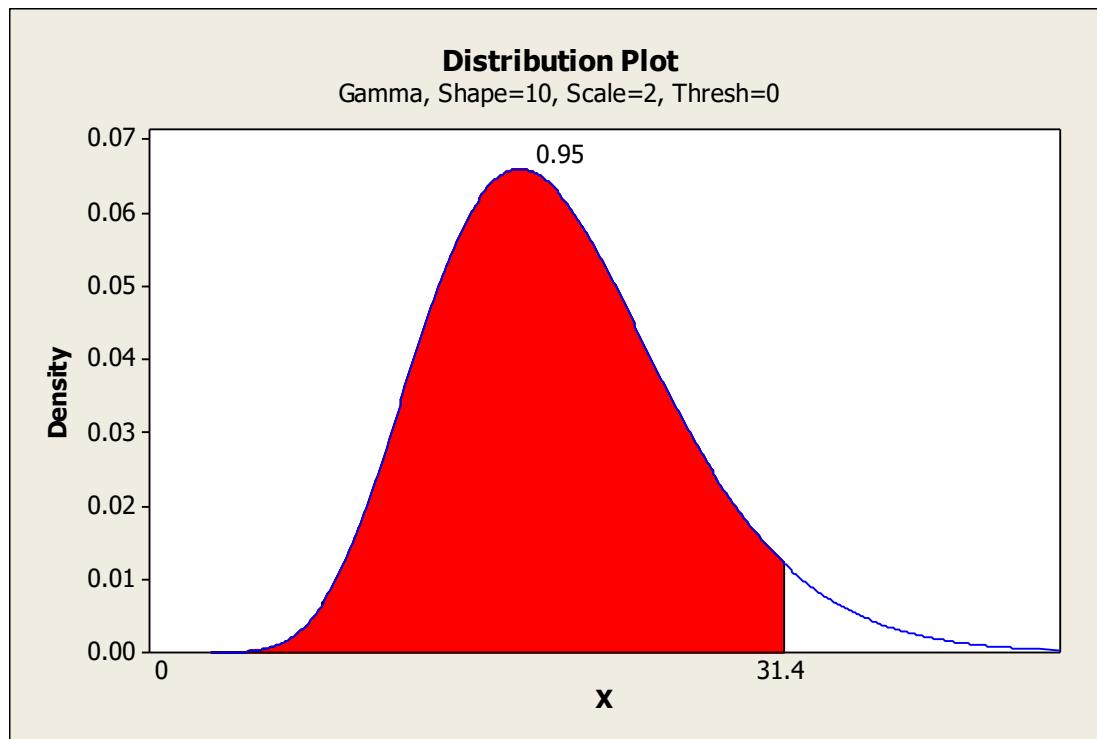
$$E(X) = \frac{r}{\lambda} = \frac{10}{0.5} = 20 \text{ hours}$$

$$V(X) = \frac{r}{\lambda^2} = \frac{10}{0.25} = 40 \text{ hours}^2$$

$$SD(X) = \frac{\sqrt{10}}{\lambda} = \sqrt{40} = 6.32 \text{ hours}$$

Example 4-24: Gamma Application-3

The slides will be completed by what length of time with 95% probability? That is: $P(X \leq x) = 0.95$



Minitab: Graph > Probability Distribution Plot >
View Probability

Using Excel

31.41 = GAMMAINV(0.95, 10, 2)

Chi-Squared Distribution

- The chi-squared distribution is a special case of the gamma distribution with
 - $\lambda = 1/2$
 - $r = v/2$ where v (nu) = 1, 2, 3, ...
 - v is called the “degrees of freedom”.
- The chi-squared distribution is used in interval estimation and hypothesis tests as discussed in Chapter 7.

Weibull Distribution

- The Weibull distribution is often used to model the time until failure for physical systems in which failures:
 - Increase over time (bearings)
 - Decrease over time (some semiconductors)
 - Remain constant over time (subject to external shock)
- Parameters provide flexibility to reflect an item's failure experience or expectation.

Weibull PDF

The random variable X with probability density function

$$f(x) = \frac{\beta}{\delta^\beta} x^{\beta-1} e^{-(x/\delta)^\beta} \quad \text{for } x > 0 \quad (4-20)$$

is a Weibull random variable with scale parameter $\delta > 0$ and shape parameter $\beta > 0$.

The cumulative density function is:

$$F(x) = 1 - e^{-(x/\delta)^\beta} \quad (4-21)$$

$$\begin{aligned} \mu &= E(X) = \delta \cdot \Gamma\left(1 + \frac{1}{\beta}\right) \\ \sigma^2 &= V(X) = \delta^2 \left[\Gamma\left(1 + \frac{2}{\beta}\right) \right] - \delta^2 \left[\Gamma\left(1 + \frac{1}{\beta}\right) \right]^2 \end{aligned} \quad (4-21a)$$

Weibull Distribution Graphs

Added slide

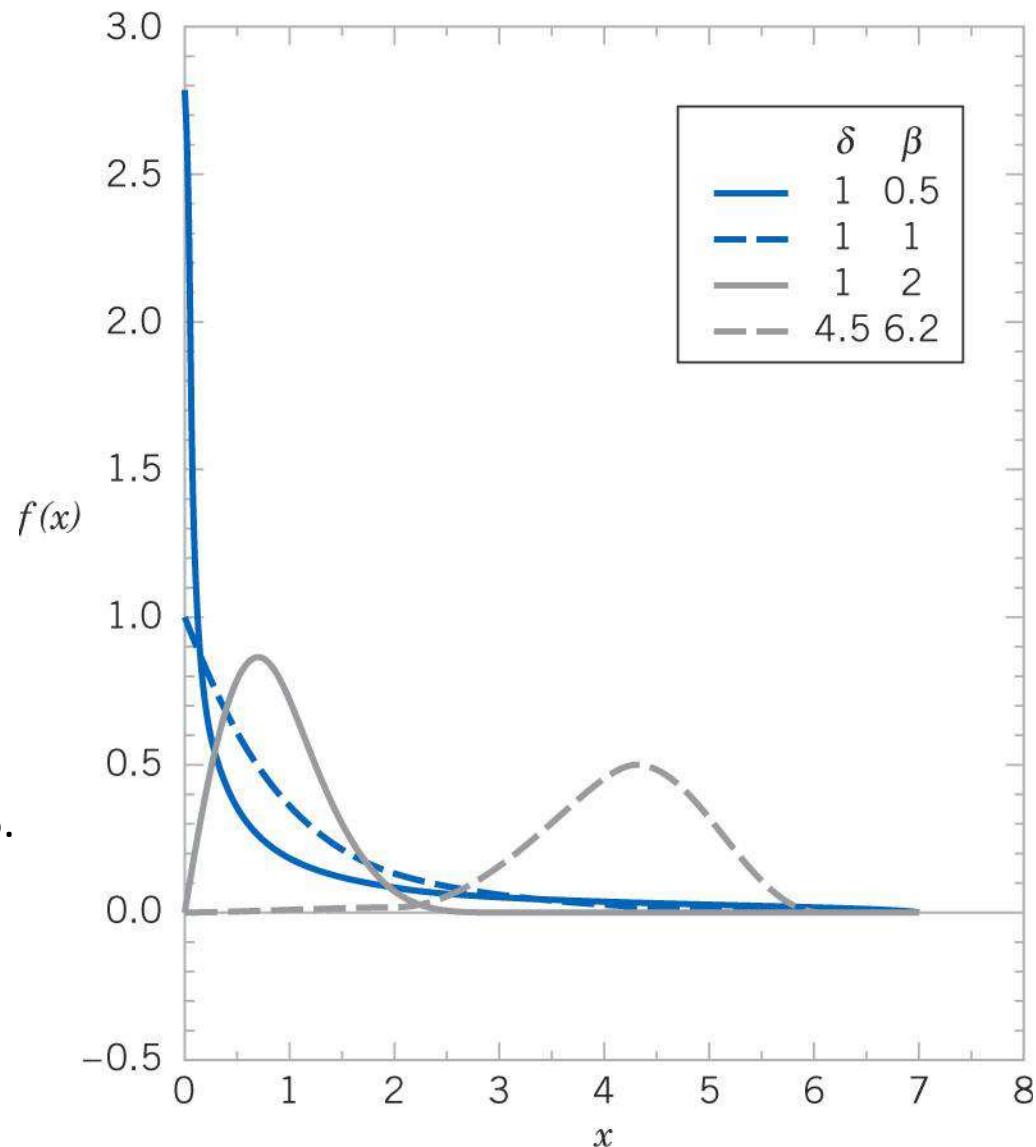


Figure 4-26 Weibull probability density function for selected values of δ and β .

Example 4-25: Bearing Wear

- The time to failure (in hours) of a bearing in a mechanical shaft is modeled as a Weibull random variable with $\beta = \frac{1}{2}$ and $\delta = 5,000$ hours.
- What is the mean time until failure?

$$\begin{aligned}E(X) &= 5000 \cdot \Gamma(1 + 1/2) = 5000 \cdot \Gamma(1.5) \\&= 5000 \cdot 0.5\sqrt{\pi} = 4,431.1 \text{ hours}\end{aligned}$$

Using Excel

4,431.1 = 5000 * EXP(GAMMALN(1.5))

- What is the probability that a bearing will last at least 6,000 hours? (error in text solution)

$$\begin{aligned}P(X > 6,000) &= 1 - F(6,000) = e^{-\left(\frac{6000}{5000}\right)^{0.5}} \\&= e^{-1.0954} = 0.334\end{aligned}$$

Using Excel

0.334 = 1 - WEIBULL(6000, 1/2, 5000, TRUE)

Lognormal Distribution

- Let W denote a normal random variable with mean of θ and variance of ω^2 , i.e., $E(W) = \theta$ and $V(W) = \omega^2$
- As a change of variable, let $X = e^W = \exp(W)$ and $W = \ln(X)$
- Now X is a lognormal random variable.

$$F(x) = P[X \leq x] = P[\exp(W) \leq x] = P[W \leq \ln(x)]$$

$$= P\left[Z \leq \frac{\ln(x) - \theta}{\omega}\right] = \Phi\left[\frac{\ln(x) - \theta}{\omega}\right] = \text{for } x > 0$$

$$= 0 \text{ for } x \leq 0$$

$$f(x) = \frac{1}{x\omega\sqrt{2\pi}} e^{-\left[\frac{\ln(x)-\theta}{2\omega}\right]^2} \quad \text{for } 0 < x < \infty$$

$$E(X) = e^{\theta + \omega^2/2} \quad \text{and} \quad V(X) = e^{2\theta + \omega^2} (e^{\omega^2} - 1) \quad (4-22)$$

Lognormal Graphs

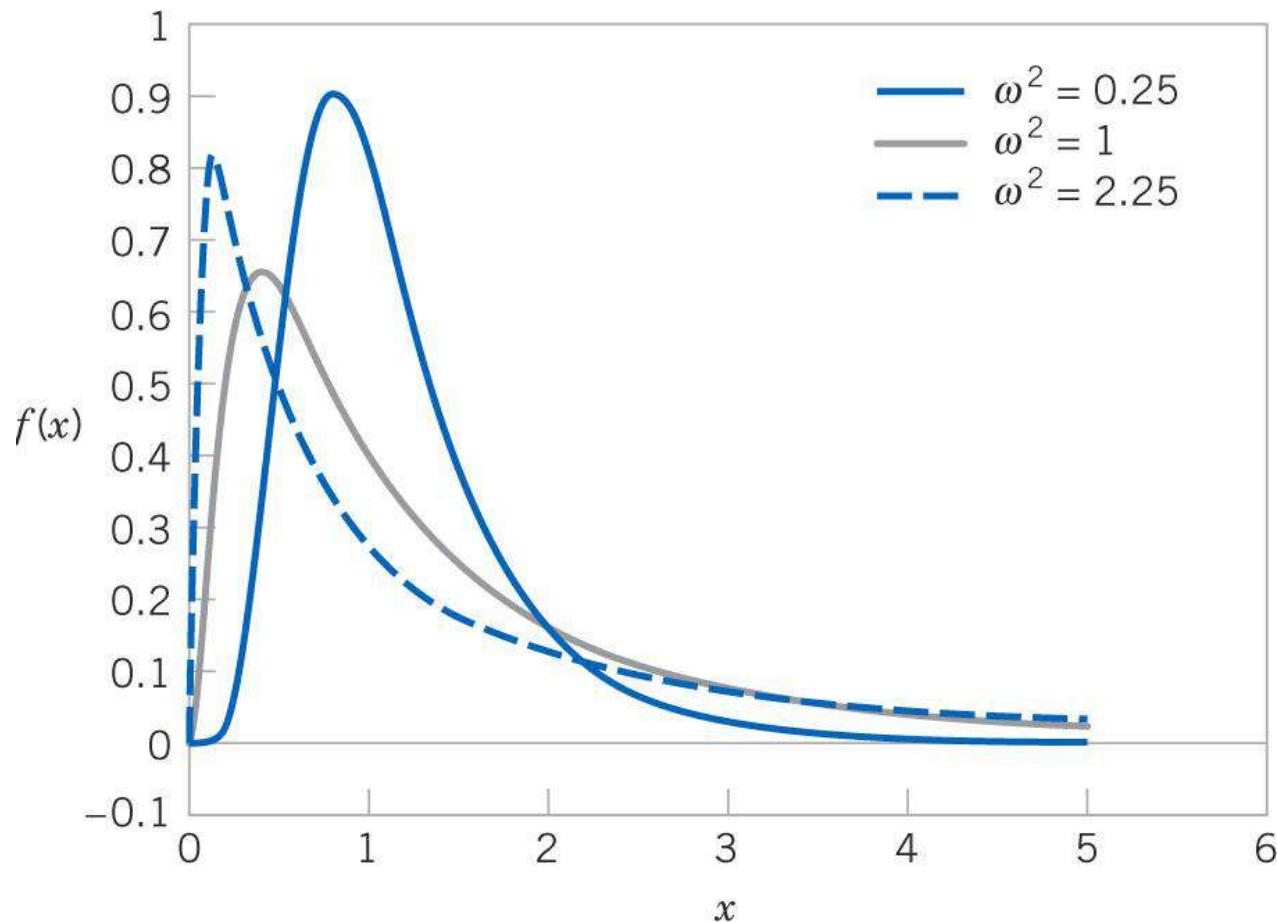


Figure 4-27 Lognormal probability density functions with $\theta = 0$ for selected values of ω^2 .

Example 4-27: Semiconductor Laser-1

The lifetime of a semiconductor laser has a lognormal distribution with $\theta = 10$ and $\omega = 1.5$ hours.

- What is the probability that the lifetime exceeds 10,000 hours?

$$\begin{aligned} P(X > 10,000) &= 1 - P[\exp(W) \leq 10,000] \\ &= 1 - P[W \leq \ln(10,000)] \\ &= 1 - \Phi\left(\frac{\ln(10,000) - 10}{1.5}\right) \\ &= 1 - \Phi(-0.5264) = 0.701 \end{aligned}$$

1 - NORMDIST(LN(10000), 10, 1.5, TRUE) =	0.701
--	-------

Example 4-27: Semiconductor Laser-2

- What lifetime is exceeded by 99% of lasers?

$$P(X > x) = P(\exp(W) > x) = P(W > \ln(x))$$

$$= 1 - \Phi\left(\frac{\ln(x) - 10}{1.5}\right) = 0.99$$

$$= 1 - \Phi(z) = 0.99 \text{ therefore } z = -2.33$$

$$\frac{\ln(x) - 10}{1.5} = -2.33 \text{ and } x = \exp(6.505) = 668.48 \text{ hours}$$

-2.3263	= NORMSINV(0.99)
6.5105	= -2.3263 * 1.5 + 10 = ln(x)
672.15	= EXP(6.5105)
(difference due to round-off error)	

- What is the mean and variance of the lifetime?

$$E(X) = e^{\theta + \sigma^2/2} = e^{10 + 1.5^2/2}$$
$$= \exp(11.125) = 67,846.29$$

$$V(X) = e^{2\theta + 2\sigma^2} (e^{\sigma^2} - 1) = e^{2 \cdot 10 + 1.5^2} (e^{1.5^2} - 1)$$
$$= \exp(22.25) \cdot [\exp(2.25) - 1] = 39,070,059,886.6$$

$$SD(X) = 197,661.5$$

Beta Distribution

A continuous distribution that is flexible, but bounded over the $[0, 1]$ interval is useful for probability models. Examples are:

- Proportion of solar radiation absorbed by a material.
- Proportion of the max time to complete a task.

The random variable X with probability density function

$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \cdot \Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \quad \text{for } 0 \leq x \leq 1$$

is a beta random variable with parameters $\alpha > 0$ and $\beta > 0$.

Beta Shapes Are Flexible

Distribution shape guidelines:

1. If $\alpha = \beta$, symmetrical about $x = 0.5$.
2. If $\alpha = \beta = 1$, uniform.
3. If $\alpha = \beta < 1$, symmetric & U-shaped.
4. If $\alpha = \beta > 1$, symmetric & mound-shaped.
5. If $\alpha \neq \beta$, skewed.

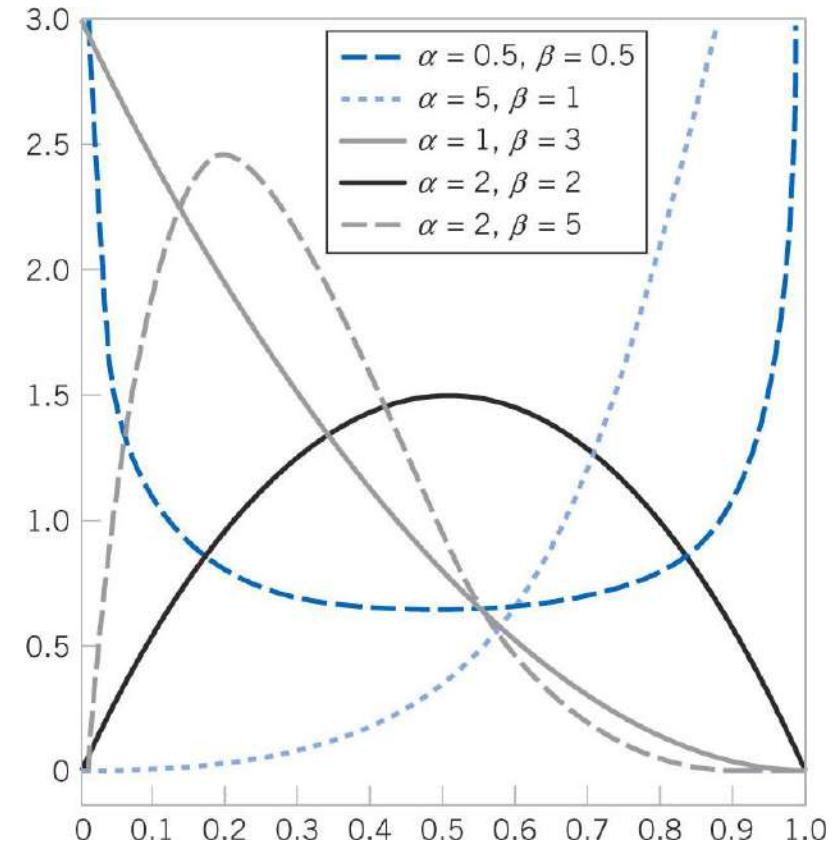


Figure 4-28 Beta probability density functions for selected values of the parameters α and β .

Example 4-27: Beta Computation-1

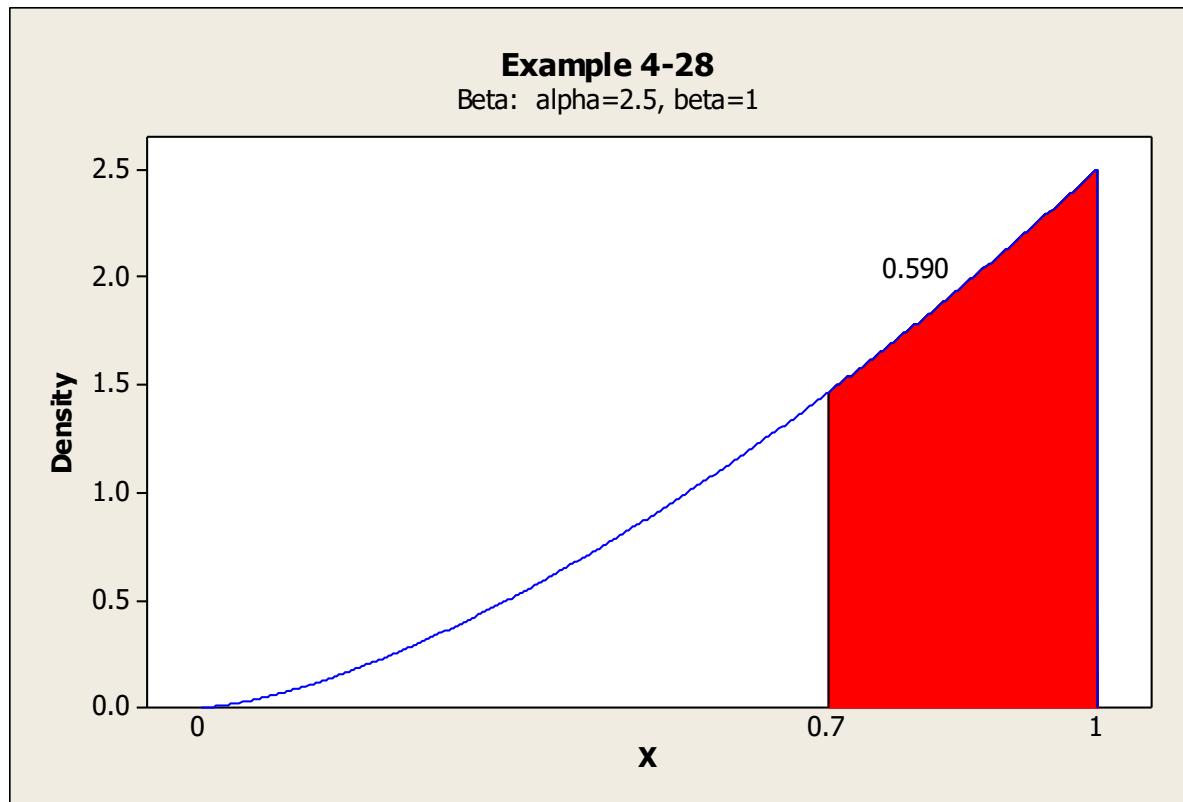
Consider the completion time of a large commercial real estate development. The proportion of the maximum allowed time to complete a task is a beta random variable with $\alpha = 2.5$ and $\beta = 1$. What is the probability that the proportion of the max time exceeds 0.7? Let X denote that proportion.

$$\begin{aligned} P(X > 0.7) &= \int_{0.7}^1 \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \cdot \Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} dx \\ &= \frac{\Gamma(3.5)}{\Gamma(2.5) \cdot \Gamma(1)} \int_{0.7}^1 x^{1.5} dx \\ &= \frac{2.5(1.5)(0.5)\sqrt{\pi}}{1.5(0.5)\sqrt{\pi} \cdot 1} \left[\frac{x^{2.5}}{2.5} \right]_{0.7}^1 \\ &= 1 - (0.7)^{2.5} = 0.59 \end{aligned}$$

Using Excel
0.590 = 1 - BETADIST(0.7,2.5,1,0,1)

Example 4-27: Beta Computation-2

This Minitab graph illustrates the prior calculation. **FIX**



Mean & Variance of the Beta Distribution

If X has a beta distribution with parameters α and β ,

$$\mu = E(X) = \frac{\alpha}{\alpha + \beta}$$

$$\sigma^2 = V(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

Example 4-28: In the prior example, $\alpha = 2.5$ and $\beta = 1$. What are the mean and variance of this distribution?

$$\mu = \frac{2.5}{2.5+1} = \frac{2.5}{3.5} = 0.71$$

$$\sigma^2 = \frac{2.5(1)}{(2.5+1)^2(2.5+1+1)} = \frac{2.5}{3.5^2(4.5)} = 0.045$$

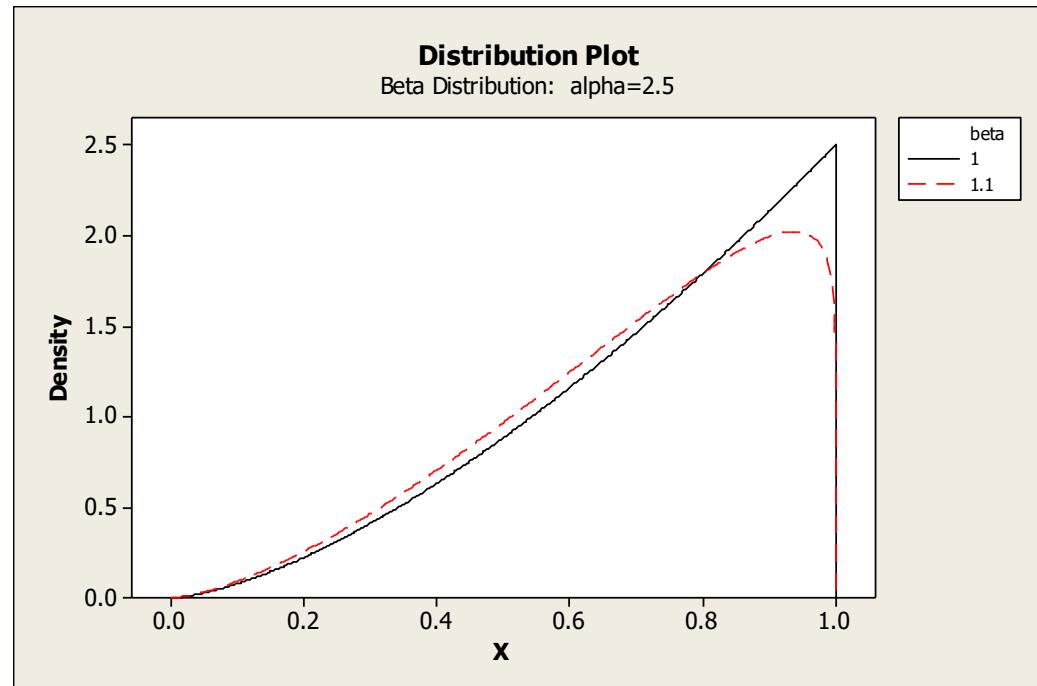
Mode of the Beta Distribution

If $\alpha > 1$ and $\beta > 1$, then the beta distribution is mound-shaped and has an interior peak, called the mode of the distribution. Otherwise, the mode occurs at an endpoint.

General formula:

$$\text{Mode} = \frac{\alpha - 1}{\alpha + \beta - 2}$$

for $\alpha > 0$ and $\beta > 0$.



case	alpha	beta	mode		
Example 4-28	2.25	1	1.00	= (2.5-1) / (2.5+1.0-2)	
Alternate	2.25	1.1	0.94	= (2.5-1) / (2.5+1.1-2)	

Extended Range for the Beta Distribution

The beta random variable X is defined for the $[0, 1]$ interval. That interval can be changed to $[a, b]$. Then the random variable W is defined as a linear function of X :

$$W = a + (b - a)X$$

With mean and variance:

$$E(W) = a + (b - a)E(X)$$

$$V(W) = (b-a)^2 V(X)$$

Important Terms & Concepts of Chapter 4

Beta distribution

Chi-squared distribution

Continuity correction

Continuous uniform distribution

Cumulative probability distribution
for a continuous random
variable

Erlang distribution

Exponential distribution

Gamma distribution

Lack of memory property of a
continuous random variable

Lognormal distribution

Mean for a continuous random
variable

Mean of a function of a continuous
random variable

Normal approximation to binomial
& Poisson probabilities

Normal distribution

Probability density function

Probability distribution of a
continuous random variable

Standard deviation of a continuous
random variable

Standardizing

Standard normal distribution

Variance of a continuous random
variable

Weibull distribution

6

Descriptive Statistics

CHAPTER OUTLINE

6-1 Numerical Summaries of
Data

6-2 Stem-and-Leaf Diagrams

6-3 Frequency Distributions
and Histograms

6-4 Box Plots

6-5 Time Sequence Plots

6-6 Probability Plots

Learning Objective for Chapter 6

After careful study of this chapter, you should be able to do the following:

1. Compute and interpret the sample mean, sample variance, sample standard deviation, sample median, and sample range.
2. Explain the concepts of sample mean, sample variance, population mean, and population variance.
3. Construct and interpret visual data displays, including the stem-and-leaf display, the histogram, and the box plot.
4. Explain the concept of random sampling.
5. Construct and interpret normal probability plots.
6. Explain how to use box plots, and other data displays, to visually compare two or more samples of data.
7. Know how to use simple time series plots to visually display the important features of time-oriented data.

Numerical Summaries of Data

- Data are the numeric observations of a phenomenon of interest. The totality of all observations is a **population**. A portion used for analysis is a random **sample**.
- We gain an understanding of this collection, possibly massive, by describing it numerically and graphically, usually with the sample data.
- We describe the collection in terms of shape, outliers, center, and spread (SOCS).
- The center is measured by the mean.
- The spread is measured by the variance.

Populations & Samples

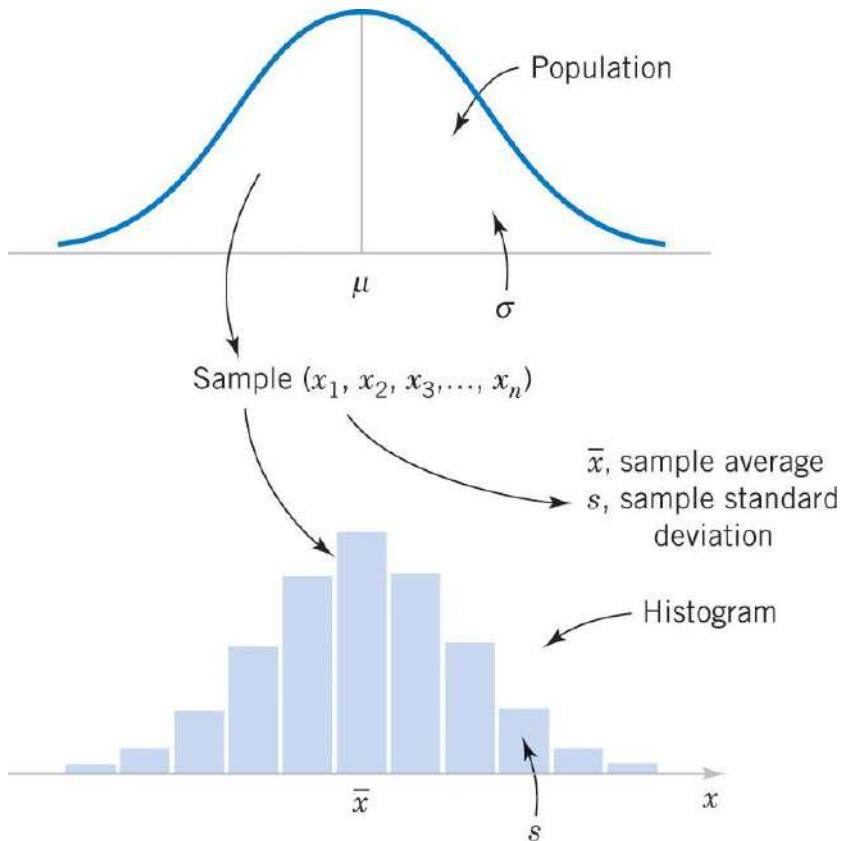


Figure 6-3 (out of order) A population is described, in part, by its **parameters**, i.e., mean (μ) and standard deviation (σ). A random sample of size n is drawn from a population and is described, in part, by its **statistics**, i.e., mean ($x\text{-bar}$) and standard deviation (s). **The statistics are used to estimate the parameters.**

Mean

If the n observations in a random sample are denoted by x_1, x_2, \dots, x_n , the **sample mean** is

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n} \quad (6-1)$$

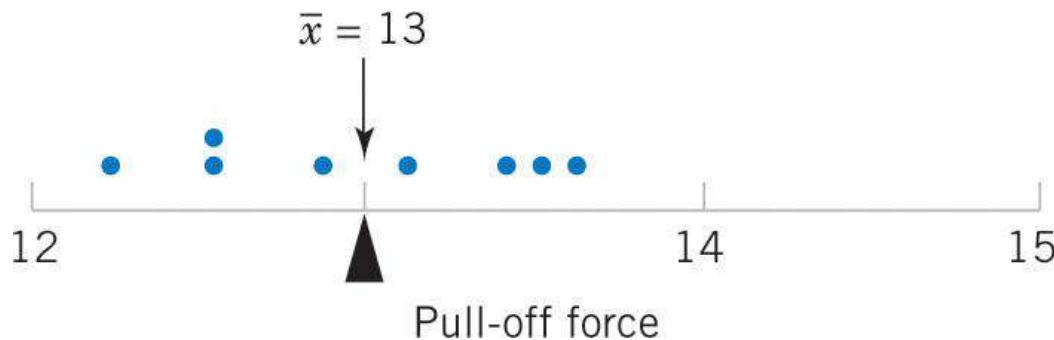
For the N observations in a population denoted by x_1, x_2, \dots, x_N , the **population mean** is analogous to a probability distribution as

$$\mu = \sum_{i=1}^N x_i \cdot f(x) = \frac{\sum_{i=1}^N x_i}{N} \quad (6-2)$$

Exercise 6-1: Sample Mean

Consider 8 observations (x_i) of pull-off force from engine connectors from Chapter 1 as shown in the table.

$$\bar{x} = \text{average} = \frac{\sum_{i=1}^8 x_i}{8} = \frac{12.6 + 12.9 + \dots + 13.1}{8}$$
$$= \frac{104}{8} = 13.0 \text{ pounds}$$



i	x_i
1	12.6
2	12.9
3	13.4
4	12.2
5	13.6
6	13.5
7	12.6
8	13.1
<hr/>	
12.99	
<hr/>	
$= \text{AVERAGE}(\$B2:\$B9)$	

Figure 6-1 The sample mean is the balance point.

Variance Defined

If the n observations in a sample are denoted by x_1, x_2, \dots, x_n , the **sample variance** is

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \quad (6-3)$$

For the N observations in a population denoted by x_1, x_2, \dots, x_N , the **population variance**, analogous to the variance of a probability distribution, is

$$\sigma^2 = \sum_{i=1}^N (x_i - \mu)^2 \cdot f(x) = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N} \quad (6-5)$$

Standard Deviation Defined

- The standard deviation is the square root of the variance.
- σ is the population standard deviation symbol.
- s is the sample standard deviation symbol.
- The units of the standard deviation are the same as:
 - The data.
 - The mean.

Rationale for the Variance

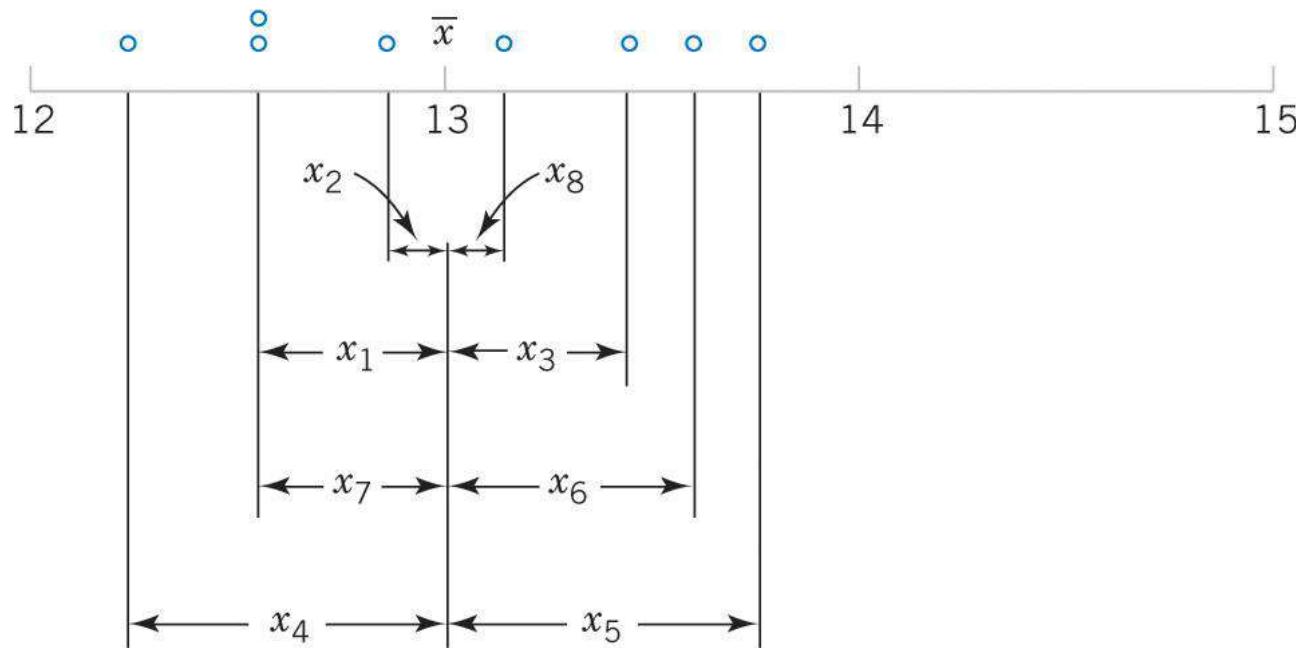


Figure 6-2 The x_i values above are the deviations from the mean. Since the mean is the balance point, the sum of the left deviations (negative) equals the sum of the right deviations (positive). If the deviations are squared, they become a measure of the data spread. The variance is the average data spread.

Example 6-2: Sample Variance

Table 6-1 displays the quantities needed to calculate the summed squared deviations, the numerator of the variance.

Dimension of:

x_i is pounds

Mean is pounds.

Variance is pounds².

Standard deviation is pounds.

Desired accuracy is generally accepted to be **one more place** than the data.

i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	12.6	-0.40	0.1600
2	12.9	-0.10	0.0100
3	13.4	0.40	0.1600
4	12.3	-0.70	0.4900
5	13.6	0.60	0.3600
6	13.5	0.50	0.2500
7	12.6	-0.40	0.1600
8	13.1	0.10	0.0100
sums =	104.00	0.00	1.6000
	divide by 8		divide by 7
mean =	13.00	variance =	0.2286
	standard deviation =		0.48

Computation of s^2

The prior calculation is definitional and tedious. A shortcut is derived here and involves just 2 sums.

$$\begin{aligned}s^2 &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^n (x_i^2 + \bar{x}^2 - 2x_i\bar{x})}{n-1} \\&= \frac{\sum_{i=1}^n x_i^2 + n\bar{x}^2 - 2\bar{x}\sum_{i=1}^n x_i}{n-1} = \frac{\sum_{i=1}^n x_i^2 + n\bar{x}^2 - 2\bar{x} \cdot n\bar{x}}{n-1} \\&= \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{n-1} = \frac{\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2 / n}{n-1} \quad (6-4)\end{aligned}$$

Example 6-3: Variance by Shortcut

$$s^2 = \frac{\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2 / n}{n-1}$$

$$= \frac{1,353.60 - (104.0)^2 / 8}{7}$$

$$= \frac{1.60}{7} = 0.2286 \text{ pounds}^2$$

$$s = \sqrt{0.2286} = 0.48 \text{ pounds}$$

i	x_i	x_i^2
1	12.6	158.76
2	12.9	166.41
3	13.4	179.56
4	12.3	151.29
5	13.6	184.96
6	13.5	182.25
7	12.6	158.76
8	13.1	171.61
sums =	104.0	1,353.60

What is this “n–1”?

- The population variance is calculated with N , the population size. Why isn't the sample variance calculated with n , the sample size?
- The true variance is based on data deviations from the true mean, μ .
- The sample calculation is based on the data deviations from $x\text{-bar}$, not μ . $X\text{-bar}$ is an **estimator** of μ ; close but not the same. So the $n-1$ divisor is used to compensate for the error in the mean estimation.

Degrees of Freedom

- The sample variance is calculated with the quantity $n-1$.
- This quantity is called the “degrees of freedom”.
- Origin of the term:
 - There are n deviations from $x\text{-bar}$ in the sample.
 - The sum of the deviations is zero. (Balance point)
 - $n-1$ of the observations can be freely determined, but the n^{th} observation is fixed to maintain the zero sum.

Sample Range

If the n observations in a sample are denoted by x_1, x_2, \dots, x_n , the sample range is:

$$r = \max(x_i) - \min(x_i)$$

It is the largest observation in the sample less the smallest observation.

From Example 6-3:

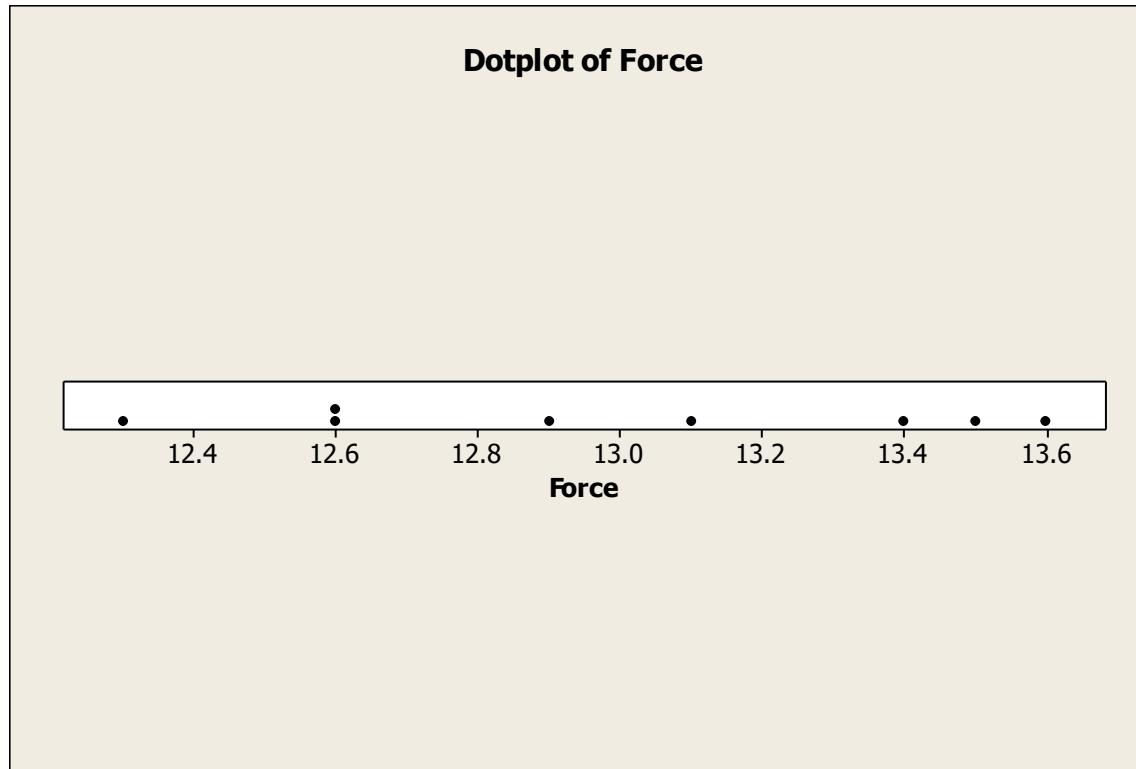
$$r = 13.6 - 12.3 = 1.30$$

Note that: population range \geq sample range

Intro to Stem & Leaf Diagrams

First, let's discuss dot diagrams – dots representing data on the number line.

Minitab produces this graphic using the Example 6-1 data.



Stem-and-Leaf Diagrams

- Dot diagrams (dotplots) are useful for small data sets. Stem & leaf diagrams are better for large sets.
- Steps to construct a stem-and-leaf diagram:
 - 1) Divide each number (x_i) into two parts: a **stem**, consisting of the leading digits, and a **leaf**, consisting of the remaining digit.
 - 2) List the stem values in a vertical column (no skips).
 - 3) Record the leaf for each observation beside its stem.
 - 4) Write the units for the stems and leaves on the display.

Example 6-4: Alloy Strength

Table 6-2 Compressive Strength (psi) of Aluminum-Lithium Specimens

105	221	183	186	121	181	180	143
97	154	153	174	120	168	167	141
245	228	174	199	181	158	176	110
163	131	154	115	160	208	158	133
207	180	190	193	194	133	156	123
134	178	76	167	184	135	229	146
218	157	101	171	165	172	158	169
199	151	142	163	145	171	148	158
160	175	149	87	160	237	150	135
196	201	200	176	150	170	118	149

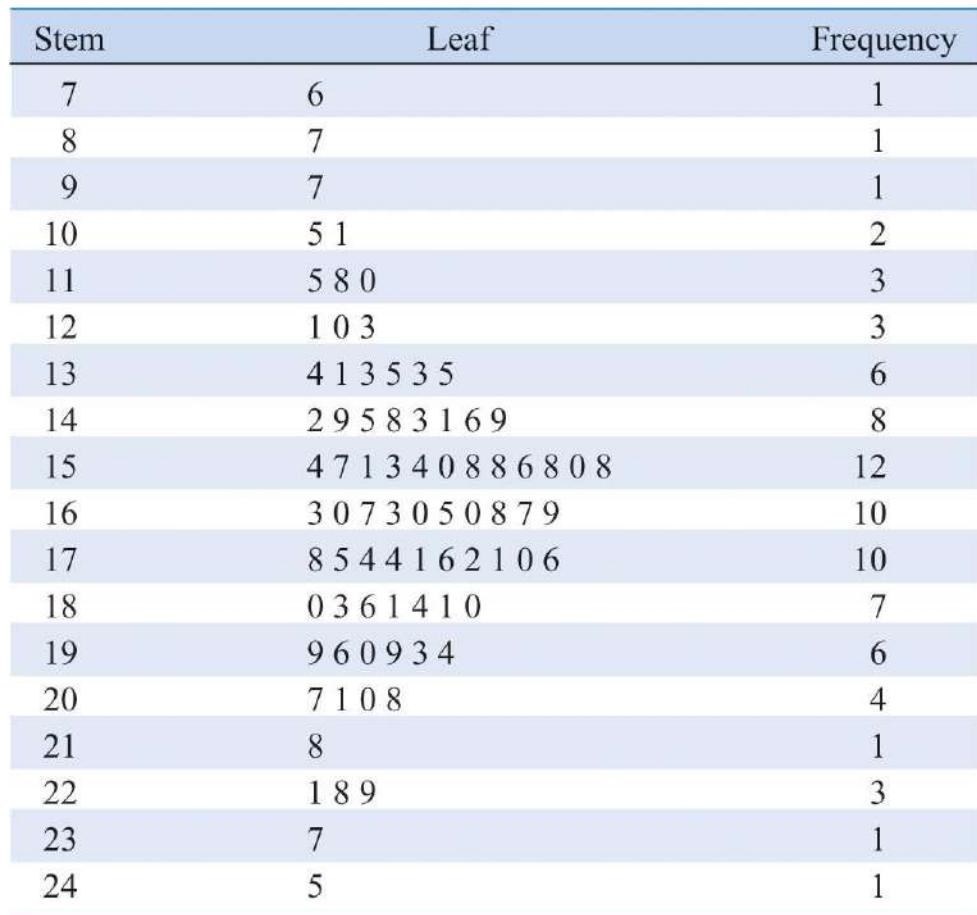


Figure 6-4 Stem-and-leaf diagram for Table 6-2 data. Center is about 155 and most data is between 110 and 200. Leaves are unordered.

Split Stems

- The purpose of the stem-and-leaf is to describe the data distribution graphically.
- If the data are too clustered, we can split and have multiple stems, thereby increasing the number of stems.
 - Split 2 for 1:
 - Lower stem for leaves 0, 1, 2, 3, 4
 - Upper stem for leaves 5, 6, 7, 8, 9
 - Split 5 for 1:
 - 1st stem for leaves 0, 1
 - 2nd stem for leaves 2, 3
 - 3rd stem for leaves 4, 5
 - 4th stem for leaves 6, 7
 - 5th stem for leaves 8, 9

Example 6-5: Chemical Yield Displays

Stem	Leaf	Stem	Leaf	Stem	Leaf
6	1 3 4 5 5 6	6L	1 3 4	6z	1
7	0 1 1 3 5 7 8 8 9	6U	5 5 6	6t	3
8	1 3 4 4 7 8 8	7L	0 1 1 3	6f	4 5 5
9	2 3 5	7U	5 7 8 8 9	6s	6
(a)		8L	1 3 4 4	6e	
		8U	7 8 8	7z	0 1 1
		9L	2 3	7t	3
		9U	5	7f	5
(b)				7s	7
				7e	8 8 9
				8z	1
				8t	3
				8f	4 4
				8s	7
				8e	8 8
				9z	
				9t	2 3
				9f	5
				9s	
				9e	
				(c)	

Figure 6-5 (a) Stems not split; too compact
(b) Stems split 2-for-1; nice shape
(c) Stems split 5-for-1; too spread out

Stem-and-Leaf by Minitab

- Table 6-2 data: Leaves are ordered, hence the data is sorted.
- Median is the middle of the sorted observations.
 - If n is odd, the middle value.
 - If n is even, the average or midpoint of the two middle values. Median is 161.5.
- Mode is 158, the most frequent value.

Figure 6-6

Stem-and-leaf of Strength

Count	Stem	Leaves
1	7	6
2	8	7
3	9	7
5	10	15
8	11	058
11	12	013
17	13	133455
25	14	12356899
37	15	001344678888
(10)	16	0003357789
33	17	0112445668
23	18	0011346
16	19	034699
10	20	0178
6	21	8
5	22	189
2	23	7
1	24	5

Quartiles

- The three quartiles partition the data into four equally sized counts or segments.
 - 25% of the data is less than q_1 .
 - 50% of the data is less than q_2 , the median.
 - 75% of the data is less than q_3 .
- Calculated as $\text{Index} = f(n+1)$ where:
 - $\text{Index } (I)$ is the I^{th} item (interpolated) of the sorted data list.
 - f is the fraction associated with the quartile.
 - n is the sample size.
- For the Table 6-2 data:

		Value of indexed item			
f	Index	I^{th}	$(I+1)^{\text{th}}$	quartile	
0.25	20.25	143	144	143.25	
0.50	40.50	160	163	161.50	
0.75	60.75	181	181	181.00	

Percentiles

- Percentiles are a special case of the quartiles.
- Percentiles partition the data into 100 segments.
- The $\text{Index} = f(n+1)$ methodology is the same.
- The 37%ile is calculated as follows:
 - Refer to the Table 6-2 stem-and-leaf diagram.
 - $\text{Index} = 0.37(81) = 29.97$
 - $37\text{\%ile} = 153 + 0.97(154 - 153) = 153.97$

Interquartile Range

- The interquartile range (IQR) is defined as:

$$\text{IQR} = q_1 - q_3.$$

- From Table 6-2:

$$\text{IQR} = 181.00 - 143.25 = 37.75 = 37.8$$

- Impact of outlier data:
 - IQR is not affected
 - Range is directly affected.

Minitab Descriptives

- The Minitab selection menu:
Stat > Basic Statistics > Display Descriptive Statistics
calculates the descriptive statistics for a data set.
- For the Table 6-2 data, Minitab produces:

Variable	N	Mean	StDev		
Strength	80	162.66	33.77		
	Min	Q1	Median	Q3	Max
	76.00	143.50	161.50	181.00	245.00
5-number summary					

Frequency Distributions

- A frequency distribution is a compact summary of data, expressed as a table, graph, or function.
- The data is gathered into **bins** or **cells**, defined by **class intervals**.
- The **number of classes**, multiplied by the class interval, should exceed the range of the data. The square root of the sample size is a guide.
- The boundaries of the class intervals should be convenient values, as should the **class width**.

Frequency Distribution Table

Considerations:

$$\text{Range} = 245 - 76 = 169$$

$$\sqrt{80} = 8.9$$

Trial class width = 18.9

Decisions:

Number of classes = 9

Class width = 20

Range of classes =
 $20 * 9 = 180$

Starting point = 70

Table 6-4 Frequency Distribution of Table 6-2 Data

Class	Frequency	Relative Frequency	Cumulative Relative Frequency
$70 \leq x < 90$	2	0.0250	0.0250
$90 \leq x < 110$	3	0.0375	0.0625
$110 \leq x < 130$	6	0.0750	0.1375
$130 \leq x < 150$	14	0.1750	0.3125
$150 \leq x < 170$	22	0.2750	0.5875
$170 \leq x < 190$	17	0.2125	0.8000
$190 \leq x < 210$	10	0.1250	0.9250
$210 \leq x < 230$	4	0.0500	0.9750
$230 \leq x < 250$	2	0.0250	1.0000
	80	1.0000	

Histograms

- A histogram is a visual display of a frequency distribution, similar to a bar chart or a stem-and-leaf diagram.
- Steps to build one with equal bin widths:
 - 1) Label the bin boundaries on the horizontal scale.
 - 2) Mark & label the vertical scale with the frequencies or relative frequencies.
 - 3) Above each bin, draw a rectangle whose height is equal to the frequency or relative frequency.

Histogram of the Table 6-2 Data

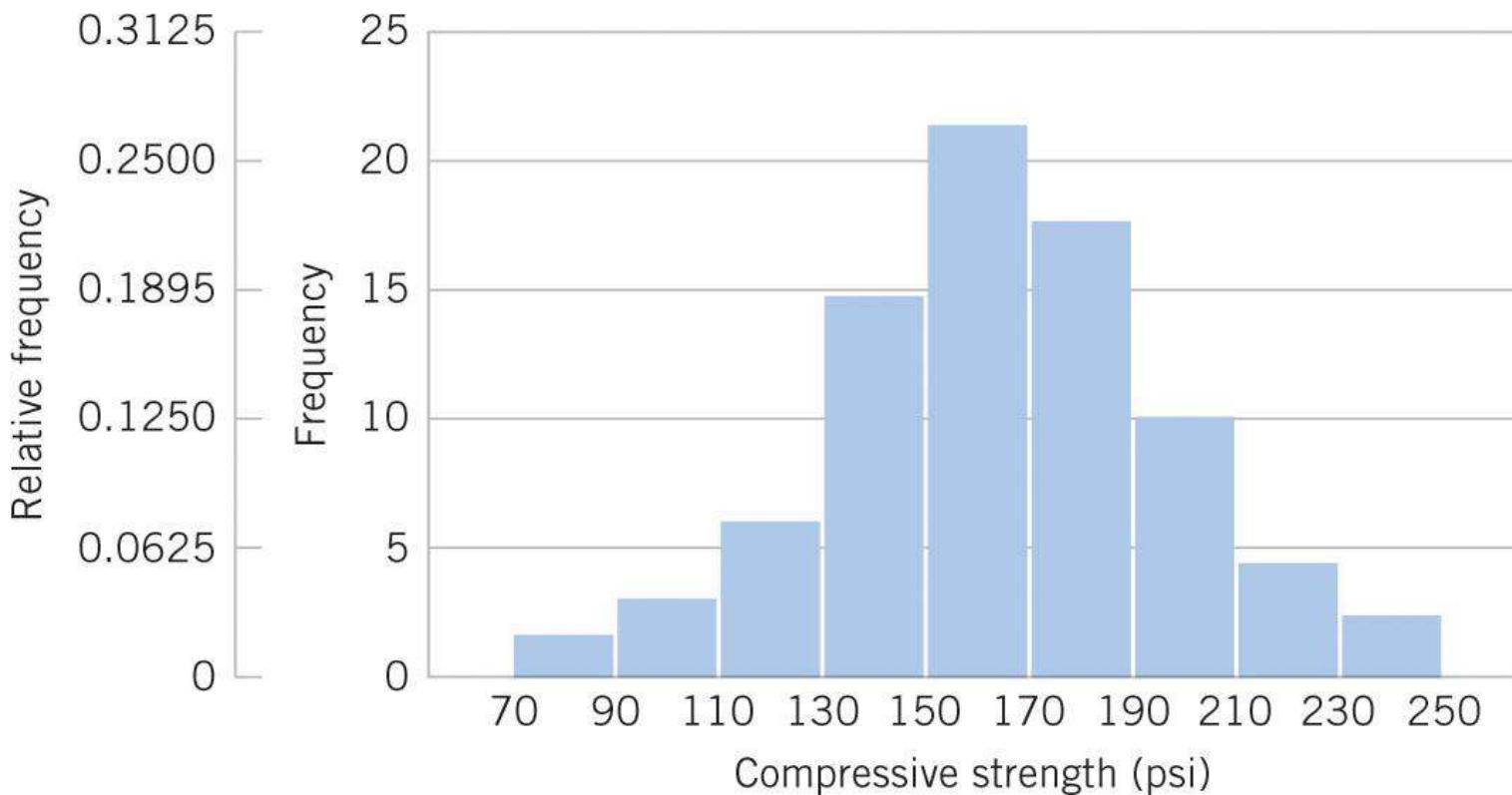


Figure 6-7 Histogram of compressive strength of 80 aluminum-lithium alloy specimens. Note these features – (1) horizontal scale bin boundaries & labels with units, (2) vertical scale measurements and labels, (3) histogram title at top or in legend.

Histograms with Unequal Bin Widths

- If the data is tightly clustered in some regions and scattered in others, it is visually helpful to use narrow class widths in the clustered region and wide class widths in the scattered areas.
- In this approach, the rectangle **area**, not the height, must be proportional to the class frequency.

$$\text{Rectangle height} = \frac{\text{bin frequency}}{\text{bin width}}$$

Poor Choices in Drawing Histograms-1

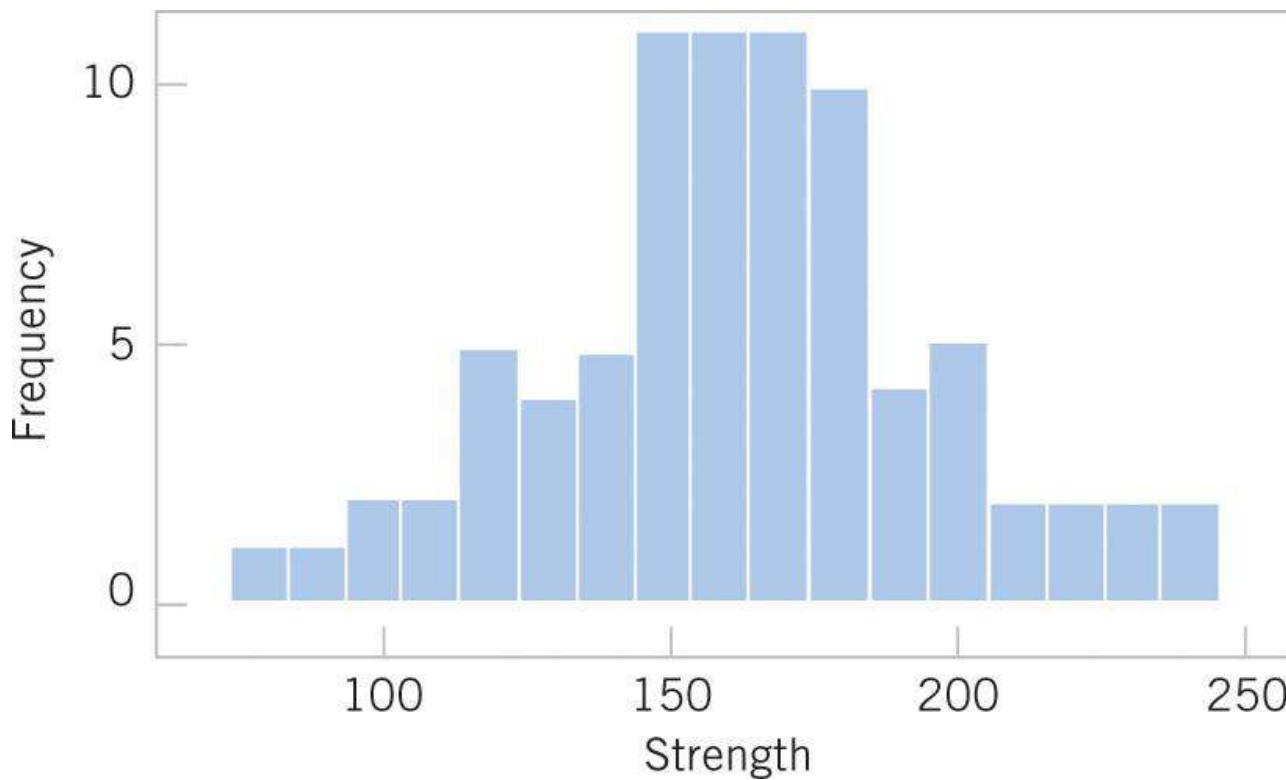


Figure 6-8 Histogram of compressive strength of 80 aluminum-lithium alloy specimens. Errors: too many bins (17) create jagged shape, horizontal scale not at class boundaries, horizontal axis label does not include units.

Poor Choices in Drawing Histograms-2

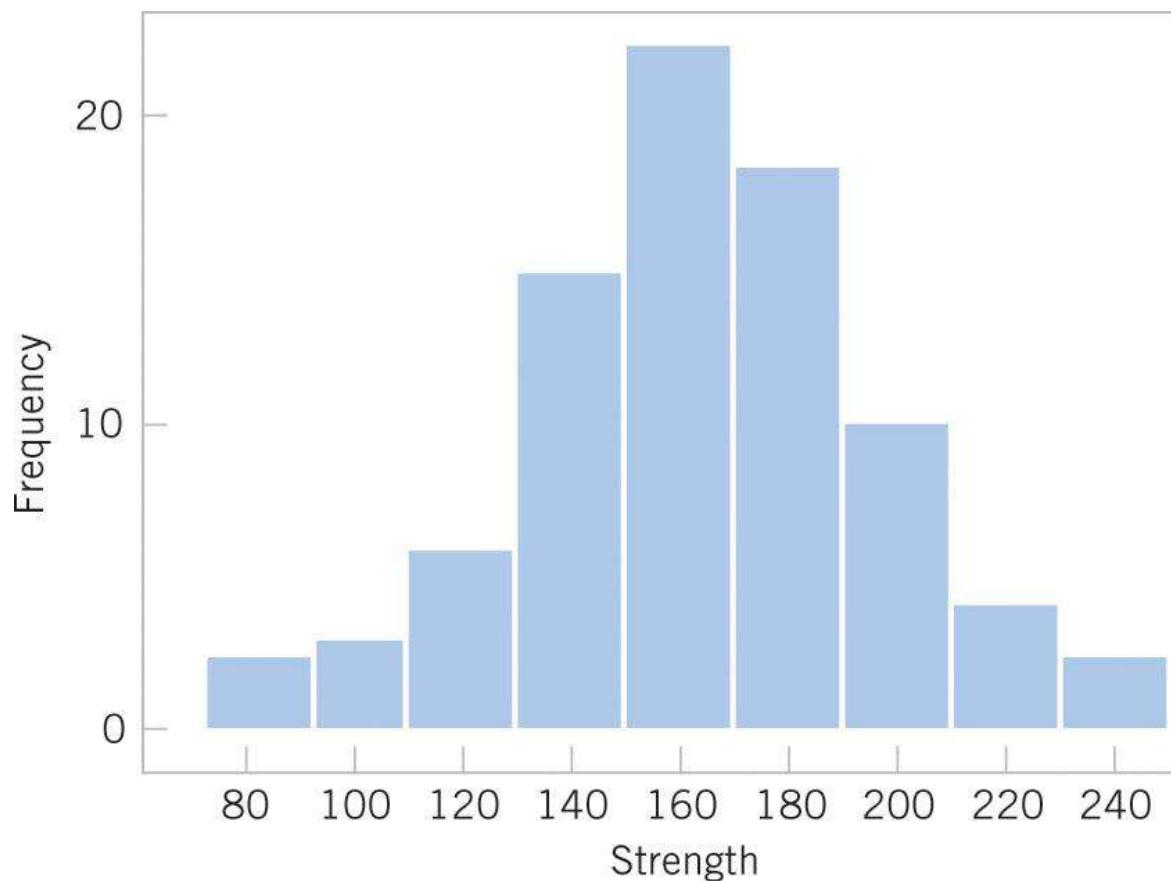


Figure 6-9 Histogram of compressive strength of 80 aluminum-lithium alloy specimens. Errors: horizontal scale not at class boundaries (cutpoints), horizontal axis label does not include units.

Cumulative Frequency Plot

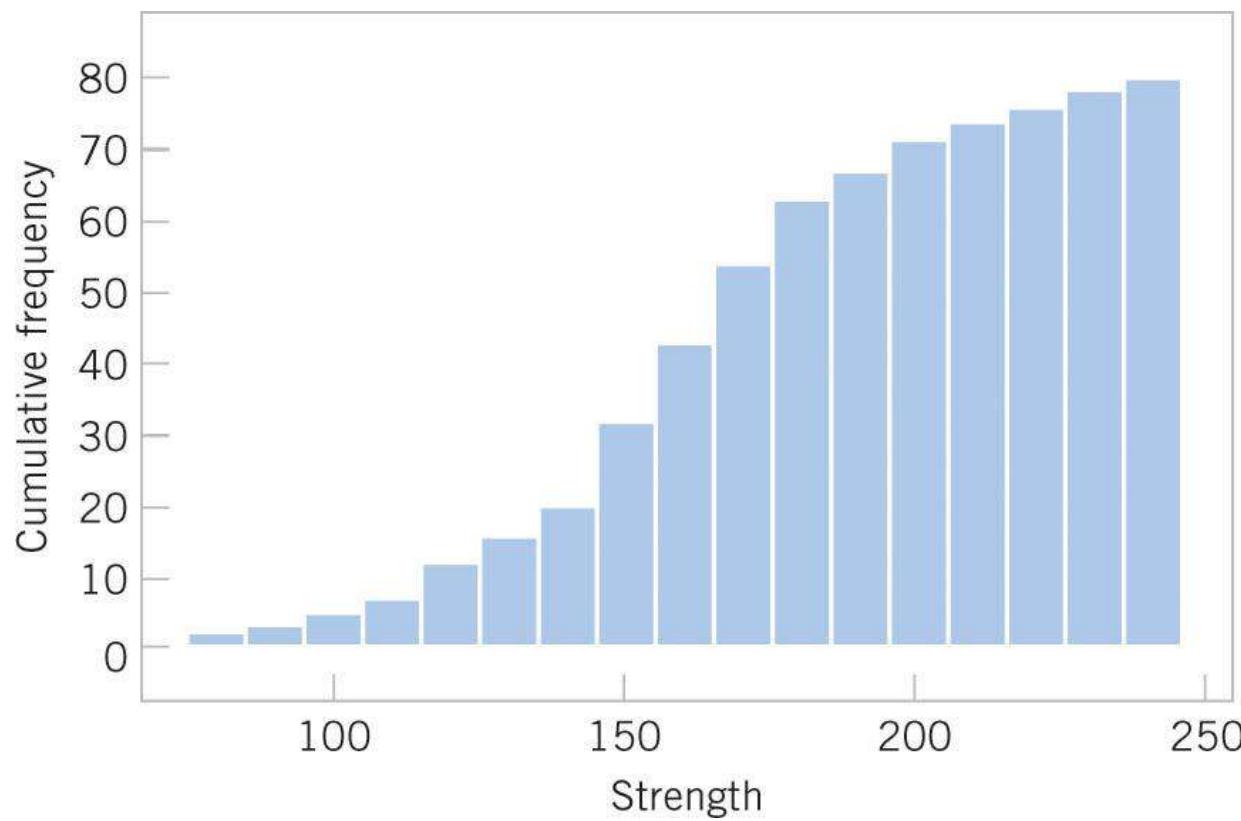


Figure 6-10 Cumulative histogram of compressive strength of 80 aluminum-lithium alloy specimens. Comment: Easy to see cumulative probabilities, hard to see distribution shape.

Shape of a Frequency Distribution

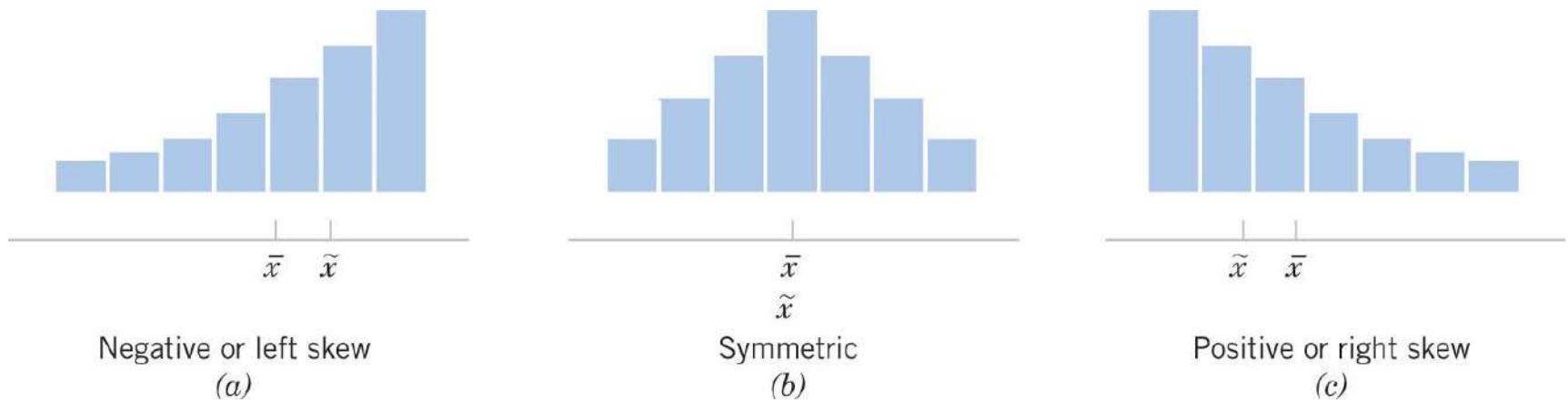


Figure 6-11 Histograms of symmetric and skewed distributions.

(b) Symmetric distribution has identical mean, median and mode measures.

(a & c) Skewed distributions are positive or negative, depending on the direction of the long tail. Their measures occur in alphabetical order as the distribution is approached from the long tail. ☺

Histograms for Categorical Data

- Categorical data is of two types:
 - Ordinal: categories have a natural order, e.g., year in college, military rank.
 - Nominal: Categories are simply different, e.g., gender, colors.
- Histogram bars are for each category, are of equal width, and have a height equal to the category's frequency or relative frequency.
- A Pareto chart is a histogram in which the categories are sequenced in decreasing order. This approach emphasizes the most and least important categories.

Example 6-6: Categorical Data Histogram

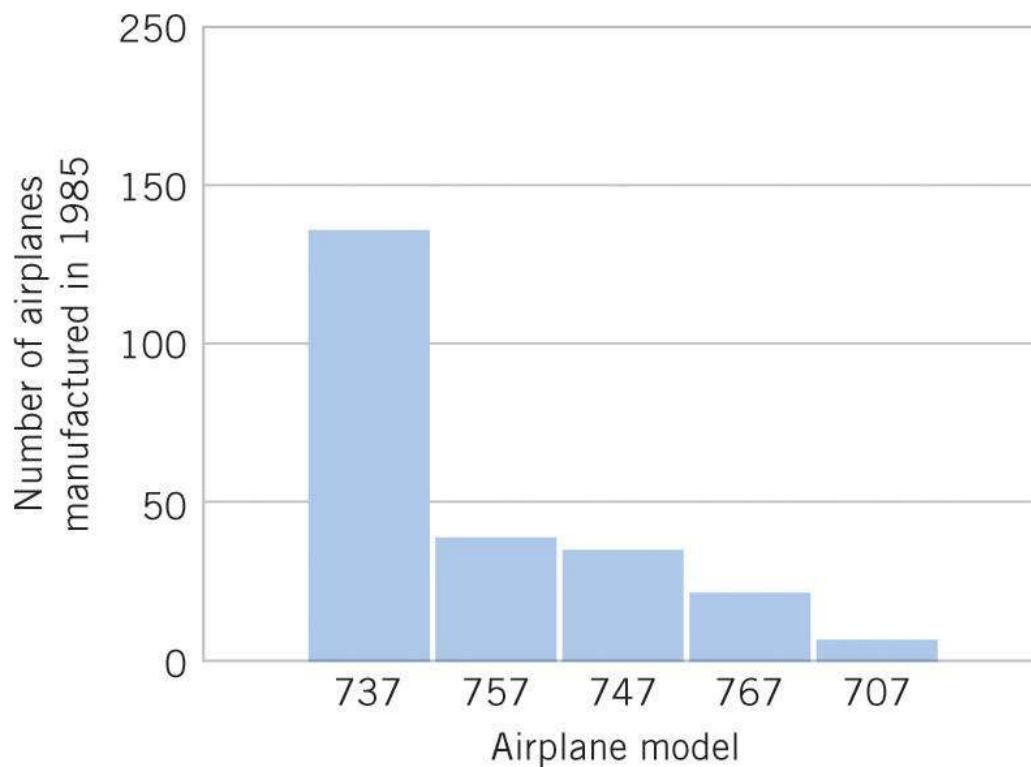


Figure 6-12 Airplane production in 1985. (Source: Boeing Company) Comment: Illustrates nominal data in spite of the numerical names, categories are shown at the bin's midpoint, a Pareto chart since the categories are in decreasing order.

Box Plot or Box-and-Whisker Chart

- A box plot is a graphical display showing **center**, **spread**, **shape**, and **outliers** (SOCS).
- It displays the **5-number summary**: \min , q_1 , **median**, q_3 , and \max .

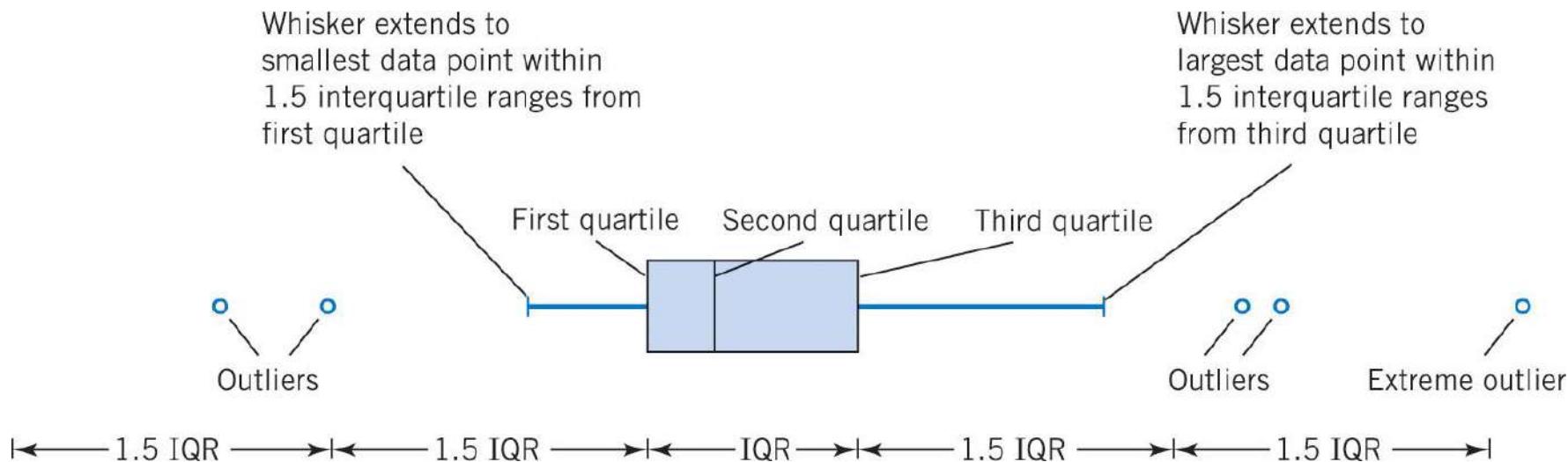


Figure 6-13 Description of a box plot.

Box Plot of Table 6-2 Data

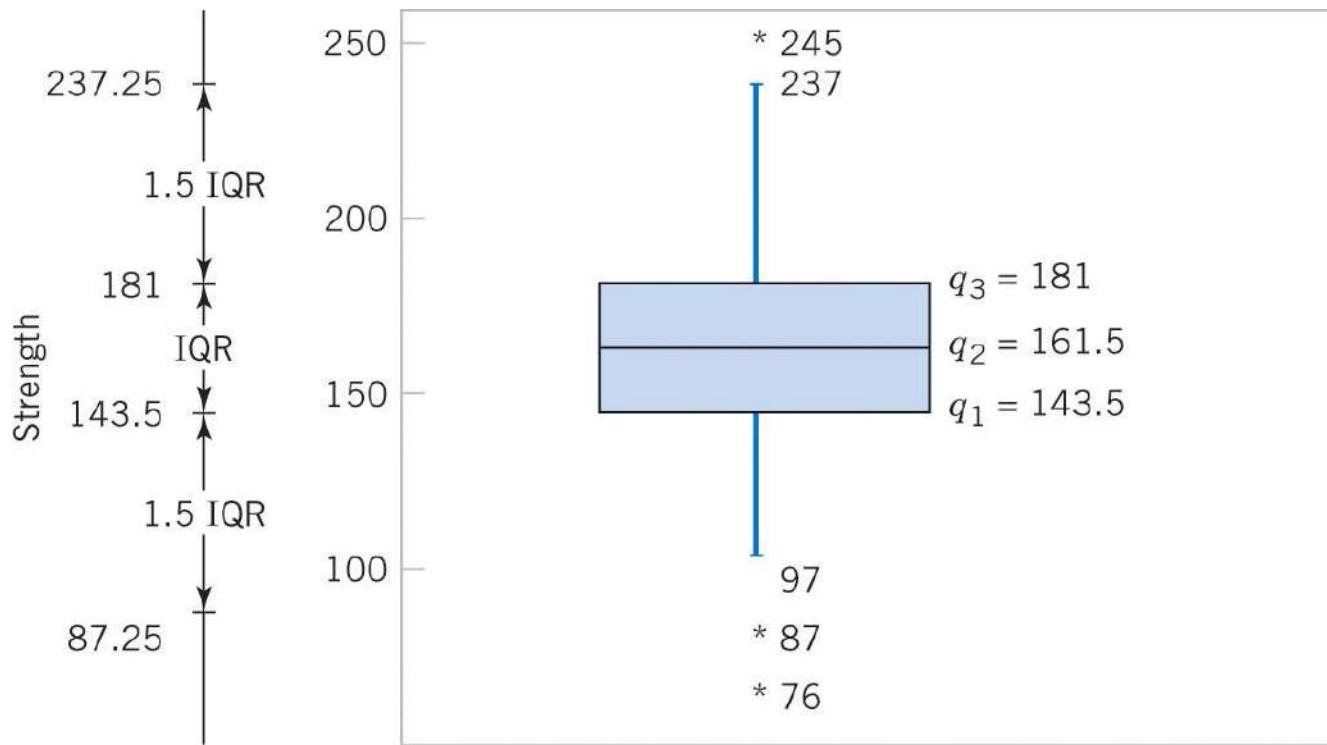


Figure 6-14 Box plot of compressive strength of 80 aluminum-lithium alloy specimens. Comment: Box plot may be shown vertically or horizontally, data reveals three outliers and no extreme outliers. Lower outlier limit is: $143.5 - 1.5 * (181.0 - 143.5) = 87.25$.

Comparative Box Plots

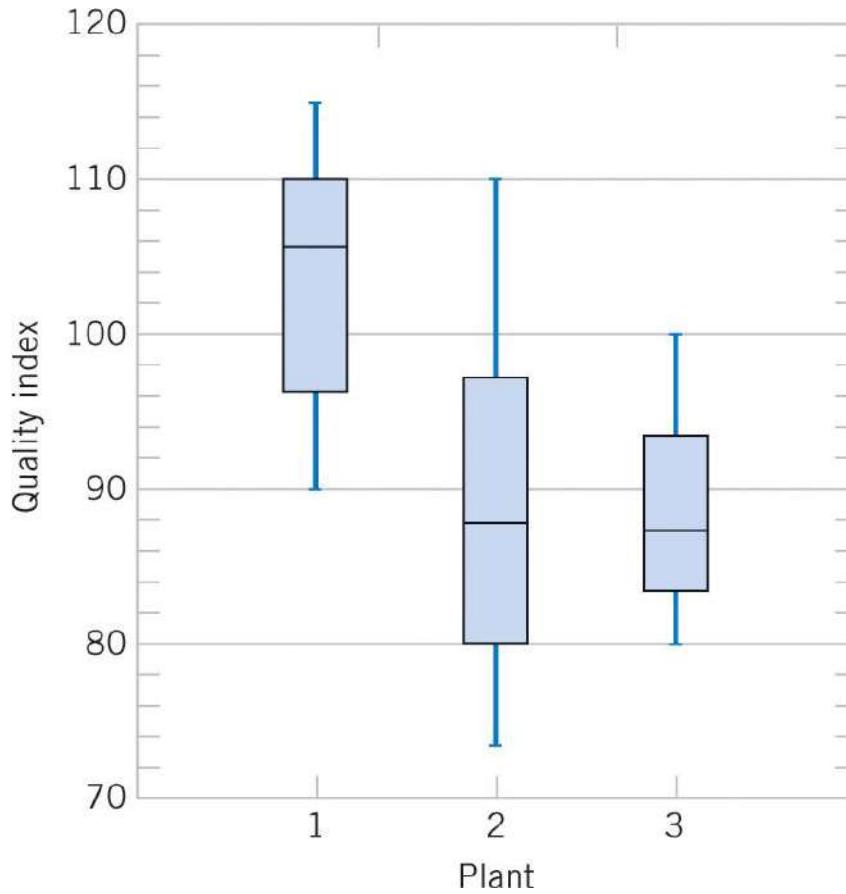
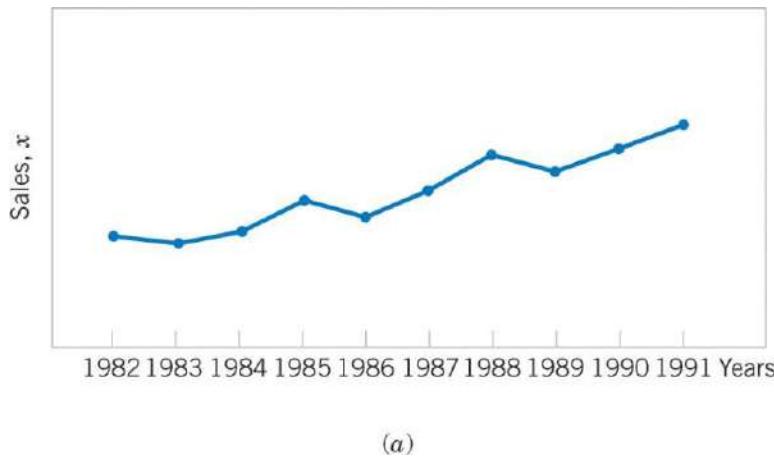


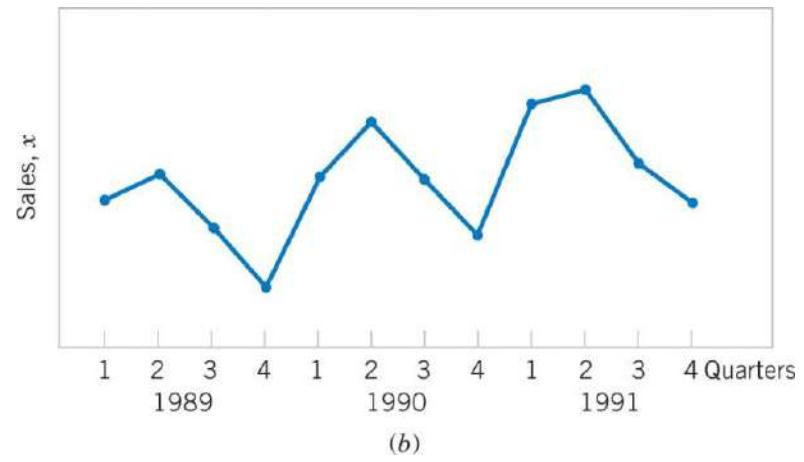
Figure 6-15 Comparative box plots of a quality index at three manufacturing plants. Comment: Plant 2 has too much variability. Plants 2 & 3 need to raise their quality index performance.

Time Sequence Plots

- A time series plot shows the data value, or statistic, on the vertical axis with time on the horizontal axis.
- A time series plot reveals trends, cycles or other time-oriented behavior that could not be otherwise seen in the data.



(a)



(b)

Figure 6-16 Company sales by year (a) & by quarter (b). The annual time interval masks cyclical quarterly variation, but shows consistent progress.

Digidot Plot of Table 6-2 Data

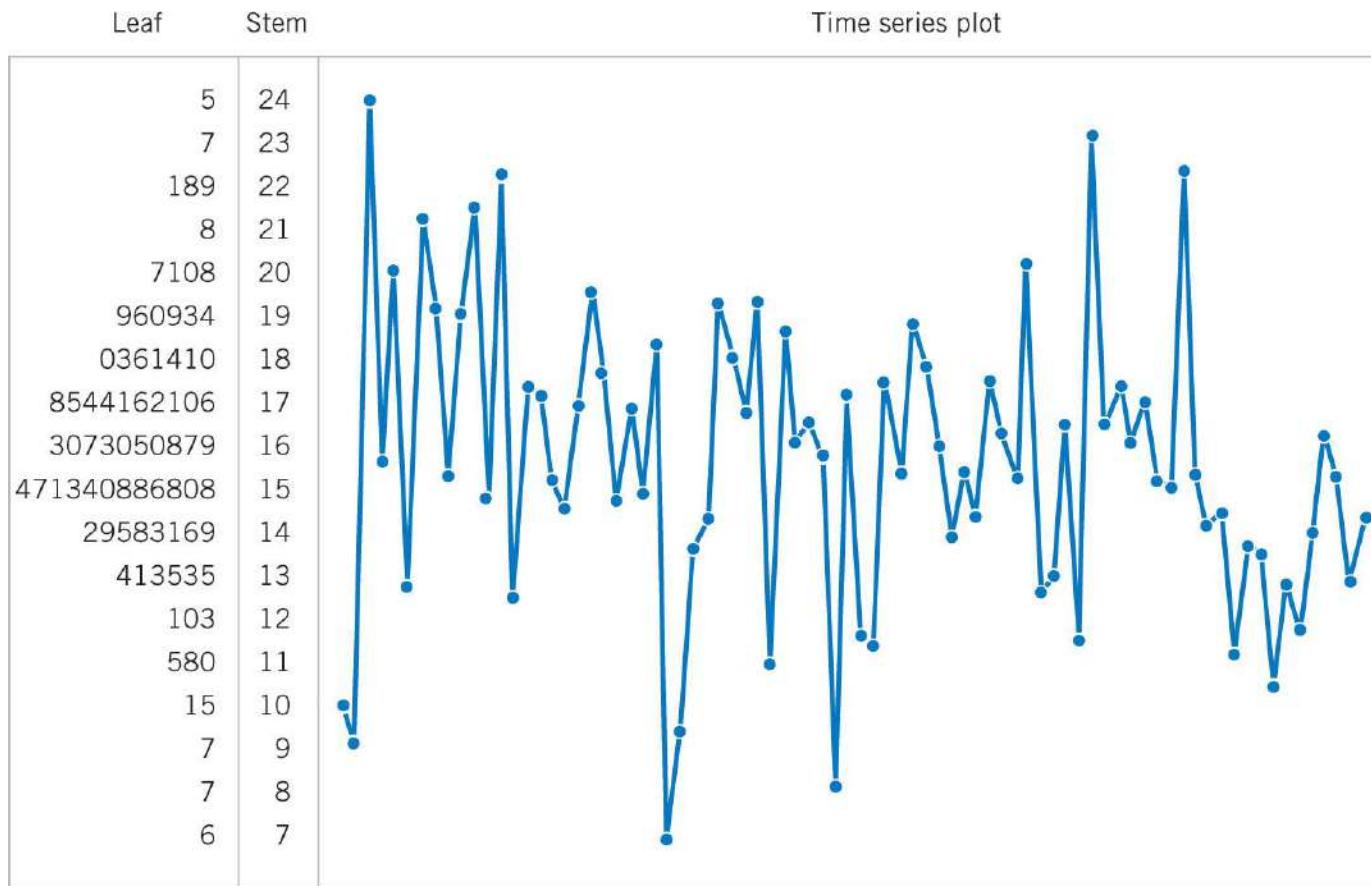
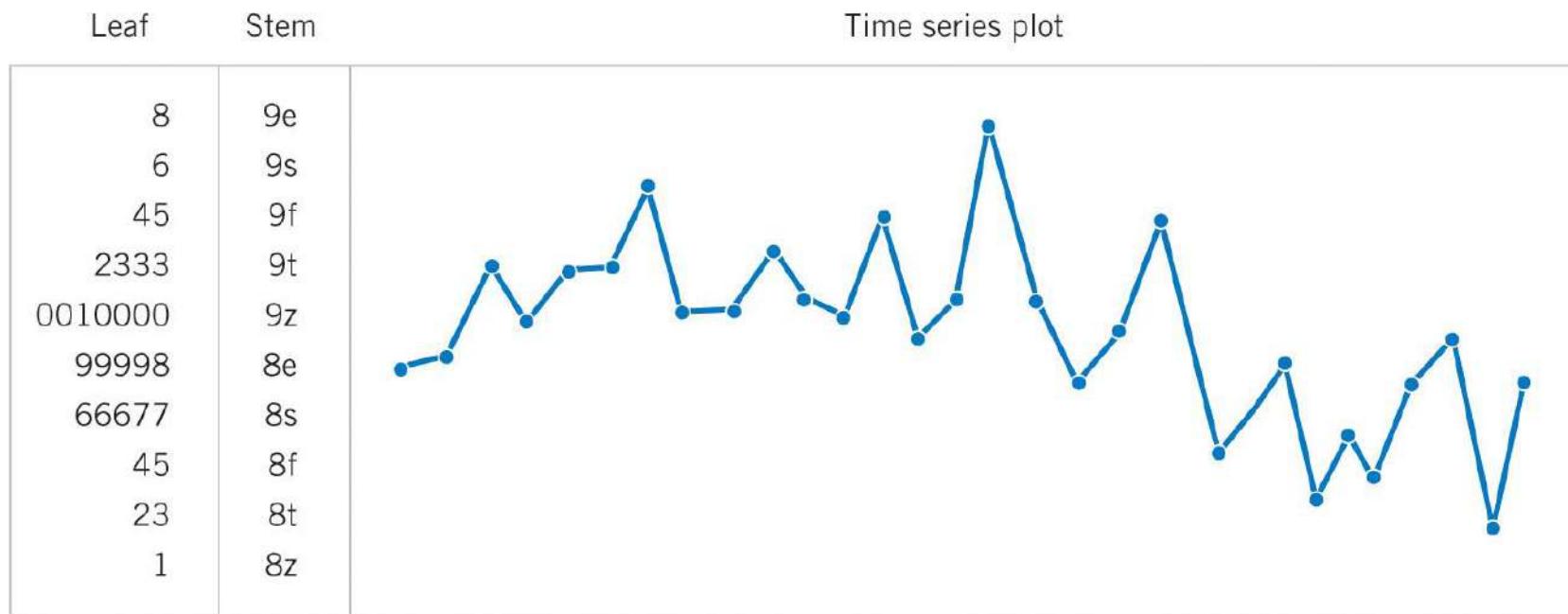


Figure 6-17 A digidot plot of the compressive strength data in Table 6-2. It combines a time series with a stem-and-leaf plot. The variability in the frequency distribution, as shown by the stem-and-leaf plot, is distorted by the apparent trend in the time series data.

Digiplot of Chemical Concentration Data



Probability Plots

- How do we know if a particular probability distribution is a reasonable model for a data set?
- We use a **probability plot** to verify such an assumption using a subjective visual examination.
- A histogram of a large data set reveals the shape of a distribution. The histogram of a small data set would not provide such a clear picture.
- A probability plot is helpful for all data set sizes.

How To Build a Probability Plot

- To construct a probability plot:
 - Sort the data observations in ascending order: $x_{(1)}$, $x_{(2)}, \dots, x_{(n)}$.
 - The observed value $x_{(j)}$ is plotted against the cumulative distribution $(j - 0.5)/n$.
 - The paired numbers are plotted on the probability paper of the proposed distribution.
 - If the paired numbers form a straight line, it is reasonable to assume that the data follows the proposed distribution.

Example 6-7: Battery Life

The effective service life (minutes) of batteries used in a laptop are given in the table. We hypothesize that battery life is adequately modeled by a normal distribution. The probability plot is shown on normal probability vertical scale.

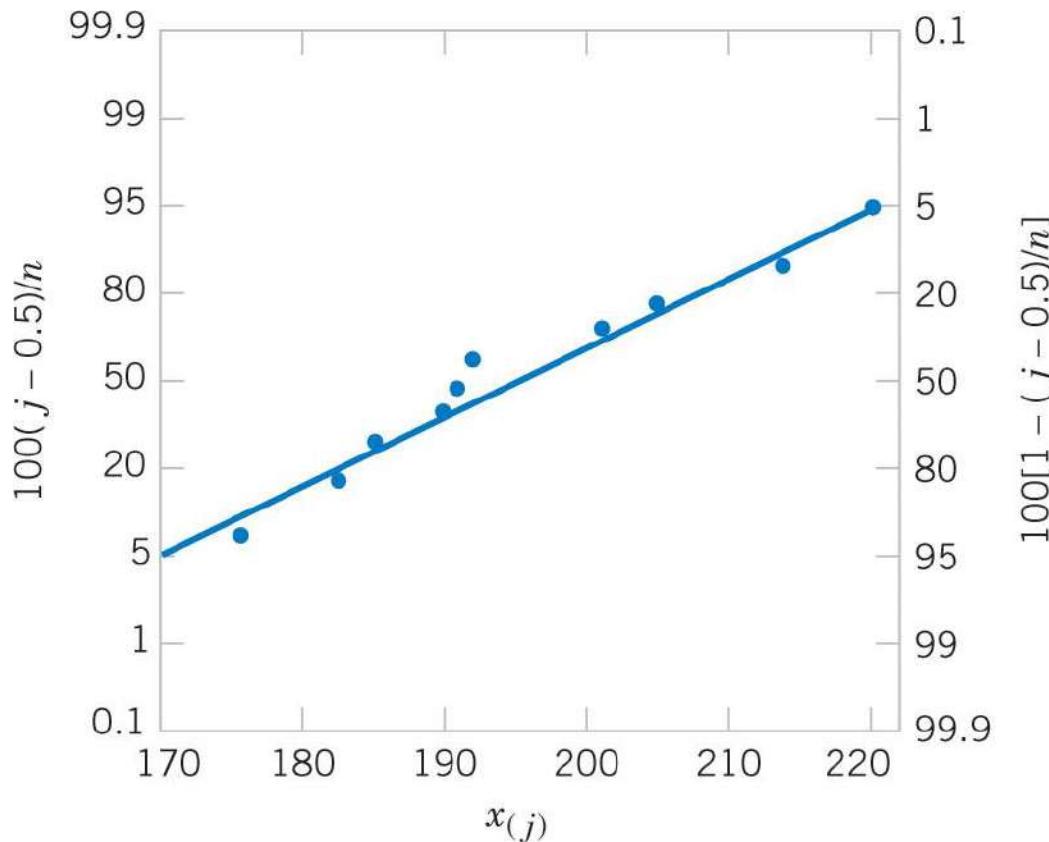


Table 6-6 Calculations
for Constructing a
Normal Probability Plot

j	$x_{(j)}$	$(j - 0.5)/10$
1	176	0.05
2	183	0.15
3	185	0.25
4	190	0.35
5	191	0.45
6	192	0.55
7	201	0.65
8	205	0.75
9	214	0.85
10	220	0.95

Figure 6-19 Normal probability plot for battery life.

Probability Plot on Ordinary Axes

A normal probability plot can be plotted on ordinary axes using z-values. The normal probability scale is not used.

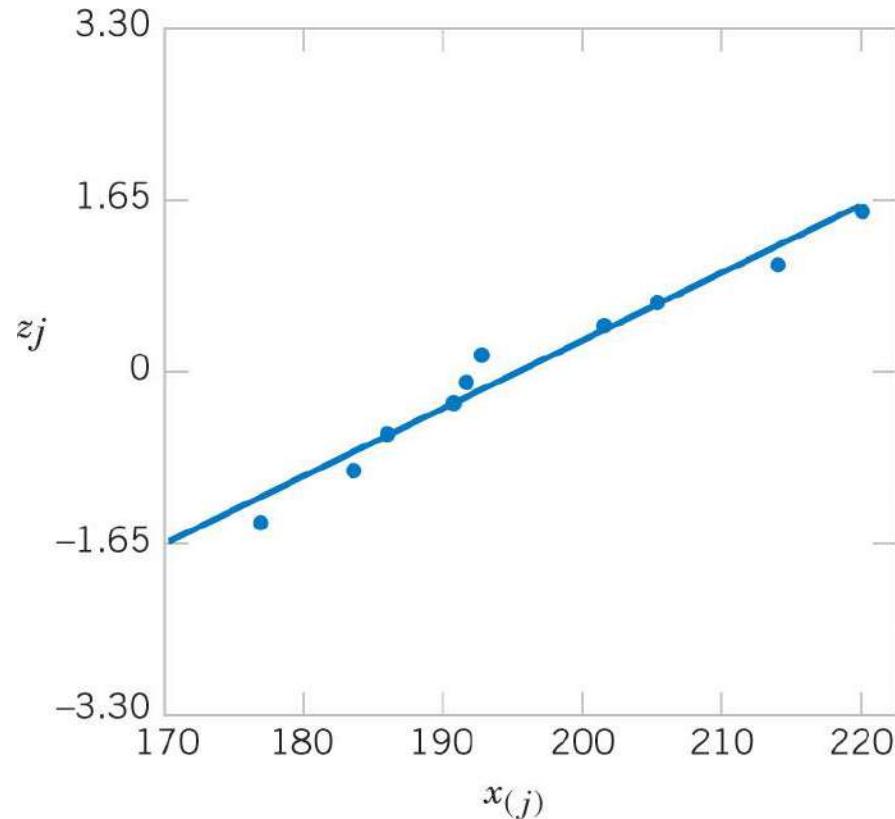


Figure 6-20 Normal Probability plot obtained from standardized normal scores. This is equivalent to Figure 6-19.

Table 6-6 Calculations for Constructing a Normal Probability Plot

j	$x_{(j)}$	$(j - 0.5)/10$	z_j
1	176	0.05	-1.64
2	183	0.15	-1.04
3	185	0.25	-0.67
4	190	0.35	-0.39
5	191	0.45	-0.13
6	192	0.55	0.13
7	201	0.65	0.39
8	205	0.75	0.67
9	214	0.85	1.04
10	220	0.95	1.64

Use of the Probability Plot

- The probability plot can identify variations from a normal distribution shape.
 - Light tails of the distribution – more peaked.
 - Heavy tails of the distribution – less peaked.
 - Skewed distributions.
- Larger samples increase the clarity of the conclusions reached.

Probability Plot Variations

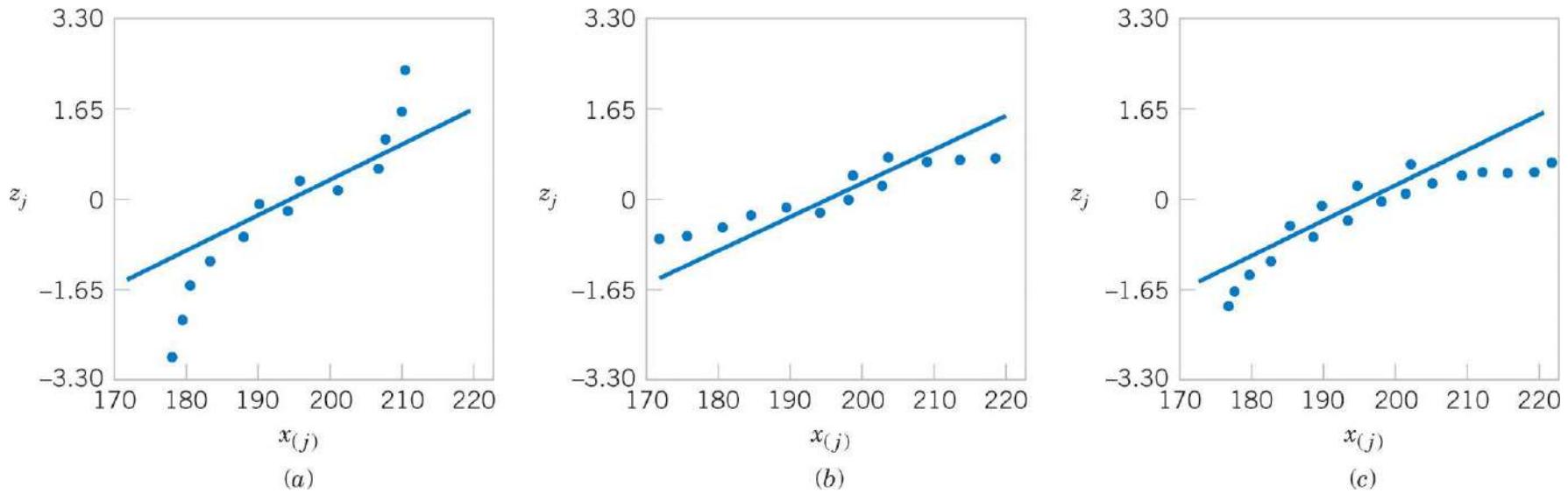
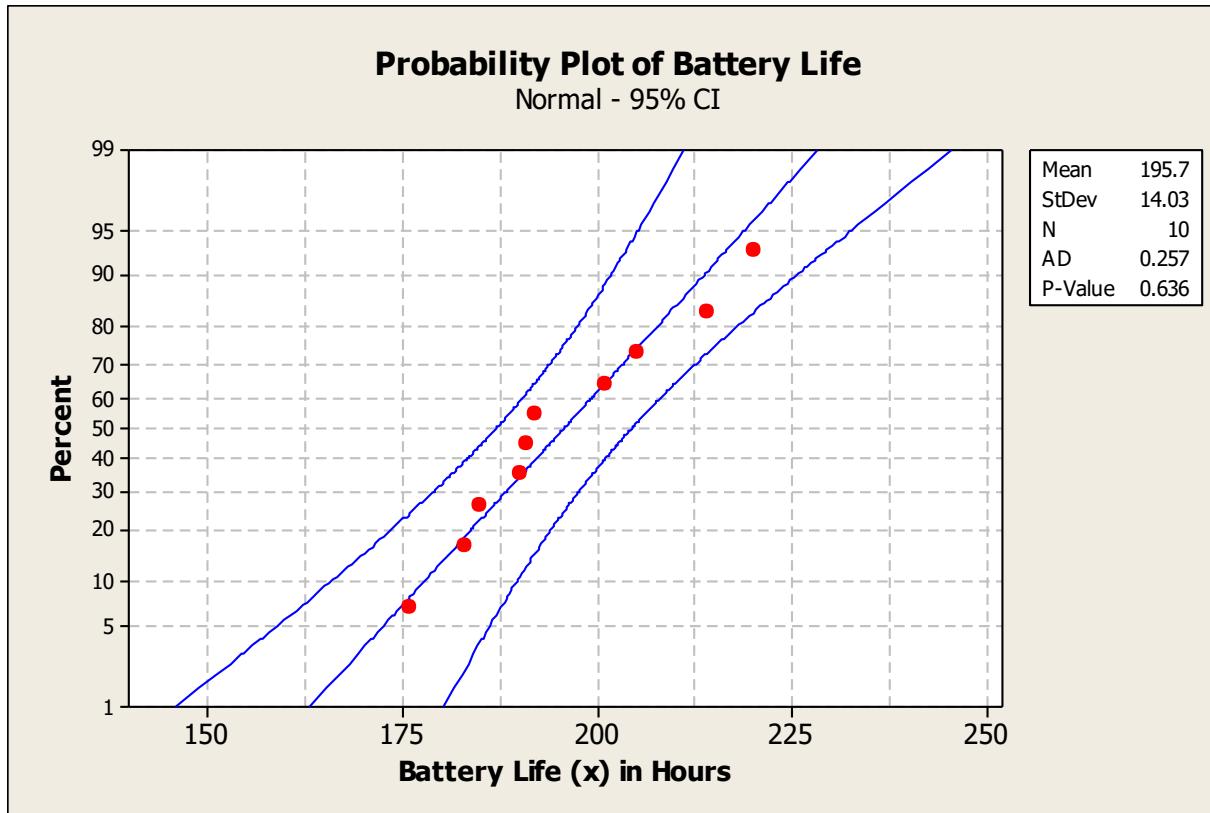


Figure 6-21 Normal probability plots indicating a non-normal distribution.

- (a) Light tailed distribution (squeezed together)
- (b) Heavy tailed distribution (stretched out)
- (c) Right skewed distribution (one end squeezed, other end stretched)

Probability Plots with Minitab

- Obtained using Minitab menu: Graphics > Probability Plot. 14 different distributions can be used.
- The curved bands provide guidance whether the proposed distribution is acceptable – all observations within the bands is good.





Search

Repository Web



[View ALL Data Sets](#)

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

Welcome to the UC Irvine Machine Learning Repository!

We currently maintain 559 data sets as a service to the machine learning community. You may [view all data sets](#) through our searchable interface. For a general overview of the Repository, please visit our [About](#) page. For information about citing data sets in publications, please read our [citation policy](#). If you wish to donate a data set, please consult our [donation policy](#). For any other questions, feel free to [contact the Repository librarians](#).

Supported By:



In Collaboration With:



Latest News:

- 09-24-2018: Welcome to the new Repository
admins Dheeru Dua and Efi Karra Taniskidou!
- 04-04-2013: Welcome to the new Repository
admins Kevin Bache and Moshe Lichman!
- 03-01-2010: Note from donor regarding Netflix data
- 10-16-2009: Two new data sets have been added.
- 09-14-2009: Several data sets have been added.
- 03-24-2008: New data sets have been added!
- 06-25-2007: Two new data sets have been added: UJI Pen Characters, MAGIC Gamma Telescope

Newest Data Sets:

- 10-03-2020: Codon usage
- 09-03-2020: Intelligent Media Accelerometer and Gyroscope (IM-AccGyro) Dataset
- 07-22-2020: Facebook Large Page-Page Network
- 07-17-2020: Amphibians
- 07-12-2020: Early stage diabetes risk prediction dataset.
- 06-

Most Popular Data Sets (hits since 2007):

- 3689053: Iris
- 2003574: Adult
- 1547963: Wine
- 1383155: Breast Cancer Wisconsin (Diagnostic)
- 1376731: Heart Disease

Assignment

- UCI: <https://archive.ics.uci.edu/ml/index.php>
- Choose 2 different Datasets, which is to be fit for Regression and Classification)
- Choose 2 Attributes of each data set
- Summary about the data
- Plot
 - Histogram
 - Stem-and-Leaf
 - Q-Q plot
 - Box plot
 - Probability Plot

Important Terms & Concepts of Chapter 6

Box plot

Standard deviation

Frequency distribution &
histogram

Variance

Median, quartiles &
percentiles

Probability plot

Multivariable data

Relative frequency
distribution

Normal probability plot

Sample:

Pareto chart

Mean

Population:

Standard deviation

Mean

Variance

Stem-and-leaf diagram

Time series plots

Foundation of Data Science and Analytics

Sampling Distribution and Point Estimation of Parameters

Arun K. Timalsina

Point Estimation

- A **point estimate** is a reasonable value of a population parameter.
- Data collected, X_1, X_2, \dots, X_n are random variables.
- Functions of these random variables, \bar{x} and s^2 , are also random variables called **statistics**.
- Statistics have their unique distributions that are called **sampling distributions**.

Point Estimator

A point estimate of some population parameter θ is a single numerical value Θ .

The statistic Θ is called the **point estimator**.

As an example, suppose the random variable X is normally distributed with an unknown mean μ . The sample mean is a point estimator of the unknown population mean μ . That is, $\mu = \bar{X}$. After the sample has been selected, the numerical value \bar{x} is the point estimate of μ .

Thus if $x_1 = 25, x_2 = 30, x_3 = 29, x_4 = 31$, the point estimate of μ is

$$\bar{x} = \frac{25 + 30 + 29 + 31}{4} = 28.75$$

Some Parameters & Their Statistics

Parameter	Measure	Statistic
μ	Mean of a single population	\bar{x}
σ^2	Variance of a single population	s^2
σ	Standard deviation of a single population	s
p	Proportion of a single population	\hat{p}
$\mu_1 - \mu_2$	Difference in means of two populations	$\bar{x}_1 - \bar{x}_2$
$p_1 - p_2$	Difference in proportions of two populations	$\hat{p}_1 - \hat{p}_2$

- There could be choices for the point estimator of a parameter.
- To estimate the mean of a population, we could choose the:
 - Sample mean.
 - Sample median.
 - Average of the largest & smallest observations of the sample.
- We need to develop criteria to compare estimates using statistical properties.

Some Definitions

- The random variables X_1, X_2, \dots, X_n are a **random sample** of size n if:
 - a) The X_i are independent random variables.
 - b) Every X_i has the same probability distribution.
- A **statistic** is any function of the observations in a random sample.
- The probability distribution of a statistic is called a **sampling distribution**.

Sampling Distribution of the Sample Mean

- A random sample of size n is taken from a normal population with mean μ and variance σ^2 .
- The observations, X_1, X_2, \dots, X_n , are normally and independently distributed.
- A linear function (X -bar) of normal and independent random variables is itself normally distributed.

$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$ has a normal distribution

with mean $\mu_{\bar{X}} = \frac{\mu + \mu + \dots + \mu}{n} = \mu$

and variance $\sigma_{\bar{X}}^2 = \frac{\sigma^2 + \sigma^2 + \dots + \sigma^2}{n^2}$

Central Limit Theorem

If X_1, X_2, \dots, X_n is a random sample of size n is taken from a population (either finite or infinite) with mean μ and finite variance σ^2 , and if \bar{X} is the sample mean, then the limiting form of the distribution of

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \quad (7-1)$$

as $n \rightarrow \infty$, is the **standard normal distribution**.

Sampling Distributions of Sample Means

Figure 7-1 Distributions of average scores from throwing dice. Mean = 3.5

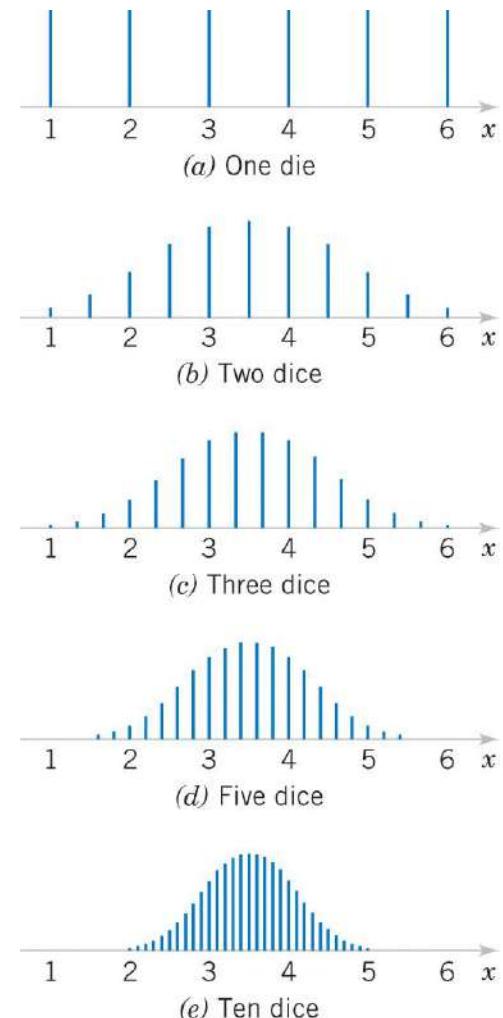
Formulas

$$\mu = \frac{b-a}{2}$$

$$\sigma_x^2 = \frac{(b-a+1)^2 - 1}{12}$$

$$\sigma_{\bar{X}}^2 = \sigma_x^2 / n$$

	n dice	var	std dev
a)	1	2.9	1.7
b)	2	1.5	1.2
c)	3	1.0	1.0
d)	5	0.6	0.8
e)	10	0.3	0.5
	a =	1	
	b =	6	



Example 7-1: Resistors

An electronics company manufactures resistors having a mean resistance of 100 ohms and a standard deviation of 10 ohms. The distribution of resistance is normal. What is the probability that a random sample of $n = 25$ resistors will have an average resistance of less than 95 ohms?

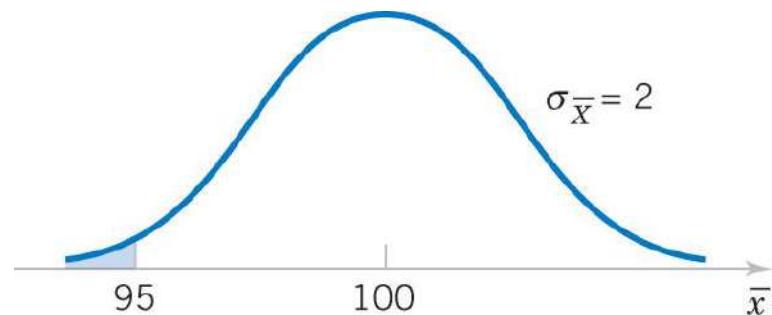


Figure 7-2 Desired probability is shaded

Answer:

$$\sigma_{\bar{X}} = \frac{\sigma_x}{\sqrt{n}} = \frac{10}{\sqrt{25}} = 2.0$$

$$\begin{aligned}\Phi\left(\frac{\bar{X} - \mu}{\sigma_{\bar{X}}}\right) &= \Phi\left(\frac{95 - 200}{2}\right) \\ &= \Phi(-2.5) = 0.0062\end{aligned}$$

0.0062 = NORMSDIST(-2.5)

A rare event at less than 1%.

Example 7-2: Central Limit Theorem

Suppose that a random variable X has a continuous uniform distribution:

$$f(x) = \begin{cases} 1/2, & 4 \leq x \leq 6 \\ 0, & \text{otherwise} \end{cases}$$

Find the distribution of the sample mean of a random sample of size $n = 40$.

Distribution is normal by the CLT.

$$\mu = \frac{b+a}{2} = \frac{6+4}{2} = 5.0$$

$$\sigma^2 = \frac{(b-a)^2}{12} = \frac{(6-4)^2}{12} = 1/3$$

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} = \frac{1/3}{40} = \frac{1}{120}$$

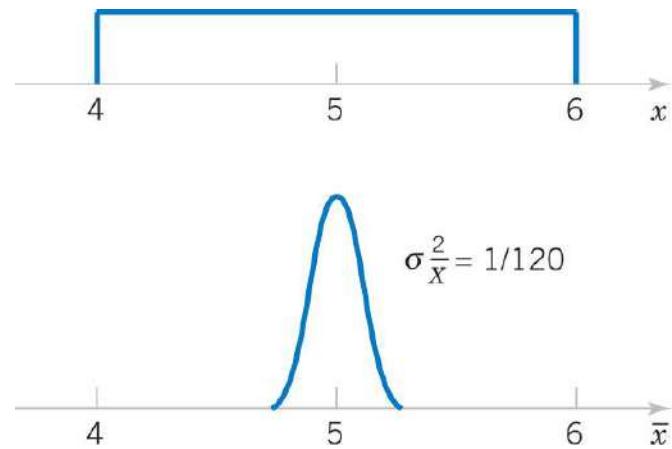


Figure 7-3 Distributions of X and \bar{X}

Two Populations

We have two independent normal populations. What is the distribution of the difference of the sample means?

The sampling distribution of $\bar{X}_1 - \bar{X}_2$ is:

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} = \mu_1 - \mu_2$$

$$\sigma_{\bar{X}_1 - \bar{X}_2}^2 = \sigma_{\bar{X}_1}^2 - \sigma_{\bar{X}_2}^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

The distribution of $\bar{X}_1 - \bar{X}_2$ is normal if:

- (1) n_1 and n_2 are both greater than 30,
regardless of the distributions of X_1 and X_2 .
- (2) n_1 and n_2 are less than 30,
while the distributions are somewhat normal.

Sampling Distribution of a Difference in Sample Means

- If we have two independent populations with means μ_1 and μ_2 , and variances σ_1^2 and σ_2^2 ,
- And if $X\bar{1}$ and $X\bar{2}$ are the sample means of two independent random samples of sizes n_1 and n_2 from these populations:
- Then the sampling distribution of:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (7-4)$$

is approximately standard normal, if the conditions of the central limit theorem apply.

- If the two populations are normal, then the sampling distribution is exactly standard normal.

Example 7-3: Aircraft Engine Life

The effective life of a component used in jet-turbine aircraft engines is a normal-distributed random variable with parameters shown (old). The engine manufacturer introduces an improvement into the manufacturing process for this component that changes the parameters as shown (new).

Random samples are selected from the “old” process and “new” process as shown.

What is the probability the difference in the two sample means is at least 25 hours?

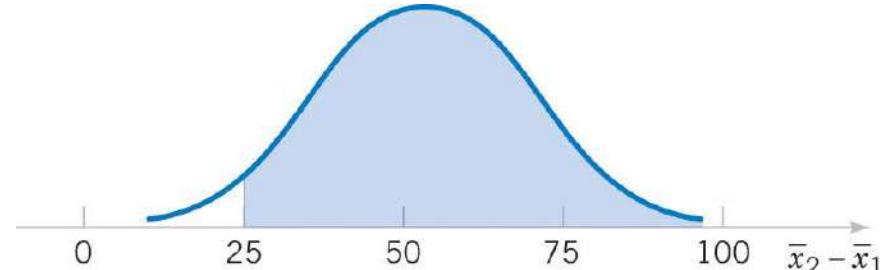


Figure 7-4 Sampling distribution of the sample mean difference.

	Process		
	Old (1)	New (2)	Diff (2-1)
\bar{x} -bar =	5,000	5,050	50
s =	40	30	50
n =	16	25	
Calculations			
$s / \sqrt{n} =$	10	6	11.7
		$z =$	-2.14
$P(\bar{x}_{\text{bar}}_2 - \bar{x}_{\text{bar}}_1 > 25) = P(Z > z) =$	0.9840		
	$= 1 - \text{NORMSDIST}(z)$		

General Concepts of Point Estimation

- We want point estimators that are:
 - Are **unbiased**.
 - Have a **minimal variance**.
- We use the **standard error of the estimator** to calculate its **mean square error**.

Unbiased Estimators Defined

The point estimator Θ is an **unbiased estimator** for the parameter θ if:

$$E(\Theta) = \theta \quad (7-5)$$

If the estimator is not unbiased, then the difference:

$$E(\Theta) - \theta \quad (7-6)$$

is called the **bias** of the estimator Θ .

The mean of the sampling distribution of Θ
is equal to θ .

Example 7-4: Sample Mean & Variance Are Unbiased-1

- X is a random variable with mean μ and variance σ^2 . Let X_1, X_2, \dots, X_n be a random sample of size n .
- Show that the sample mean (\bar{X}) is an unbiased estimator of μ .

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) \\ &= \frac{1}{n}[E(X_1) + E(X_2) + \dots + E(X_n)] \\ &= \frac{1}{n}[\mu + \mu + \dots + \mu] = \frac{n\mu}{n} = \mu \end{aligned}$$

Example 7-4: Sample Mean & Variance Are Unbiased-2

Show that the sample variance (S^2) is a unbiased estimator of σ^2 .

$$\begin{aligned} E(S^2) &= E\left(\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}\right) = \frac{1}{n-1} E\left[\sum_{i=1}^n (X_i^2 + \bar{X}^2 - 2\bar{X}X_i)\right] \\ &= \frac{1}{n-1} \left[E\left(\sum_{i=1}^n X_i^2 - n\bar{X}^2\right) \right] = \frac{1}{n-1} \left[\sum_{i=1}^n E(X_i^2) - nE(\bar{X}^2) \right] \\ &= \frac{1}{n-1} \left[\sum_{i=1}^n (\mu^2 + \sigma^2) - n(\mu^2 + \sigma^2/n) \right] \\ &= \frac{1}{n-1} \left[n\mu^2 + n\sigma^2 - n\mu^2 - \sigma^2 \right] = \frac{1}{n-1} [(n-1)\sigma^2] = \sigma^2 \end{aligned}$$

Other Unbiased Estimators of the Population Mean

$$\text{Mean} = \bar{X} = \frac{110.4}{10} = 11.04$$

$$\text{Median} = X = \frac{10.3 + 11.6}{2} = 10.95$$

$$\text{Trimmed mean} = \frac{110.4 - 8.5 - 14.1}{8} = 10.81$$

- All three statistics are unbiased.
 - Do you see why?
- Which is best?
 - We want the most reliable one.

i	x_i	x_i'
1	12.8	8.5
2	9.4	8.7
3	8.7	9.4
4	11.6	9.8
5	13.1	10.3
6	9.8	11.6
7	14.1	12.1
8	8.5	12.8
9	12.1	13.1
10	10.3	14.1
Σ	110.4	

Choosing Among Unbiased Estimators

Suppose that Θ_1 and Θ_2 are unbiased estimators of θ .

The variance of Θ_1 is less than the variance of Θ_2 .

$\therefore \Theta_1$ is preferable.

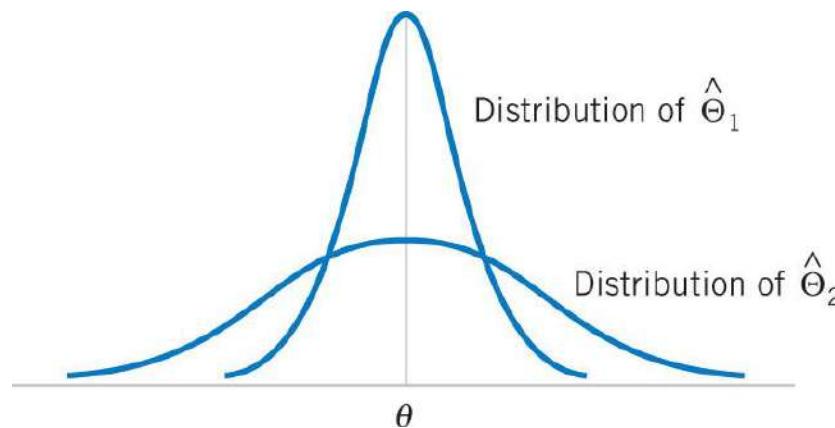


Figure 7-5 The sampling distributions of two unbiased estimators.

Minimum Variance Unbiased Estimators

- If we consider all unbiased estimators of θ , the one with the smallest variance is called the **minimum variance unbiased estimator (MVUE)**.
- If X_1, X_2, \dots, X_n is a random sample of size n from a normal distribution with mean μ and variance σ^2 , then the sample $X\text{-bar}$ is the MVUE for μ .
- The sample mean and a single observation are unbiased estimators of μ . The variance of the:
 - Sample mean is σ^2/n
 - Single observation is σ^2
 - Since $\sigma^2/n \leq \sigma^2$, the sample mean is preferred.

Standard Error of an Estimator

The **standard error** of an estimator Θ is its standard deviation, given by

$$\sigma_{\Theta} = \sqrt{V(\Theta)}.$$

If the standard error involves unknown parameters that can be estimated, substitution of these values into σ_{Θ}

produces an **estimated standard error**, denoted by $\hat{\sigma}_{\Theta}$

Equivalent notation: $\hat{\sigma}_{\Theta} = se(\hat{\Theta})$

If the X_i are $\sim N(\mu, \sigma^2)$, then \bar{X} is normally distributed,

and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$. If σ is not known, then $\hat{\sigma}_{\bar{X}} = \frac{s}{\sqrt{n}}$.

Example 7-5: Thermal Conductivity

- These observations are 10 measurements of thermal conductivity of Armco iron.
- Since σ is not known, we use s to calculate the standard error.
- Since the standard error is 0.2% of the mean, the mean estimate is fairly precise. We can be very confident that the true population mean is $41.924 \pm 2(0.0898)$.

x_i	
41.60	
41.48	
42.34	
41.95	
41.86	
42.18	
41.72	
42.26	
41.81	
42.04	
41.924	= Mean
0.284	= Std dev (s)
0.0898	= Std error

Mean Squared Error

The mean squared error of an estimator Θ of the parameter θ is defined as:

$$\text{MSE}(\Theta) = E(\Theta - \theta)^2 \quad (7-7)$$

Can be rewritten as

$$\begin{aligned} &= E[\Theta - E(\Theta)]^2 + [\theta - E(\Theta)]^2 \\ &= V(\Theta) + (\text{bias})^2 \end{aligned}$$

Conclusion: The mean squared error (MSE) of the estimator is equal to the variance of the estimator plus the bias squared. It measures both characteristics.

Relative Efficiency

- The MSE is an important criterion for comparing two estimators.

$$\text{Relative efficiency} = \frac{\text{MSE}(\Theta_1)}{\text{MSE}(\Theta_2)}$$

- If the relative efficiency is less than 1, we conclude that the 1st estimator is superior to the 2nd estimator.

Optimal Estimator

- A biased estimator can be preferred to an unbiased estimator if it has a smaller MSE.
- Biased estimators are occasionally used in linear regression.
- An estimator whose MSE is smaller than that of any other estimator is called an **optimal estimator**.

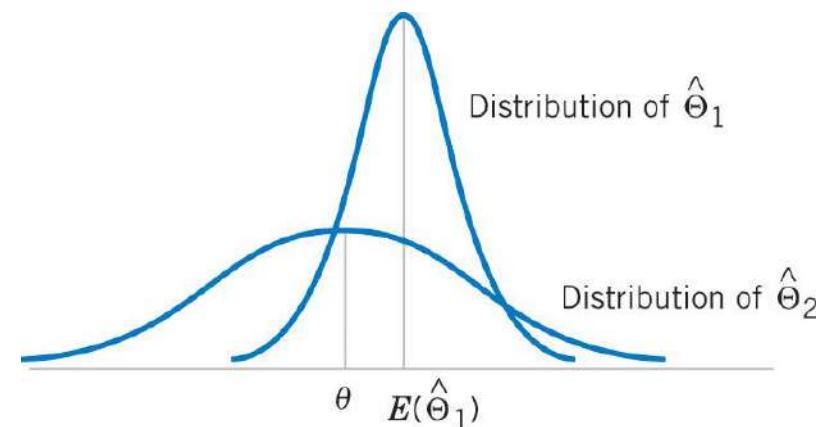


Figure 7-6 A biased estimator has a smaller variance than the unbiased estimator.

Methods of Point Estimation

- There are three methodologies to create point estimates of a population parameter.
 - Method of moments
 - Method of maximum likelihood
 - Bayesian estimation of parameters
- Each approach can be used to create estimators with varying degrees of biasedness and relative MSE efficiencies.

Method of Moments

- A “**moment**” is a kind of an expected value of a random variable.
- A **population moment** relates to the entire population or its representative function.
- A **sample moment** is calculated like its associated population moments.

Moments Defined

- Let X_1, X_2, \dots, X_n be a random sample from the probability $f(x)$, where $f(x)$ can be either a:
 - Discrete probability mass function, or
 - Continuous probability density function
- The k^{th} population moment (or distribution moment) is $E(X^k)$, $k = 1, 2, \dots$.
- The k^{th} sample moment is $(1/n)\sum X^k$, $k = 1, 2, \dots$.
- If $k = 1$ (called the first moment), then:
 - Population moment is μ .
 - Sample moment is $x\text{-bar}$.
- The sample mean is the moment estimator of the population mean.

Moment Estimators

Let X_1, X_2, \dots, X_n be a random sample from either a probability mass function or a probability density function with m unknown parameters $\theta_1, \theta_2, \dots, \theta_m$.

The **moment estimators** $\Theta_1, \Theta_2, \dots, \Theta_m$ are found by equating the first m population moments to the first m sample moments and solving the resulting simultaneous equations for the unknown parameters.

Example 7-6: Exponential Moment Estimator-1

- Suppose that X_1, X_2, \dots, X_n is a random sample from an exponential distribution with parameter λ .
- There is only one parameter to estimate, so equating population and sample first moments, we have $E(X) = \bar{X}$.
- $E(X) = 1/\lambda = \bar{X}$
- $\lambda = 1/\bar{X}$ is the moment estimator.

Example 7-6: Exponential Moment Estimator-2

- As an example, the time to failure of an electronic module is exponentially distributed.
- Eight units are randomly selected and tested. Their times to failure are shown.
- The moment estimate of the λ parameter is 0.04620.

x_i	
11.96	
5.03	
67.40	
16.07	
31.50	
7.73	
11.10	
22.38	
21.646	= Mean
0.04620	= λ est

Example 7-7: Normal Moment Estimators

Suppose that X_1, X_2, \dots, X_n is a random sample from a normal distribution with parameter μ and σ^2 . So $E(X) = \mu$ and $E(X^2) = \mu^2 + \sigma^2$.

$$\mu = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{and} \quad \mu^2 + \sigma^2 = \frac{1}{n} \sum_{i=1}^n X_i^2$$

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2 = \frac{\sum_{i=1}^n X_i^2 - n \left(\frac{1}{n} \sum_{i=1}^n X_i \right)^2}{n}$$

$$= \frac{1}{n} \left[\sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i \right)^2}{n} \right] = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n} \quad (\text{biased})$$

Example 7-8: Gamma Moment Estimators-1

Parameters = Statistics

$$\frac{r}{\lambda} = E(X) = \bar{X} \text{ is the mean}$$

$$\frac{r}{\lambda^2} = E(X^2) - E(X)^2 \text{ is the variance or}$$

$$\frac{r(r+1)}{\lambda^2} = E(X^2) \text{ and now solving for } r \text{ and } \lambda :$$

$$\hat{r} = \frac{\bar{X}^2}{(1/n) \sum_{i=1}^n X_i^2 - \bar{X}^2}$$

$$\lambda = \frac{\bar{X}}{(1/n) \sum_{i=1}^n X_i^2 - \bar{X}^2}$$

Example 7-8: Gamma Moment Estimators-2

Using the exponential example data shown, we can estimate the parameters of the gamma distribution.

\bar{x} =	21.646
$\sum X^2$ =	6645.4247

x_i	x_i^2
11.96	143.0416
5.03	25.3009
67.40	4542.7600
16.07	258.2449
31.50	992.2500
7.73	59.7529
11.10	123.2100
22.38	500.8644

$$\hat{r} = \frac{\bar{X}^2}{(1/n) \sum_{i=1}^n X_i^2 - \bar{X}^2} = \frac{21.646^2}{(1/8)6645.4247 - 21.646^2} = 1.29$$

$$\lambda = \frac{\bar{X}}{(1/n) \sum_{i=1}^n X_i^2 - \bar{X}^2} = \frac{21.646}{(1/8)6645.4247 - 21.646^2} = 0.0598$$

Maximum Likelihood Estimators

- Suppose that X is a random variable with probability distribution $f(x;\theta)$, where θ is a single unknown parameter. Let x_1, x_2, \dots, x_n be the observed values in a random sample of size n . Then the **likelihood function** of the sample is:

$$L(\theta) = f(x_1; \theta) \cdot f(x_2; \theta) \cdot \dots \cdot f(x_n; \theta) \quad (7-9)$$

- Note that the likelihood function is now a function of only the unknown parameter θ . The **maximum likelihood estimator** (MLE) of θ is the value of θ that maximizes the likelihood function $L(\theta)$.
- If X is a discrete random variable, then $L(\theta)$ is the probability of obtaining those sample values. The MLE is the θ that maximizes that probability.

Example 7-9: Bernoulli MLE

Let X be a Bernoulli random variable. The probability mass function is $f(x;p) = p^x(1-p)^{1-x}$, $x = 0, 1$ where P is the parameter to be estimated. The likelihood function of a random sample of size n is:

$$L(p) = p^{x_1} (1-p)^{1-x_1} \cdot p^{x_2} (1-p)^{1-x_2} \cdots \cdot p^{x_n} (1-p)^{1-x_n}$$

$$= \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i} (1-p)^{n - \sum_{i=1}^n x_i}$$

$$\ln L(p) = \left(\sum_{i=1}^n x_i \right) \ln p + \left(n - \sum_{i=1}^n x_i \right) \ln (1-p)$$

$$\frac{d \ln L(p)}{dp} = \frac{\sum_{i=1}^n x_i}{p} - \frac{n - \sum_{i=1}^n x_i}{(1-p)} = 0$$

$$p = \frac{\sum_{i=1}^n x_i}{n}$$

Example 7-10: Normal MLE for μ

Let X be a normal random variable with unknown mean μ and known variance σ^2 . The likelihood function of a random sample of size n is:

$$L(\mu) = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i - \mu)^2/(2\sigma^2)}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{\frac{-1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2}$$

$$\ln L(\mu) = \frac{-n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$

$$\frac{d \ln L(\mu)}{d \mu} = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0$$

$$\mu = \frac{\sum_{i=1}^n x_i}{n} = \bar{X} \text{ (same as moment estimator)}$$

Example 7-11: Exponential MLE

Let X be a exponential random variable with parameter λ . The likelihood function of a random sample of size n is:

$$L(\lambda) = \prod_{i=1}^n \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^n x_i}$$

$$\ln L(\lambda) = n \ln(\lambda) - \lambda \sum_{i=1}^n x_i$$

$$\frac{d \ln L(\lambda)}{d \lambda} = \frac{n}{\lambda} - \sum_{i=1}^n x_i = 0$$

$$\lambda = n \sqrt[n]{\sum_{i=1}^n x_i} = 1/\bar{X} \quad (\text{same as moment estimator})$$

Why Does MLE Work?

- From Examples 7-6 & 11 using the 8 data observations, the plot of the $\ln L(\lambda)$ function maximizes at $\lambda = 0.0462$. The curve is flat near max indicating estimator not precise.
- As the sample size increases, while maintaining the same $x\bar{x}$, the curve maximums are the same, but sharper and more precise.
- Large samples are better ☺

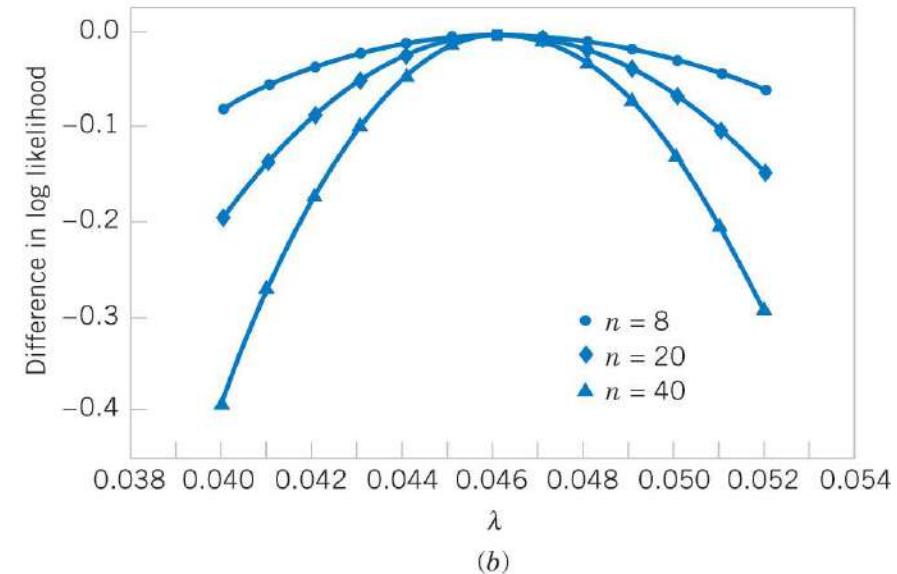
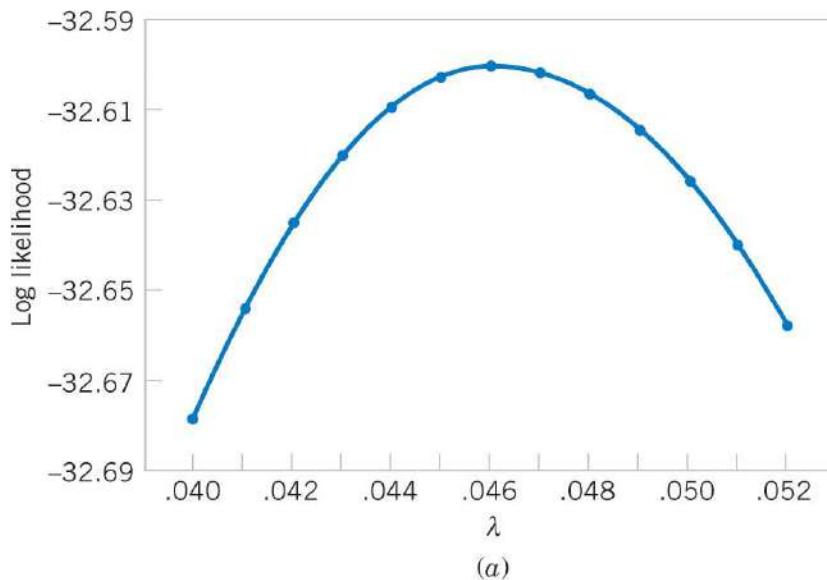


Figure 7-7 Log likelihood for exponential distribution. (a) $n = 8$, (b) $n = 8, 20, 40$.

Example 7-12: Normal MLEs for μ & σ^2

Let X be a normal random variable with both unknown mean μ and variance σ^2 . The likelihood function of a random sample of size n is:

$$L(\mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sigma \sqrt{2\pi}} e^{-(x_i - \mu)^2 / (2\sigma^2)}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{\frac{-1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2}$$

$$\ln L(\mu, \sigma^2) = \frac{-n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$

$$\frac{\partial \ln L(\mu, \sigma^2)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0$$

$$\frac{\partial \ln L(\mu, \sigma^2)}{\partial (\sigma^2)} = \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2 = 0$$

$$\mu = \bar{X} \quad \text{and} \quad \sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n}$$

Properties of an MLE

Under very general and non-restrictive conditions,
when the sample size n is large and if Θ is the MLE of the parameter ,

- (1) Θ is an approximately unbiased estimator for θ , i.e., $[E(\Theta) = \theta]$
- (2) The variance of Θ is nearly as small as the variance
that could be obtained with any other estimator, and
- (3) Θ has an approximate normal distribution.

Notes:

- Mathematical statisticians will often prefer MLEs because of these properties. Properties (1) and (2) state that MLEs are MVUEs.
- To use MLEs, the distribution of the population must be known or assumed.

Importance of Large Sample Sizes

- Consider the MLE for σ^2 shown in Example 7-12:

$$E(\hat{\sigma}^2) = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n} = \frac{n-1}{n} \sigma^2$$

Then the bias is:

$$E(\hat{\sigma}^2) - \sigma^2 = \frac{n-1}{n} \sigma^2 - \sigma^2 = \frac{-\sigma^2}{n}$$

- Since the bias is negative, the MLE underestimates the true variance σ^2 .
- The MLE is an asymptotically (large sample) unbiased estimator. The bias approaches zero as n increases.

Invariance Property

Let $\Theta_1, \Theta_2, \dots, \Theta_k$ be the maximum likelihood estimators (MLEs) of the parameters $\theta_1, \theta_2, \dots, \theta_k$.

Then the MLEs for any function $h(\theta_1, \theta_2, \dots, \theta_k)$ of these parameters is the same function $h(\Theta_1, \Theta_2, \dots, \Theta_k)$ of the estimators $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$.

This property is illustrated in Example 7-13.

Example 7-13: Invariance

For the normal distribution, the MLEs were:

$$\mu = \bar{X} \quad \text{and} \quad \sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n}$$

To obtain the MLE of the function $h(\mu, \sigma^2) = \sqrt{\sigma^2} = \sigma$, substitute the estimators μ and σ^2 into the function h :

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n}}$$

which is **not** the sample standard deviation s .

Complications of the MLE Method

The method of maximum likelihood is an excellent technique, however there are two complications:

1. It may not be easy to maximize the likelihood function because the derivative function set to zero may be difficult to solve algebraically.
2. The likelihood function may be impossible to solve, so numerical methods must be used.

The following two examples illustrate.

Example 7-14: Uniform Distribution MLE

Let X be uniformly distributed on the interval 0 to a .

$$f(x) = 1/a \text{ for } 0 \leq x \leq a$$

$$L(a) = \prod_{i=1}^n \frac{1}{a} = \frac{1}{a^n} = a^{-n} \text{ for } 0 \leq x_i \leq a$$

$$\frac{dL(a)}{da} = \frac{-n}{a^{n+1}} = -na^{-(n+1)}$$

$$a = \max(x_i)$$

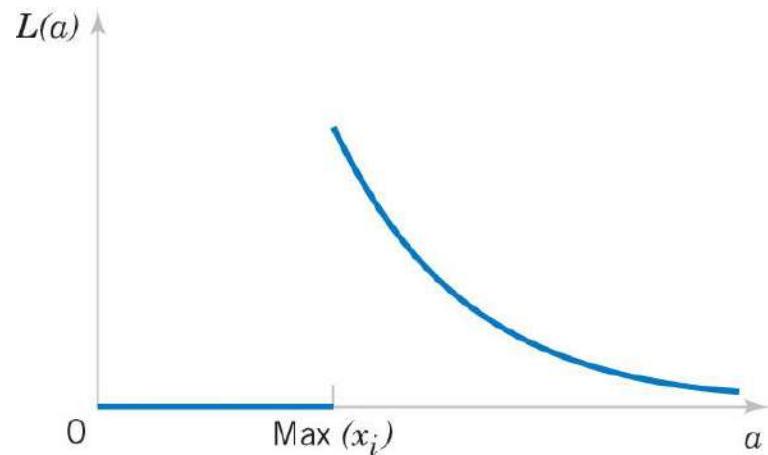


Figure 7-8 The likelihood function for this uniform distribution

Calculus methods don't work here because $L(a)$ is maximized at the discontinuity.

Clearly, a cannot be smaller than $\max(x_i)$, thus the MLE is $\max(x_i)$.

Example 7-15: Gamma Distribution MLE-1

Let X_1, X_2, \dots, X_n be a random sample from a gamma distribution. The log of the likelihood function is:

$$\begin{aligned}\ln L(r, \lambda) &= \ln \left(\prod_{i=1}^n \frac{\lambda^r x_i^{r-1} e^{-\lambda x_i}}{\Gamma(r)} \right) \\ &= nr \ln(\lambda) + (r-1) \sum_{i=1}^n \ln(x_i) - n \ln[\Gamma(r)] - \lambda \sum_{i=1}^n x_i\end{aligned}$$

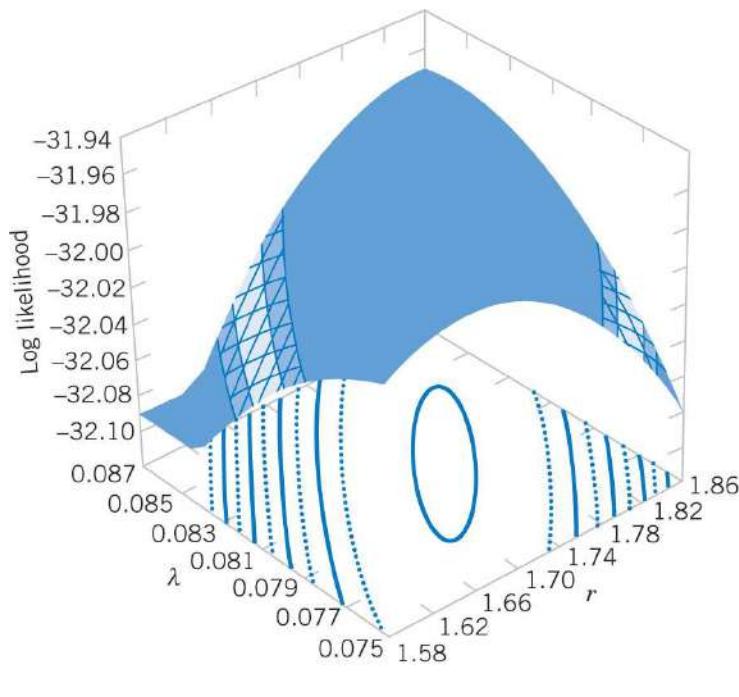
$$\frac{\partial \ln L(r, \lambda)}{\partial r} = n \ln(\lambda) + \sum_{i=1}^n \ln(x_i) - n \frac{\Gamma'(r)}{\Gamma(r)} = 0$$

$$\frac{\partial \ln L(r, \lambda)}{\partial \lambda} = \frac{nr}{\lambda} - \sum_{i=1}^n x_i = 0$$

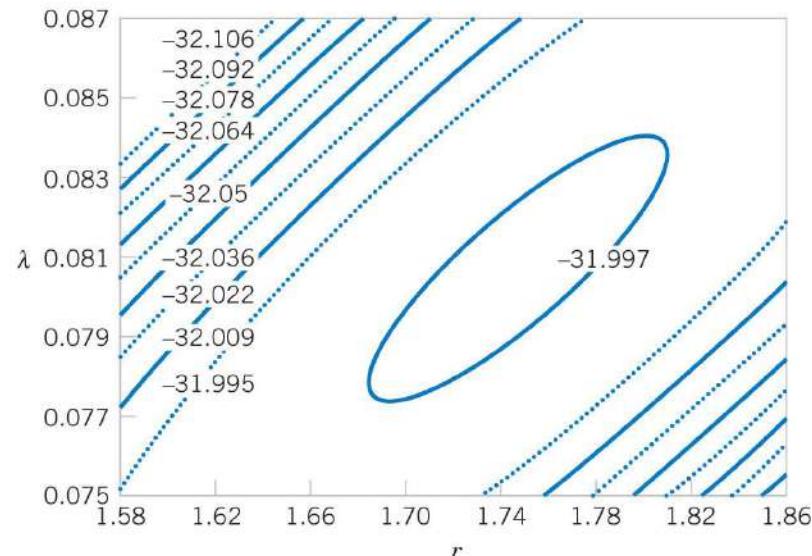
$$\lambda = \frac{\hat{r}}{x} \quad \text{and} \quad n \ln(\lambda) + \sum_{i=1}^n \ln(x_i) = n \frac{\Gamma'(r)}{\Gamma(r)}$$

There is no closed solution for \hat{r} and $\hat{\lambda}$.

Example 7-15: Gamma Distribution MLE-2



(a)



(b)

Figure 7-9 Log likelihood for the gamma distribution using the failure time data ($n=8$). (a) is the log likelihood surface. (b) is the contour plot. The log likelihood function is maximized at $r = 1.75$, $\lambda = 0.08$ using numerical methods. Note the imprecision of the MLEs inferred by the flat top of the function.

Bayesian Estimation of Parameters-1

- The **moment** and **likelihood** methods interpret probabilities as relative frequencies and are called **objective frequencies**.
- The Bayesian method combines sample information with prior information.
- The random variable X has a probability distribution of parameter θ called $f(x|\theta)$. θ could be determined by classical methods.
- Additional information about θ can be expressed as $f(\theta)$, the **prior distribution**, with mean μ_0 and variance σ_0^2 , with θ as the random variable. Probabilities associated with $f(\theta)$ are **subjective probabilities**.
- The **joint distribution** is $f(x_1, x_2, \dots, x_n, \theta)$
- The **posterior distribution** is $f(\theta|x_1, x_2, \dots, x_n)$ is our degree of belief regarding θ after gathering data

Bayesian Estimation of Parameters-2

- Now putting these together, the joint is:

- $f(x_1, x_2, \dots, x_n, \theta) = f(x_1, x_2, \dots, x_n | \theta) \cdot f(\theta)$

- The marginal is:

$$f(x_1, x_2, \dots, x_n) = \begin{cases} \sum_{\theta} f(x_1, x_2, \dots, x_n, \theta), & \text{for } \theta \text{ discrete} \\ \int_{-\infty}^{\infty} f(x_1, x_2, \dots, x_n, \theta) d\theta, & \text{for } \theta \text{ continuous} \end{cases}$$

- The desired posterior distribution is:

$$f(\theta | x_1, x_2, \dots, x_n) = \frac{f(x_1, x_2, \dots, x_n, \theta)}{f(x_1, x_2, \dots, x_n)}$$

- And the Bayesian estimator of θ is the expected value of the posterior distribution

Example 7-16: Bayes Estimator for a Normal Mean-1

Let X_1, X_2, \dots, X_n be a random sample from a normal distribution unknown mean μ and known variance σ^2 . Assume that the prior distribution for μ is:

$$f(\mu) = \frac{1}{\sqrt{2\pi}\sigma_0} e^{-(\mu-\mu_0)^2/2\sigma_0^2} = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\left(\mu^2 - 2\mu\mu_0 + \mu_0^2\right)/2\sigma_0^2}$$

The joint distribution of the sample is:

$$\begin{aligned} f(x_1, x_2, \dots, x_n | \mu) &= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\left(1/2\sigma^2\right)\sum_{i=1}^n (x_i - \mu)^2} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\left(1/2\sigma^2\right)\left(\sum_{i=1}^n x_i^2 - 2\mu\sum_{i=1}^n x_i + n\mu^2\right)} \end{aligned}$$

Example 7-16: Bayes Estimator for a Normal Mean-2

Now the joint distribution of the sample and μ is:

$$f(x_1, x_2, \dots, x_n, \mu) = f(x_1, x_2, \dots, x_n | \mu) \cdot f(\mu)$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2} \sqrt{2\pi\sigma_0^2}} e^u$$

$$\text{where } u = \left(\frac{-1}{2} \right) \left[\mu^2 \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right) - 2\mu \left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum x_i}{\sigma^2} \right) + \frac{\sum x_i^2}{\sigma^2} + \frac{\mu_0^2}{\sigma_0^2} \right]$$

$$= h_1(\cdot) e^{-(1/2) \left[\mu^2 \left(\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2/n} \right) - 2\mu \left(\frac{\mu_0}{\sigma_0^2} + \frac{\bar{x}}{\sigma^2/n} \right) \right]} \quad \& \text{ completing the square}$$

$$= h_2(\cdot) e^{-\left(1/2\right) \left(\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2/n} \right) \left[\mu^2 - \left(\frac{(\sigma^2/n)\mu_0}{\sigma_0^2 + \sigma^2/n} + \frac{\bar{x}\sigma_0^2}{\sigma_0^2 + \sigma^2/n} \right) \right]^2}$$

$$f(\mu | x_1, x_2, \dots, x_n) = h_3(\cdot) e^{-\left(1/2\right) \left(\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2/n} \right) \left[\mu^2 - \left(\frac{(\sigma^2/n)\mu_0 + \sigma_0^2 \bar{x}}{\sigma_0^2 + \sigma^2/n} \right) \right]} \quad \text{is the posterior distribution}$$

$h_i(\cdot)$ = function to collect unneeded components (not μ)

Example 7-16: Bayes Estimator for a Normal Mean-3

- After all that algebra, the bottom line is:

$$E(\mu) = \mu = \frac{(\sigma^2/n)\mu_0 + \sigma_0^2 \bar{x}}{\sigma_0^2 + \sigma^2/n}$$

$$V(\mu) = \left(\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2/n} \right)^{-1} = \frac{\sigma_0^2 (\sigma^2/n)}{\sigma_0^2 + \sigma^2/n}$$

- Observations:
 - Estimator is a weighted average of μ_0 and \bar{x} .
 - \bar{x} is the MLE for μ .
 - The importance of μ_0 decreases as n increases.

Example 7-16: Bayes Estimator for a Normal Mean-4

To illustrate:

- The prior parameters: $\mu_0 = 0, \sigma_0^2 = 1$
- Sample: $n = 10, \bar{x} = 0.75, \sigma^2 = 4$

$$\begin{aligned}\mu &= \frac{(\sigma^2/n)\mu_0 + \sigma_0^2\bar{x}}{\sigma_0^2 + \sigma^2/n} \\ &= \frac{(4/10)0 + 1(0.75)}{1 + (4/10)} = 0.536\end{aligned}$$

Important Terms & Concepts of Chapter 7

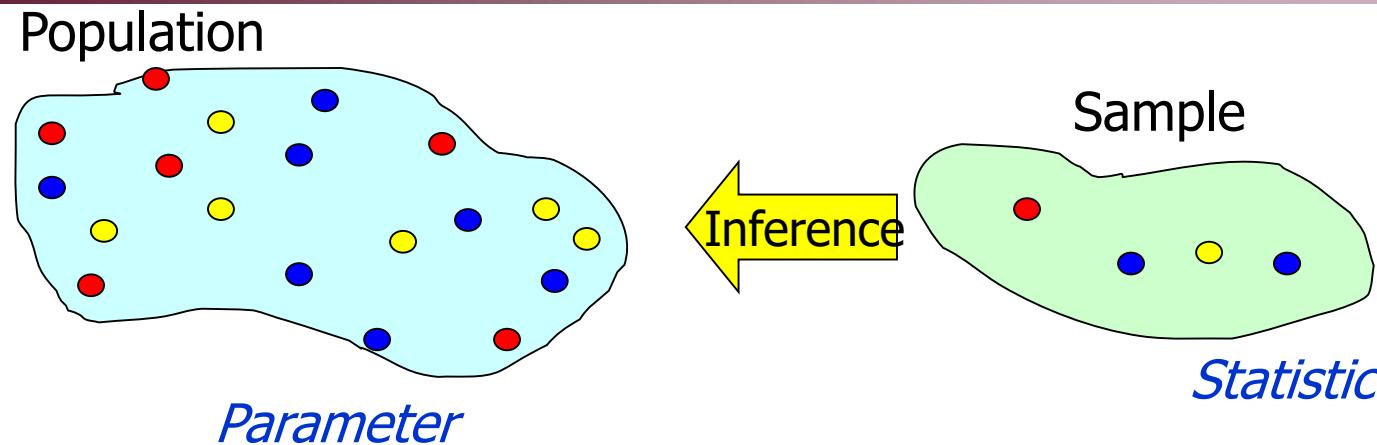
Bayes estimator	Parameter estimation
Bias in parameter estimation	Point estimator
Central limit theorem	Population or distribution moments
Estimator vs. estimate	Posterior distribution
Likelihood function	Prior distribution
Maximum likelihood estimator	Sample moments
Mean square error of an estimator	Sampling distribution
Minimum variance unbiased estimator	An estimator has a:
Moment estimator	– Standard error
Normal distribution as the sampling distribution of the:	– Estimated standard error
– sample mean	Statistic
– difference in two sample means	Statistical inference
	Unbiased estimator

Foundation of Data Science and Analytics

Estimation using Confidence Intervals

Arun K. Timalsina

Estimation



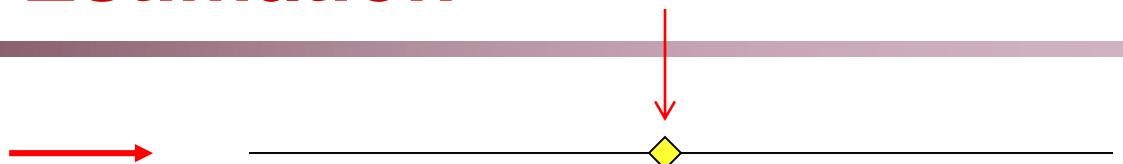
- There are two types of inference: estimation and hypothesis testing; **estimation** is introduced first.

E.g., the sample mean (\bar{x}) is used to **estimate** the population mean (μ).

- The objective of estimation is to determine the **approximate value** of a population parameter on the basis of a sample statistic.

Point & Interval Estimation

1) Point Estimate



2) Interval Estimate



- **Point estimate:**

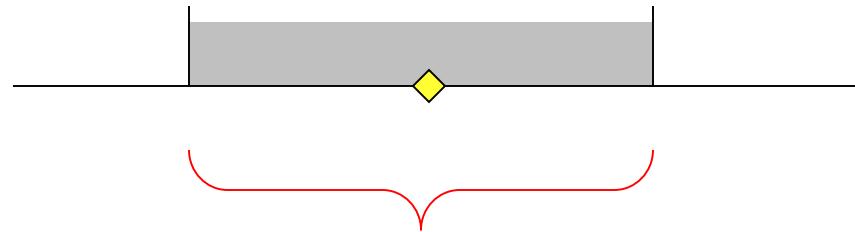
A point estimate is a single number,

- **Interval estimate**

Provides additional information about variability

Point & Interval Estimation

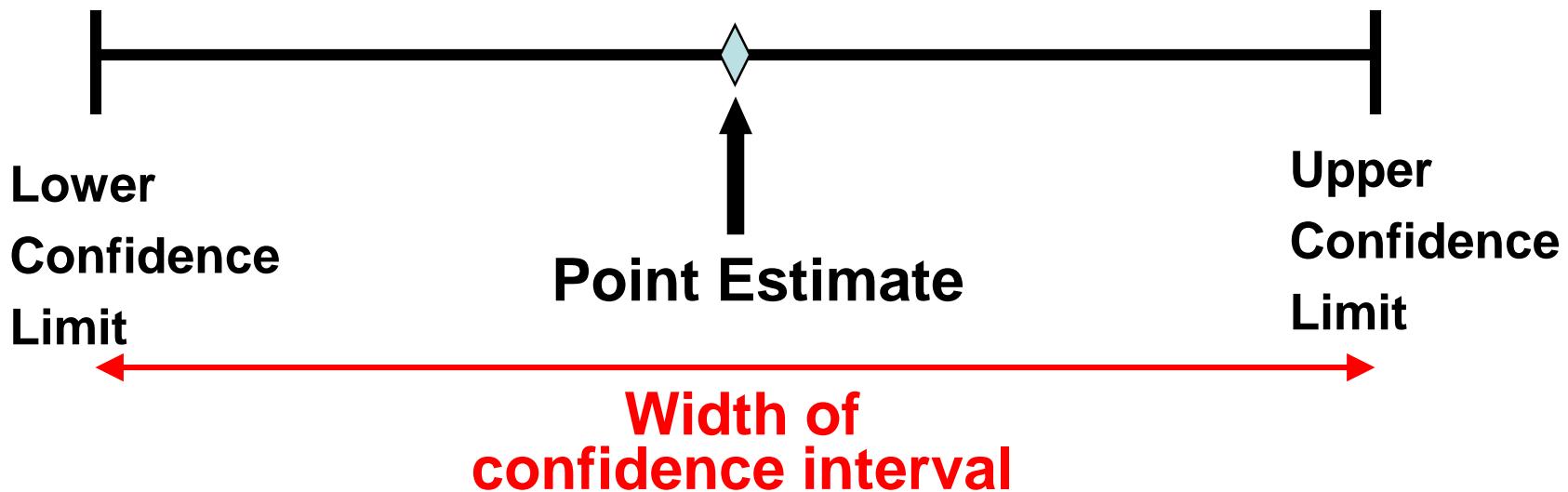
- An ***interval estimator*** draws inferences about a population by estimating the value of an unknown parameter using an interval.



- That is we say (with some ___% certainty) that the population parameter of interest is between some lower and upper bounds.

Point & Interval Estimation

- A point estimate is a rule or formula that tells us how to calculate a numerical estimate based on the measurements contained in the sample
 - A point estimate is a single number,
- An interval estimator is a formula that tell us how to use sample data to calculate an interval that estimates a population
 - A confidence interval provides additional information about variability



Point & Interval Estimation

For example, suppose we want to estimate the mean summer income of a class of business students. For n=25 students, \bar{x} is calculated to be 400 \$/week.

point estimate

interval estimate

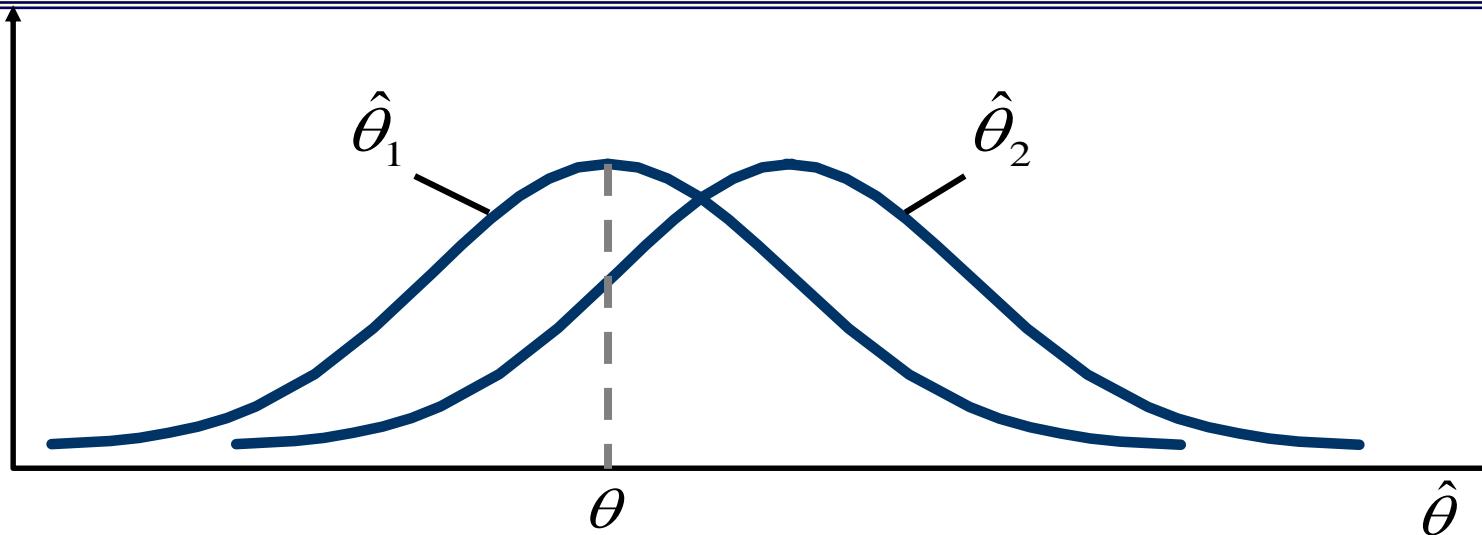
An alternative statement is:

The mean income is ***between*** 380 and 420 \$/week.

Unbiasedness

The point estimator $\hat{\theta}$ is said to be an unbiased estimator of the parameter θ if the expected value, or mean, of the sampling distribution of $\hat{\theta}$ is θ ; that is,

$$E(\hat{\theta}) = \theta$$



Minimum Variance Unbiased Estimator (MVUE)

Suppose there are several unbiased estimators of θ .

Then the unbiased estimator with the smallest variance is said to be **the most efficient estimator** or to be **the minimum variance unbiased estimator** of θ .

Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two unbiased estimators of θ , based on the same number of sample observations. Then,

a) $\hat{\theta}_1$ is said to be more efficient than $\hat{\theta}_2$ if

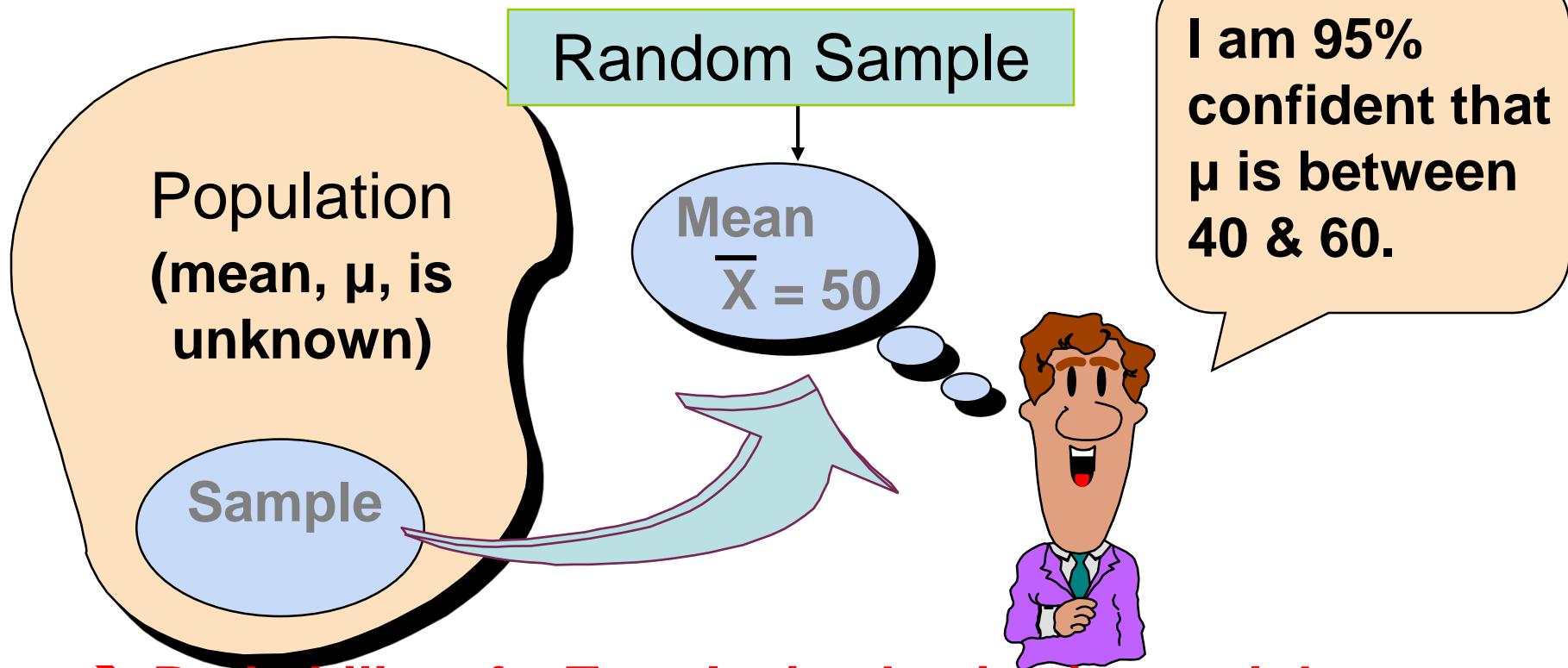
$$Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$$

b) The relative efficiency of $\hat{\theta}_1$ with respect to $\hat{\theta}_2$ is the ratio of their variances; that is,

$$\text{Relative Efficiency} = \frac{Var(\hat{\theta}_2)}{Var(\hat{\theta}_1)}$$

Confidence Coefficient

- Confidence coefficient: The probability that the random interval, prior to sampling, will contain the estimated parameter $(1 - \alpha)$.
- You specify α → $(1 - \alpha)$ 100% confidence that the constructed interval will contain the estimate.

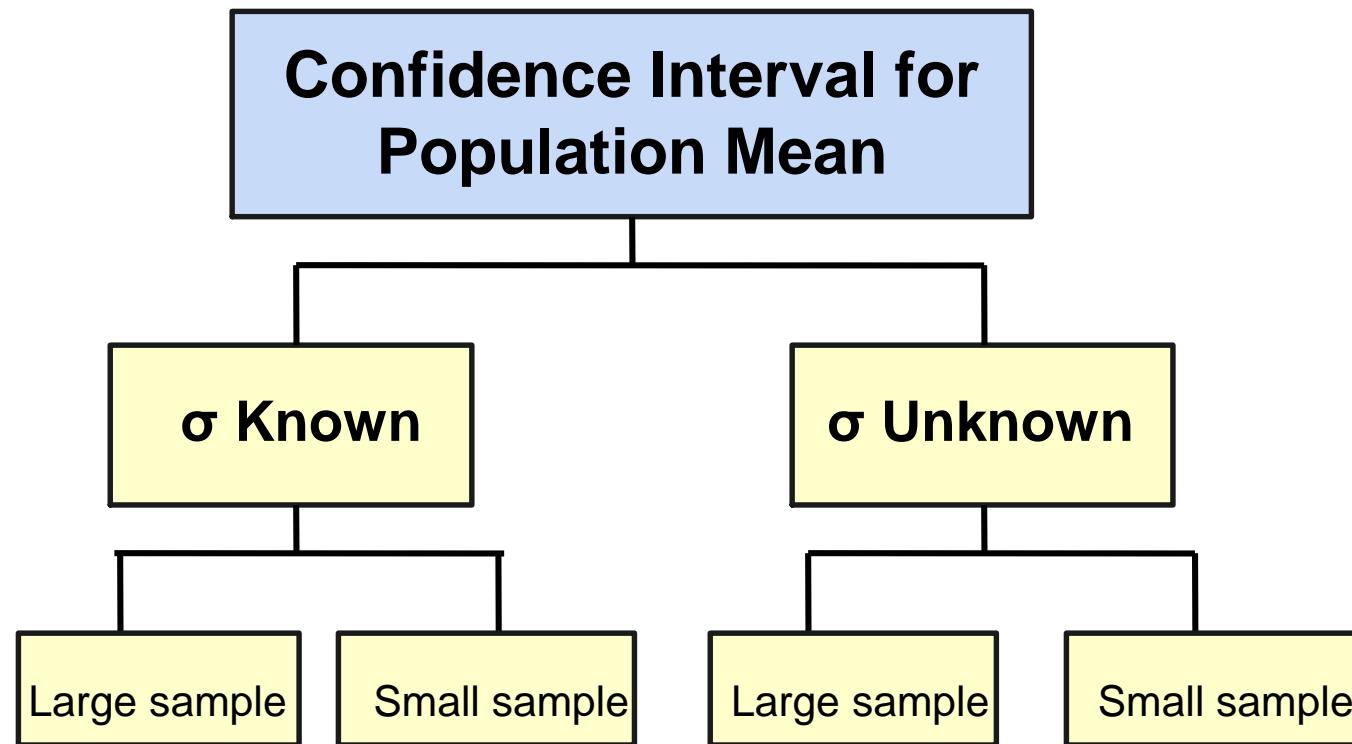


α → Probability of a Type I mistake: i.e. interval does not actually contain parameter.

Confidence Coefficient

- Suppose **confidence level** = 95%
- Also written $(1 - \alpha)$ = 0.95
- A relative frequency interpretation:
 - In the long run, 95% of all the confidence intervals that can be constructed will contain the unknown true parameter
- A specific interval either will contain or will not contain the true parameter

Confidence Intervals for Population Mean



Confidence Interval for Population Mean

- We can calculate an interval estimator from a sampling distribution, by:

Drawing a sample of size n from the population

Calculating its mean,

And, by the central limit theorem, we know that X is normally (or approximately normally) distributed so...

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

- ...will have a standard normal (or approximately normal) distribution.

Confidence Interval for Population Mean

Looking at this in more detail...

Known, i.e. standard normal distribution

Known, i.e. sample mean

Unknown, i.e. we want to **estimate** the population mean

Known, i.e. its **assumed** we know the population standard deviation...

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

Known, i.e. the number of items sampled

Confidence Interval for Population Mean

$$P\left(\mu - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

the confidence interval

- Thus, the **probability** that the interval:

the sample mean is in the center of the interval...

$$\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = \left\{ \bar{X} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right\}$$

contains the population mean μ is $1 - \alpha$. This is a **confidence interval estimator for μ** .

Confidence Interval for μ (σ known)

For samples of size > 30 , the confidence interval is expressed as

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

$$\bar{x} \pm z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right)$$

Large Sample

For samples of size < 30 , the confidence interval is expressed as

$$t = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

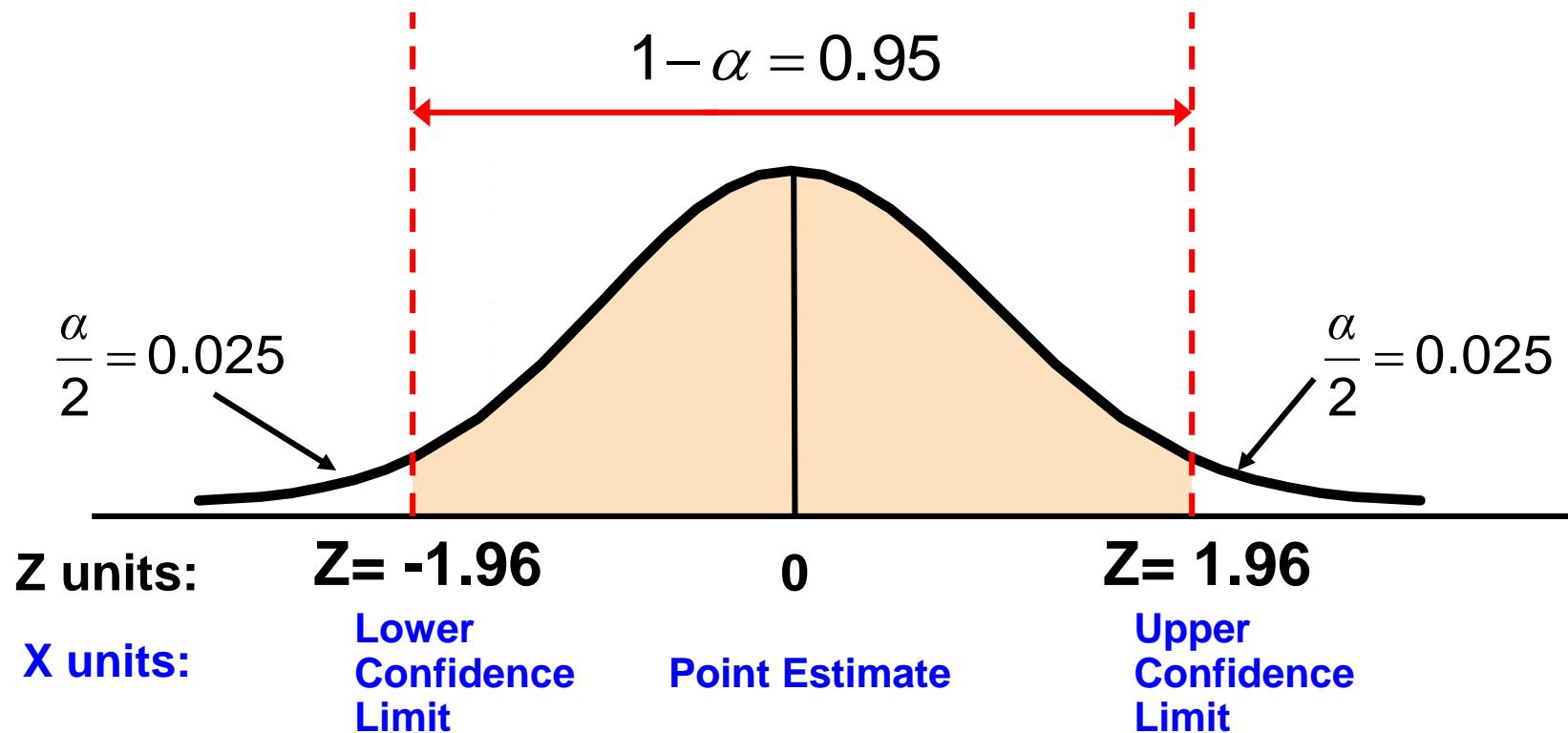
$$\bar{x} \pm t_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right)$$

Small Sample

Finding the Critical Value, Z (Example)

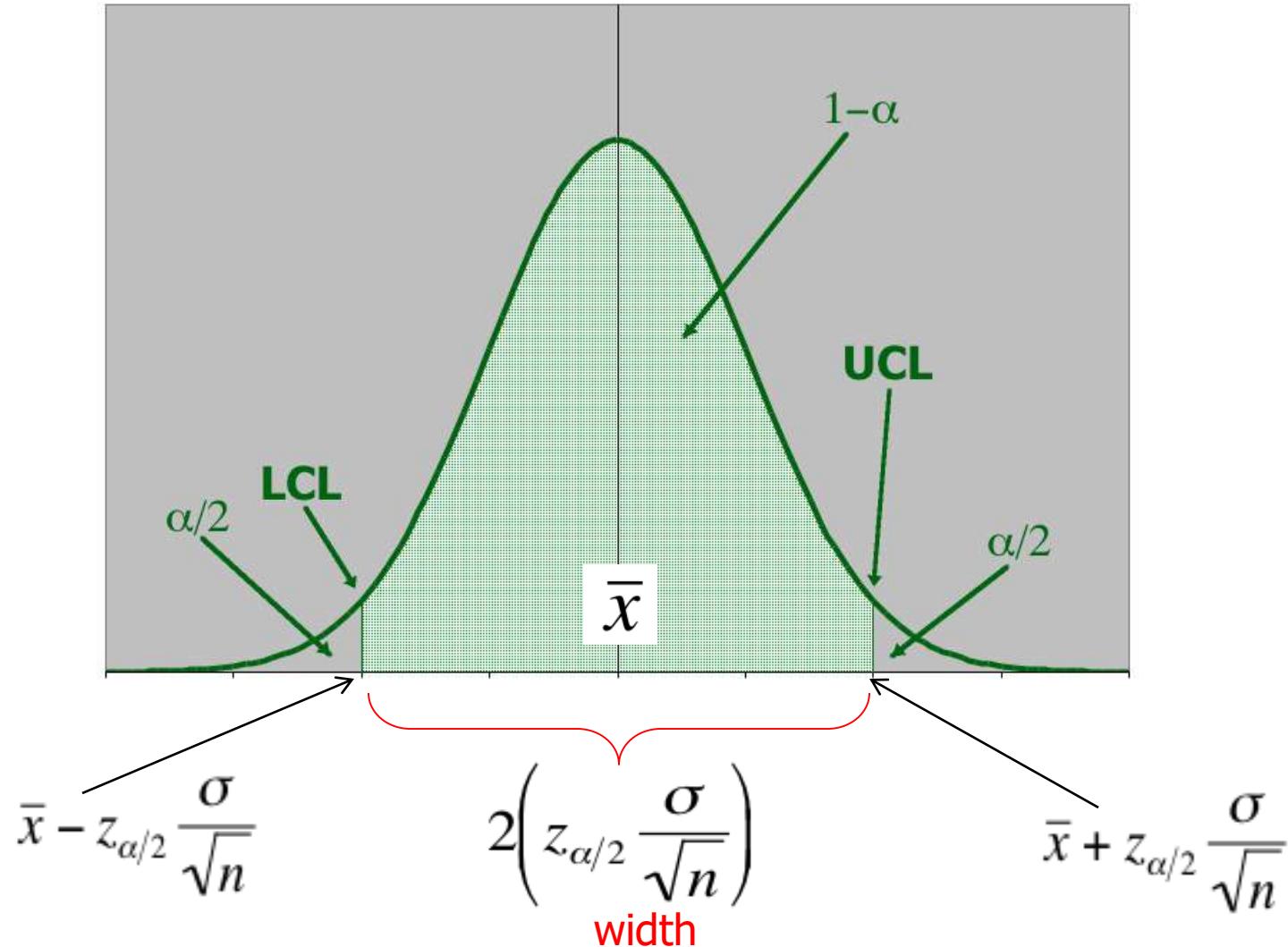
α	0.01	0.02	0.05	0.10
$Z_{\alpha/2}$	2.58	2.33	1.96	1.645
Confidence Level	99%	98%	95%	90%

Consider a 95% confidence interval: $Z = \pm 1.96$



Graphically...

...here is the confidence interval for μ :



Example

- A sample of 40 circuits from a large normal population has a mean resistance of 2.20 ohms. We know from past testing that the population standard deviation is 0.35 ohms.
- Determine a 95% confidence interval for the true mean resistance of the population.

Example

$$\bar{X} \pm Z \frac{\sigma}{\sqrt{n}}$$

$$= 2.20 \pm 1.96 (0.35/\sqrt{40})$$

$$= 2.20 \pm 0.108$$

$$2.092 \leq \mu \leq 2.308$$

We are 95% confident that the true mean resistance is between 2.092 and 2.308 ohms

Although the true mean may or may not be in this interval, 95% of intervals formed in this manner will contain the true mean

Confidence Interval for μ (σ Unknown)

For samples of size > 30 , the confidence interval is expressed as

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$\bar{x} \pm z_{\alpha/2} \left(\frac{s}{\sqrt{n}} \right)$$

Large Sample

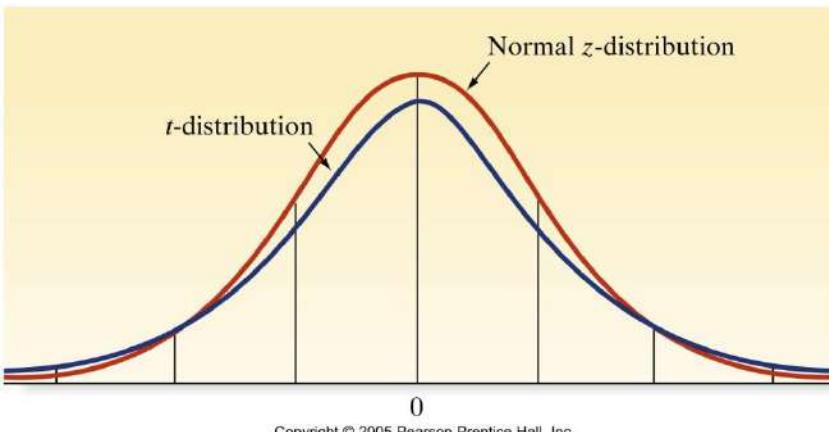
For samples of size < 30 , the confidence interval is expressed as

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

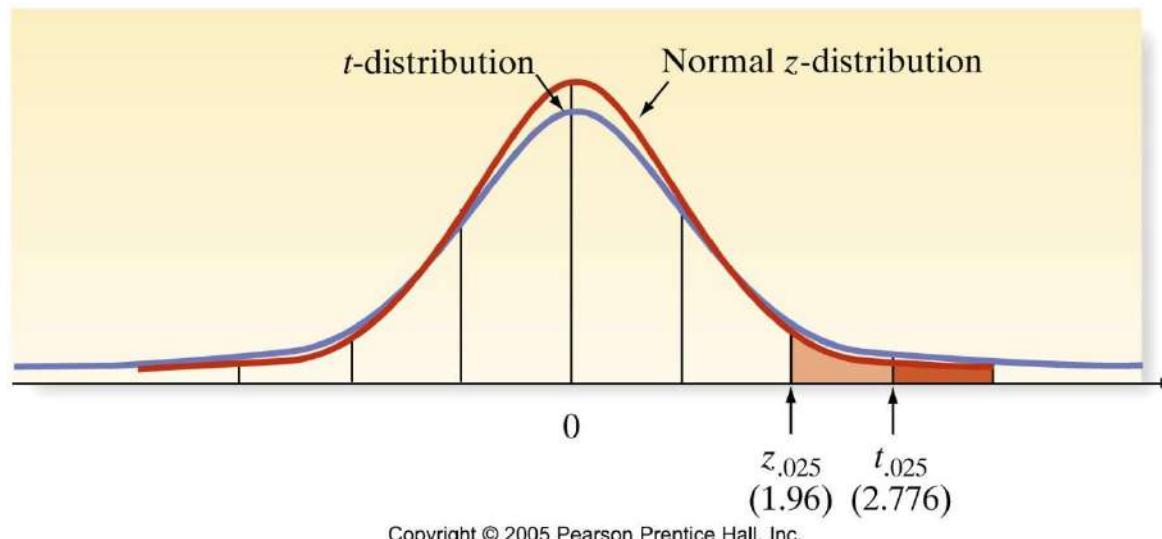
$$\bar{x} \pm t_{\alpha/2} \left(\frac{s}{\sqrt{n}} \right)$$

Small Sample

Small-Sample Confidence Interval for a Population Mean



- The t-statistic has: a sampling distribution very similar to z
- Variability dependent on n , or sample size.
- Variability is expressed as **(n-1) degrees of freedom (df)**. As (df) gets smaller, variability increases

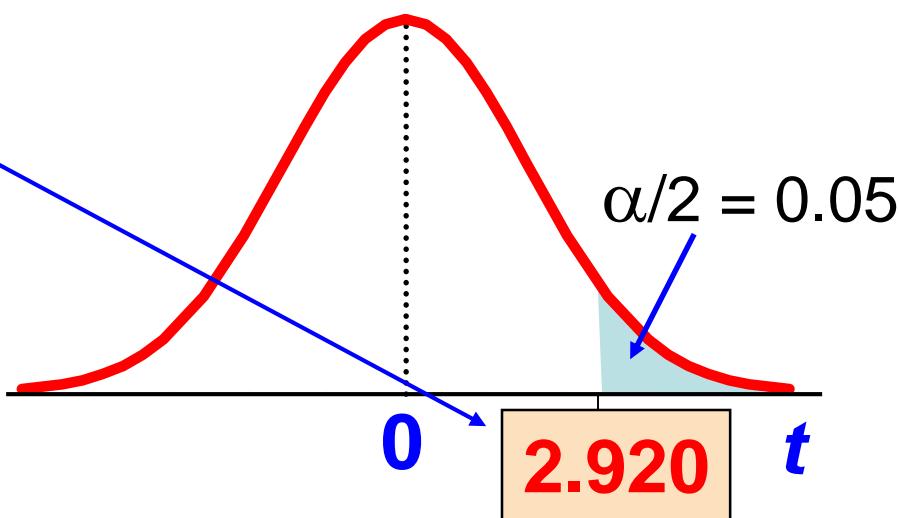


Student's t Table

		Upper Tail Area		
		.25	.10	.05
df				
1		1.000	3.078	6.314
2		0.817	1.886	2.920
3		0.765	1.638	2.353

The body of the table contains t values, not probabilities

Let: $n = 3$
 $df = n - 1 = 2$
 $\alpha = 0.10$
 $\alpha/2 = 0.05$



Example

A random sample of $n = 25$ has $\bar{X} = 50$ and $S = 8$. Form a 95% confidence interval for μ

$$- \text{d.f.} = n - 1 = 24, \text{ so } t_{\alpha/2, n-1} = t_{0.025, 24} = 2.0639$$

The confidence interval is

$$\bar{X} \pm t_{\alpha/2, n-1} \frac{S}{\sqrt{n}} = 50 \pm (2.0639) \frac{8}{\sqrt{25}}$$

$$46.698 \leq \mu \leq 53.302$$

Confidence Interval for a Proportion

$$Z = \frac{X - np}{\sqrt{np(1 - p)}} = \frac{\hat{P} - p}{\sqrt{\frac{p(1 - p)}{n}}}$$

$$-z_{\alpha/2} \leq \frac{\hat{p} - p}{\sqrt{\hat{p}(1 - \hat{p})/n}} \leq z_{\alpha/2}$$

$$\frac{\hat{p} - p}{\sqrt{\hat{p}(1 - \hat{p})/n}} \sim N(0,1)$$

$$\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

Example

In a random sample of 85 automobile engine crankshaft bearings, 10 have a surface finish that is rougher than the specifications allow. Therefore, a point estimate of the proportion of bearings in the population that exceeds the roughness specification is $\hat{p} = x/n = 10/85 = 0.12$. A 95% two-sided confidence interval for p is computed from

$$\hat{p} - z_{0.025} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \leq p \leq \hat{p} + z_{0.025} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

$$0.12 - 1.96 \sqrt{\frac{0.12(0.88)}{85}} \leq p \leq 0.12 + 1.96 \sqrt{\frac{0.12(0.88)}{85}}$$

$$0.05 \leq p \leq 0.19$$

Example

- In a random sample of 80 automotive crankshaft bearings, 15 of the bearings have a surface finish that is rougher than the specifications will allow.
 - The point estimate of the fraction nonconforming in the process: $15/80 = 0.1875$
 - Assuming that the normal approximation is appropriate
 - 95% two-sided confidence interval: $\alpha = 0.05 \quad z_{\alpha/2} = 1.96$

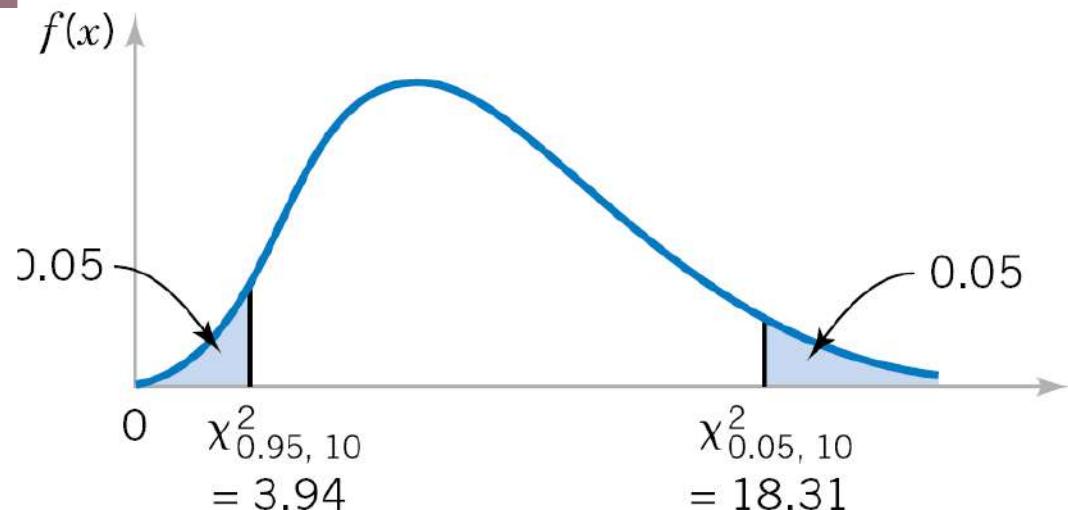
$$\hat{p} = 0.1875, z_{\alpha/2} = 1.96$$

$$\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

$$0.1020 \leq p \leq 0.2730$$

Confidence Interval for the Variance

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$



$$\chi^2_{1-\alpha/2, n-1} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{\alpha/2, n-1}$$

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2, n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha/2, n-1}}$$

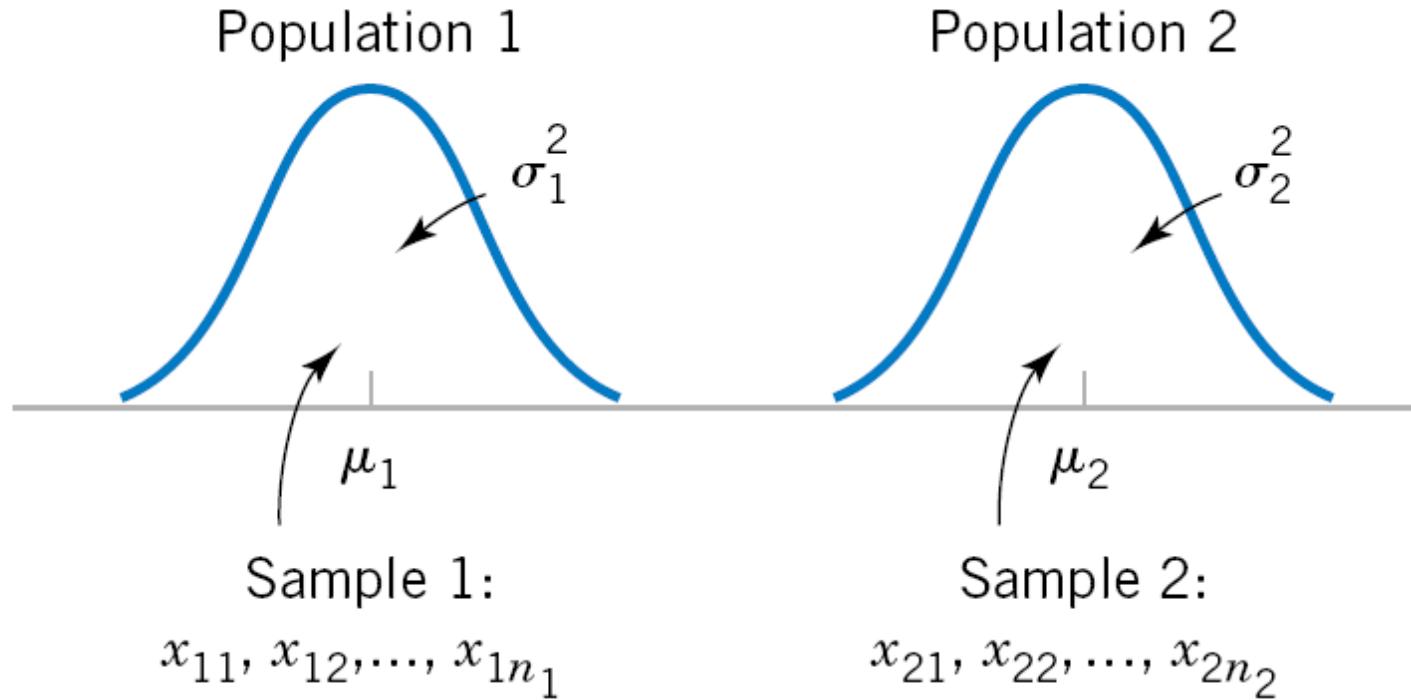
Example

An automatic filling machine is used to fill bottles with liquid detergent. A random sample of 20 bottles results in a sample variance of fill volume of $s^2 = 0.0153$ (fluid ounces) 2 . If the variance of fill volume is too large, an unacceptable proportion of bottles will be under- or overfilled. We will assume that the fill volume is approximately normally distributed. A 95% upper-confidence interval is found from

$$\sigma^2 \leq \frac{(n - 1)s^2}{\chi_{0.95, 19}^2}$$

$$\sigma^2 \leq \frac{(19)0.0153}{10.117} = 0.0287 \text{ (fluid ounce)}^2$$

Confidence Interval Between Two Populations



Confidence Interval for the Difference Between Two Means

Large Sample

σ Known

$$\bar{x}_1 - \bar{x}_2 - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Large Sample

σ Unknown

$$(\bar{X}_1 - \bar{X}_2) - z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \leq (\mu_1 - \mu_2) \leq (\bar{X}_1 - \bar{X}_2) + z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Confidence Interval for the Difference Between Two Means (σ unknown and equal)

$$\sigma_1^2 = \sigma_2^2$$

Small Sample

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$(\bar{X}_1 - \bar{X}_2) - t_{\alpha/2, n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq (\mu_1 - \mu_2)$$

$$\leq (\bar{X}_1 - \bar{X}_2) - t_{\alpha/2, n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Example

An article in the journal *Hazardous Waste and Hazardous Materials* (Vol. 6, 1989) reported the results of an analysis of the weight of calcium in standard cement and cement doped with lead. Reduced levels of calcium would indicate that the hydration mechanism in the cement is blocked and would allow water to attack various locations in the cement structure. Ten samples of standard cement had an average weight percent calcium of $\bar{x}_1 = 90.0$, with a sample standard deviation of $s_1 = 5.0$, while 15 samples of the lead-doped cement had an average weight percent calcium of $\bar{x}_2 = 87.0$, with a sample standard deviation of $s_2 = 4.0$.

$$\begin{aligned}s_p^2 &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \\&= \frac{9(5.0)^2 + 14(4.0)^2}{10 + 15 - 2} \\&= 19.52\end{aligned}$$

Example (cont.)

$$\bar{x}_1 - \bar{x}_2 - t_{0.025, 23} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + t_{0.025, 23} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$



$$90.0 - 87.0 - 2.069(4.4) \sqrt{\frac{1}{10} + \frac{1}{15}} \leq \mu_1 - \mu_2$$

$$\leq 90.0 - 87.0 + 2.069(44) \sqrt{\frac{1}{10} + \frac{1}{15}}$$



$$-0.72 \leq \mu_1 - \mu_2 \leq 6.72$$

Confidence Interval for the Difference Between Two Means (σ unknown and unequal)

Small Sample

$$\sigma_1^2 \neq \sigma_2^2$$

$$\bar{x}_1 - \bar{x}_2 - t_{\alpha/2, v} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + t_{\alpha/2, v} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$n_1 = n_2 = n \Rightarrow \nu = n_1 + n_2 - 2$$

$$n_1 \neq n_2 \Rightarrow \nu = \frac{(s_1^2 / n_1 + s_2^2 / n_2)^2}{(s_1^2 / n_1)^2 / (n_1 - 1) + (s_2^2 / n_2)^2 / (n_2 - 1)}$$

Example

Tensile strength tests were performed on two different grades of aluminum spars used in manufacturing the wing of a commercial transport aircraft. From past experience with the spar manufacturing process and the testing procedure, the standard deviations of tensile strengths are assumed to be known. The data obtained are as follows: $n_1 = 10$, $\bar{x}_1 = 87.6$, $\sigma_1 = 1$, $n_2 = 12$, $\bar{x}_2 = 74.5$, and $\sigma_2 = 1.5$. If μ_1 and μ_2 denote the true mean tensile strengths for the two grades of spars, we may find a 90% confidence interval on the difference in mean strength $\mu_1 - \mu_2$ as follows:

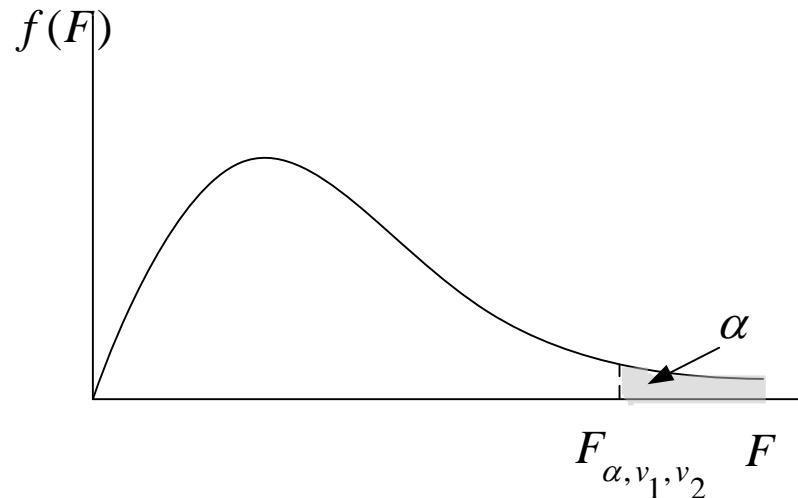
$$\bar{x}_1 - \bar{x}_2 - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$87.6 - 74.5 - 1.645 \sqrt{\frac{(1)^2}{10} + \frac{(1.5)^2}{12}} \leq \mu_1 - \mu_2 \leq 87.6 - 74.5 + 1.645 \sqrt{\frac{(1)^2}{10} + \frac{(1.5)^2}{12}}$$

$$12.22 \leq \mu_1 - \mu_2 \leq 13.98$$

Confidence Interval for the Ratio of Two Variance

$$\frac{s_1^2}{s_2^2} / \frac{\sigma_1^2}{\sigma_2^2} \sim F_{n_1-1, n_2-1}$$



$$F_{1-\alpha/2, n_1-1, n_2-1} \leq \frac{s_1^2}{s_2^2} / \frac{\sigma_1^2}{\sigma_2^2} \leq F_{\alpha/2, n_1-1, n_2-1}$$

$$\frac{s_1^2}{s_2^2} \left(\frac{1}{F_{\alpha/2, n_1-1, n_2-1}} \right) \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{s_1^2}{s_2^2} \left(\frac{1}{F_{1-\alpha/2, n_1-1, n_2-1}} \right)$$

Confidence Interval for the Difference Between Two Binomial Proportions

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}_1(1 - \hat{p}_1)/n_1 + \hat{p}_2(1 - \hat{p}_2)/n_2}} \sim N(0,1)$$

$$-z_{\alpha/2} \leq \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}_1(1 - \hat{p}_1)/n_1 + \hat{p}_2(1 - \hat{p}_2)/n_2}} \leq z_{\alpha/2}$$

$$(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} \leq (p_1 - p_2)$$
$$\leq (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

MATCHED pairs

- In a matched pairs study, subjects are matched in pairs and the outcomes are compared within each matched pair.
 - Example: before and after studies

Modern Language Association listening scores for French teachers							
Teacher	Pretest	Posttest	Gain	Teacher	Pretest	Posttest	Gain
1	32	34	2	11	30	36	6
2	31	31	0	12	20	26	6
3	29	35	6	13	24	27	3
4	10	16	6	14	24	24	0
5	30	33	3	15	31	32	1
6	33	36	3	16	30	31	1
7	22	24	2	17	15	15	0
8	25	28	3	18	32	34	2
9	32	26	-6	19	23	26	3
10	20	26	6	20	23	26	3

MATCHED pairs

Large Sample

$$\bar{d} \pm z_{\alpha/2} \left(\frac{\sigma_d}{\sqrt{n}} \right)$$

Small Sample

$$\bar{d} \pm t_{\alpha/2} \left(\frac{s_d}{\sqrt{n}} \right)$$

Example

Time in Seconds to Parallel Park Two Automobiles

Subject	Automobile		Difference (d_j)
	1(x_{1j})	2(x_{2j})	
1	37.0	17.8	19.2
2	25.8	20.2	5.6
3	16.2	16.8	-0.6
4	24.2	41.4	-17.2
5	22.0	21.4	0.6
6	33.4	38.4	-5.0
7	23.8	16.8	7.0
8	58.2	32.2	26.0
9	33.6	27.8	5.8
10	24.4	23.2	1.2
11	23.4	29.6	-6.2
12	21.2	20.6	0.6
13	36.2	32.2	4.0
14	29.8	53.8	-24.0

Example Cont.

$$\bar{d} =$$

$$1.21$$

$$s_D = 12.68.$$

90% confidence interval

$$\bar{d} - t_{0.05,13}s_D/\sqrt{n} \leq \mu_D \leq \bar{d} + t_{0.05,13}s_D/\sqrt{n}$$

$$1.21 - 1.771(12.68)/\sqrt{14} \leq \mu_D \leq 1.21 + 1.771(12.68)/\sqrt{14}$$
$$-4.79 \leq \mu_D \leq 7.21$$

Selecting the Sample Size...

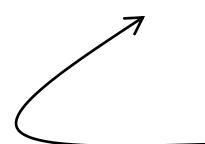
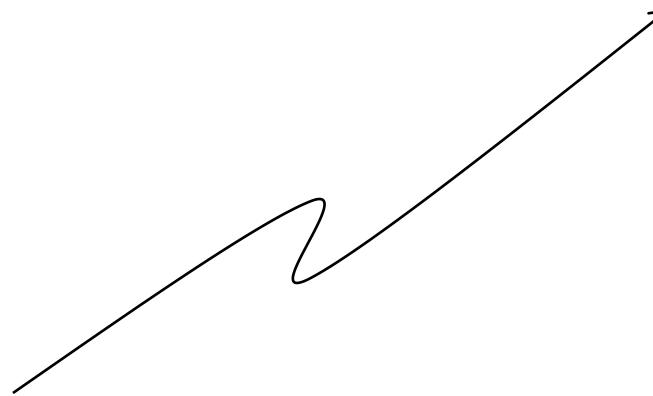
- We can control the width of the interval by determining the sample size necessary to produce narrow intervals.
- Suppose we want to estimate the mean demand “to within 5 units”; i.e. we want the interval estimate to be: $\bar{x} \pm 5$

- Since:

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

- It follows that

$$z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 5$$



Solve for **n** to get requisite sample size!

Selecting the Sample Size...

- Solving the equation...

$$n = \left(\frac{z_{\alpha/2} \sigma}{5} \right)^2 = \left(\frac{(1.96)(75)}{5} \right)^2 = 865$$

- that is, to produce a 95% confidence interval estimate of the mean (± 5 units), we need to sample 865 lead time periods (vs. the 25 data points we have currently).

Sample Size to Estimate a Mean...

- The general formula for the sample size needed to estimate a population mean with an interval estimate of:

$$\bar{x} \pm W$$

- Requires a sample size of at least this large:

$$n = \left(\frac{z_{\alpha/2} \sigma}{W} \right)^2$$

Example 10.2...

- A lumber company must estimate the mean diameter of trees to determine whether or not there is sufficient lumber to harvest an area of forest. They need to estimate this to within 1 inch at a confidence level of 99%. The tree diameters are normally distributed with a standard deviation of 6 inches.
- How many trees need to be sampled?

- Things we know:
- Confidence level = 99%, therefore $\alpha = .01$

$1 - \alpha$	α	$\alpha / 2$	$z_{\alpha/2}$
.90	.10	.05	$z_{.05} = 1.645$
.95	.05	.025	$z_{.025} = 1.96$
.98	.02	.01	$z_{.01} = 2.33$
.99	.01	.005	$z_{.005} = 2.575$

- We want $\bar{x} \pm 1$ hence $W=1$. $z_{\alpha/2} = z_{.005} = 2.575$
- We are given that $\sigma = 6$.

Foundation of Data Science and Analytics

Test of Hypotheses

Arun K. Timalsina

Introduction

- A criminal trial is an example of hypothesis testing without the statistics.
- In a trial a jury must decide between two hypotheses. The null hypothesis is
 H_0 : The defendant is innocent
- The alternative hypothesis or research hypothesis is
 H_1 : The defendant is guilty
- The jury does not know which hypothesis is true. They must make a decision on the basis of evidence presented.

Nonstatistical Hypothesis Testing...

- In the language of statistics **convicting the defendant** is called *rejecting the null hypothesis in favor of the alternative hypothesis*. That is, the jury is saying that there is enough evidence to conclude that the defendant is guilty (i.e., there is enough evidence to support the alternative hypothesis).
- If the **jury acquits** it is stating that *there is not enough evidence to support the alternative hypothesis*. Notice that the jury is not saying that the defendant is innocent, only that there is not enough evidence to support the alternative hypothesis. That is why we never say that we accept the null hypothesis.

Nonstatistical Hypothesis Testing

- There are two possible errors:
 - A Type I error occurs when we reject a true null hypothesis. That is, a Type I error occurs when the jury convicts an innocent person.
 - A Type II error occurs when we don't reject a false null hypothesis. That occurs when a guilty defendant is acquitted.

Type I error: Reject a true null hypothesis

Type II error: Do not reject a false null hypothesis.

$$P(\text{Type I error}) = \alpha$$

$$P(\text{Type II error}) = \beta$$

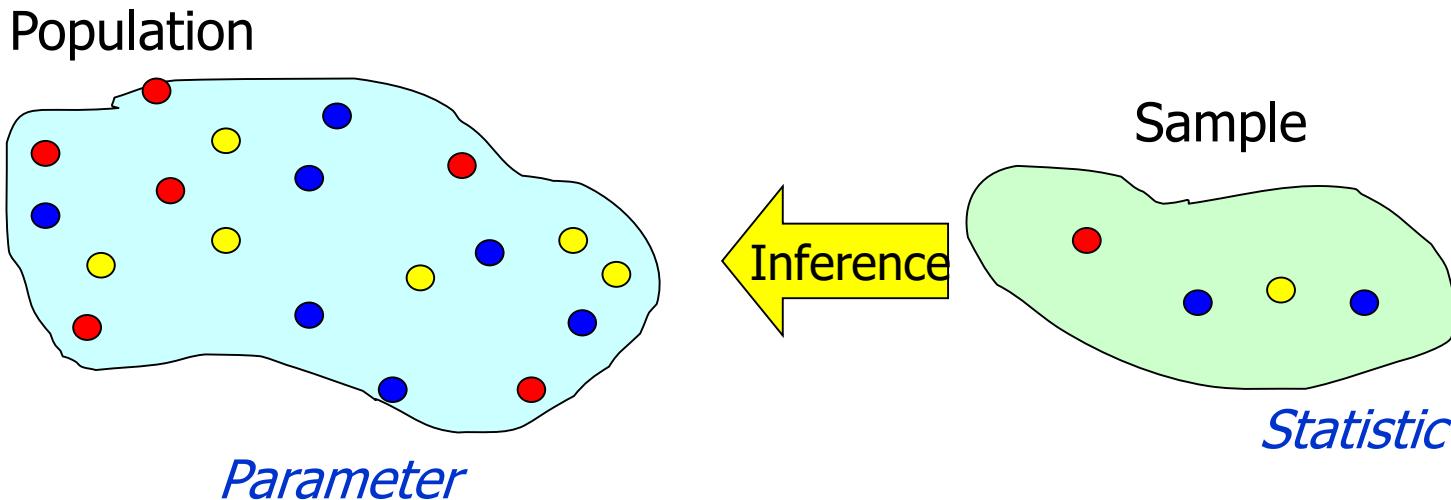
Outcomes and Probabilities

Possible Hypothesis Test Outcomes

	Actual Situation	
Decision	H_0 True	H_0 False
Do Not Reject H_0	No error $(1 - \alpha)$	Type II Error (β)
Reject H_0	Type I Error (α)	No Error $(1 - \beta)$

Key:
Outcome
(Probability)

Estimation



- Hypothesis testing allows us to determine whether enough statistical evidence exists to conclude that a **belief** (i.e. ***hypothesis***) about a parameter is supported by the data.

Concepts of Hypothesis Testing

- There are **two** hypotheses. One is called the ***null hypothesis*** and the other the ***alternative*** or ***research hypothesis***. The usual notation is:

pronounced
H “nought”

H_0 : — *the ‘null’ hypothesis*

H_1 : — *the ‘alternative’ or ‘research’ hypothesis*

- The null hypothesis (H_0) will always state that the ***parameter equals the value*** specified in the alternative hypothesis (H_1)

Concepts of Hypothesis Testing

- Rather than estimate the mean demand, if operations manager wants to know ***whether the mean is different from 350 units.*** We can rephrase this request into a test of the hypothesis:

$$H_0: \bar{X} = 350$$

Thus, our research hypothesis becomes:

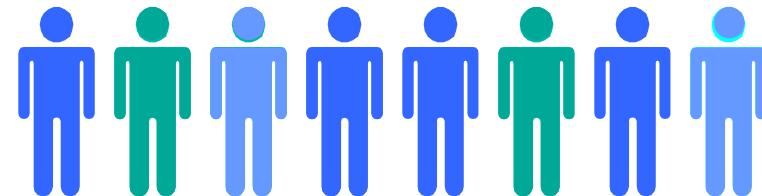
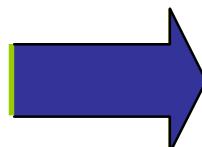
$$H_1: \bar{X} \neq 350$$

This is what we are interested
in determining...

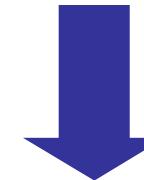
- The **goal** of the process is to determine ***whether there is enough evidence*** to infer that the alternative hypothesis is true.

Hypothesis Testing Process

Claim: the population mean age is 50.
(Null Hypothesis:
 $H_0: \mu = 50$)



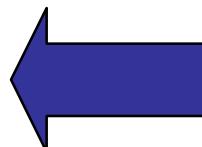
Population



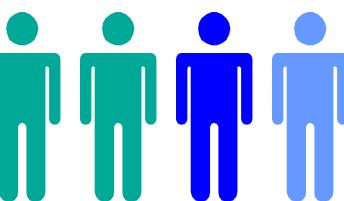
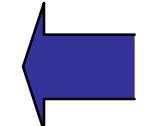
Now select a random sample

Is $\bar{X}=20$ likely if $\mu = 50$?

If not likely,
REJECT
Null Hypothesis



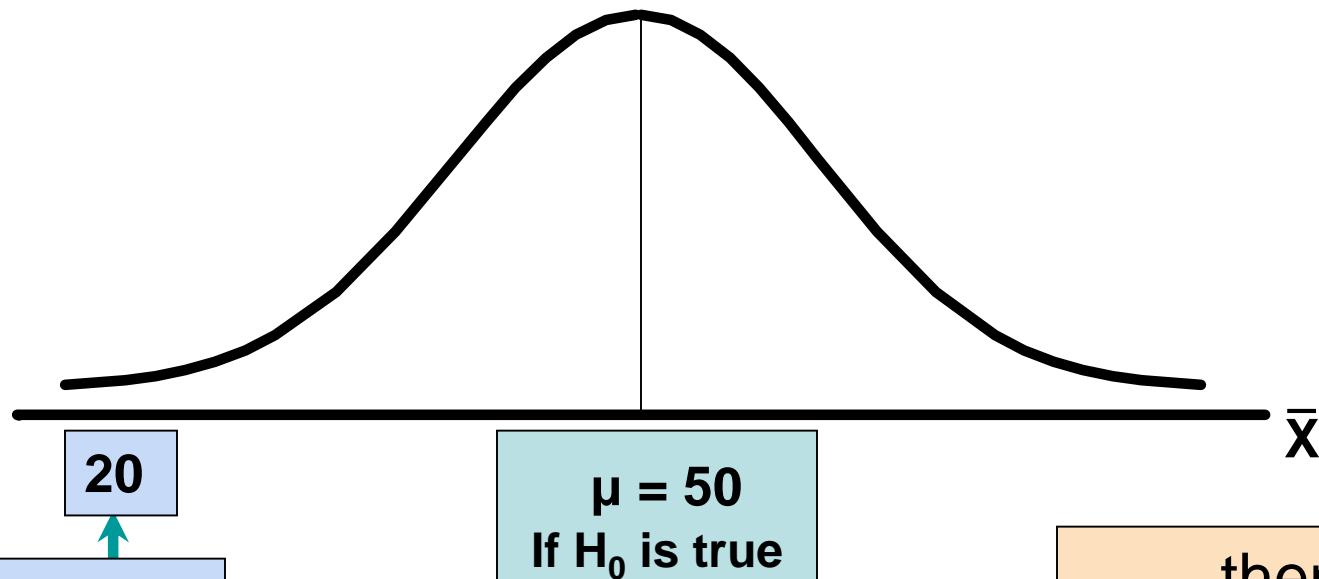
Suppose the sample mean age is 20: $\bar{X} = 20$



Sample

Reason for Rejecting H_0

Sampling Distribution of \bar{X}



If it is unlikely that we would get a sample mean of this value ...

... if in fact this were the population mean...

$\mu = 50$
If H_0 is true

... then we reject the null hypothesis that $\mu = 50$.

Level of Significance, α

- Defines the unlikely values of the sample statistic if the null hypothesis is true
 - Defines rejection region of the sampling distribution
- Is designated by α , (level of significance)
 - Typical values are 0.01, 0.05, or 0.10
- Is selected by the researcher at the beginning
- Provides the critical value(s) of the test

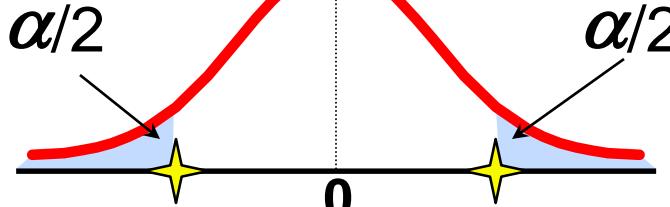
Level of Significance and the Rejection Region

Level of significance = α

$$H_0: \mu = 3$$

$$H_1: \mu \neq 3$$

Two-tail test

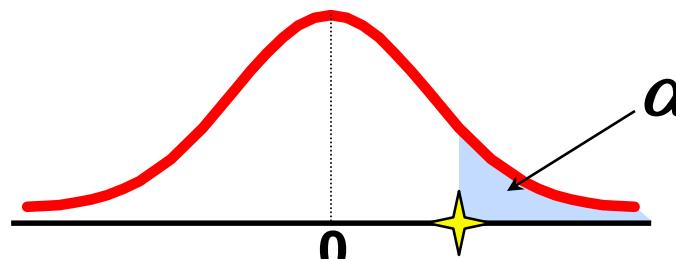


★ Represents critical value

$$H_0: \mu \leq 3$$

$$H_1: \mu > 3$$

Upper-tail test

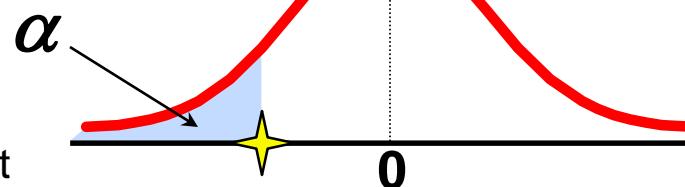


Rejection region is shaded

$$H_0: \mu \geq 3$$

$$H_1: \mu < 3$$

Lower-tail test



Hypothesis Tests for the Mean

For samples of size > 30

$$Z = \frac{\bar{X} - \mu}{\left(\frac{s}{\sqrt{n}} \right)}$$

Two tailed

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_a &: \mu \neq \mu_0 \end{aligned}$$

One tailed

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_a &: \mu > \mu_0 \text{ or} \\ &\quad \mu < \mu_0 \end{aligned}$$

For samples of size < 30

$$T = \frac{\bar{X} - \mu_0}{\left(\frac{s}{\sqrt{n}} \right)}$$

Two tailed

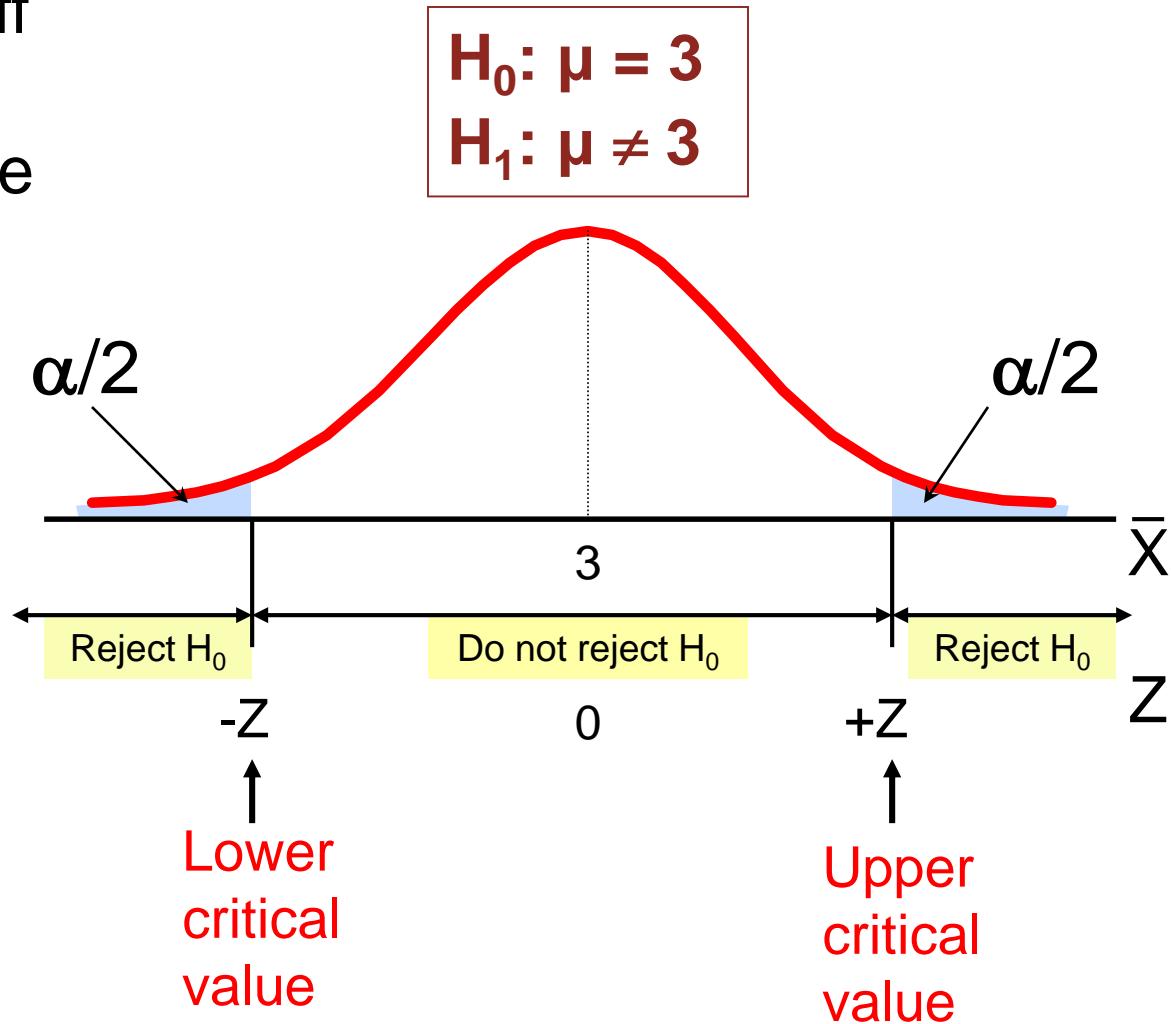
$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_a &: \mu \neq \mu_0 \end{aligned}$$

One tailed

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_a &: \mu > \mu_0 \text{ or} \\ &\quad \mu < \mu_0 \end{aligned}$$

Two-Tail Tests

There are two cutoff values (critical values), defining the regions of rejection



Example

**Test the claim that the true mean # of TV sets
in US homes *is different from 3 units***
(Assume $s = 0.8$)

1. State the appropriate null and alternative hypotheses
 - $H_0: \mu = 3$ $H_1: \mu \neq 3$ (This is a two-tail test)
2. Specify the desired level of significance and the sample size
 - Suppose that $\alpha = 0.05$ and $n = 100$ are chosen for this test

Example

3. Determine the appropriate technique
 - σ is known so this is a Z test.
4. Determine the critical values
 - For $\alpha = 0.05$ the critical Z values are ± 1.96
5. Collect the data and compute the test statistic
 - Suppose the sample results are
 - $n = 100, X = 2.84$ ($\sigma = 0.8$ is assumed known)

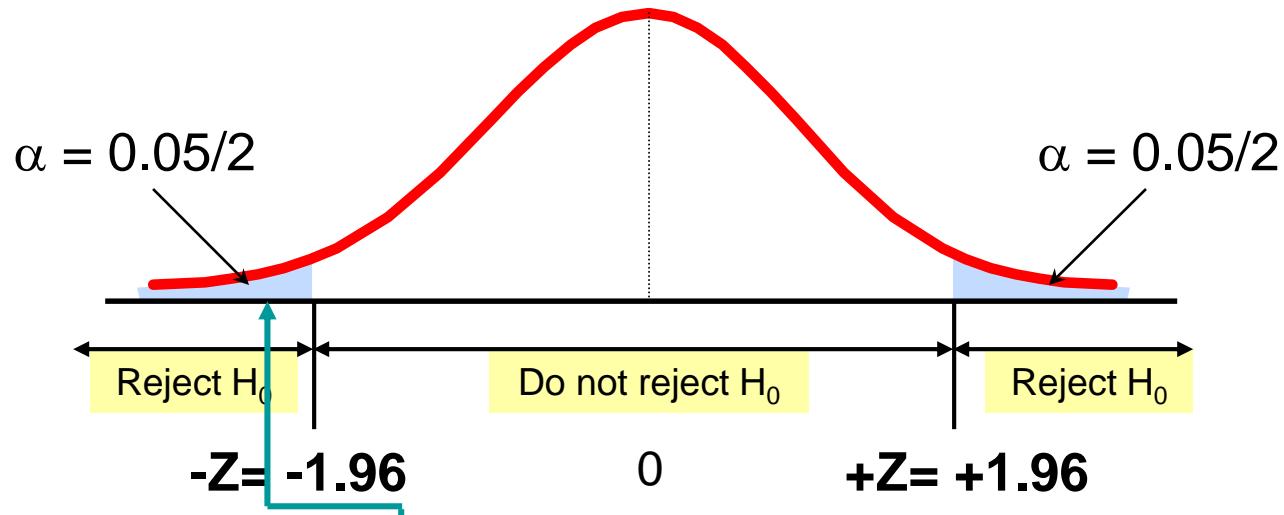
So the test statistic is:

$$Z = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} = \frac{2.84 - 3}{\frac{0.8}{\sqrt{100}}} = \frac{- .16}{.08} = -2.0$$

Example

6. Is the test statistic in the rejection region?

Reject H_0 if
 $Z < -1.96$ or
 $Z > 1.96$;
otherwise
do not reject
 H_0



Here, $Z = -2.0 < -1.96$, so the test statistic is in the rejection region

Since $Z = -2.0 < -1.96$, we reject the null hypothesis and conclude that there is sufficient evidence that the mean number of TVs in US homes is not equal to 3

p-Value Approach to Testing

- p-value: Probability of obtaining a test statistic more extreme (\leq or \geq) than the observed sample value given H_0 is true
 - Also called observed level of significance
 - Smallest value of α for which H_0 can be rejected

If p-value $< \alpha$, reject H_0

If p-value $\geq \alpha$, do not reject H_0

Example

- Example: How likely is it to see a sample mean of 2.84 (or something further from the mean, in either direction) if the true mean is $\mu = 3.0$?

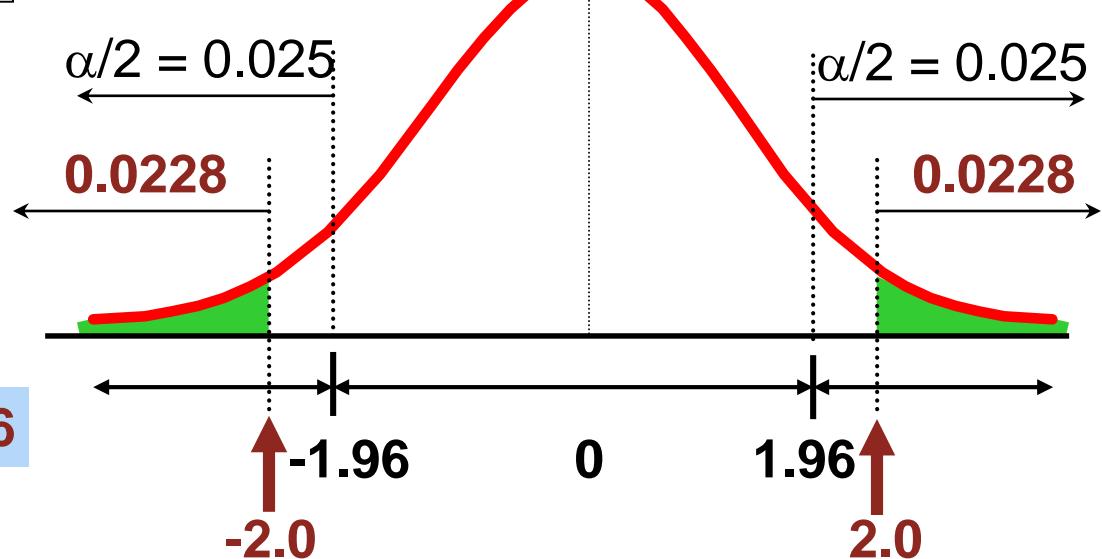
$\bar{X} = 2.84$ is translated
to a Z score of $Z = -2.0$

$$P(Z < -2.0) = 0.0228$$

$$P(Z > 2.0) = 0.0228$$

p-value

$$= 0.0228 + 0.0228 = 0.0456$$



Connection to Confidence Intervals

- For $\bar{X} = 2.84$, $s = 0.8$ and $n = 100$, the 95% confidence interval is:

$$2.84 - (1.96) \frac{0.8}{\sqrt{100}} \text{ to } 2.84 + (1.96) \frac{0.8}{\sqrt{100}}$$

$$2.6832 \leq \mu \leq 2.9968$$

- Since this interval does not contain the hypothesized mean (3.0), we reject the null hypothesis at $\alpha = 0.05$

Example

The average cost of a hotel room in New York is said to be \$168 per night. A random sample of 25 hotels resulted in $\bar{X} = \$172.50$ and

$s = \$15.40$. Test at the $\alpha = 0.05$ level.

(Assume the population distribution is normal)

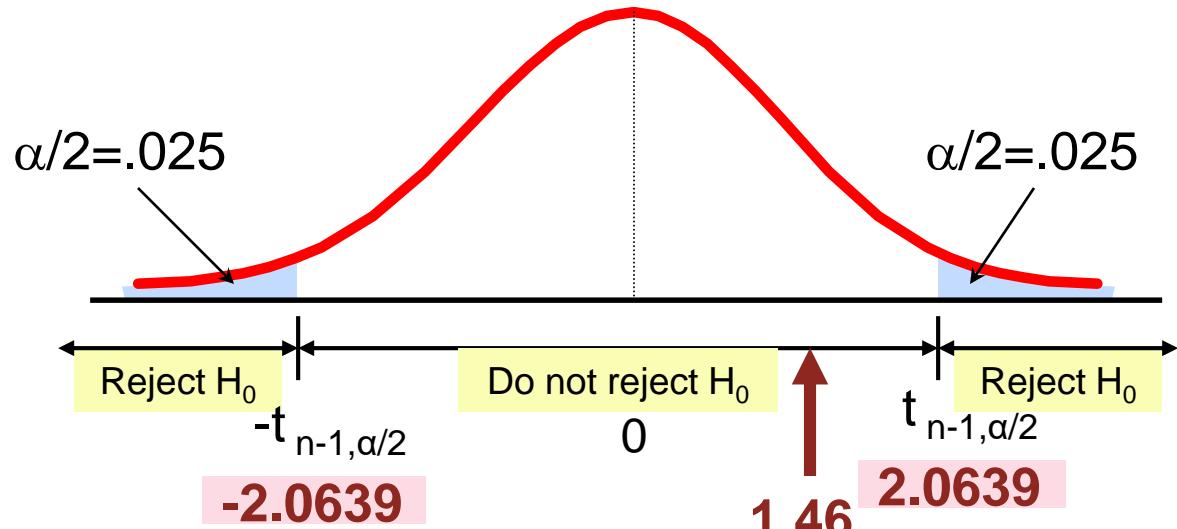
$$\begin{aligned}H_0: \mu &= 168 \\H_1: \mu &\neq 168\end{aligned}$$

Example (cont.)

$$\begin{aligned} H_0: \mu &= 168 \\ H_1: \mu &\neq 168 \end{aligned}$$

$$\alpha = 0.05$$

- $n = 25$
- σ is unknown, so use a **t statistic**
- Critical Value:
 $t_{24} = \pm 2.0639$



$$t_{n-1} = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{172.50 - 168}{\frac{15.40}{\sqrt{25}}} = 1.46$$

Do not reject H_0 : not sufficient evidence that true mean cost is different than \$168

Connection to Confidence Intervals

For $\bar{X} = 172.5$, $s = 15.4$ and $n = 25$,
the 95% confidence interval is :

$$166.14 \leq \mu \leq 178.86$$

Since this interval contains the Hypothesized mean (168),
we do not reject the null hypothesis at $\alpha = 0.05$

Hypothesis Tests for the Proportion

$$p = \frac{X}{n} = \frac{\text{number of successes in sample}}{\text{sample size}}$$

Two tailed

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

$$\begin{aligned}H_0 &: p = p_0 \\H_a &: p \neq p_0\end{aligned}$$

One tailed

$$\begin{aligned}H_0 &: p = p_0 \\H_a &: p > p_0 \text{ or} \\&\quad p < p_0\end{aligned}$$

Example

A marketing company claims that it receives 8% responses from its mailing. To test this claim, a random sample of 500 were surveyed with 25 responses. Test at the $\alpha = 0.05$ significance level.

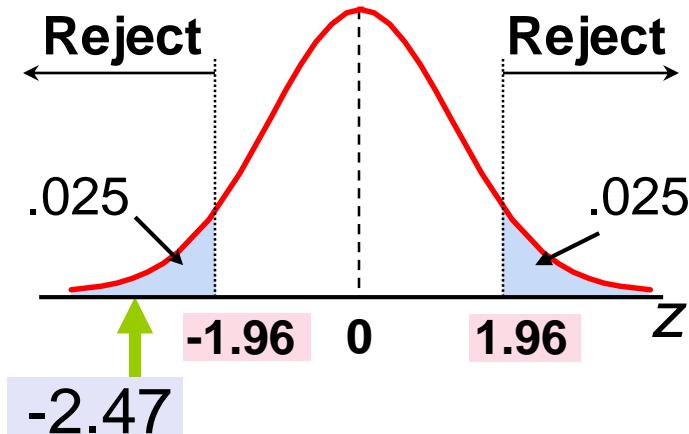
$$\begin{aligned} H_0: p &= 0.08 \\ H_1: p &\neq 0.08 \end{aligned}$$

Example (cont.)

Test Statistic:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{.05 - .08}{\sqrt{\frac{.08(1-.08)}{500}}} = -2.47$$

Critical Values: ± 1.96



Decision:

Reject H_0 at $\alpha = 0.05$

Conclusion:

There is sufficient evidence to reject the company's claim of 8% response rate.

Testing the Two Population Means: Independent Samples

Large Sample,
 σ Unknown

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Large Sample,
 σ known

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Small Sample

$$\sigma_1^2 = \sigma_2^2$$

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Testing the Two Population Means: Independent Samples

Small Sample

$$\sigma_1^2 \neq \sigma_2^2$$

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$n_1 = n_2 = n \Rightarrow v = n_1 + n_2 - 2$$

$$n_1 \neq n_2 \Rightarrow v = \frac{(s_1^2 / n_1 + s_2^2 / n_2)^2}{(s_1^2 / n_1)^2 / (n_1 - 1) + (s_2^2 / n_2)^2 / (n_2 - 1)}$$

Testing the Two Population Means: Independent Samples

Lower-tail test:

$$H_0: \mu_1 \geq \mu_2$$

$$H_1: \mu_1 < \mu_2$$

i.e.,

$$H_0: \mu_1 - \mu_2 \geq 0$$

$$H_1: \mu_1 - \mu_2 < 0$$

Upper-tail test:

$$H_0: \mu_1 \leq \mu_2$$

$$H_1: \mu_1 > \mu_2$$

i.e.,

$$H_0: \mu_1 - \mu_2 \leq 0$$

$$H_1: \mu_1 - \mu_2 > 0$$

Two-tail test:

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

i.e.,

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$

Hypothesis tests for $\mu_1 - \mu_2$

Two Population Means, Independent Samples

Lower-tail test:

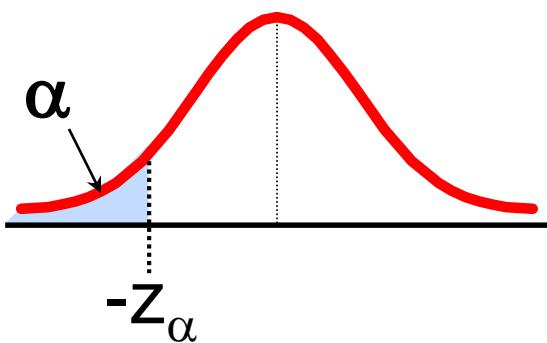
$$\begin{aligned} H_0: \mu_1 - \mu_2 &\geq 0 \\ H_1: \mu_1 - \mu_2 &< 0 \end{aligned}$$

Upper-tail test:

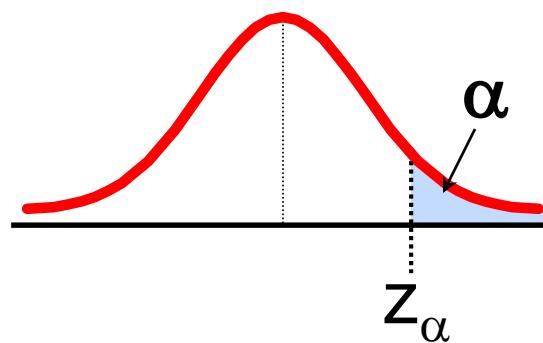
$$\begin{aligned} H_0: \mu_1 - \mu_2 &\leq 0 \\ H_1: \mu_1 - \mu_2 &> 0 \end{aligned}$$

Two-tail test:

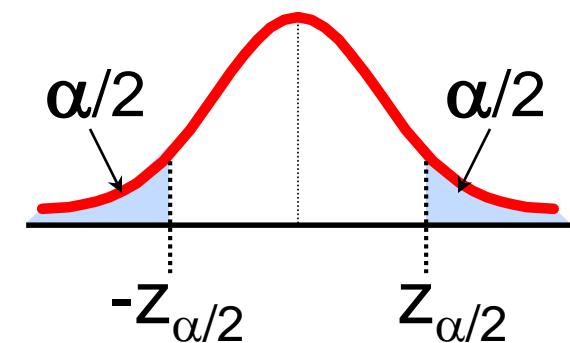
$$\begin{aligned} H_0: \mu_1 - \mu_2 &= 0 \\ H_1: \mu_1 - \mu_2 &\neq 0 \end{aligned}$$



Reject H_0 if $Z < -Z_\alpha$



Reject H_0 if $Z > Z_\alpha$



Reject H_0 if $Z < -Z_{\alpha/2}$
or $Z > Z_{\alpha/2}$

Example

- You are a financial analyst for a brokerage firm. Is there a difference in dividend yield between stocks listed on the NYSE & NASDAQ? You collect the following data:

	<u>NYSE</u>	<u>NASDAQ</u>
Number	21	25
•Sample mean	3.27	2.53
•Sample std dev	1.30	1.16

Assuming both populations are approximately normal with equal variances, is there a difference in average yield ($\alpha = 0.05$)?

Example (cont.)

The test statistic is:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{(3.27 - 2.53) - 0}{\sqrt{1.5021 \left(\frac{1}{21} + \frac{1}{25} \right)}} = 2.040$$

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)} = \frac{(21 - 1)1.30^2 + (25 - 1)1.16^2}{(21 - 1) + (25 - 1)} = 1.5021$$

Example (cont.)

$H_0: \mu_1 - \mu_2 = 0$ i.e. ($\mu_1 = \mu_2$)

$H_1: \mu_1 - \mu_2 \neq 0$ i.e. ($\mu_1 \neq \mu_2$)

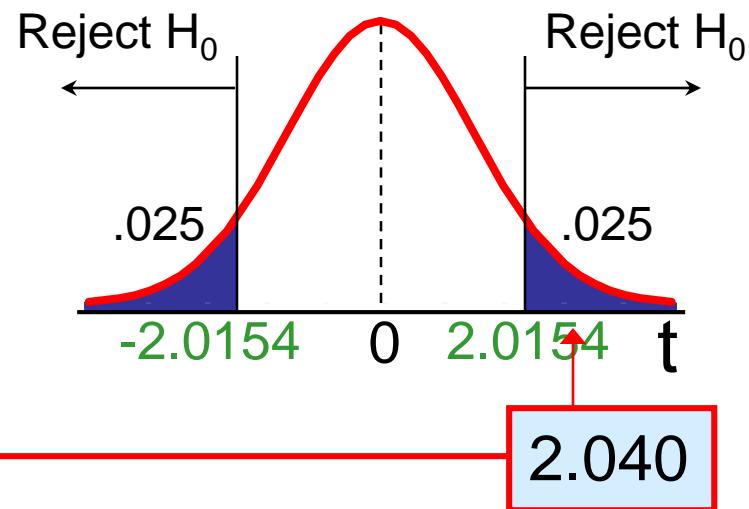
$$\alpha = 0.05$$

$$df = 21 + 25 - 2 = 44$$

Critical Values: $t = \pm 2.0154$

Test Statistic:

$$t = \frac{3.27 - 2.53}{\sqrt{1.5021 \left(\frac{1}{21} + \frac{1}{25} \right)}} = 2.040$$



Decision:

Reject H_0 at $\alpha = 0.05$

Conclusion:

There is evidence of a difference in means.

Testing the Difference Between Two Population Means: Matched Pairs

Large Sample,

Small Sample

$$Z = \frac{\bar{d} - \mu_D}{\frac{s_D}{\sqrt{n}}}$$

$$t = \frac{\bar{d} - \mu_D}{\frac{s_D}{\sqrt{n}}}$$

$$\begin{aligned}H_0 : \mu_1 - \mu_2 &= 0 \\H_a : \mu_1 - \mu_2 &\neq 0\end{aligned}$$

$$\begin{aligned}H_0 : \mu_1 - \mu_2 &= 0 \\H_a : \mu_1 - \mu_2 &> 0 \\&\mu_1 - \mu_2 < 0\end{aligned}$$

Sample Mean

$$\bar{d} = \frac{\sum_{i=1}^n D_i}{n}$$

Sample Standard Deviation

$$S_D = \sqrt{\frac{\sum_{i=1}^n (D_i - \bar{d})^2}{n - 1}}$$

Example

Assume you send your salespeople to a “customer service” training workshop. Has the training made a difference in the number of complaints? You collect the following data:

<u>Salesperson</u>	<u>Number of Complaints:</u>		<u>(2) - (1)</u> <u>Difference, D,</u>
	<u>Before (1)</u>	<u>After (2)</u>	
C.B.	6	4	- 2
T.F.	20	6	-14
M.H.	3	2	- 1
R.K.	0	0	0
M.O.	4	0	- 4 -21

$$\bar{d} = \frac{\sum D_i}{n}$$
$$= -4.2$$

$$S_D = \sqrt{\frac{\sum (D_i - \bar{D})^2}{n-1}}$$
$$= 5.67$$

Example (cont.)

Has the training made a difference in the number of complaints (at the 0.01 level)?

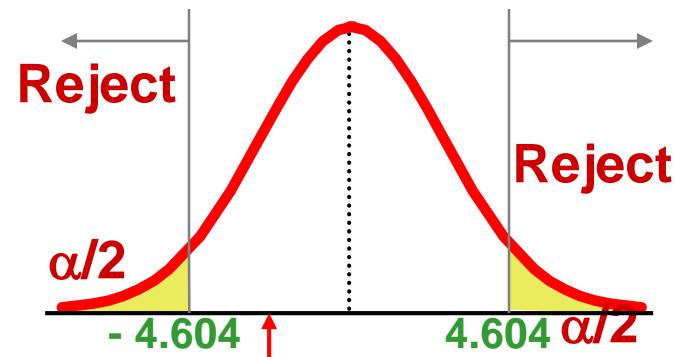
$$\begin{aligned} H_0: \mu_D &= 0 \\ H_1: \mu_D &\neq 0 \end{aligned}$$

$$\alpha = .01 \quad \bar{D} = -4.2$$

$$\begin{aligned} \text{Critical Value} &= \pm 4.604 \\ \text{d.f.} &= n - 1 = 4 \end{aligned}$$

Test Statistic:

$$t = \frac{\bar{D} - \mu_D}{S_D / \sqrt{n}} = \frac{-4.2 - 0}{5.67 / \sqrt{5}} = \boxed{-1.66}$$



Decision: Do not reject H_0
(t stat is not in the reject region)

Conclusion: There is not a significant change in the number of complaints.

Example

- You're a marketing research analyst. You want to compare a client's calculator to a competitor's. You sample 8 retail stores.

At the **.01** level, does your client's calculator sell for ***less than*** their competitor's

<u>Store</u>	(1) <u>Client</u>	(2) <u>Competitor</u>
1	\$ 10	\$ 11
2	8	11
3	7	10
4	9	12
5	11	11
6	10	13
7	9	12
8	8	10

Example

$H_0: \mu_D = 0$ ($\mu_D = \mu_1 - \mu_2$)

$H_a: \mu_D < 0$

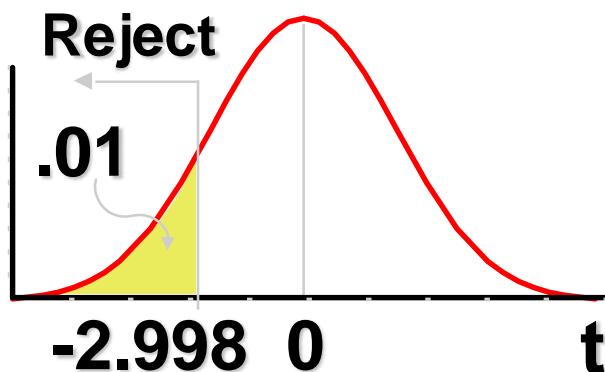
$\alpha = .01$

$df = 8 - 1 = 7$

Critical Value(s):

Test Statistic:

$$t = \frac{\bar{x}_D - D_0}{S_D / \sqrt{n_D}} = \frac{-2.25 - 0}{1.16 / \sqrt{8}} = -5.486$$



Decision:

Reject at $\alpha = .01$

Conclusion:

There Is Evidence
Client's Brand (1) Sells
for Less

Hypothesis Tests for Two Population Proportions

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}, \quad \hat{p}_1 = \frac{x_1}{n_1}, \quad$$

$$\hat{p}_2 = \frac{x_2}{n_2}$$

Lower-tail test:

$$H_0: p_1 \geq p_2$$

$$H_1: p_1 < p_2$$

i.e.,

$$H_0: p_1 - p_2 \geq 0$$

$$H_1: p_1 - p_2 < 0$$

Upper-tail test:

$$H_0: p_1 \leq p_2$$

$$H_1: p_1 > p_2$$

i.e.,

$$H_0: p_1 - p_2 \leq 0$$

$$H_1: p_1 - p_2 > 0$$

Two-tail test:

$$H_0: p_1 = p_2$$

$$H_1: p_1 \neq p_2$$

i.e.,

$$H_0: p_1 - p_2 = 0$$

$$H_1: p_1 - p_2 \neq 0$$

Hypothesis Tests for Two Population Proportions

Population proportions

Lower-tail test:

$$H_0: p_1 - p_2 \geq 0$$

$$H_1: p_1 - p_2 < 0$$

Upper-tail test:

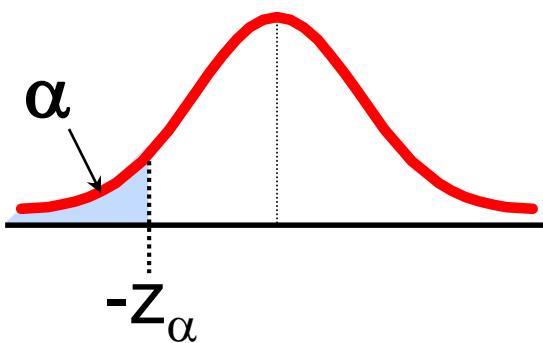
$$H_0: p_1 - p_2 \leq 0$$

$$H_1: p_1 - p_2 > 0$$

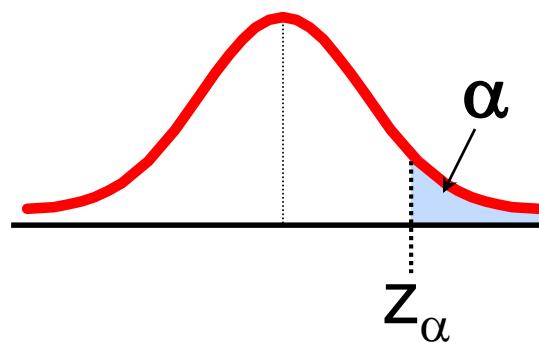
Two-tail test:

$$H_0: p_1 - p_2 = 0$$

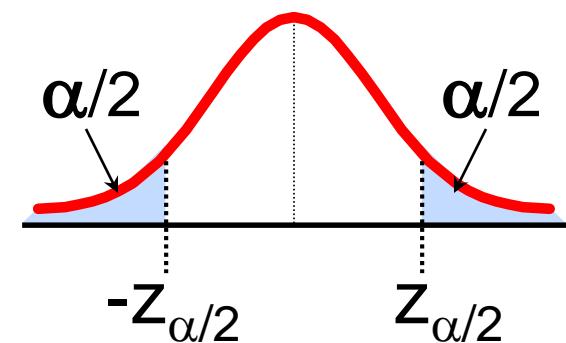
$$H_1: p_1 - p_2 \neq 0$$



Reject H_0 if $Z < -Z_\alpha$



Reject H_0 if $Z > Z_\alpha$



Reject H_0 if $Z < -Z_{\alpha/2}$
or $Z > Z_{\alpha/2}$

Example

Is there a significant difference between the proportion of men and the proportion of women who will vote Yes on Proposition A?

- In a random sample, 36 of 72 men and 31 of 50 women indicated they would vote Yes
- Test at the .05 level of significance

Example (cont.)

- The hypothesis test is:

$H_0: p_1 - p_2 = 0$ (the two proportions are equal)

$H_1: p_1 - p_2 \neq 0$ (there is a significant difference between proportions)

- The sample proportions are:

– Men: $\hat{p}_1 = 36/72 = .50$

– Women: $\hat{p}_2 = 31/50 = .62$

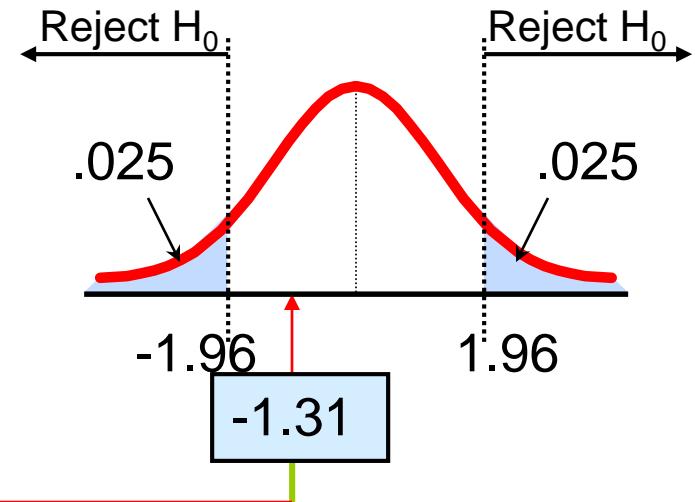
- The pooled estimate for the overall proportion is:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{36 + 31}{72 + 50} = \frac{67}{122} = .549$$

Example (cont.)

The test statistic for $p_1 - p_2$ is:

$$\begin{aligned} z &= \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \\ &= \frac{(.50 - .62) - (0)}{\sqrt{.549(1-.549)\left(\frac{1}{72} + \frac{1}{50}\right)}} = \boxed{-1.31} \end{aligned}$$



Critical Values = ± 1.96
For $\alpha = .05$

Decision: Do not reject H_0
Conclusion: There is not significant evidence of a difference in proportions who will vote yes between men and women.

Testing a Population Variance

The test statistic

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

$$H_0 : \sigma^2 = \sigma_0^2$$

$$H_a : \sigma^2 \neq \sigma_0^2$$

$$H_0 : \sigma^2 = \sigma_0^2$$

$$H_a : \sigma^2 > \sigma_0^2$$

$$\sigma^2 < \sigma_0^2$$

Example

- Consider a container filling machine. Management wants a machine to fill 1 liter (1,000 cc's) so that the variance of the fills is less than 1 cc². A random sample of n=25 1 liter fills were taken. Does the machine perform as it should at the 5% significance level?
- We want to show that:

$$H_1: \sigma^2 < 1$$

Variance is less than 1 cc²

(so our null hypothesis becomes: $H_0: \sigma^2 = 1$). We will use this test statistic:

$$\chi^2 = \frac{(n - 1)s^2}{\sigma^2}$$

Example (cont.)

- Since our alternative hypothesis is phrased as:

$$H_a: \sigma^2 < 1$$

- We will reject H_0 in favor of H_1 if our test statistic falls into this rejection region:

$$\chi^2 < \chi^2_{1-\alpha, n-1} = \chi^2_{1-.05, 25-1} = \chi^2_{.95, 24} = 13.8484$$

- We compute the sample variance to be: $s^2 = .8088$
- And thus our test statistic takes on this value...

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(25-1)(.8088)}{1} = 19.41$$

compare

There is not enough evidence to infer that the claim is true.

$$\chi^2 = 19.41 < \chi^2_{1-\alpha, n-1} = 13.8484$$

Testing the Ratio of Two Populations Variances

$$H_0: \sigma_1^2 / \sigma_2^2 = 1$$
$$H_1: \sigma_1^2 / \sigma_2^2 \neq 1$$

Two-tail test

One-tail test

$$H_0: \sigma_1^2 / \sigma_2^2 = 1$$
$$H_1: \sigma_1^2 / \sigma_2^2 > 1$$

Covers both
Read book
pages 379 - 382

Lower-tail test

$$H_0: \sigma_1^2 \geq \sigma_2^2$$
$$H_1: \sigma_1^2 < \sigma_2^2$$

Upper-tail test

$$H_0: \sigma_1^2 \leq \sigma_2^2$$
$$H_1: \sigma_1^2 > \sigma_2^2$$

The F test statistic is (two-tail) :

$$F = \frac{s_1^2}{s_2^2}$$

if $s_1^2 > s_2^2$

$$F = \frac{s_2^2}{s_1^2}$$

if $s_2^2 > s_1^2$

For one-tail test, always perform Upper tail test with adjustment on making proper H_1 :

e.g. $H_1: \sigma_1^2 < \sigma_2^2$ should be changed to $H_1: \sigma_2^2 / \sigma_1^2 > 1$

Example

You are a financial analyst for a brokerage firm. You want to compare dividend yields between stocks listed on the NYSE & NASDAQ. You collect the following data:

	<u>NYSE</u>	<u>NASDAQ</u>
Number	21	25
Mean	3.27	2.53
Std dev	1.30	1.16

Is there a difference in the variances between the NYSE & NASDAQ at the $\alpha = 0.05$ level?

F Test: Example Solution

- Form the hypothesis test:

$$H_0: \sigma^2_1/\sigma^2_2 = 1 \quad (\text{there is no difference between variances})$$
$$H_1: \sigma^2_1/\sigma^2_2 \neq 1 \quad (\text{there is a difference between variances})$$

Find the F critical values for $\alpha = 0.05$:

Here, NYSE s > NASDAQ s ,

– Numerator:

$$\bullet n_1 - 1 = 21 - 1 = 20 \text{ d.f.}$$

– Denominator:

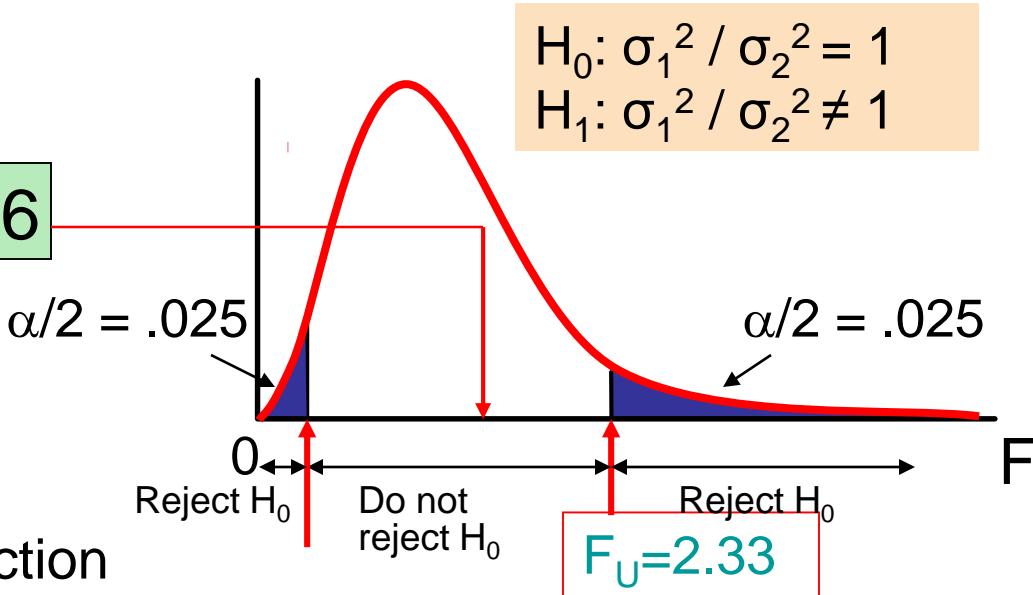
$$\bullet n_2 - 1 = 25 - 1 = 24 \text{ d.f.}$$

$$F_U = F_{.025, 20, 24} = 2.33$$

Example (cont.)

The test statistic is:

$$F = \frac{S_1^2}{S_2^2} = \frac{1.30^2}{1.16^2} = 1.256$$



- $F = 1.256$ is not in the rejection region, so we **do not reject H_0**
- **Conclusion:** There is not sufficient evidence of a difference in variances at $\alpha = .05$

Foundation of Data Science and Analytics

Simple Linear Regression

Arun K. Timalsina

11-1: Empirical Models

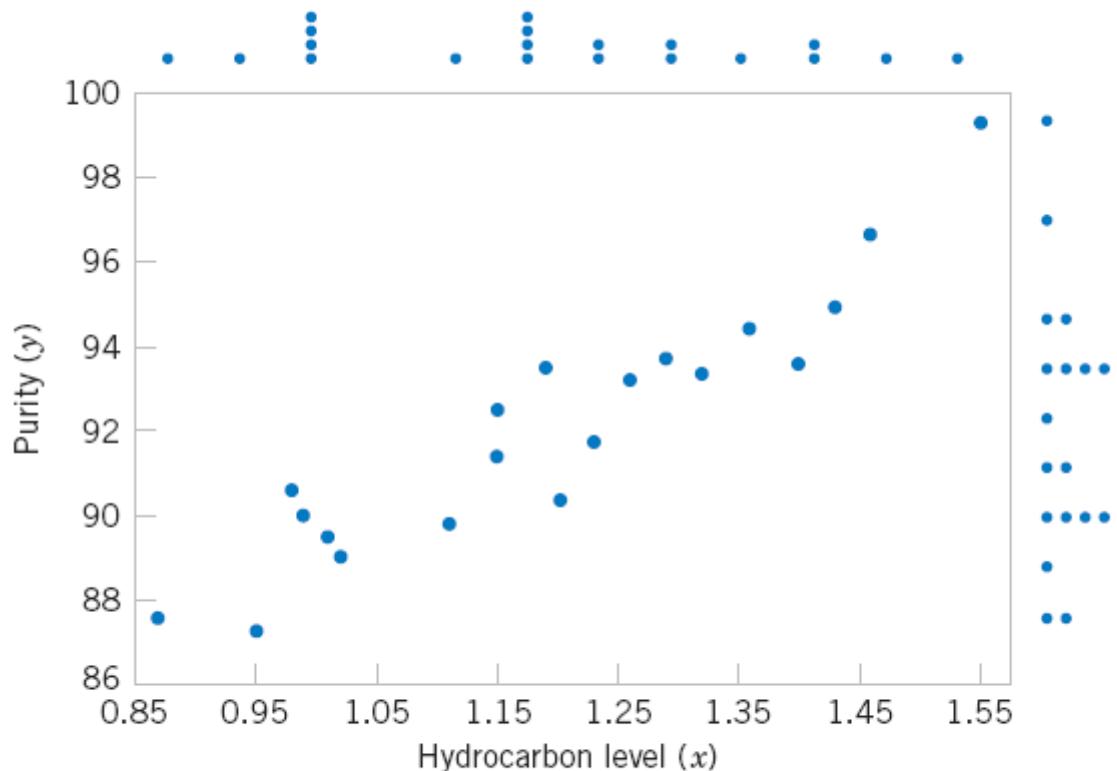
- Many problems in engineering and science involve exploring the relationships between two or more variables.
- **Regression analysis** is a statistical technique that is very useful for these types of problems.
- For example, in a chemical process, suppose that the yield of the product is related to the process-operating temperature.
- Regression analysis can be used to build a model to predict yield at a given temperature level.

11-1: Empirical Models

Table 11-1 Oxygen and Hydrocarbon Levels

Observation Number	Hydrocarbon Level x (%)	Purity y (%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33

11-1: Empirical Models



11-1: Empirical Models

Based on the scatter diagram, it is probably reasonable to assume that the mean of the random variable Y is related to x by the following straight-line relationship:

$$E(Y|x) = \mu_{Y|x} = \beta_0 + \beta_1 x$$

where the slope and intercept of the line are called **regression coefficients**.
The **simple linear regression model** is given by

$$Y = \beta_0 + \beta_1 x + \epsilon$$

where ϵ is the random error term.

11-1: Empirical Models

We think of the regression model as an empirical model. Suppose that the mean and variance of ϵ are 0 and σ^2 , respectively, then

$$E(Y|x) = E(\beta_0 + \beta_1x + \epsilon) = \beta_0 + \beta_1x + E(\epsilon) = \beta_0 + \beta_1x$$

The variance of Y given x is

$$V(Y|x) = V(\beta_0 + \beta_1x + \epsilon) = V(\beta_0 + \beta_1x) + V(\epsilon) = 0 + \sigma^2 = \sigma^2$$

11-1: Empirical Models

- The true regression model is a line of mean values:

$$\mu_{Y|x} = \beta_0 + \beta_1 x$$

where β_1 can be interpreted as the change in the mean of Y for a unit change in x .

- Also, the variability of Y at a particular value of x is determined by the error variance, σ^2 .
- This implies there is a distribution of Y -values at each x and that the variance of this distribution is the same at each x .

11-1: Empirical Models

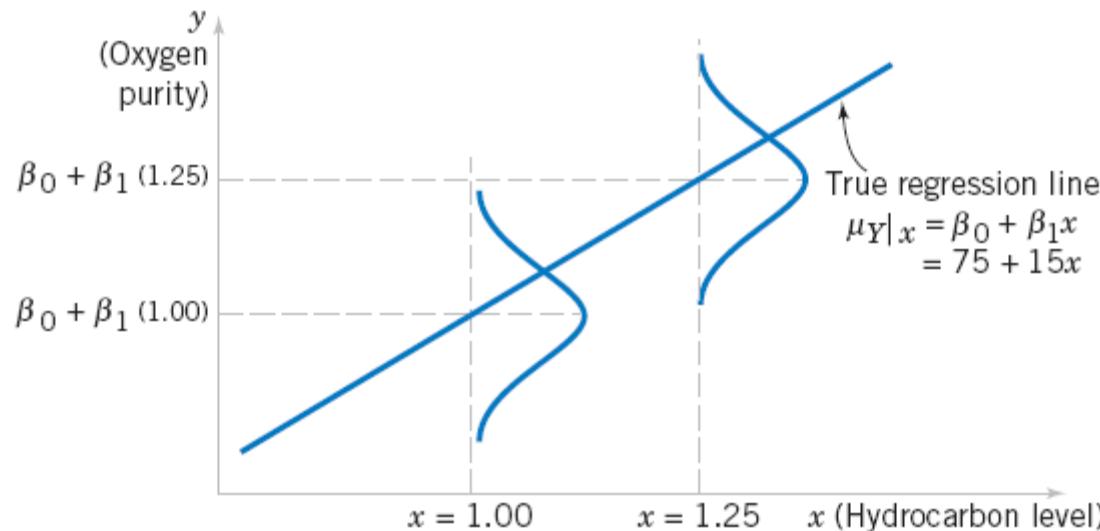


Figure 11-2 The distribution of Y for a given value of x for the oxygen purity–hydrocarbon data.

Figure 11-2 The distribution of Y for a given value of x for the oxygen purity–hydrocarbon data.

11-2: Simple Linear Regression

- The case of **simple linear regression** considers a single **regressor** or **predictor** x and a **dependent** or **response variable** Y .
- The expected value of Y at each level of x is a random variable:

$$E(Y|x) = \beta_0 + \beta_1 x$$

- We assume that each observation, Y , can be described by the model

$$Y = \beta_0 + \beta_1 x + \epsilon$$

11-2: Simple Linear Regression

- Suppose that we have n pairs of observations $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$.

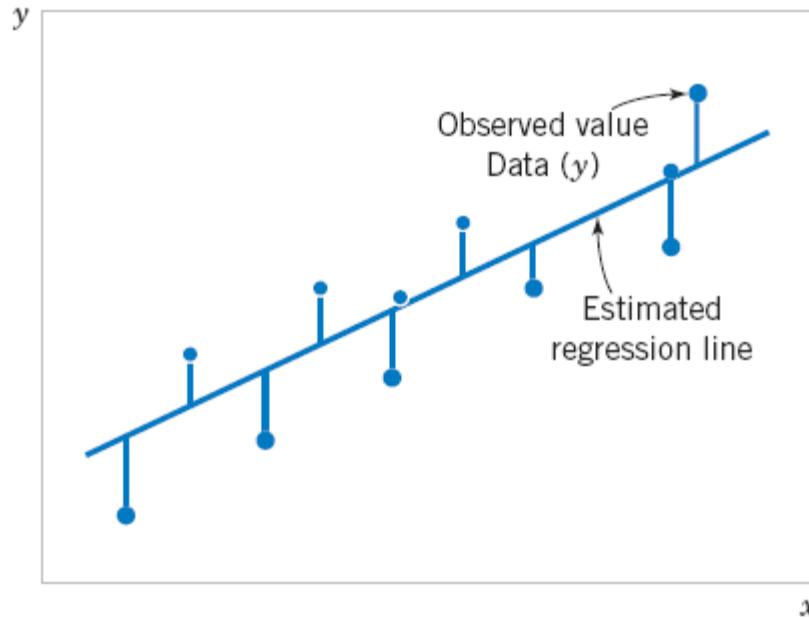


Figure 11-3 Deviations of the data from the estimated regression model.

Figure 11-3 Deviations of the data from the estimated regression model.

11-2: Simple Linear Regression

- The **method of least squares** is used to estimate the parameters, β_0 and β_1 by minimizing the sum of the squares of the vertical deviations in Figure 11-3.

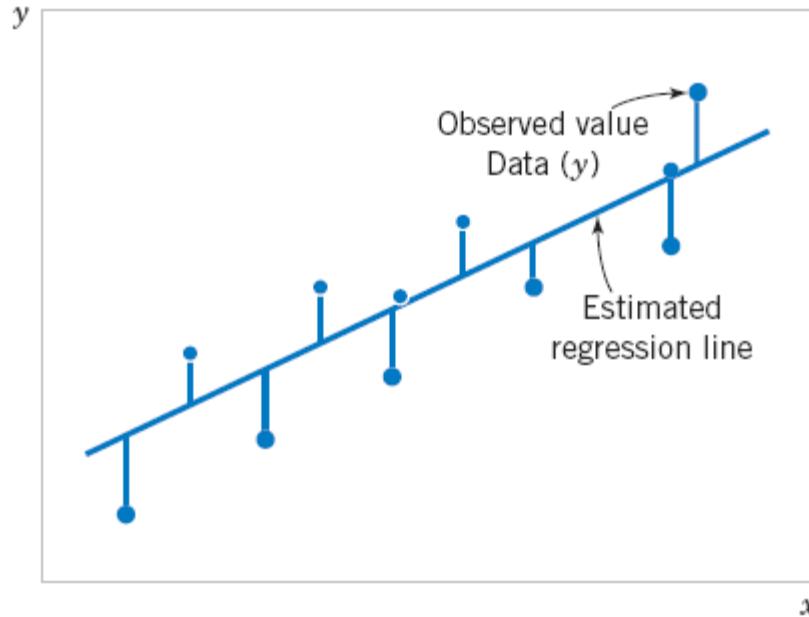


Figure 11-3 Deviations of the data from the estimated regression model.

Figure 11-3 Deviations of the data from the estimated regression model.

11-2: Simple Linear Regression

- Using Equation 11-2, the n observations in the sample can be expressed as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \dots, n$$

- The sum of the squares of the deviations of the observations from the true regression line is

$$L = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

11-2: Simple Linear Regression

$$L = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

The least squares estimators of β_0 and β_1 , say, $\hat{\beta}_0$ and $\hat{\beta}_1$, must satisfy

$$\frac{\partial L}{\partial \beta_0} \Big|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\frac{\partial L}{\partial \beta_1} \Big|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

11-2: Simple Linear Regression

Simplifying these two equations yields

$$\begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_i &= \sum_{i=1}^n y_i \\ \hat{\beta}_0 \sum_{i=1}^n x_i + \hat{\beta}_1 \sum_{i=1}^n x_i^2 &= \sum_{i=1}^n y_i x_i \end{aligned} \tag{11-6}$$

Equations 11-6 are called the **least squares normal equations**. The solution to the normal equations results in the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$.

11-2: Simple Linear Regression

Definition

The **least squares estimates** of the intercept and slope in the simple linear regression model are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (11-7)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n y_i x_i - \left(\sum_{i=1}^n y_i \right) \left(\sum_{i=1}^n x_i \right)}{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i \right)^2}{n}} \quad (11-8)$$

where $\bar{y} = (1/n) \sum_{i=1}^n y_i$ and $\bar{x} = (1/n) \sum_{i=1}^n x_i$.

11-2: Simple Linear Regression

The **fitted or estimated regression line** is therefore

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad (11-9)$$

Note that each pair of observations satisfies the relationship

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i, \quad i = 1, 2, \dots, n$$

where $e_i = y_i - \hat{y}_i$ is called the **residual**. The residual describes the error in the fit of the model to the i th observation y_i . Later in this chapter we will use the residuals to provide information about the adequacy of the fitted model.

11-2: Simple Linear Regression

Notation

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}$$

$$S_{xy} = \sum_{i=1}^n y_i(x_i - \bar{x})^2 = \sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n}$$

11-2: Simple Linear Regression

Example 11-1

EXAMPLE 11-1 Oxygen Purity

We will fit a simple linear regression model to the oxygen purity data in Table 11-1. The following quantities may be computed:

$$n = 20 \quad \sum_{i=1}^{20} x_i = 23.92 \quad \sum_{i=1}^{20} y_i = 1,843.21$$

$$\bar{x} = 1.1960 \quad \bar{y} = 92.1605$$

$$\sum_{i=1}^{20} y_i^2 = 170,044.5321 \quad \sum_{i=1}^{20} x_i^2 = 29.2892$$

$$\sum_{i=1}^{20} x_i y_i = 2,214.6566$$

$$S_{xx} = \sum_{i=1}^{20} x_i^2 - \frac{\left(\sum_{i=1}^{20} x_i\right)^2}{20} = 29.2892 - \frac{(23.92)^2}{20} \\ = 0.68088$$

and

$$S_{xy} = \sum_{i=1}^{20} x_i y_i - \frac{\left(\sum_{i=1}^{20} x_i\right)\left(\sum_{i=1}^{20} y_i\right)}{20} \\ = 2,214.6566 - \frac{(23.92)(1,843.21)}{20} = 10.17744$$

11-2: Simple Linear Regression

Example 11-1

Therefore, the least squares estimates of the slope and intercept are

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{10.17744}{0.68088} = 14.94748$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 92.1605 - (14.94748)1.196 = 74.28331$$

The fitted simple linear regression model (with the coefficients reported to three decimal places) is

$$\hat{y} = 74.283 + 14.947x$$

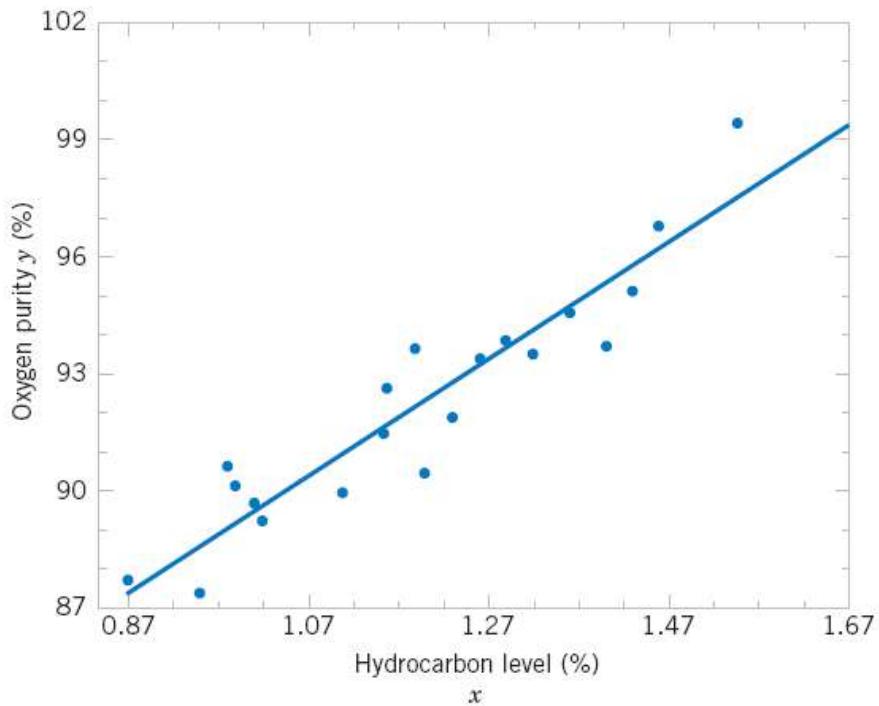
This model is plotted in Fig. 11-4, along with the sample data.

11-2: Simple Linear Regression

Example 11-1

Figure 11-4 Scatter plot of oxygen purity y versus hydrocarbon level x and regression model $\hat{y} = 74.20 + 14.97x$.

Figure 11-4 Scatter plot of oxygen purity y versus hydrocarbon level x and regression model $\hat{y} = 74.283 + 14.947x$.



11-2: Simple Linear Regression

Example 11-1

Computer software programs are widely used in regression modeling. These programs typically carry more decimal places in the calculations. Table 11-2 shows a portion of the output from Minitab for this problem. The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are highlighted. In subsequent sections we will provide explanations for the information provided in this computer output.

11-1: Empirical Models

Table 11-1 Oxygen and Hydrocarbon Levels

Observation Number	Hydrocarbon Level x (%)	Purity y (%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33

Table 11-2 Minitab Output for the Oxygen Purity Data in Example 11-1

Regression Analysis

The regression equation is

$$\text{Purity} = 74.3 + 14.9 \text{ HC Level}$$

Predictor	Coef	SE Coef	T	P
Constant	74.283 $\leftarrow \hat{\beta}_0$	1.593	46.62	0.000
HC Level	14.947 $\leftarrow \hat{\beta}_1$	1.317	11.35	0.000

$$S = 1.087$$

$$R\text{-Sq} = 87.7\%$$

$$R\text{-Sq (adj)} = 87.1\%$$

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	152.13	152.13	128.86	0.000
Residual Error	18	21.25 $\leftarrow SS_E$	1.18 $\leftarrow \hat{\sigma}^2$		
Total	19	173.38			

Predicted Values for New Observations

New Obs	Fit	SE Fit	95.0% CI	95.0% PI
1	89.231	0.354	(88.486, 89.975)	(86.830, 91.632)

Values of Predictors for New Observations

New Obs	HC Level
1	1.00

11-2: Simple Linear Regression

Estimating σ^2

The error sum of squares is

$$SS_E = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

It can be shown that the expected value of the error sum of squares is $E(SS_E) = (n - 2)\sigma^2$.

11-2: Simple Linear Regression

Estimating σ^2

An **unbiased estimator** of σ^2 is

$$\hat{\sigma}^2 = \frac{SS_E}{n - 2} \quad (11-13)$$

where SS_E can be easily computed using

$$SS_E = SS_T - \hat{\beta}_1 S_{xy} \quad (11-14)$$

11-3: Properties of the Least Squares Estimators

- Slope Properties

$$E(\hat{\beta}_1) = \beta_1 \quad V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

- Intercept Properties

$$E(\hat{\beta}_0) = \beta_0 \quad \text{and} \quad V(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]$$

11-1: Empirical Models

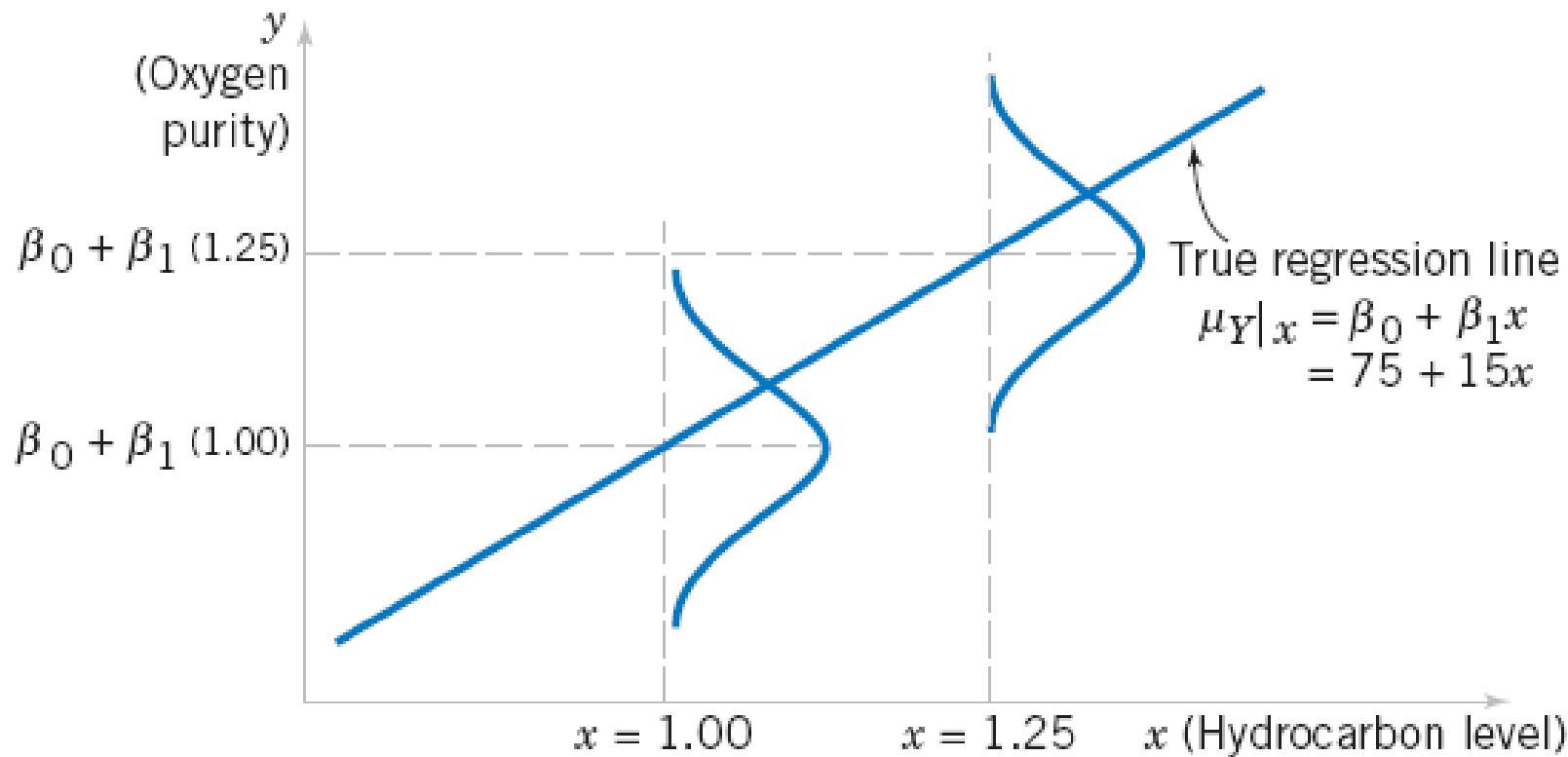


Figure 11-2 The distribution of Y for a given value of x for the oxygen purity–hydrocarbon data.

11-4: Hypothesis Tests in Simple Linear Regression

11-4.1 Use of *t*-Tests

Suppose we wish to test

$$H_0: \beta_1 = \beta_{1,0}$$

$$H_1: \beta_1 \neq \beta_{1,0}$$

An appropriate test statistic would be

$$T_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2 / S_{xx}}}$$

11-4: Hypothesis Tests in Simple Linear Regression

11-4.1 Use of *t*-Tests

The test statistic could also be written as:

$$T_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{se(\hat{\beta}_1)}$$

We would reject the null hypothesis if

$$|t_0| > t_{\alpha/2,n-2}$$

11-4: Hypothesis Tests in Simple Linear Regression

11-4.1 Use of *t*-Tests

Suppose we wish to test

$$H_0: \beta_0 = \beta_{0,0}$$

$$H_1: \beta_0 \neq \beta_{0,0}$$

An appropriate test statistic would be

$$T_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]}} = \frac{\hat{\beta}_0 - \beta_{0,0}}{se(\hat{\beta}_0)}$$

11-4: Hypothesis Tests in Simple Linear Regression

11-4.1 Use of t -Tests

We would reject the null hypothesis if

$$|t_0| > t_{\alpha/2, n-2}$$

11-4: Hypothesis Tests in Simple Linear Regression

11-4.1 Use of *t*-Tests

An important special case of the hypotheses of Equation 11-18 is

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

These hypotheses relate to the **significance of regression**. *Failure* to reject H_0 is equivalent to concluding that there is no linear relationship between x and Y .

11-4: Hypothesis Tests in Simple Linear Regression

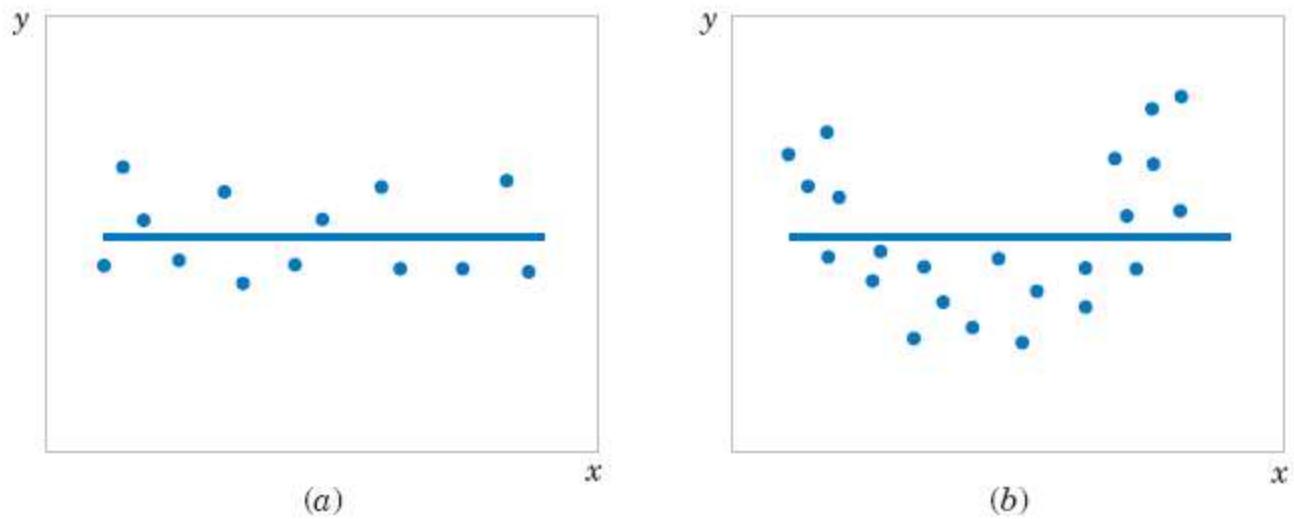


Figure 11-5 The hypothesis $H_0: \beta_1 = 0$ is not rejected.

Figure 11-5 The hypothesis $H_0: \beta_1 = 0$ is not rejected.

11-4: Hypothesis Tests in Simple Linear Regression

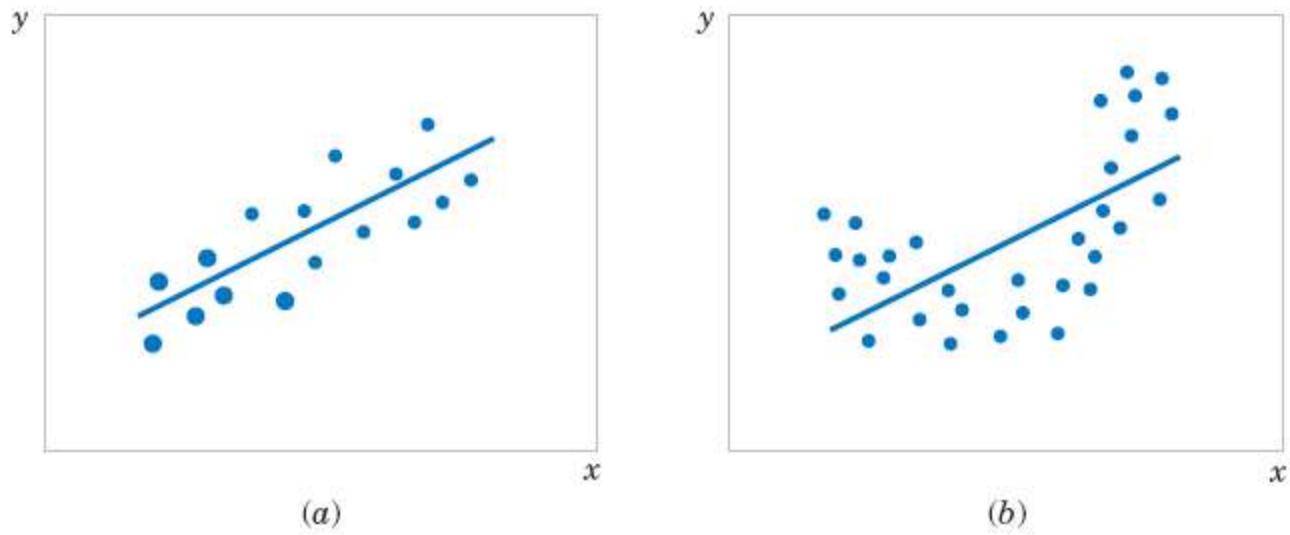


Figure 11-6 The hypothesis $H_0: \beta_1 = 0$ is rejected.

Figure 11-6 The hypothesis $H_0: \beta_1 = 0$ is rejected.

11-4: Hypothesis Tests in Simple Linear Regression

Example 11-2

EXAMPLE 11-2 Oxygen Purity Tests of Coefficients

We will test for significance of regression using the model for the oxygen purity data from Example 11-1. The hypotheses are

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

and we will use $\alpha = 0.01$. From Example 11-1 and Table 11-2 we have

$$\hat{\beta}_1 = 14.947 \quad n = 20, \quad S_{xx} = 0.68088, \quad \hat{\sigma}^2 = 1.18$$

so the t -statistic in Equation 10-20 becomes

$$t_0 = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/S_{xx}}} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{14.947}{\sqrt{1.18/0.68088}} = 11.35$$

Practical Interpretation: Since the reference value of t is $t_{0.005,18} = 2.88$, the value of the test statistic is very far into the critical region, implying that $H_0: \beta_1 = 0$ should be rejected. There is strong evidence to support this claim. The P -value for this test is $P \approx 1.23 \times 10^{-9}$. This was obtained manually with a calculator.

Table 11-2 presents the Minitab output for this problem. Notice that the t -statistic value for the slope is computed as 11.35 and that the reported P -value is $P = 0.000$. Minitab also reports the t -statistic for testing the hypothesis $H_0: \beta_0 = 0$. This statistic is computed from Equation 11-22, with $\beta_{0,0} = 0$, as $t_0 = 46.62$. Clearly, then, the hypothesis that the intercept is zero is rejected.

11-4: Hypothesis Tests in Simple Linear Regression

11-4.2 Analysis of Variance Approach to Test Significance of Regression

The analysis of variance identity is

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (11-24)$$

Symbolically,

$$SS_T = SS_R + SS_E \quad (11-25)$$

11-4: Hypothesis Tests in Simple Linear Regression

11-4.2 Analysis of Variance Approach to Test Significance of Regression

If the null hypothesis, $H_0: \beta_1 = 0$ is true, the statistic

$$F_0 = \frac{SS_R/1}{SS_E/(n - 2)} = \frac{MS_R}{MS_E} \quad (11-26)$$

follows the $F_{1,n-2}$ distribution and we would reject if $f_0 > f_{\alpha,1,n-2}$.

11-4: Hypothesis Tests in Simple Linear Regression

11-4.2 Analysis of Variance Approach to Test Significance of Regression

The quantities, MS_R and MS_E are called **mean squares**.
Analysis of variance table:

Table 11-3 Analysis of Variance for Testing Significance of Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	$SS_R = \hat{\beta}_1 S_{xy}$	1	MS_R	MS_R/MS_E
Error	$SS_E = SS_T - \hat{\beta}_1 S_{xy}$	$n - 2$	MS_E	
Total	SS_T	$n - 1$		

Note that $MS_E = \hat{\sigma}^2$.

11-4: Hypothesis Tests in Simple Linear Regression

Example 11-3

EXAMPLE 11-3 Oxygen Purity ANOVA

We will use the analysis of variance approach to test for significance of regression using the oxygen purity data model from Example 11-1. Recall that $SS_T = 173.38$, $\hat{\beta}_1 = 14.947$, $S_{xy} = 10.17744$, and $n = 20$. The regression sum of squares is

$$SS_R = \hat{\beta}_1 S_{xy} = (14.947)10.17744 = 152.13$$

and the error sum of squares is

$$SS_E = SS_T - SS_R = 173.38 - 152.13 = 21.25$$

The analysis of variance for testing $H_0: \beta_1 = 0$ is summarized in the Minitab output in Table 11-2. The test statistic is $f_0 = MS_R/MS_E = 152.13/1.18 = 128.86$, for which we find that the P -value is $P = 1.23 \times 10^{-9}$, so we conclude that β_1 is not zero.

There are frequently minor differences in terminology among computer packages. For example, sometimes the regression sum of squares is called the “model” sum of squares, and the error sum of squares is called the “residual” sum of squares.

11-4: Hypothesis Tests in Simple Linear Regression

Note that the analysis of variance procedure for testing for significance of regression is equivalent to the t -test in Section 11-5.1. That is, either procedure will lead to the same conclusions. This is easy to demonstrate by starting with the t -test statistic in Equation 11-19 with $\beta_{1,0} = 0$, say

$$T_0 = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/S_{xx}}} \quad (11-27)$$

Squaring both sides of Equation 11-27 and using the fact that $\hat{\sigma}^2 = MS_E$ results in

$$T_0^2 = \frac{\hat{\beta}_1^2 S_{xx}}{MS_E} = \frac{\hat{\beta}_1 S_{xY}}{MS_E} = \frac{MS_R}{MS_E} \quad (11-28)$$

Note that T_0^2 in Equation 11-28 is identical to F_0 in Equation 11-26. It is true, in general, that the square of a t random variable with v degrees of freedom is an F random variable, with one and v degrees of freedom in the numerator and denominator, respectively. Thus, the test using T_0 is equivalent to the test based on F_0 . Note, however, that the t -test is somewhat more flexible in that it would allow testing against a one-sided alternative hypothesis, while the F -test is restricted to a two-sided alternative.

11-5: Confidence Intervals

11-5.1 Confidence Intervals on the Slope and Intercept

Definition

Under the assumption that the observations are normally and independently distributed, a $100(1 - \alpha)\%$ **confidence interval on the slope** β_1 in simple linear regression is

$$\hat{\beta}_1 - t_{\alpha/2, n-2} \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2, n-2} \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \quad (11-29)$$

Similarly, a $100(1 - \alpha)\%$ **confidence interval on the intercept** β_0 is

$$\begin{aligned} \hat{\beta}_0 - t_{\alpha/2, n-2} \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]} \\ \leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2, n-2} \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]} \end{aligned} \quad (11-30)$$

11-5: Confidence Intervals

Example 11-4

EXAMPLE 11-4 Oxygen Purity Confidence Interval on the Slope

We will find a 95% confidence interval on the slope of the regression line using the data in Example 11-1. Recall that $\hat{\beta}_1 = 14.947$, $S_{xx} = 0.68088$, and $\hat{\sigma}^2 = 1.18$ (see Table 11-2). Then, from Equation 11-29 we find

$$\hat{\beta}_1 - t_{0.025,18} \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{0.025,18} \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

or

$$14.947 - 2.101 \sqrt{\frac{1.18}{0.68088}} \leq \beta_1 \leq 14.947 + 2.101 \sqrt{\frac{1.18}{0.68088}}$$

This simplifies to

$$12.181 \leq \beta_1 \leq 17.713$$

Practical Interpretation: This CI does not include zero, so there is strong evidence (at $\alpha = 0.05$) that the slope is not zero. The CI is reasonably narrow (± 2.766) because the error variance is fairly small.

11-5: Confidence Intervals

11-5.2 Confidence Interval on the Mean Response

$$\hat{\mu}_{Y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

Definition

A $100(1 - \alpha)\%$ **confidence interval about the mean response** at the value of $x = x_0$, say $\mu_{Y|x_0}$, is given by

$$\begin{aligned}\hat{\mu}_{Y|x_0} - t_{\alpha/2, n-2} \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]} \\ \leq \mu_{Y|x_0} \leq \hat{\mu}_{Y|x_0} + t_{\alpha/2, n-2} \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]}\end{aligned}\quad (11-31)$$

where $\hat{\mu}_{Y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$ is computed from the fitted regression model.

11-5: Confidence Intervals

Example 11-5

EXAMPLE 11-5 Oxygen Purity Confidence Interval on the Mean Response

We will construct a 95% confidence interval about the mean response for the data in Example 11-1. The fitted model is $\hat{\mu}_{Y|x_0} = 74.283 + 14.947x_0$, and the 95% confidence interval on $\mu_{Y|x_0}$ is found from Equation 11-31 as

$$\hat{\mu}_{Y|x_0} \pm 2.101 \sqrt{1.18 \left[\frac{1}{20} + \frac{(x_0 - 1.1960)^2}{0.68088} \right]}$$

Suppose that we are interested in predicting mean oxygen purity when $x_0 = 1.00\%$. Then

$$\hat{\mu}_{Y|x_{1.00}} = 74.283 + 14.947(1.00) = 89.23$$

and the 95% confidence interval is

$$89.23 \pm 2.101 \sqrt{1.18 \left[\frac{1}{20} + \frac{(1.00 - 1.1960)^2}{0.68088} \right]}$$

11-5: Confidence Intervals

Example 11-5

or

$$89.23 \pm 0.75$$

Therefore, the 95% CI on $\mu_{Y|1.00}$ is

$$88.48 \leq \mu_{Y|1.00} \leq 89.98$$

This is a reasonable narrow CI.

Minitab will also perform these calculations. Refer to Table 11-2. The predicted value of y at $x = 1.00$ is shown along with the 95% CI on the mean of y at this level of x .

11-5: Confidence Intervals

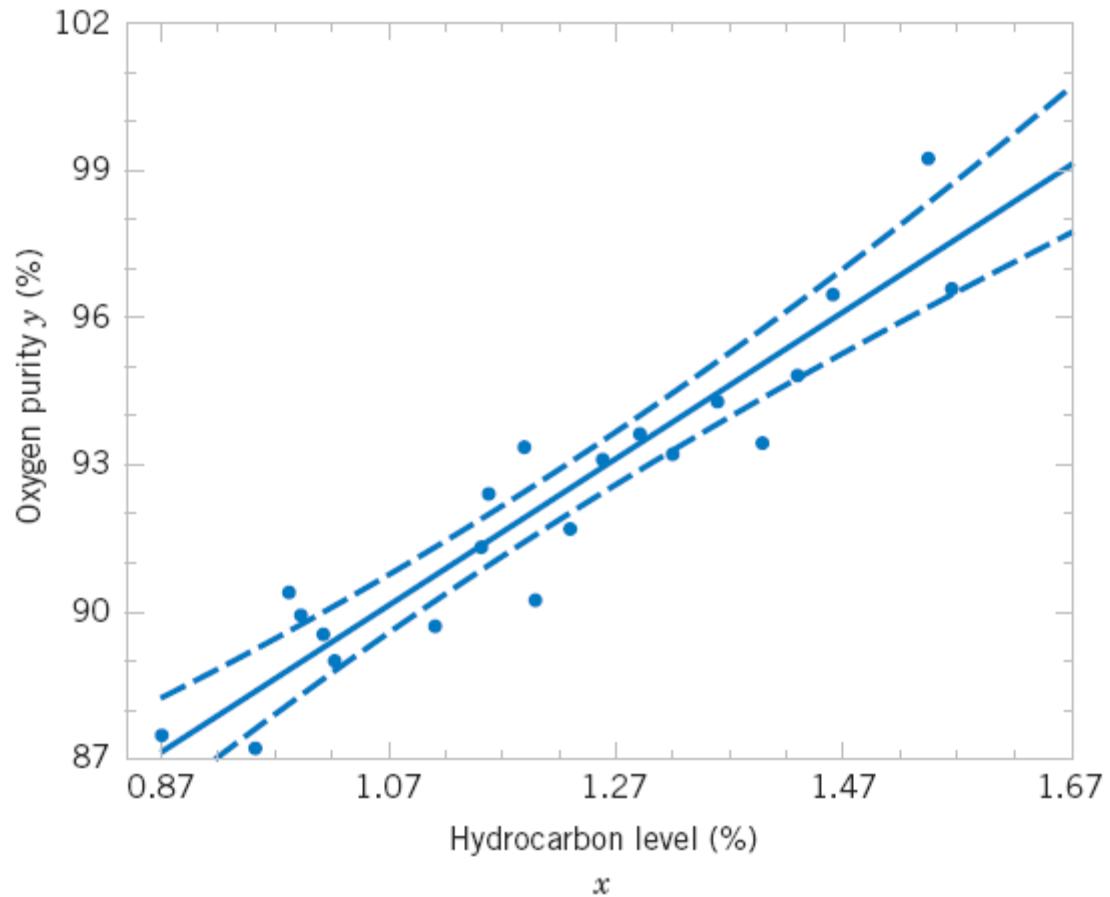
Example 11-5

By repeating these calculations for several different values for x_0 , we can obtain confidence limits for each corresponding value of $\mu_{Y|x_0}$. Figure 11-7 displays the scatter diagram with the fitted model and the corresponding 95% confidence limits plotted as the upper and lower lines. The 95% confidence level applies only to the interval obtained at one value of x and not to the entire set of x -levels. Notice that the width of the confidence interval on $\mu_{Y|x_0}$ increases as $|x_0 - \bar{x}|$ increases.

11-5: Confidence Intervals

Figure 11-7

Figure 11-7 Scatter diagram of oxygen purity data from Example 11-1 with fitted regression line and 95 percent confidence limits on $\mu_{Y|x_0}$.



11-6: Prediction of New Observations

If x_0 is the value of the regressor variable of interest,

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

is the point estimator of the new or future value of the response, Y_0 .

11-6: Prediction of New Observations

Definition

A $100(1 - \alpha)$ % prediction interval on a future observation Y_0 at the value x_0 is given by

$$\begin{aligned}\hat{y}_0 - t_{\alpha/2, n-2} \sqrt{\hat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]} \\ \leq Y_0 \leq \hat{y}_0 + t_{\alpha/2, n-2} \sqrt{\hat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]}\end{aligned}\quad (11-33)$$

The value \hat{y}_0 is computed from the regression model $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

11-6: Prediction of New Observations

Example 11-6

EXAMPLE 11-6 Oxygen Purity Prediction Interval

To illustrate the construction of a prediction interval, suppose we use the data in Example 11-1 and find a 95% prediction interval on the next observation of oxygen purity at $x_0 = 1.00\%$. Using Equation 11-33 and recalling from Example 11-5 that $\hat{y}_0 = 89.23$, we find that the prediction interval is

$$\begin{aligned} & 89.23 - 2.101 \sqrt{1.18 \left[1 + \frac{1}{20} + \frac{(1.00 - 1.1960)^2}{0.68088} \right]} \\ & \leq Y_0 \leq 89.23 + 2.101 \sqrt{1.18 \left[1 + \frac{1}{20} + \frac{(1.00 - 1.1960)^2}{0.68088} \right]} \end{aligned}$$

11-6: Prediction of New Observations

Example 11-6

which simplifies to

$$86.83 \leq y_0 \leq 91.63$$

This is a reasonably narrow prediction interval.

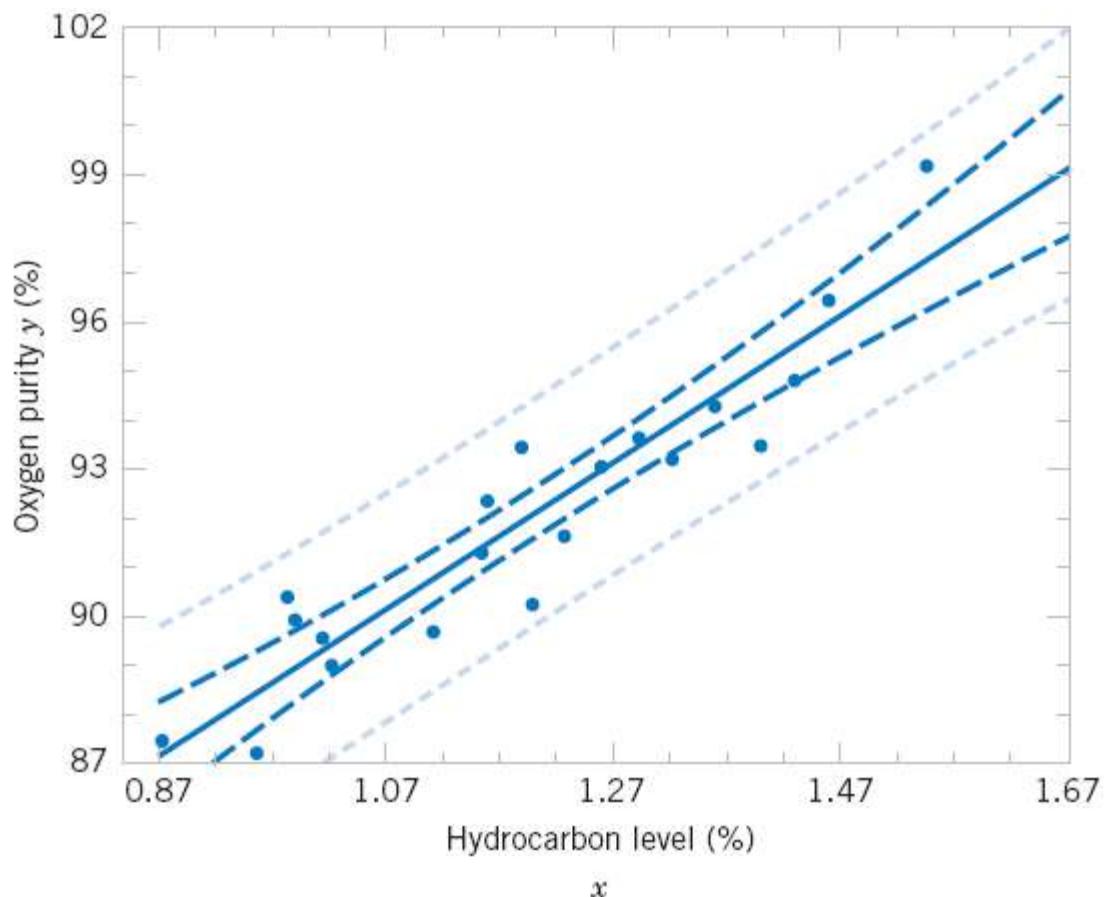
Minitab will also calculate prediction intervals. Refer to the output in Table 11-2. The 95% PI on the future observation at $x_0 = 1.00$ is shown in the display.

By repeating the foregoing calculations at different levels of x_0 , we may obtain the 95% prediction intervals shown graphically as the lower and upper lines about the fitted regression model in Fig. 11-8. Notice that this graph also shows the 95% confidence limits on $\mu_{Y|x_0}$ calculated in Example 11-5. It illustrates that the prediction limits are always wider than the confidence limits.

11-6: Prediction of New Observations

Figure 11-8

Figure 11-8 Scatter diagram of oxygen purity data from Example 11-1 with fitted regression line, 95% prediction limits (outer lines), and 95% confidence limits on $\mu_{Y|x_0}$.



11-7: Adequacy of the Regression Model

- Fitting a regression model requires several **assumptions**.
 1. Errors are uncorrelated random variables with mean zero;
 2. Errors have constant variance; and,
 3. Errors be normally distributed.
- The analyst should always consider the validity of these assumptions to be doubtful and conduct analyses to examine the adequacy of the model

11-7: Adequacy of the Regression Model

11-7.1 Residual Analysis

- The **residuals** from a regression model are $e_i = y_i - \hat{y}_i$, where y_i is an actual observation and \hat{y}_i is the corresponding fitted value from the regression model.
- Analysis of the residuals is frequently helpful in checking the assumption that the errors are approximately normally distributed with constant variance, and in determining whether additional terms in the model would be useful.

11-7: Adequacy of the Regression Model

11-7.1 Residual Analysis

Figure 11-9 Patterns for residual plots. (a) satisfactory, (b) funnel, (c) double bow, (d) nonlinear.

[Adapted from Montgomery, Peck, and Vining (2006).]

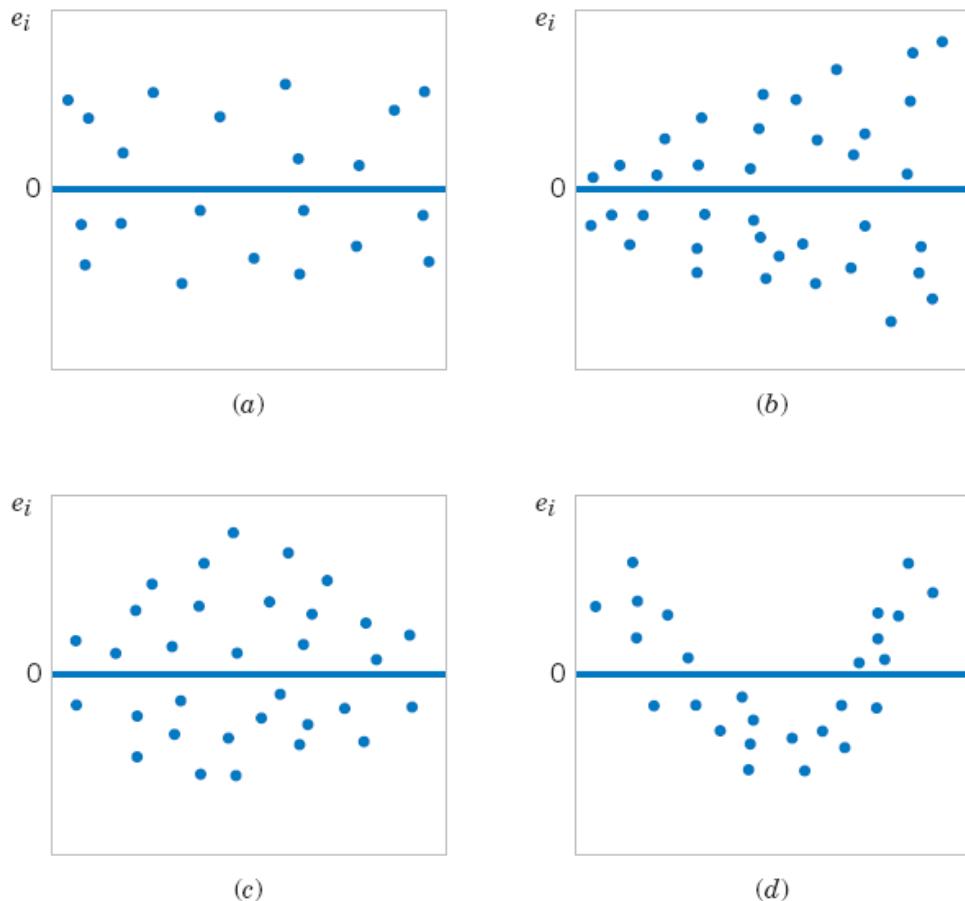


Figure 11-9 Patterns for residual plots. (a) Satisfactory, (b) Funnel, (c) Double bow, (d) Nonlinear. [Adapted from Montgomery, Peck, and Vining (2006).]

11-7: Adequacy of the Regression Model

Example 11-7

EXAMPLE 11-7 Oxygen Purity Residuals

The regression model for the oxygen purity data in Example 11-1 is $\hat{y} = 74.283 + 14.947x$. Table 11-4 presents the observed and predicted values of y at each value of x from this data set, along with the corresponding residual. These values were computed using Minitab and show the number of decimal places typical of computer output. A normal probability plot of the residuals is shown in Fig. 11-10. Since the residuals fall approximately along a straight line in the figure, we conclude that there is no severe departure from normality. The residuals are also plotted against the predicted value \hat{y}_i in Fig. 11-11 and against the hydrocarbon levels x_i in Fig. 11-12. These plots do not indicate any serious model inadequacies.

11-7: Adequacy of the Regression Model

Example 11-7

Table 11-4 Oxygen Purity Data from Example 11-1, Predicted Values, and Residuals

	Hydrocarbon Level, x	Oxygen Purity, y	Predicted Value, \hat{y}	Residual $e = y - \hat{y}$		Hydrocarbon Level, x	Oxygen Purity, y	Predicted Value, \hat{y}	Residual $e = y - \hat{y}$
1	0.99	90.01	89.081	0.929	11	1.19	93.54	92.071	1.469
2	1.02	89.05	89.530	-0.480	12	1.15	92.52	91.473	1.047
3	1.15	91.43	91.473	-0.043	13	0.98	90.56	88.932	1.628
4	1.29	93.74	93.566	0.174	14	1.01	89.54	89.380	0.160
5	1.46	96.73	96.107	0.623	15	1.11	89.85	90.875	-1.025
6	1.36	94.45	94.612	-0.162	16	1.20	90.39	92.220	-1.830
7	0.87	87.59	87.288	0.302	17	1.26	93.25	93.117	0.133
8	1.23	91.77	92.669	-0.899	18	1.32	93.41	94.014	-0.604
9	1.55	99.42	97.452	1.968	19	1.43	94.98	95.658	-0.678
10	1.40	93.65	95.210	-1.560	20	0.95	87.33	88.483	-1.153

11-7: Adequacy of the Regression Model

Example 11-7

Figure 11-10 Normal probability plot of residuals, Example 11-7.

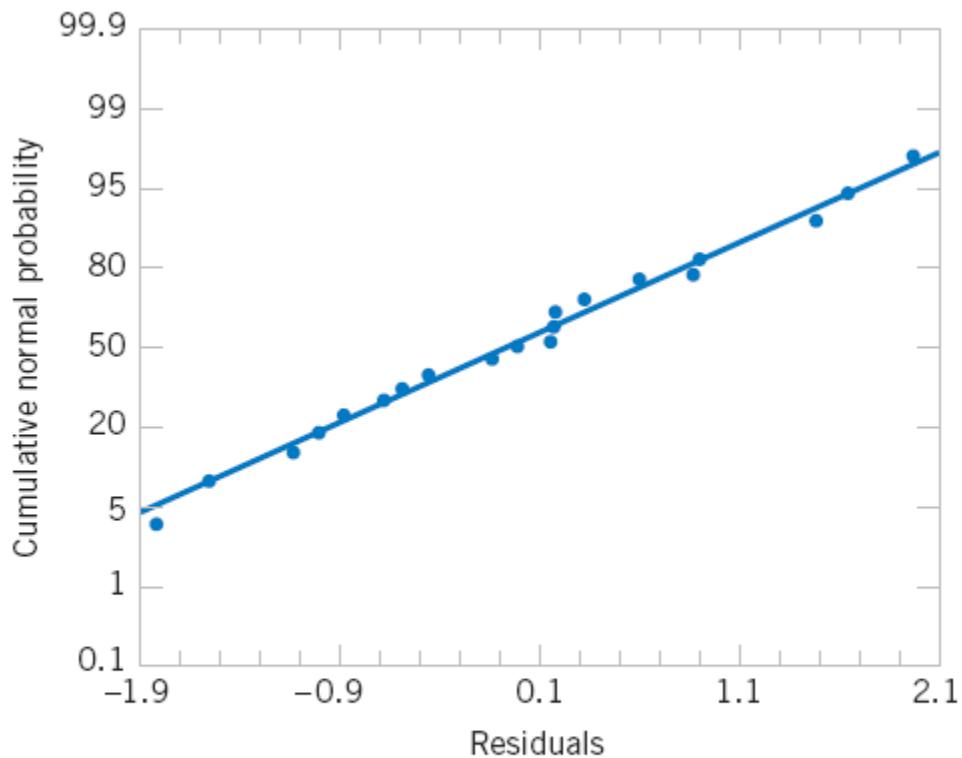


Figure 11-10 Normal probability plot of residuals, Example 11-7.

11-7: Adequacy of the Regression Model

Example 11-7

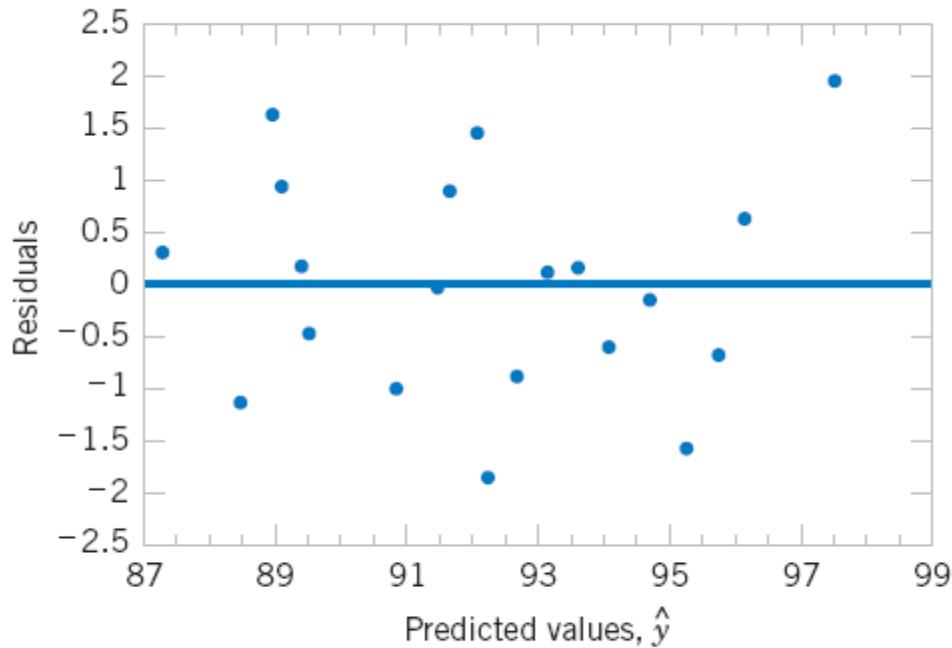


Figure 11-11 Plot of residuals versus predicted oxygen purity, \hat{y} , Example 11-7.

Figure 11-11 Plot of residuals versus predicted oxygen purity \hat{y} , Example 11-7.

11-7: Adequacy of the Regression Model

11-7.2 Coefficient of Determination (R^2)

- The quantity

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

is called the **coefficient of determination** and is often used to judge the adequacy of a regression model.

- $0 \leq R^2 \leq 1$;
- We often refer (loosely) to R^2 as the amount of variability in the data explained or accounted for by the regression model.

11-7: Adequacy of the Regression Model

11-7.2 Coefficient of Determination (R^2)

- For the oxygen purity regression model,

$$\begin{aligned} R^2 &= SS_R/SS_T \\ &= 152.13/173.38 \\ &= 0.877 \end{aligned}$$

- Thus, the model accounts for 87.7% of the variability in the data.

11-8: Correlation

We assume that the joint distribution of X_i and Y_i is the bivariate normal distribution presented in Chapter 5, and μ_Y and σ_Y^2 are the mean and variance of Y , μ_X and σ_X^2 are the mean and variance of X , and ρ is the **correlation coefficient** between Y and X . Recall that the correlation coefficient is defined as

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \quad (11-35)$$

where σ_{XY} is the covariance between Y and X .

The conditional distribution of Y for a given value of $X = x$ is

$$f_{Y|x}(y) = \frac{1}{\sqrt{2\pi}\sigma_{Y|x}} \exp\left[-\frac{1}{2}\left(\frac{y - \beta_0 - \beta_1 x}{\sigma_{Y|x}}\right)^2\right] \quad (11-36)$$

where

$$\beta_0 = \mu_Y - \mu_X \rho \frac{\sigma_Y}{\sigma_X} \quad (11-37)$$

$$\beta_1 = \frac{\sigma_Y}{\sigma_X} \rho \quad (11-38)$$

11-8: Correlation

It is possible to draw inferences about the correlation coefficient ρ in this model. The estimator of ρ is the **sample correlation coefficient**

$$R = \frac{\sum_{i=1}^n Y_i(X_i - \bar{X})}{\left[\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2 \right]^{1/2}} = \frac{S_{XY}}{(S_{XX} S_{YY})^{1/2}} \quad (11-43)$$

Note that

$$\hat{\beta}_1 = \left(\frac{S_{YY}}{S_{XX}} \right)^{1/2} R \quad (11-44)$$

We may also write:

$$R^2 = \hat{\beta}_1^2 \frac{S_{XX}}{S_{YY}} = \frac{\hat{\beta}_1 S_{XY}}{S_{YY}} = \frac{S_{XY}}{S_{YY}}$$

11-8: Correlation

It is often useful to test the hypotheses

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

The appropriate test statistic for these hypotheses is

$$T_0 = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \quad (11-46)$$

Reject H_0 if $|t_0| > t_{\alpha/2,n-2}$.

11-8: Correlation

The test procedure for the hypothesis

$$H_0: \rho = \rho_0$$

$$H_1: \rho \neq \rho_0$$

where $\rho_0 \neq 0$ is somewhat more complicated. In this case, the appropriate test statistic is

$$Z_0 = (\operatorname{arctanh} R - \operatorname{arctanh} \rho_0)(n - 3)^{1/2} \quad (11-49)$$

Reject H_0 if $|z_0| > z_{\alpha/2}$.

11-8: Correlation

The approximate $100(1 - \alpha)\%$ confidence interval is

$$\tanh\left(\operatorname{arctanh} r - \frac{z_{\alpha/2}}{\sqrt{n-3}}\right) \leq \rho \leq \tanh\left(\operatorname{arctanh} r + \frac{z_{\alpha/2}}{\sqrt{n-3}}\right) \quad (11-50)$$

11-8: Correlation

Example 11-8

EXAMPLE 11-8 Wire Bond Pull Strength

In Chapter 1 (Section 1-3) an application of regression analysis is described in which an engineer at a semiconductor assembly plant is investigating the relationship between pull strength of a wire bond and two factors: wire length and die height. In this example, we will consider only one of the factors, the wire length. A random sample of 25 units is selected and tested, and the wire bond pull strength and wire length are observed for each unit. The data are shown in Table 1-2. We assume that pull strength and wire length are jointly normally distributed.

Figure 11-13 shows a scatter diagram of wire bond strength versus wire length. We have used the Minitab option of displaying box plots of each individual variable on the scatter diagram. There is evidence of a linear relationship between the two variables.

The Minitab output for fitting a simple linear regression model to the data is shown below.

11-8: Correlation

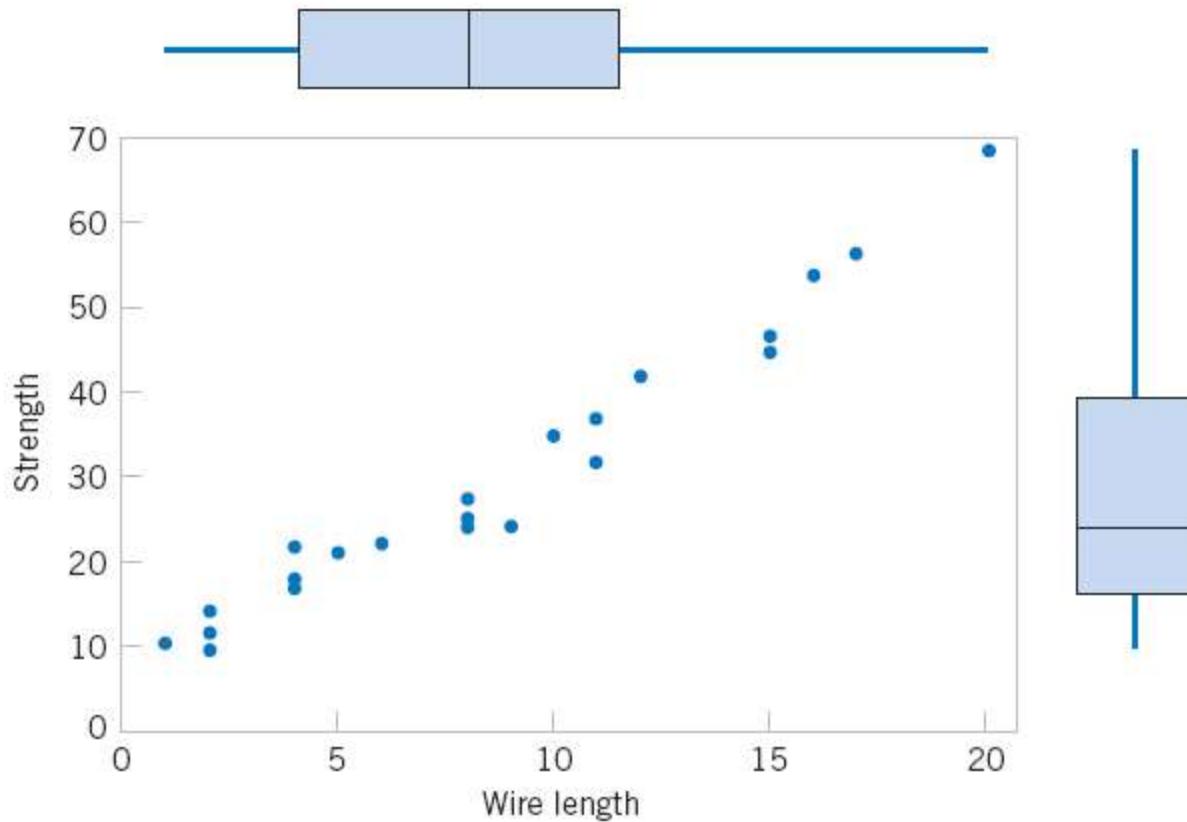


Figure 11-13 Scatter plot of wire bond strength versus wire length, Example 11-8.

Figure 11-13 Scatter plot of wire bond strength versus wire length, Example 11-8.

11-8: Correlation

Minitab Output for Example 11-8

Regression Analysis: Strength versus Length

The regression equation is

$$\text{Strength} = 5.11 + 2.90 \text{ Length}$$

Predictor	Coef	SE Coef	T	P
Constant	5.115	1.146	4.46	0.000
Length	2.9027	0.1170	24.80	0.000

$$S = 3.093$$

$$R-\text{Sq} = 96.4\%$$

$$R-\text{Sq}(\text{adj}) = 96.2\%$$

$$\text{PRESS} = 272.144$$

$$R-\text{Sq}(\text{pred}) = 95.54\%$$

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	5885.9	5885.9	615.08	0.000
Residual Error	23	220.1	9.6		
Total	24	6105.9			

11-8: Correlation

Example 11-8 (continued)

Now $S_{xx} = 698.56$ and $S_{xy} = 2027.7132$, and the sample correlation coefficient is

$$r = \frac{S_{xy}}{[S_{xx}SS_T]^{1/2}} = \frac{2027.7132}{[(698.560)(6105.9)]^{1/2}} = 0.9818$$

Note that $r^2 = (0.9818)^2 = 0.9640$ (which is reported in the Minitab output), or that approximately 96.40% of the variability in pull strength is explained by the linear relationship to wire length.

11-8: Correlation

Example 11-8 (continued)

Now suppose that we wish to test the hypotheses

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

with $\alpha = 0.05$. We can compute the t -statistic of Equation 11-46 as

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.9818\sqrt{23}}{\sqrt{1-0.9640}} = 24.8$$

This statistic is also reported in the Minitab output as a test of $H_0: \beta_1 = 0$. Because $t_{0.025,23} = 2.069$, we reject H_0 and conclude that the correlation coefficient $\rho \neq 0$.

11-8: Correlation

Example 11-8 (continued)

Finally, we may construct an approximate 95% confidence interval on ρ from Equation 11-50. Since $\text{arctanh } r = \text{arctanh } 0.9818 = 2.3452$, Equation 11-50 becomes

$$\tanh\left(2.3452 - \frac{1.96}{\sqrt{22}}\right) \leq \rho \leq \tanh\left(2.3452 + \frac{1.96}{\sqrt{22}}\right)$$

which reduces to

$$0.9585 \leq \rho \leq 0.9921$$

11-9: Transformation and Logistic Regression

We occasionally find that the straight-line regression model $Y = \beta_0 + \beta_1x + \epsilon$ is inappropriate because the true regression function is nonlinear. Sometimes nonlinearity is visually determined from the scatter diagram, and sometimes, because of prior experience or underlying theory, we know in advance that the model is nonlinear. Occasionally, a scatter diagram will exhibit an apparent nonlinear relationship between Y and x . In some of these situations, a nonlinear function can be expressed as a straight line by using a suitable transformation. Such nonlinear models are called **intrinsically linear**.

11-9: Transformation and Logistic Regression

Example 11-9

EXAMPLE 11-9 Windmill Power

A research engineer is investigating the use of a windmill to generate electricity and has collected data on the DC output from this windmill and the corresponding wind velocity. The data are plotted in Figure 11-14 and listed in Table 11-5 (p.439).

Table 11-5 Observed Values y_i and Regressor Variable x_i for Example 11-9.

Table 11-5 Observed Values y_i and Regressor Variable x_i for Example 11-9

Observation Number, i	Wind Velocity (mph), x_i	DC Output, y_i
1	5.00	1.582
2	6.00	1.822
3	3.40	1.057
4	2.70	0.500
5	10.00	2.236
6	9.70	2.386
7	9.55	2.294
8	3.05	0.558
9	8.15	2.166
10	6.20	1.866
11	2.90	0.653
12	6.35	1.930
13	4.60	1.562
14	5.80	1.737
15	7.40	2.088
16	3.60	1.137
17	7.85	2.179
18	8.80	2.112
19	7.00	1.800
20	5.45	1.501
21	9.10	2.303
22	10.20	2.310
23	4.10	1.194
24	3.95	1.144
25	2.45	0.123

11-9: Transformation and Logistic Regression

Example 11-9 (Continued)

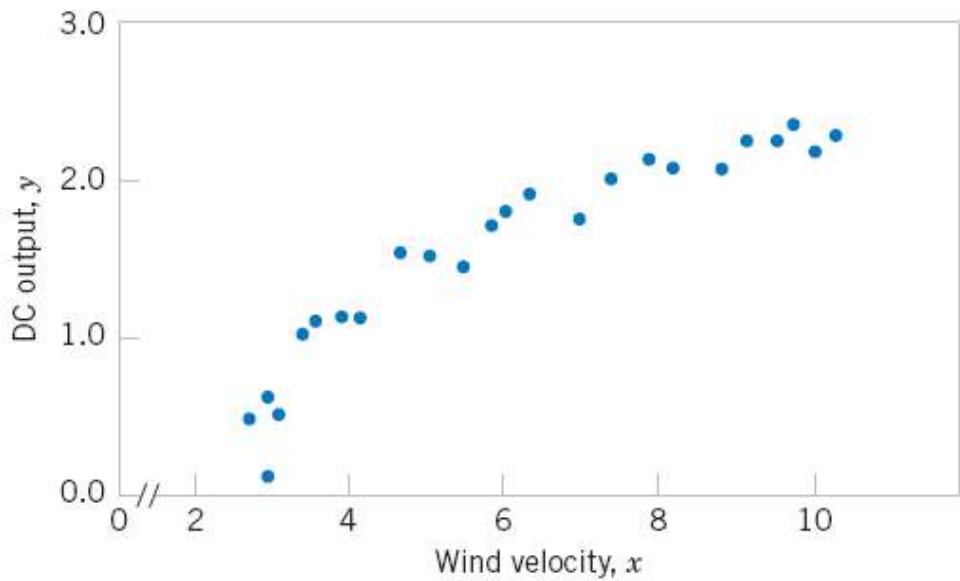


Figure 11-14 Plot of DC output y versus wind velocity x for the windmill data.

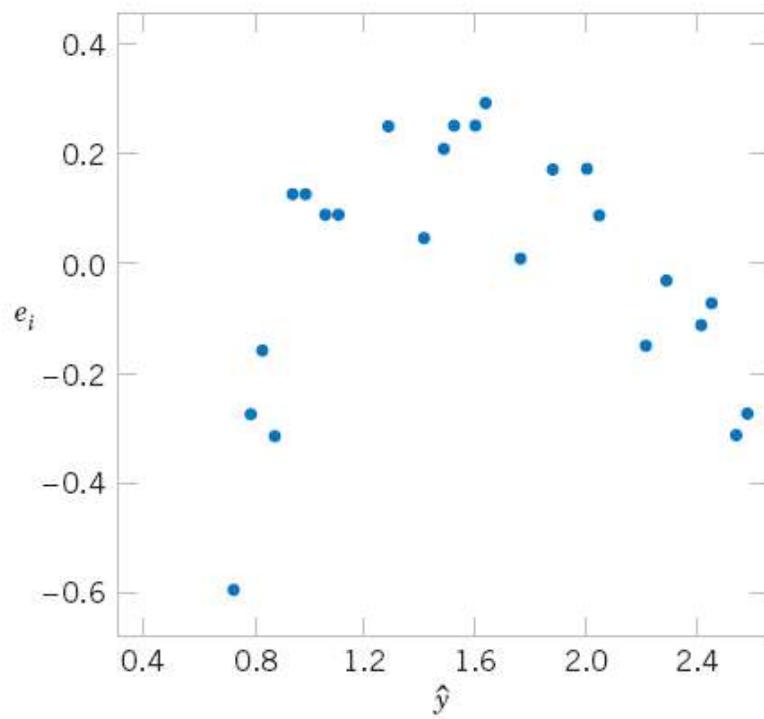


Figure 11-15 Plot of residuals e_i versus fitted values \hat{y}_i for the windmill data.

11-9: Transformation and Logistic Regression

Example 11-9 (Continued)

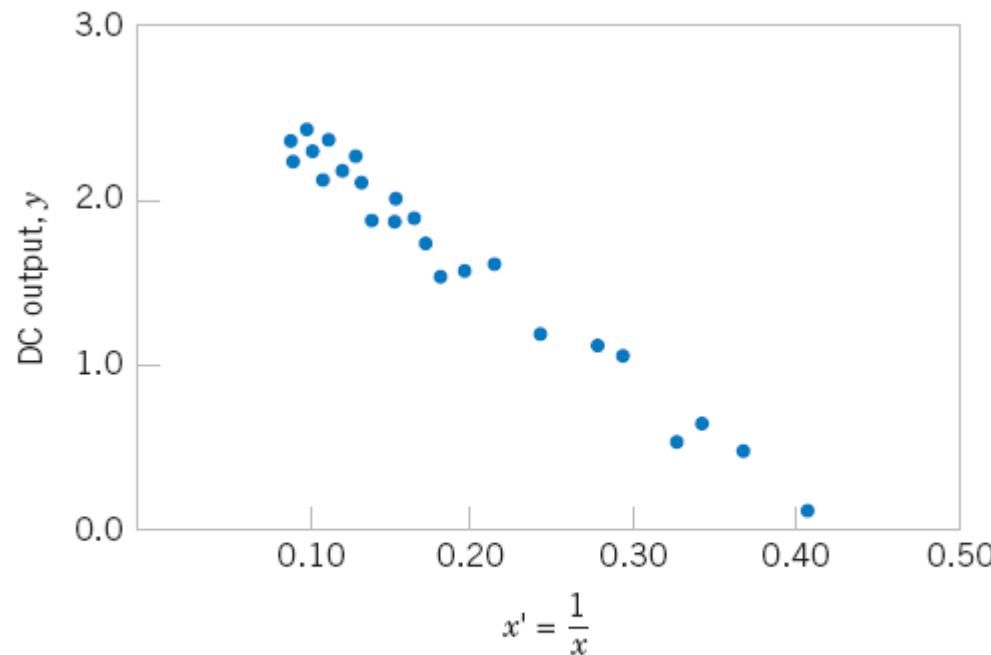


Figure 11-16 Plot of DC output versus $x' = 1/x$ for the windmill data.

11-9: Transformation and Logistic Regression

Example 11-9 (Continued)

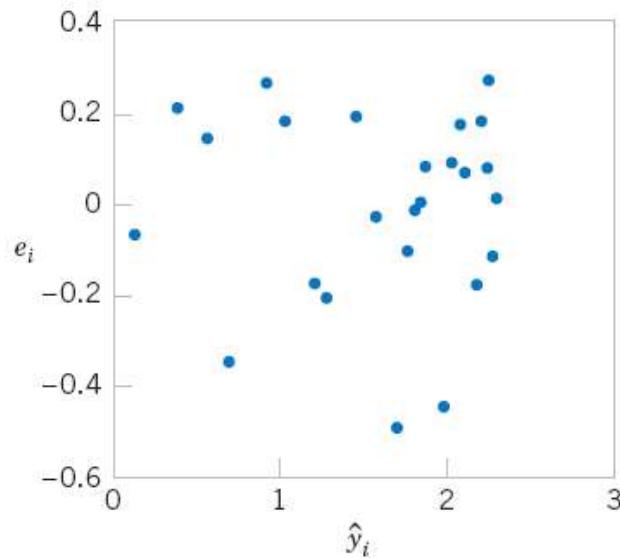


Figure 11-17 Plot of residuals versus fitted values \hat{y}_i for the transformed model for the windmill data.

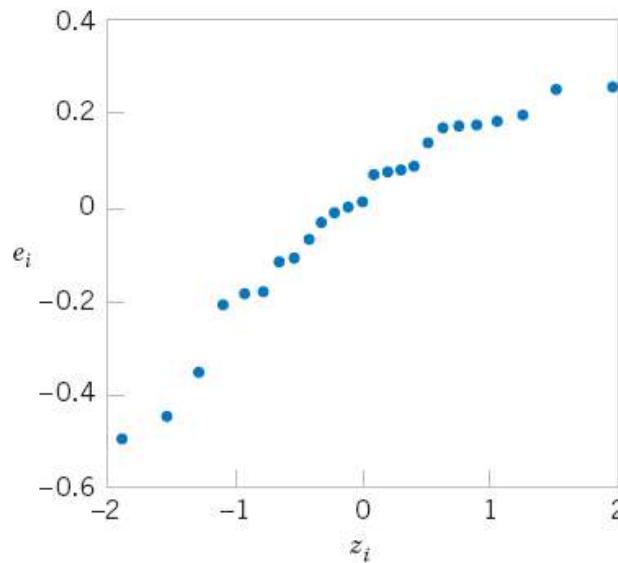


Figure 11-18 Normal probability plot of the residuals for the transformed model for the windmill data.

A plot of the residuals from the transformed model versus \hat{y} is shown in Figure 11-17. This plot does not reveal any serious problem with inequality of variance. The normal probability plot, shown in Figure 11-18, gives a mild indication that the errors come from a distribution with heavier tails than the normal (notice the slight upward and downward curve at the extremes). This normal probability plot has the z-score value plotted on the horizontal axis. Since there is no strong signal of model inadequacy, we conclude that the transformed model is satisfactory.

Important Terms & Concepts of Chapter 11

Analysis of variance test in regression

Confidence interval on mean response

Correlation coefficient

Empirical model

Confidence intervals on model parameters

Intrinsically linear model

Least squares estimation of regression model parameters

Logistics regression

Model adequacy checking

Odds ratio

Prediction interval on a future observation

Regression analysis

Residual plots

Residuals

Scatter diagram

Simple linear regression model standard error

Statistical test on model parameters

Transformations

Foundation of Data Science and Analytics

Multiple Linear Regression

Arun K. Timalsina

12-1: Multiple Linear Regression Models

12-1.1 Introduction

- Many applications of regression analysis involve situations in which there are more than one regressor variable.
- A regression model that contains more than one regressor variable is called a **multiple regression model**.

12-1: Multiple Linear Regression Models

12-1.1 Introduction

- For example, suppose that the effective life of a cutting tool depends on the cutting speed and the tool angle. A possible multiple regression model could be

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

where

Y – tool life

x_1 – cutting speed

x_2 – tool angle

12-1: Multiple Linear Regression Models

12-1.1 Introduction

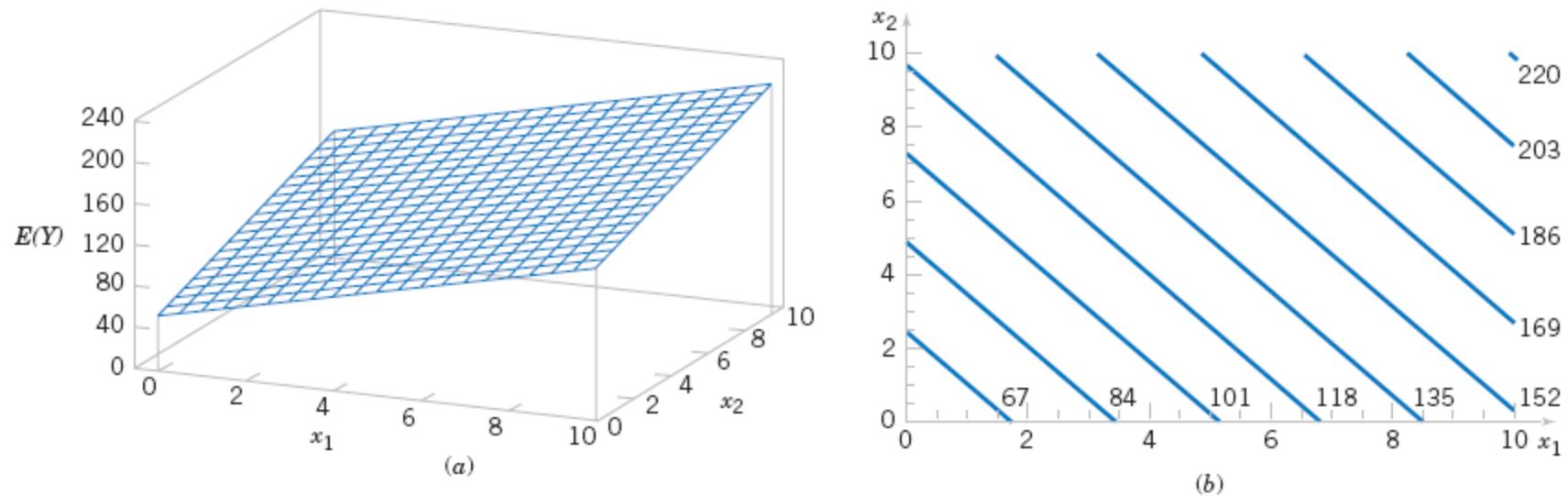


Figure 12-1 (a) The regression plane for the model $E(Y) = 50 + 10x_1 + 7x_2$. (b) The contour plot.

Figure 12-1 (a) The regression plane for the model $E(Y) = 50 + 10x_1 + 7x_2$. (b) The contour plot

12-1: Multiple Linear Regression Models

12-1.1 Introduction

In general, the **dependent variable** or **response** Y may be related to k **independent** or **regressor variables**. The model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon \quad (12-2)$$

is called a multiple linear regression model with k regressor variables. The parameters $\beta_j, j = 0, 1, \dots, k$, are called the regression coefficients. This model describes a hyperplane in the k -dimensional space of the regressor variables $\{x_j\}$. The parameter β_j represents the expected change in response Y per unit change in x_j when all the remaining regressors x_i ($i \neq j$) are held constant.

12-1: Multiple Linear Regression Models

12-1.1 Introduction

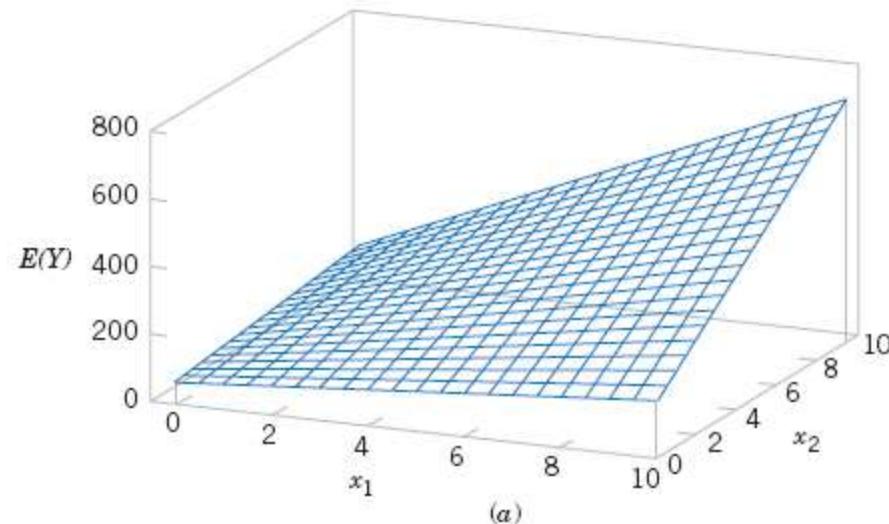
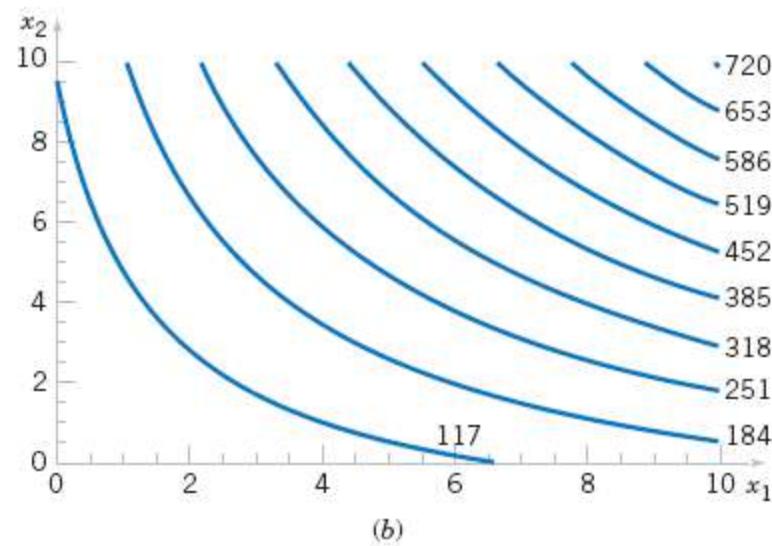


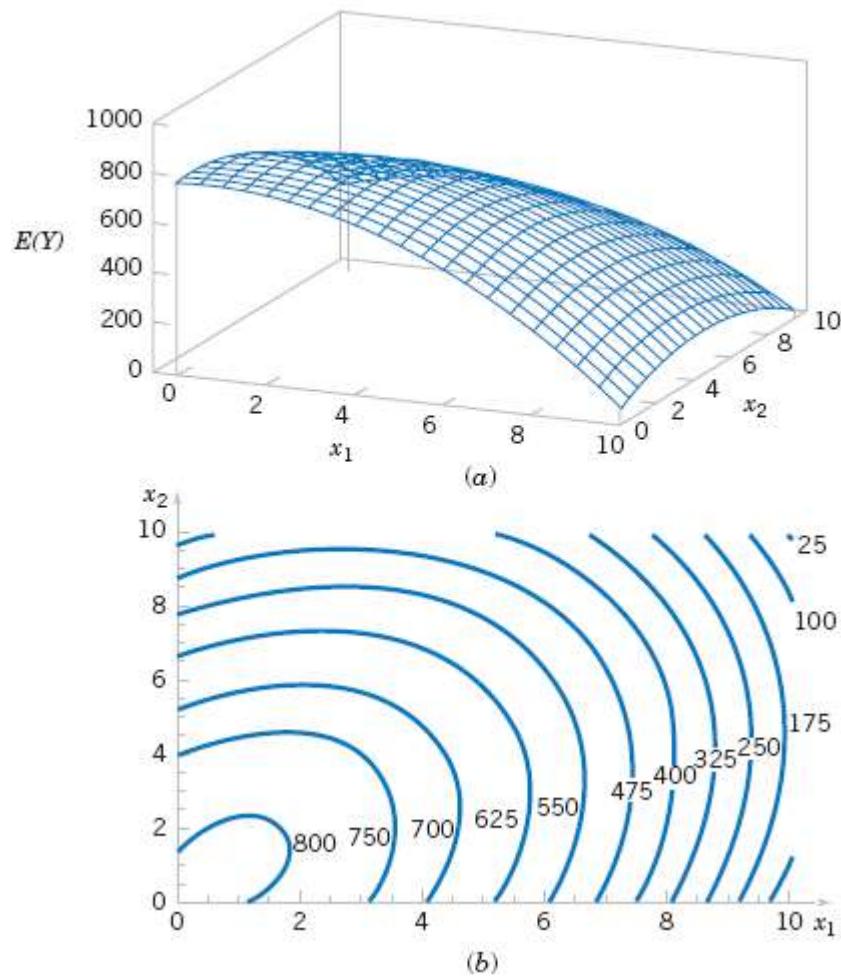
Figure 12-2 (a) Three-dimensional plot of the regression model $E(Y) = 50 + 10x_1 + 7x_2 + 5x_1x_2$. (b) The contour plot



12-1: Multiple Linear Regression Models

12-1.1 Introduction

Figure 12-3 (a) Three-dimensional plot of the regression model $E(Y) = 800 + 10x_1 + 7x_2 - 8.5x_1^2 - 5x_2^2 + 4x_1x_2$. (b) The contour plot



12-1: Multiple Linear Regression Models

12-1.2 Least Squares Estimation of the Parameters

The **method of least squares** may be used to estimate the regression coefficients in the multiple regression model, Equation 12-2. Suppose that $n > k$ observations are available, and let x_{ij} denote the i th observation or level of variable x_j . The observations are

$$(x_{i1}, x_{i2}, \dots, x_{ik}, y_i), \quad i = 1, 2, \dots, n \quad \text{and} \quad n > k$$

It is customary to present the data for multiple regression in a table such as Table 12-1.

Table 12-1 Data for Multiple Linear Regression

y	x_1	x_2	...	x_k
y_1	x_{11}	x_{12}	...	x_{1k}
y_2	x_{21}	x_{22}	...	x_{2k}
\vdots	\vdots	\vdots		\vdots
y_n	x_{n1}	x_{n2}	...	x_{nk}

12-1: Multiple Linear Regression Models

12-1.2 Least Squares Estimation of the Parameters

- The least squares function is given by

$$L = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^k \beta_j x_{ij} \right)^2$$

- The least squares estimates must satisfy

$$\frac{\partial L}{\partial \beta_0} \Bigg|_{\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k} = -2 \sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \sum_{j=1}^k \hat{\beta}_j x_{ij} \right) = 0$$

and

$$\frac{\partial L}{\partial \beta_j} \Bigg|_{\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k} = -2 \sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \sum_{j=1}^k \hat{\beta}_j x_{ij} \right) x_{ij} = 0 \quad j = 1, 2, \dots, k$$

12-1: Multiple Linear Regression Models

12-1.2 Least Squares Estimation of the Parameters

- The **least squares normal Equations** are

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_{i1} + \hat{\beta}_2 \sum_{i=1}^n x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^n x_{ik} = \sum_{i=1}^n y_i$$

$$\begin{aligned}\hat{\beta}_0 \sum_{i=1}^n x_{i1} + \hat{\beta}_1 \sum_{i=1}^n x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^n x_{i1}x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^n x_{i1}x_{ik} &= \sum_{i=1}^n x_{i1}y_i \\ \vdots &\quad \vdots & \vdots & \vdots & \vdots\end{aligned}$$

$$\hat{\beta}_0 \sum_{i=1}^n x_{ik} + \hat{\beta}_1 \sum_{i=1}^n x_{ik}x_{i1} + \hat{\beta}_2 \sum_{i=1}^n x_{ik}x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^n x_{ik}^2 = \sum_{i=1}^n x_{ik}y_i$$

- The solution to the normal Equations are the **least squares estimators** of the regression coefficients.

12-1: Multiple Linear Regression Models

Example 12-1

EXAMPLE 12-1 Wire Bond Strength

In Chapter 1, we used data on pull strength of a wire bond in a semiconductor manufacturing process, wire length, and die height to illustrate building an empirical model. We will use the same data, repeated for convenience in Table 12-2, and show the details of estimating the model parameters. A three-dimensional scatter plot of the data is presented in Fig. 1-15. Figure 12-4 shows a matrix of two-dimensional scatter plots of the data. These displays can be helpful in visualizing the relationships among variables in a multivariable data set. For example, the plot indicates that there is a strong linear relationship between strength and wire length.

12-1: Multiple Linear Regression Models

Example 12-1

Table 12-2 Wire Bond Data for Example 12-1

Observation Number	Pull Strength y	Wire Length x_1	Die Height x_2	Observation Number	Pull Strength y	Wire Length x_1	Die Height x_2
1	9.95	2	50	14	11.66	2	360
2	24.45	8	110	15	21.65	4	205
3	31.75	11	120	16	17.89	4	400
4	35.00	10	550	17	69.00	20	600
5	25.02	8	295	18	10.30	1	585
6	16.86	4	200	19	34.93	10	540
7	14.38	2	375	20	46.59	15	250
8	9.60	2	52	21	44.88	15	290
9	24.35	9	100	22	54.12	16	510
10	27.50	8	300	23	56.63	17	590
11	17.08	4	412	24	22.13	6	100
12	37.00	11	400	25	21.15	5	400
13	41.95	12	500				

12-1: Multiple Linear Regression Models

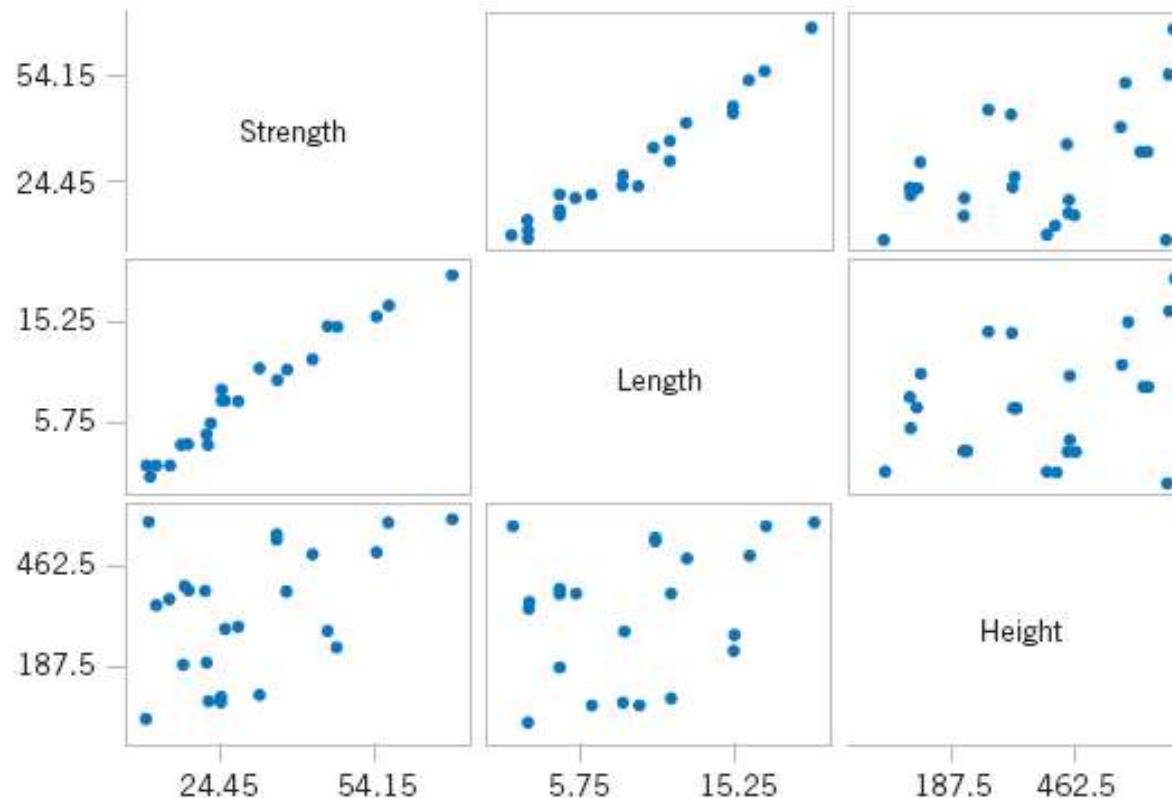


Figure 12-4 Matrix of scatter plots (from Minitab) for the wire bond pull strength data in Table 12-2.

Figure 12-4 Matrix of scatter plots (from Minitab) for the wire bond pull strength data in Table 12-2.

12-1: Multiple Linear Regression Models

Example 12-1

Specifically, we will fit the multiple linear regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

where Y = pull strength, x_1 = wire length, and x_2 = die height. From the data in Table 12-2 we calculate

$$n = 25, \sum_{i=1}^{25} y_i = 725.82$$

$$\sum_{i=1}^{25} x_{i1} = 206, \sum_{i=1}^{25} x_{i2} = 8,294$$

$$\sum_{i=1}^{25} x_{i1}^2 = 2,396, \sum_{i=1}^{25} x_{i2}^2 = 3,531,848$$

$$\sum_{i=1}^{25} x_{i1}x_{i2} = 77,177, \sum_{i=1}^{25} x_{i1}y_i = 8,008.47,$$

$$\sum_{i=1}^{25} x_{i2}y_i = 274,816.71$$

12-1: Multiple Linear Regression Models

Example 12-1

For the model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, the normal equations 12-10 are

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_{i1} + \hat{\beta}_2 \sum_{i=1}^n x_{i2} = \sum_{i=1}^n y_i$$

$$\hat{\beta}_0 \sum_{i=1}^n x_{i1} + \hat{\beta}_1 \sum_{i=1}^n x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^n x_{i1}x_{i2} = \sum_{i=1}^n x_{i1}y_i$$

$$\hat{\beta}_0 \sum_{i=1}^n x_{i2} + \hat{\beta}_1 \sum_{i=1}^n x_{i1}x_{i2} + \hat{\beta}_2 \sum_{i=1}^n x_{i2}^2 = \sum_{i=1}^n x_{i2}y_i$$

Inserting the computed summations into the normal equations, we obtain

$$25\hat{\beta}_0 + 206\hat{\beta}_1 + 8294\hat{\beta}_2 = 725.82$$

$$206\hat{\beta}_0 + 2396\hat{\beta}_1 + 77,177\hat{\beta}_2 = 8,008.47$$

$$8294\hat{\beta}_0 + 77,177\hat{\beta}_1 + 3,531,848\hat{\beta}_2 = 274,816.71$$

12-1: Multiple Linear Regression Models

Example 12-1

The solution to this set of equations is

$$\hat{\beta}_0 = 2.26379, \quad \hat{\beta}_1 = 2.74427, \quad \hat{\beta}_2 = 0.01253$$

Therefore, the fitted regression equation is

$$\hat{y} = 2.26379 + 2.74427x_1 + 0.01253x_2$$

Practical Interpretation: This equation can be used to predict pull strength for pairs of values of the regressor variables wire length (x_1) and die height (x_2). This is essentially the same regression model given in Section 1-3. Figure 1-16 shows a three-dimensional plot of the plane of predicted values \hat{y} generated from this equation.

12-1: Multiple Linear Regression Models

12-1.3 Matrix Approach to Multiple Linear Regression

Suppose the model relating the regressors to the response is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i \quad i = 1, 2, \dots, n$$

In matrix notation this model can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

12-1: Multiple Linear Regression Models

12-1.3 Matrix Approach to Multiple Linear Regression

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \text{and} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

12-1: Multiple Linear Regression Models

12-1.3 Matrix Approach to Multiple Linear Regression

We wish to find the vector of least squares estimators that minimizes:

$$L = \sum_{i=1}^n \epsilon_i^2 = \epsilon' \epsilon = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

The resulting least squares estimate is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} \quad (12-13)$$

12-1: Multiple Linear Regression Models

12-1.3 Matrix Approach to Multiple Linear Regression

The fitted regression model is

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^k \hat{\beta}_j x_{ij} \quad i = 1, 2, \dots, n \quad (12-14)$$

In matrix notation, the fitted model is

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

The difference between the observation y_i and the fitted value \hat{y}_i is a **residual**, say, $e_i = y_i - \hat{y}_i$. The $(n \times 1)$ vector of residuals is denoted by

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} \quad (12-15)$$

12-1: Multiple Linear Regression Models

Example 12-2

In Example 12-1, we illustrated fitting the multiple regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

where y is the observed pull strength for a wire bond, x_1 is the wire length, and x_2 is the die height. The 25 observations are in Table 12-2. We will now use the matrix approach to fit the regression model above to these data. The model matrix \mathbf{X} and y vector for this model are

Example 12-2

X =	$\begin{bmatrix} 1 & 2 & 50 \\ 1 & 8 & 110 \\ 1 & 11 & 120 \\ 1 & 10 & 550 \\ 1 & 8 & 295 \\ 1 & 4 & 200 \\ 1 & 2 & 375 \\ 1 & 2 & 52 \\ 1 & 9 & 100 \\ 1 & 8 & 300 \\ 1 & 4 & 412 \\ 1 & 11 & 400 \\ 1 & 12 & 500 \\ 1 & 2 & 360 \\ 1 & 4 & 205 \\ 1 & 4 & 400 \\ 1 & 20 & 600 \\ 1 & 1 & 585 \\ 1 & 10 & 540 \\ 1 & 15 & 250 \\ 1 & 15 & 290 \\ 1 & 16 & 510 \\ 1 & 17 & 590 \\ 1 & 6 & 100 \\ 1 & 5 & 400 \end{bmatrix}$	y =	$\begin{bmatrix} 9.95 \\ 24.45 \\ 31.75 \\ 35.00 \\ 25.02 \\ 16.86 \\ 14.38 \\ 9.60 \\ 24.35 \\ 27.50 \\ 17.08 \\ 37.00 \\ 41.95 \\ 11.66 \\ 21.65 \\ 17.89 \\ 69.00 \\ 10.30 \\ 34.93 \\ 46.59 \\ 44.88 \\ 54.12 \\ 56.63 \\ 22.13 \\ 21.15 \end{bmatrix}$
-----	---	-----	---

12-1: Multiple Linear Regression Models

Example 12-2

The $\mathbf{X}'\mathbf{X}$ matrix is

$$\begin{aligned}\mathbf{X}'\mathbf{X} &= \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 2 & 8 & \cdots & 5 \\ 50 & 110 & \cdots & 400 \end{bmatrix} \begin{bmatrix} 1 & 2 & 50 \\ 1 & 8 & 110 \\ \vdots & \vdots & \vdots \\ 1 & 5 & 400 \end{bmatrix} \\ &= \begin{bmatrix} 25 & 206 & 8,294 \\ 206 & 2,396 & 77,177 \\ 8,294 & 77,177 & 3,531,848 \end{bmatrix}\end{aligned}$$

and the $\mathbf{X}'\mathbf{y}$ vector is

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 2 & 8 & \cdots & 5 \\ 50 & 110 & \cdots & 400 \end{bmatrix} \begin{bmatrix} 9.95 \\ 24.45 \\ \vdots \\ 21.15 \end{bmatrix} = \begin{bmatrix} 725.82 \\ 8,008.47 \\ 274,816.71 \end{bmatrix}$$

The least squares estimates are found from Equation 12-13 as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

12-1: Multiple Linear Regression Models

Example 12-2

or

$$\begin{aligned}\left[\begin{array}{c} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{array} \right] &= \left[\begin{array}{ccc} 25 & 206 & 8,294 \\ 206 & 2,396 & 77,177 \\ 8,294 & 77,177 & 3,531,848 \end{array} \right]^{-1} \left[\begin{array}{c} 725.82 \\ 8,008.37 \\ 274,811.31 \end{array} \right] \\ &= \left[\begin{array}{ccc} 0.214653 & -0.007491 & -0.000340 \\ -0.007491 & 0.001671 & -0.000019 \\ -0.000340 & -0.000019 & +0.0000015 \end{array} \right] \left[\begin{array}{c} 725.82 \\ 8,008.37 \\ 274,811.31 \end{array} \right] \\ &= \left[\begin{array}{c} 2.26379143 \\ 2.74426964 \\ 0.01252781 \end{array} \right]\end{aligned}$$

Therefore, the fitted regression model with the regression coefficients rounded to five decimal places is

$$\hat{y} = 2.26379 + 2.74427x_1 + 0.01253x_2$$

This is identical to the results obtained in Example 12-1.

12-1: Multiple Linear Regression Models

Example 12-2

This regression model can be used to predict values of pull strength for various values of wire length (x_1) and die height (x_2). We can also obtain the **fitted values** \hat{y}_i by substituting each observation (x_{i1}, x_{i2}) , $i = 1, 2, \dots, n$, into the equation. For example, the first observation has $x_{11} = 2$ and $x_{12} = 50$, and the fitted value is

$$\begin{aligned}\hat{y}_1 &= 2.26379 + 2.74427x_{11} + 0.01253x_{12} \\ &= 2.26379 + 2.74427(2) + 0.01253(50) \\ &= 8.38\end{aligned}$$

The corresponding observed value is $y_1 = 9.95$. The *residual* corresponding to the first observation is

$$\begin{aligned}e_1 &= y_1 - \hat{y}_1 \\ &= 9.95 - 8.38 \\ &= 1.57\end{aligned}$$

Table 12-3 displays all 25 fitted values \hat{y}_i and the corresponding residuals. The fitted values and residuals are calculated to the same accuracy as the original data.

12-1: Multiple Linear Regression Models

Example 12-2

Table 12-3 Observations, Fitted Values, and Residuals for Example 12-2

Observation Number	y_i	\hat{y}_i	$e_i = y_i - \hat{y}_i$	Observation Number	y_i	\hat{y}_i	$e_i = y_i - \hat{y}_i$
1	9.95	8.38	1.57	14	11.66	12.26	-0.60
2	24.45	25.60	-1.15	15	21.65	15.81	5.84
3	31.75	33.95	-2.20	16	17.89	18.25	-0.36
4	35.00	36.60	-1.60	17	69.00	64.67	4.33
5	25.02	27.91	-2.89	18	10.30	12.34	-2.04
6	16.86	15.75	1.11	19	34.93	36.47	-1.54
7	14.38	12.45	1.93	20	46.59	46.56	0.03
8	9.60	8.40	1.20	21	44.88	47.06	-2.18
9	24.35	28.21	-3.86	22	54.12	52.56	1.56
10	27.50	27.98	-0.48	23	56.63	56.31	0.32
11	17.08	18.40	-1.32	24	22.13	19.98	2.15
12	37.00	37.46	-0.46	25	21.15	21.00	0.15
13	41.95	41.46	0.49				

Table 12-4 Minitab Multiple Regression Output for the Wire Bond Pull Strength Data

Regression Analysis: Strength versus Length, Height

The regression equation is

$$\text{Strength} = 2.26 + 2.74 \text{ Length} + 0.0125 \text{ Height}$$

Predictor	Coef	SE Coef	T	P	VIF
Constant	$\hat{\beta}_0 \rightarrow 2.264$	1.060	2.14	0.044	
Length	$\hat{\beta}_1 \rightarrow 2.74427$	0.09352	29.34	0.000	1.2
Height	$\hat{\beta}_2 \rightarrow 0.012528$	0.002798	4.48	0.000	1.2

$$S = 2.288$$

$$R\text{-Sq} = 98.1\%$$

$$R\text{-Sq (adj)} = 97.9\%$$

$$\text{PRESS} = 156.163$$

$$R\text{-Sq (pred)} = 97.44\%$$

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	5990.8	2995.4	572.17	0.000
Residual Error	22	115.2	5.2 $\leftarrow \hat{\sigma}^2$		
Total	24	6105.9			

Source	DF	Seq SS
Length	1	5885.9
Height	1	104.9

Values of Predictors for New Observations

New Obs	Fit	SE Fit	95.0% CI	95.0% PI
1	27.663	0.482	(26.663, 28.663)	(22.814, 32.512)

Values of Predictors for New Observations

New Obs	Length	Height
1	8.00	275

12-1: Multiple Linear Regression Models

Estimating σ^2

An unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n - p} = \frac{SS_E}{n - p} \quad (12-16)$$

12-1: Multiple Linear Regression Models

12-1.4 Properties of the Least Squares Estimators

Unbiased estimators:

$$\begin{aligned} E(\hat{\beta}) &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \epsilon)] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon] \\ &= \beta \end{aligned}$$

Covariance Matrix:

$$\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} C_{00} & C_{01} & C_{02} \\ C_{10} & C_{11} & C_{12} \\ C_{20} & C_{21} & C_{22} \end{bmatrix}$$

12-1: Multiple Linear Regression Models

12-1.4 Properties of the Least Squares Estimators

Individual variances and covariances:

$$V(\hat{\beta}_j) = \sigma^2 C_{jj}, \quad j = 0, 1, 2$$

$$\text{cov}(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 C_{ij}, \quad i \neq j$$

In general,

$$\text{cov}(\hat{\beta}) = \sigma^2 (\mathbf{X}' \mathbf{X})^{-1} = \sigma^2 \mathbf{C}$$

12-2: Hypothesis Tests in Multiple Linear Regression

12-2.1 Test for Significance of Regression

The appropriate hypotheses are

$$\begin{aligned} H_0: \beta_1 &= \beta_2 = \cdots = \beta_k = 0 \\ H_1: \beta_j &\neq 0 \quad \text{for at least one } j \end{aligned} \tag{12-18}$$

The test statistic is

$$F_0 = \frac{SS_R/k}{SS_E/(n - p)} = \frac{MS_R}{MS_E} \tag{12-19}$$

12-2: Hypothesis Tests in Multiple Linear Regression

12-2.1 Test for Significance of Regression

Table 12-9 Analysis of Variance for Testing Significance of Regression in Multiple Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	SS_R	k	MS_R	MS_R/MS_E
Error or residual	SS_E	$n - p$	MS_E	
Total	SS_T	$n - 1$		

12-2: Hypothesis Tests in Multiple Linear Regression

Example 12-3

EXAMPLE 12-3 Wire Bond Strength ANOVA

We will test for significance of regression (with $\alpha = 0.05$) using the wire bond pull strength data from Example 12-1. The total sum of squares is

$$SS_T = \mathbf{y}'\mathbf{y} - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n} = 27,178.5316 - \frac{(725.82)^2}{25}$$
$$= 6105.9447$$

12-2: Hypothesis Tests in Multiple Linear Regression

Example 12-3

The regression or model sum of squares is computed from Equation 12-20 as follows:

$$SS_R = \hat{\beta}' \mathbf{X}' \mathbf{y} - \frac{\left(\sum_{i=1}^n y_i \right)^2}{n} = 27,063.3581 - \frac{(725.82)^2}{25} \\ = 5990.7712$$

and by subtraction

$$SS_E = SS_T - SS_R = \mathbf{y}' \mathbf{y} - \hat{\beta}' \mathbf{X}' \mathbf{y} = 115.1716$$

12-2: Hypothesis Tests in Multiple Linear Regression

Example 12-3

The analysis of variance is shown in Table 12-10. To test $H_0: \beta_1 = \beta_2 = 0$, we calculate the statistic

$$f_0 = \frac{MS_R}{MS_E} = \frac{2995.3856}{5.2352} = 572.17$$

Since $f_0 > f_{0.05,2,22} = 3.44$ (or since the P -value is considerably smaller than $\alpha = 0.05$), we reject the null hypothesis and conclude that pull strength is linearly related to either wire length or die height, or both.

Practical Interpretation: Rejection of H_0 does not necessarily imply that the relationship found is an appropriate model for predicting pull strength as a function of wire length and die height. Further tests of model adequacy are required before we can be comfortable using this model in practice.

12-2: Hypothesis Tests in Multiple Linear Regression

Example 12-3

Table 12-10 Test for Significance of Regression for Example 12-3

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	f_0	P-value
Regression	5990.7712	2	2995.3856	572.17	1.08E-19
Error or residual	115.1735	22	5.2352		
Total	6105.9447	24			

12-2: Hypothesis Tests in Multiple Linear Regression

R² and Adjusted R²

The **coefficient of multiple determination**

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

- For the wire bond pull strength data, we find that $R^2 = SS_R/SS_T = 5990.7712/6105.9447 = 0.9811$.
- Thus, the model accounts for about 98% of the variability in the pull strength response.

12-2: Hypothesis Tests in Multiple Linear Regression

R² and Adjusted R²

The **adjusted R²** is

$$R_{\text{adj}}^2 = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)} \quad (12-23)$$

- The adjusted R² statistic penalizes the analyst for adding terms to the model.
- It can help guard against **overfitting** (including regressors that are not really useful)

12-2: Hypothesis Tests in Multiple Linear Regression

12-2.2 Tests on Individual Regression Coefficients and Subsets of Coefficients

The hypotheses for testing the significance of any individual regression coefficient:

$$\begin{aligned} H_0: \beta_j &= \beta_{j0} \\ H_1: \beta_j &\neq \beta_{j0} \end{aligned} \tag{12-24}$$

12-2: Hypothesis Tests in Multiple Linear Regression

12-2.2 Tests on Individual Regression Coefficients and Subsets of Coefficients

The test statistic is

$$T_0 = \frac{\hat{\beta}_j - \beta_{j0}}{\sqrt{\sigma^2 C_{jj}}} = \frac{\hat{\beta}_j - \beta_{j0}}{se(\hat{\beta}_j)} \quad (12-25)$$

- Reject H_0 if $|t_0| > t_{\alpha/2, n-p}$.
- This is called a **partial** or **marginal test**

12-2: Hypothesis Tests in Multiple Linear Regression

Example 12-4

EXAMPLE 12-4 Wire Bond Strength Coefficient Test

Consider the wire bond pull strength data, and suppose that we want to test the hypothesis that the regression coefficient for x_2 (die height) is zero. The hypotheses are

$$H_0: \beta_2 = 0$$

$$H_1: \beta_2 \neq 0$$

The main diagonal element of the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix corresponding to $\hat{\beta}_2$ is $C_{22} = 0.0000015$, so the t -statistic in Equation 12-25 is

$$t_0 = \frac{\hat{\beta}_2}{\sqrt{\hat{\sigma}^2 C_{22}}} = \frac{0.01253}{\sqrt{(5.2352)(0.0000015)}} = 4.477$$

12-2: Hypothesis Tests in Multiple Linear Regression

Example 12-4

Note that we have used the estimate of σ^2 reported to four decimal places in Table 12-10. Since $t_{0.025,22} = 2.074$, we reject $H_0: \beta_2 = 0$ and conclude that the variable x_2 (die height) contributes significantly to the model. We could also have used a P -value to draw conclusions. The P -value for $t_0 = 4.477$ is $P = 0.0002$, so with $\alpha = 0.05$ we would reject the null hypothesis.

Practical Interpretation: Note that this test measures the marginal or partial contribution of x_2 given that x_1 is in the model. That is, the t -test measures the contribution of adding the variable $x_2 = \text{die height}$ to a model that already contains $x_1 = \text{wire length}$. Table 12-4 shows the value of the t -test computed by Minitab. The Minitab t -test statistic is reported to two decimal places. Note that the computer produces a t -test for each regression coefficient in the model. These t -tests indicate that both regressors contribute to the model.

12-2: Hypothesis Tests in Multiple Linear Regression

The general regression significance test or the extra sum of squares method:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

We wish to test the hypotheses:

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

(12-28)

12-2: Hypothesis Tests in Multiple Linear Regression

A general form of the model can be written:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \epsilon$$

where \mathbf{X}_1 represents the columns of \mathbf{X} associated with $\boldsymbol{\beta}_1$ and \mathbf{X}_2 represents the columns of \mathbf{X} associated with $\boldsymbol{\beta}_2$

12-2: Hypothesis Tests in Multiple Linear Regression

For the full model:

$$SS_R(\boldsymbol{\beta}) = \hat{\boldsymbol{\beta}}' \mathbf{X}' \mathbf{y} \quad (p = k + 1 \text{ degrees of freedom})$$

$$MS_E = \frac{\mathbf{y}' \mathbf{y} - \hat{\boldsymbol{\beta}}' \mathbf{X}' \mathbf{y}}{n - p}$$

If H_0 is true, the reduced model is

$$\mathbf{y} = \mathbf{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$$

$$SS_R(\boldsymbol{\beta}_2) = \hat{\boldsymbol{\beta}}_2' \mathbf{X}_2' \mathbf{y} \quad (p - r \text{ degrees of freedom})$$

12-2: Hypothesis Tests in Multiple Linear Regression

The test statistic is:

$$F_0 = \frac{SS_R(\beta_1 | \beta_2)/r}{MS_E} \quad (12-33)$$

Reject H_0 if $f_0 > f_{\alpha, r, n-p}$

The test in Equation (12-32) is often referred to as a
partial F-test

12-2: Hypothesis Tests in Multiple Linear Regression

Example 12-6

EXAMPLE 12-6 Wire Bond Strength General Regression Test

Consider the wire bond pull-strength data in Example 12-1. We will investigate the contribution of two new variables, x_3 and x_4 , to the model using the partial F -test approach. The new variables are explained at the end of this example. That is, we wish to test

$$H_0: \beta_3 = \beta_4 = 0 \quad H_1: \beta_3 \neq 0 \text{ or } \beta_4 \neq 0$$

To test this hypothesis, we need the extra sum of squares due to β_3 and β_4 or

$$\begin{aligned} SS_R(\beta_4, \beta_3 | \beta_2, \beta_1, \beta_0) &= SS_R(\beta_4, \beta_3, \beta_2, \beta_1, \beta_0) - SS_R(\beta_2, \beta_1, \beta_0) \\ &= SS_R(\beta_4, \beta_3, \beta_2, \beta_1 | \beta_0) - SS_R(\beta_2, \beta_1 | \beta_0) \end{aligned}$$

12-2: Hypothesis Tests in Multiple Linear Regression

Example 12-6

In Example 12-3 we calculated

$$SS_R(\beta_2, \beta_1 | \beta_0) = \mathbf{\beta}' \mathbf{X}' \mathbf{y} - \frac{\left(\sum_{i=1}^n y_i \right)^2}{n} = 5990.7712 \text{ (two degrees of freedom)}$$

Also, Table 12-4 shows the Minitab output for the model with only x_1 and x_2 as predictors. In the analysis of variance table, we can see that $SS_R = 5990.8$ and this agrees with our calculation. In practice, the computer output would be used to obtain this sum of squares.

If we fit the model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$, we can use the same matrix formula. Alternatively, we can look at SS_R from computer output for this model. The analysis of variance table for this model is shown in Table 12-11 and we see that

$$SS_R(\beta_4, \beta_3, \beta_2, \beta_1 | \beta_0) = 6024.0 \text{ (four degrees of freedom)}$$

Therefore,

$$SS_R(\beta_4, \beta_3 | \beta_2, \beta_1, \beta_0) = 6024.0 - 5990.8 = 33.2 \text{ (two degrees of freedom)}$$

12-2: Hypothesis Tests in Multiple Linear Regression

Example 12-6

This is the increase in the regression sum of squares due to adding x_3 and x_4 to a model already containing x_1 and x_2 . To test H_0 , calculate the test statistic

$$f_0 = \frac{SS_R(\beta_4, \beta_3 | \beta_2, \beta_1, \beta_0)/2}{MS_E} = \frac{33.2/2}{4.1} = 4.05$$

Note that MS_E from the full model using x_1 , x_2 , x_3 and x_4 is used in the denominator of the test statistic. Because $f_{0.05, 2, 20} = 3.49$, we reject H_0 and conclude that at least one of the new variables contributes significantly to the model. Further analysis and tests will be needed to refine the model and determine if one or both of x_3 and x_4 are important.

The mystery of the new variables can now be explained. These are quadratic powers of the original predictors wire length and wire height. That is, $x_3 = x_1^2$ and $x_4 = x_2^2$. A test for quadratic terms is a common use of partial F -tests. With this information and the original data for x_1 and x_2 , you can use computer software to reproduce these calculations. Multiple regression allows models to be extended in such a simple manner that the real meaning of x_3 and x_4 did not even enter into the test procedure. Polynomial models such as this are discussed further in Section 12-6.

12-3: Confidence Intervals in Multiple Linear Regression

12-3.1 Confidence Intervals on Individual Regression Coefficients

Definition

A $100(1 - \alpha)\%$ confidence interval on the regression coefficient β_j , $j = 0, 1, \dots, k$ in the multiple linear regression model is given by

$$\hat{\beta}_j - t_{\alpha/2,n-p} \sqrt{\hat{\sigma}^2 C_{jj}} \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2,n-p} \sqrt{\hat{\sigma}^2 C_{jj}} \quad (12-35)$$

12-3: Confidence Intervals in Multiple Linear Regression

Example 12-7

EXAMPLE 12-7 Wire Bond Strength Confidence Interval

We will construct a 95% confidence interval on the parameter β_1 in the wire bond pull strength problem. The point estimate of β_1 is $\hat{\beta}_1 = 2.74427$ and the diagonal element of $(X'X)^{-1}$ corresponding to β_1 is $C_{11} = 0.001671$. The estimate of σ^2 is $\hat{\sigma}^2 = 5.2352$, and $t_{0.025,22} = 2.074$. Therefore, the 95% CI on β_1 is computed from Equation 12-35 as

$$\begin{aligned} 2.74427 - (2.074)\sqrt{(5.2352)(.001671)} &\leq \beta_1 \leq 2.74427 \\ + (2.074)\sqrt{(5.2352)(.001671)} \end{aligned}$$

which reduces to

$$2.55029 \leq \beta_1 \leq 2.93825$$

Also, computer software such as Minitab can be used to help calculate this confidence interval. From the regression output in Table 10-4, $\hat{\beta}_1 = 2.74427$ and the standard error of $\hat{\beta}_1 = 0.0935$. This standard error is the multiplier of the t -table constant in the confidence interval. That is, $0.0935 = \sqrt{(5.2352)(0.001671)}$. Consequently, all the numbers are available from the computer output to construct the interval and this is the typical method used in practice.

12-3: Confidence Intervals in Multiple Linear Regression

12-3.2 Confidence Interval on the Mean Response

The mean response at a point \mathbf{x}_0 is estimated by

$$\hat{\mu}_{Y|\mathbf{x}_0} = \mathbf{x}'_0 \hat{\beta}$$

The variance of the estimated mean response is

$$V(\hat{\mu}_{Y|\mathbf{x}_0}) = \sigma^2 \mathbf{x}'_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0$$

12-3: Confidence Intervals in Multiple Linear Regression

12-3.2 Confidence Interval on the Mean Response

Definition

For the multiple linear regression model, a $100(1 - \alpha)\%$ confidence interval on the mean response at the point $x_{01}, x_{02}, \dots, x_{0k}$ is

$$\begin{aligned}\hat{\mu}_{Y|x_0} &= t_{\alpha/2,n-p} \sqrt{\hat{\sigma}^2 \mathbf{x}'_0 (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_0} \\ &\leq \mu_{Y|x_0} \leq \hat{\mu}_{Y|x_0} + t_{\alpha/2,n-p} \sqrt{\hat{\sigma}^2 \mathbf{x}'_0 (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_0}\end{aligned}\quad (12-39)$$

12-3: Confidence Intervals in Multiple Linear Regression

EXAMPLE 12-8 Wire Bond Strength Confidence Interval on the Mean Response

Example 12-8

The engineer in Example 12-1 would like to construct a 95% CI on the mean pull strength for a wire bond with wire length $x_1 = 8$ and die height $x_2 = 275$. Therefore,

$$\mathbf{x}_0 = \begin{bmatrix} 1 \\ 8 \\ 275 \end{bmatrix}$$

The estimated mean response at this point is found from Equation 12-36 as

$$\hat{\mu}_{Y|\mathbf{x}_0} = \mathbf{x}_0' \hat{\boldsymbol{\beta}} = [1 \quad 8 \quad 275] \begin{bmatrix} 2.26379 \\ 2.74427 \\ 0.01253 \end{bmatrix} = 27.66$$

12-3: Confidence Intervals in Multiple Linear Regression

Example 12-8

The variance of $\hat{\mu}_{Y|x_0}$ is estimated by

$$\hat{\sigma}^2 x_0' (\mathbf{X}' \mathbf{X})^{-1} x_0 = 5.2352 [1 \ 8 \ 275]$$

$$\times \begin{bmatrix} .214653 & -.007491 & -.000340 \\ -.007491 & .001671 & -.000019 \\ -.000340 & -.000019 & .0000015 \end{bmatrix} \begin{bmatrix} 1 \\ 8 \\ 275 \end{bmatrix}$$

$$= 5.2352 (0.0444) = 0.23244$$

Therefore, a 95% CI on the mean pull strength at this point is found from Equation 12-39 as

$$27.66 - 2.074 \sqrt{0.23244} \leq \mu_{Y|x_0} \leq 27.66 + 2.074 \sqrt{0.23244}$$

which reduces to

$$26.66 \leq \mu_{Y|x_0} \leq 28.66$$

12-4: Prediction of New Observations

A point estimate of the future observation Y_0 is

$$\hat{y}_0 = \mathbf{x}'_0 \hat{\beta}$$

A $100(1-\alpha)\%$ **prediction interval** for this future observation is

A $100(1 - \alpha)\%$ **prediction interval** for this future observation is

$$\begin{aligned} \hat{y}_0 - t_{\alpha/2,n-p} \sqrt{\hat{\sigma}^2(1 + \mathbf{x}'_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0)} \\ \leq Y_0 \leq \hat{y}_0 + t_{\alpha/2,n-p} \sqrt{\hat{\sigma}^2(1 + \mathbf{x}'_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0)} \end{aligned} \quad (12-41)$$

12-4: Prediction of New Observations

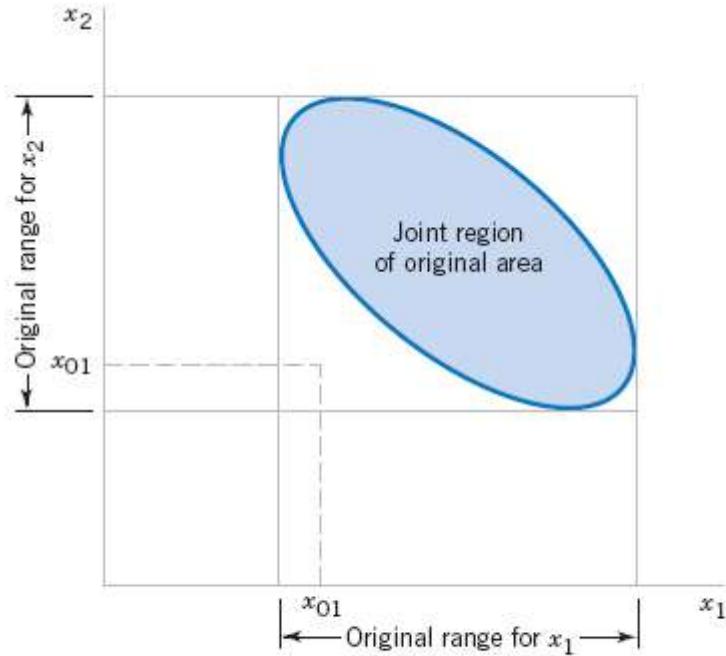


Figure 12-5 An example of extrapolation in multiple regression.

Figure 12-5 An example of extrapolation in multiple regression

12-4: Prediction of New Observations

Example 12-9

EXAMPLE 12-9 Wire Bond Strength Confidence Interval

Suppose that the engineer in Example 12-1 wishes to construct a 95% prediction interval on the wire bond pull strength when the wire length is $x_1 = 8$ and the die height is $x_2 = 275$. Note that $\mathbf{x}'_0 = [1 \ 8 \ 275]$, and the point estimate of the pull strength is $\hat{y}_0 = \mathbf{x}'_0 \hat{\beta} = 27.66$. Also, in Example 12-8 we calculated $\mathbf{x}'_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0 = 0.04444$. Therefore, from Equation 12-41 we have

$$27.66 - 2.074 \sqrt{5.2352(1 + 0.0444)} \leq Y_0 \leq 27.66 + 2.074 \sqrt{5.2352(1 + 0.0444)}$$

and the 95% prediction interval is

$$22.81 \leq Y_0 \leq 32.51$$

Notice that the prediction interval is wider than the confidence interval on the mean response at the same point, calculated in Example 12-8. The Minitab output in Table 12-4 also displays this prediction interval.

12-5: Model Adequacy Checking

12-5.1 Residual Analysis

Example 12-10

The residuals for the model from Example 12-1 are shown in Table 12-3. A normal probability plot of these residuals is shown in Fig. 12-6. No severe deviations from normality are obviously apparent, although the two largest residuals ($e_{15} = 5.84$ and $e_{17} = 4.33$) do not fall extremely close to a straight line drawn through the remaining residuals.

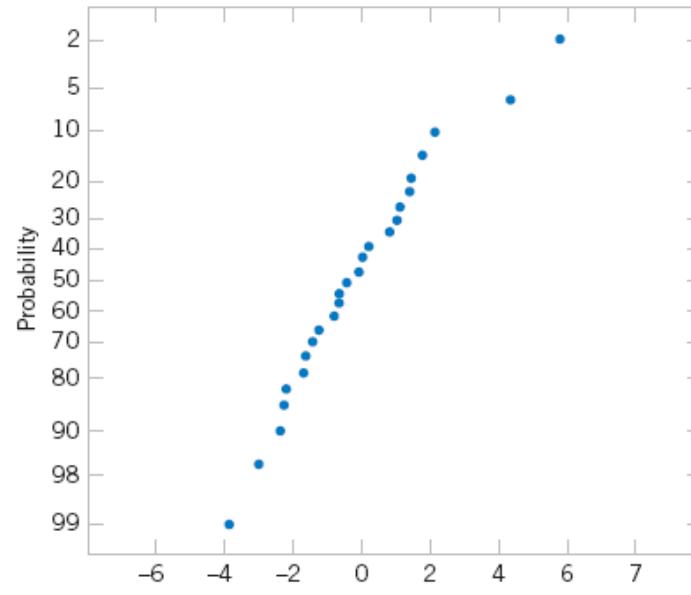


Figure 12-6 Normal probability plot of residuals

Figure 12-6 Normal probability plot of residuals.

12-5: Model Adequacy Checking

12-5.1 Residual Analysis

Example 12-10

The **standardized residuals**

$$d_i = \frac{e_i}{\sqrt{MS_E}} = \frac{e_i}{\sqrt{\sigma^2}} \quad (12-42)$$

are often more useful than the ordinary residuals when assessing residual magnitude. For the wire bond strength example, the standardized residuals corresponding to e_{15} and e_{17} are $d_{15} = 5.84/\sqrt{5.2352} = 2.55$ and $d_{17} = 4.33/\sqrt{5.2352} = 1.89$, and they do not seem unusually large. Inspection of the data does not reveal any error in collecting observations 15 and 17, nor does it produce any other reason to discard or modify these two points.

12-5: Model Adequacy Checking

12-5.1 Residual Analysis

Example 12-10

The residuals are plotted against \hat{y} in Fig. 12-7, and against x_1 and x_2 in Figs. 12-8 and 12-9, respectively.* The two largest residuals, e_{15} and e_{17} , are apparent. Figure 12-8 gives some indication that the model underpredicts the pull strength for assemblies with short wire length ($x_1 \leq 6$) and long wire length ($x_1 \geq 15$) and overpredicts the strength for assemblies with intermediate wire length ($7 \leq x_1 \leq 14$). The same impression is obtained from Fig. 12-7. Either the relationship between strength and wire length is not linear (requiring that a term involving x_1^2 , say, be added to the model), or other regressor variables not presently in the model affected the response.

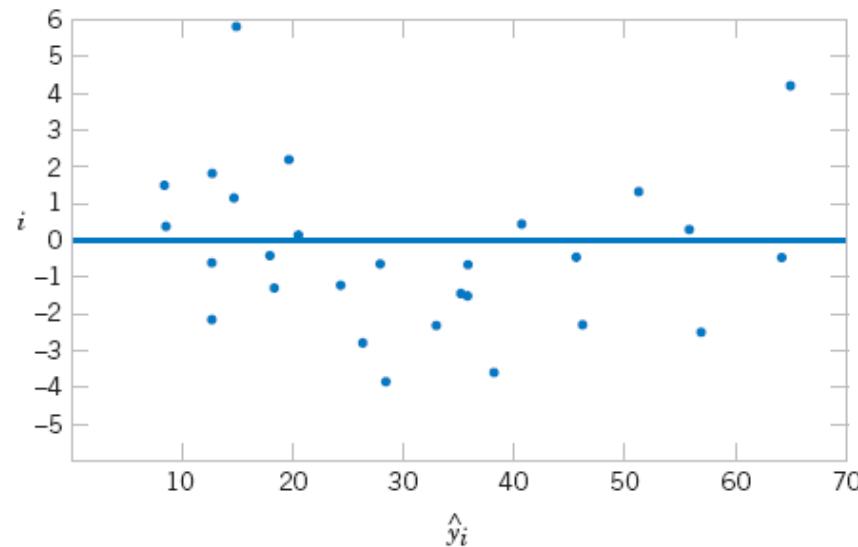


Figure 12-7 Plot of residuals

Figure 12-7 Plot of residuals against \hat{y} .

12-5: Model Adequacy Checking

12-5.1 Residual Analysis

Example 12-10

Either the relationship between strength and wire length is not linear (requiring that a term involving x_1^2 , say, be added to the model), or other regressor variables not presently in the model affected the response.

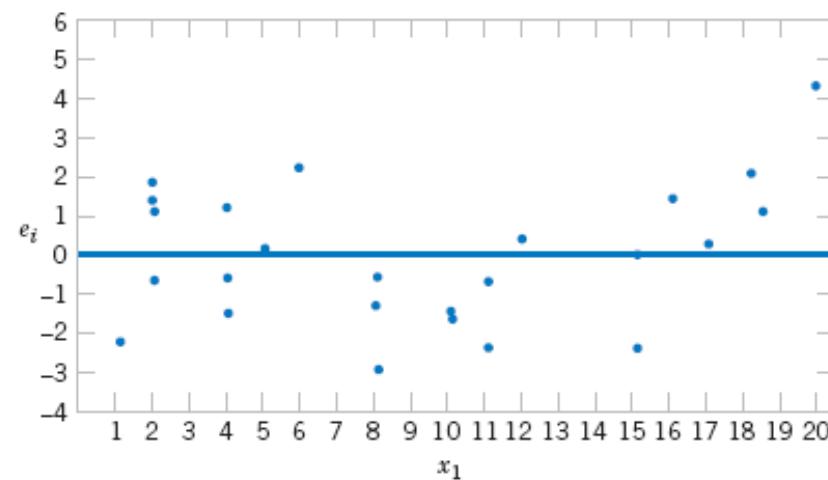


Figure 12-8 Plot of residuals against x_1 .

Figure 12-8 Plot of residuals against x_1 .

12-5: Model Adequacy Checking

12-5.1 Residual Analysis

Example 12-10

Figure 12-9 Plot of residuals against x_2 .

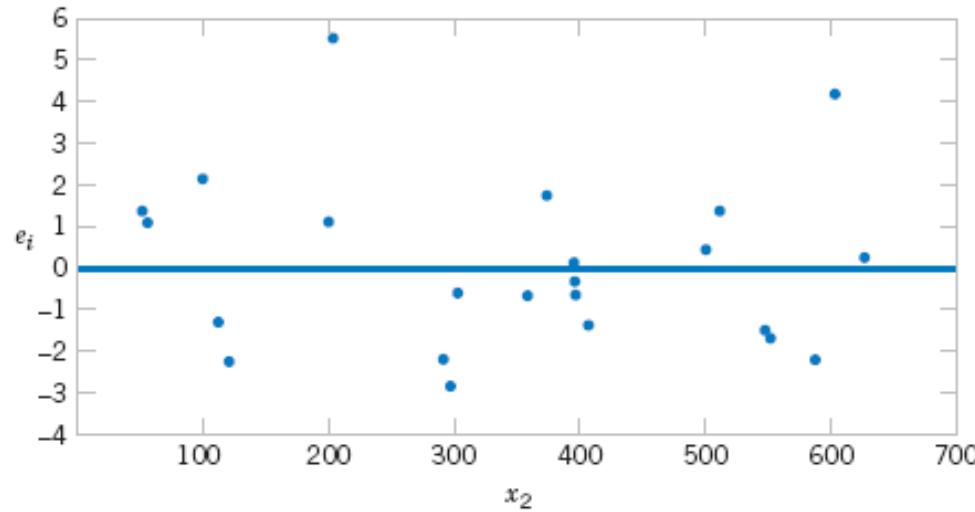


Figure 12-9 Plot of residuals against x_2 .

12-5: Model Adequacy Checking

12-5.1 Residual Analysis

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}} \quad i = 1, 2, \dots, n \quad (12-43)$$

where h_{ii} is the i th diagonal element of the matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

The \mathbf{H} matrix is sometimes called the “**hat**” matrix, since

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

12-5: Model Adequacy Checking

12-5.1 Residual Analysis

Since each row of the matrix \mathbf{X} corresponds to a vector, say $\mathbf{x}'_i = [1, x_{i1}, x_{i2}, \dots, x_{ik}]$, another way to write the diagonal elements of the hat matrix is

$$h_{ii} = \mathbf{x}'_i (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_i \quad (12-44)$$

The variance of the i th residual is

$$V(e_i) = \sigma^2(1 - h_{ii}), \quad i = 1, 2, \dots, n$$

12-5: Model Adequacy Checking

12-5.1 Residual Analysis

To illustrate, consider the two observations identified in the wire bond strength data (Example 12-10) as having residuals that might be unusually large, observations 15 and 17. The standardized residuals are

$$d_{15} = \frac{e_{15}}{\sqrt{\hat{\sigma}^2}} = \frac{5.84}{\sqrt{5.2352}} = 2.55 \quad \text{and} \quad d_{17} = \frac{e_{17}}{\sqrt{MS_E}} = \frac{4.33}{\sqrt{5.2352}} = 1.89$$

Now $h_{15,15} = 0.0737$ and $h_{17,17} = 0.2593$, so the studentized residuals are

$$r_{15} = \frac{e_{15}}{\sqrt{\hat{\sigma}^2(1 - h_{15,15})}} = \frac{5.84}{\sqrt{5.2352(1 - 0.0737)}} = 2.65$$

and

$$r_{17} = \frac{e_{17}}{\sqrt{\hat{\sigma}^2(1 - h_{17,17})}} = \frac{4.33}{\sqrt{5.2352(1 - 0.2593)}} = 2.20$$

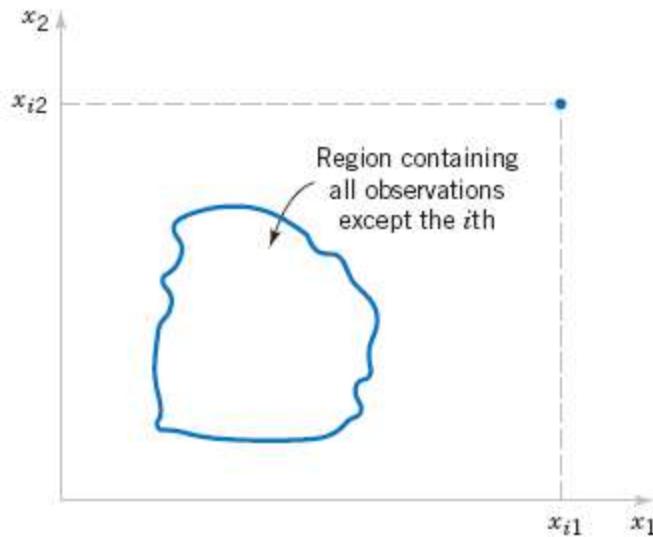
Notice that the studentized residuals are larger than the corresponding standardized residuals. However, the studentized residuals are still not so large as to cause us serious concern about possible outliers.

12-5: Model Adequacy Checking

12-5.2 Influential Observations

Figure 12-10 A point that is remote in x-space.

Figure 12-10 A point that is remote in x-space.



12-5: Model Adequacy Checking

12-5.2 Influential Observations

Cook's distance measure

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})' \mathbf{X}' \mathbf{X} (\hat{\beta}_{(i)} - \hat{\beta})}{p \hat{\sigma}^2} \quad i = 1, 2, \dots, n$$

$$D_i = \frac{r_i^2}{p} \frac{h_{ii}}{(1 - h_{ii})} \quad i = 1, 2, \dots, n \quad (12-45)$$

12-5: Model Adequacy Checking

Example 12-11

EXAMPLE 12-11 Wire Bond Strength Cook's Distances

Table 12-12 lists the values of the hat matrix diagonals h_{ii} and Cook's distance measure D_i for the wire bond pull strength data in Example 12-1. To illustrate the calculations, consider the first observation:

$$\begin{aligned} D_1 &= \frac{r_1^2}{p} \cdot \frac{h_{11}}{(1 - h_{11})} \\ &= -\frac{[e_1/\sqrt{MS_E(1 - h_{11})}]^2}{p} \cdot \frac{h_{11}}{(1 - h_{11})} \end{aligned}$$

$$\begin{aligned} &= \frac{[1.57/\sqrt{5.2352(1 - 0.1573)}]^2}{3} \cdot \frac{0.1573}{(1 - 0.1573)} \\ &= 0.035 \end{aligned}$$

The Cook distance measure D_i does not identify any potentially influential observations in the data, for no value of D_i exceeds unity.

12-5: Model Adequacy Checking

Example 12-11

Table 12-12 Influence Diagnostics for the Wire Bond Pull Strength Data 2

Observations		Cook's Distance Measure		Observations		Cook's Distance Measure	
i	h_{ii}		D_i	i	h_{ii}		D_i
1	0.1573		0.035	14	0.1129		0.003
2	0.1116		0.012	15	0.0737		0.187
3	0.1419		0.060	16	0.0879		0.001
4	0.1019		0.021	17	0.2593		0.565
5	0.0418		0.024	18	0.2929		0.155
6	0.0749		0.007	19	0.0962		0.018
7	0.1181		0.036	20	0.1473		0.000
8	0.1561		0.020	21	0.1296		0.052
9	0.1280		0.160	22	0.1358		0.028
10	0.0413		0.001	23	0.1824		0.002
11	0.0925		0.013	24	0.1091		0.040
12	0.0526		0.001	25	0.0729		0.000
13	0.0820		0.001				

12-6: Aspects of Multiple Regression Modeling

12-6.1 Polynomial Regression Models

The linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is a general model that can be used to fit any relationship that is **linear in the unknown parameters $\boldsymbol{\beta}$** . This includes the important class of **polynomial regression models**. For example, the second-degree polynomial in one variable

$$Y = \beta_0 + \beta_1 x + \beta_{11} x^2 + \epsilon \quad (12-46)$$

and the second-degree polynomial in two variables

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon \quad (12-47)$$

are linear regression models.

12-6: Aspects of Multiple Regression Modeling

Example 12-12

EXAMPLE 12-12 Airplane Sidewall Panels

Sidewall panels for the interior of an airplane are formed in a 1500-ton press. The unit manufacturing cost varies with the production lot size. The data shown below give the average cost per unit (in hundreds of dollars) for this product (y) and the production lot size (x). The scatter diagram, shown in Fig. 12-11, indicates that a second-order polynomial may be appropriate.

y	1.81	1.70	1.65	1.55	1.48	1.40
x	20	25	30	35	40	50
y	1.30	1.26	1.24	1.21	1.20	1.18
x	60	65	70	75	80	90

12-6: Aspects of Multiple Regression Modeling

Example 12-11

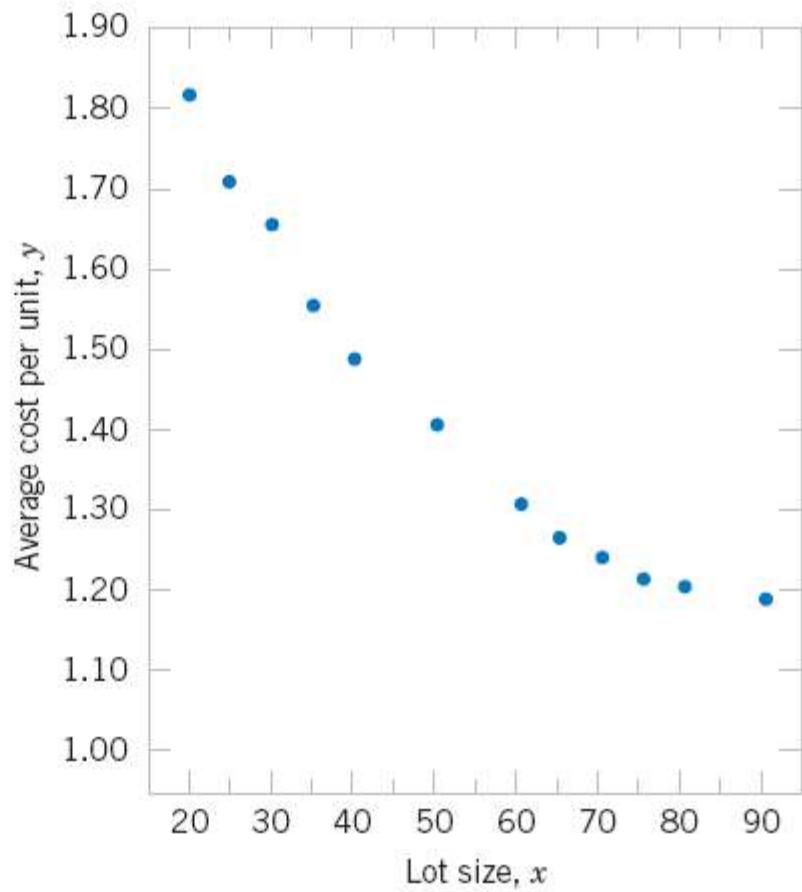


Figure 12-11 Data for Example 12-11.

Figure 12-11 Data for Example 12-11.

Example 12-12

We will fit the model

$$Y = \beta_0 + \beta_1 x + \beta_{11} x^2 + \epsilon$$

The \mathbf{y} vector, the model matrix \mathbf{X} and the $\boldsymbol{\beta}$ vector are as follows:

$$\mathbf{y} = \begin{bmatrix} 1.81 \\ 1.70 \\ 1.65 \\ 1.55 \\ 1.48 \\ 1.40 \\ 1.30 \\ 1.26 \\ 1.24 \\ 1.21 \\ 1.20 \\ 1.18 \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & 20 & 400 \\ 1 & 25 & 625 \\ 1 & 30 & 900 \\ 1 & 35 & 1225 \\ 1 & 40 & 1600 \\ 1 & 50 & 2500 \\ 1 & 60 & 3600 \\ 1 & 65 & 4225 \\ 1 & 70 & 4900 \\ 1 & 75 & 5625 \\ 1 & 80 & 6400 \\ 1 & 90 & 8100 \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_{11} \end{bmatrix}$$

12-6: Aspects of Multiple Regression Modeling

Example 12-12

Solving the normal equations $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$ gives the fitted model

$$\hat{y} = 2.19826629 - 0.02252236x + 0.00012507x^2$$

Conclusions: The test for significance of regression is shown in Table 12-13. Since $f_0 = 1762.3$ is significant at 1%, we conclude that at least one of the parameters β_1 and β_{11} is not zero. Furthermore, the standard tests for model adequacy do not reveal any unusual behavior, and we would conclude that this is a reasonable model for the sidewall panel cost data.

Table 12-13 Test for Significance of Regression for the Second-Order Model in Example 12-12

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	f_0	P-value
Regression	0.52516	2	0.26258	1762.28	2.12E-12
Error	0.00134	9	0.00015		
Total	0.5265	11			

12-6: Aspects of Multiple Regression Modeling

12-6.2 Categorical Regressors and Indicator Variables

- Many problems may involve **qualitative** or **categorical** variables.
- The usual method for the different levels of a qualitative variable is to use **indicator** variables.
- For example, to introduce the effect of two different operators into a regression model, we could define an indicator variable as follows:

$$x = \begin{cases} 0 & \text{if the observation is from operator 1} \\ 1 & \text{if the observation is from operator 2} \end{cases}$$

12-6: Aspects of Multiple Regression Modeling

Example 12-13

EXAMPLE 12-13 Surface Finish

A mechanical engineer is investigating the surface finish of metal parts produced on a lathe and its relationship to the speed (in revolutions per minute) of the lathe. The data are shown in Table 12-15. Note that the data have been collected using two different types of cutting tools. Since the type of cutting tool likely affects the surface finish, we will fit the model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

where Y is the surface finish, x_1 is the lathe speed in revolutions per minute, and x_2 is an indicator variable denoting the type of cutting tool used; that is,

$$x_2 = \begin{cases} 0, & \text{for tool type 302} \\ 1, & \text{for tool type 416} \end{cases}$$

12-6: Aspects of Multiple Regression Modeling

Example 12-13

Table 12-15 Surface Finish Data for Example 12-13

Observation Number, i	Surface Finish y_i	RPM	Type of Cutting Tool	Observation Number, i	Surface Finish y_i	RPM	Type of Cutting Tool
1	45.44	225	302	11	33.50	224	416
2	42.03	200	302	12	31.23	212	416
3	50.10	250	302	13	37.52	248	416
4	48.75	245	302	14	37.13	260	416
5	47.92	235	302	15	34.70	243	416
6	47.79	237	302	16	33.92	238	416
7	52.26	265	302	17	32.13	224	416
8	50.52	259	302	18	35.47	251	416
9	45.58	221	302	19	33.49	232	416
10	44.78	218	302	20	32.29	216	416

12-6: Aspects of Multiple Regression Modeling

Example 12-13

The parameters in this model may be easily interpreted. If $x_2 = 0$, the model becomes

$$Y = \beta_0 + \beta_1 x_1 + \epsilon$$

which is a straight-line model with slope β_1 and intercept β_0 . However, if $x_2 = 1$, the model becomes

$$Y = \beta_0 + \beta_1 x_1 + \beta_2(1) + \epsilon = (\beta_0 + \beta_2) + \beta_1 x_1 + \epsilon$$

which is a straight-line model with slope β_1 and intercept $\beta_0 + \beta_2$. Thus, the model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ implies that surface finish is linearly related to lathe speed and that the slope β_1 does not depend on the type of cutting tool used. However, the type of cutting tool does affect the intercept, and β_2 indicates the change in the intercept associated with a change in tool type from 302 to 416.

Example 12-12

The model matrix \mathbf{X} and \mathbf{y} vector for this problem are as follows:

$\mathbf{X} =$	1 225 0	45.44
	1 200 0	42.03
	1 250 0	50.10
	1 245 0	48.75
	1 235 0	47.92
	1 237 0	47.79
	1 265 0	52.26
	1 259 0	50.52
	1 221 0	45.58
	1 218 0	44.78
	1 224 1	33.50
	1 212 1	31.23
	1 248 1	37.52
	1 260 1	37.13
	1 243 1	34.70
	1 238 1	33.92
	1 224 1	32.13
	1 251 1	35.47
	1 232 1	33.49
	1 216 1	32.29

12-6: Aspects of Multiple Regression Modeling

Example 12-13

The fitted model is

$$\hat{y} = 14.27620 + 0.14115x_1 - 13.28020x_2$$

Conclusions: The analysis of variance for this model is shown in Table 12-16. Note that the hypothesis $H_0: \beta_1 = \beta_2 = 0$ (significance of regression) would be rejected at any reasonable level of significance because the P -value is very small. This table also contains the sums of squares

$$\begin{aligned} SS_R &= SS_R(\beta_1, \beta_2 | \beta_0) \\ &= SS_R(\beta_1 | \beta_0) + SS_R(\beta_2 | \beta_1, \beta_0) \end{aligned}$$

so a test of the hypothesis $H_0: \beta_2 = 0$ can be made. Since this hypothesis is also rejected, we conclude that tool type has an effect on surface finish.

12-6: Aspects of Multiple Regression Modeling

Example 12-13

Table 12-16 Analysis of Variance for Example 12-13

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	f_0	P-value
Regression	1012.0595	2	506.0297	1103.69	1.02E-18
$SS_R(\beta_1 \beta_0)$	130.6091	1	130.6091	284.87	4.70E-12
$SS_R(\beta_2 \beta_1,\beta_0)$	881.4504	1	881.4504	1922.52	6.24E-19
Error	7.7943	17	0.4585		
Total	1019.8538	19			

12-6: Aspects of Multiple Regression Modeling

12-6.3 Selection of Variables and Model Building

$$C_p = \frac{SS_E(p)}{\hat{\sigma}^2} - n + 2p \quad (12-48)$$

$$\text{PRESS} = \sum_{i=1}^n (y_i - \hat{y}_{(i)})^2 = \sum_{i=1}^n \left(\frac{e_i}{1 - h_{ii}} \right)^2$$

12-6: Aspects of Multiple Regression Modeling

12-6.3 Selection of Variables and Model Building

All Possible Regressions – Example 12-14

EXAMPLE 12-14 Wine Quality

Table 12-17 presents data on taste-testing 38 brands of pinot noir wine (the data were first reported in an article by Kwan, Kowalski, and Skogenboe in an article in the *Journal of Agricultural and Food Chemistry*, Vol. 27, 1979, and it also appears as one of the default data sets in Minitab). The response variable is $y = \text{quality}$, and we wish to find the “best” regression equation that relates quality to the other five parameters.

Figure 12-12 is the matrix of scatter plots for the wine quality data, as constructed by Minitab. We notice that there are some indications of possible linear relationships between quality and the regressors, but there is no obvious visual impression of which regressors would be appropriate. Table 12-18 lists the all possible regressions output from Minitab. In this analysis, we asked Minitab to present the best three equations for each subset size. Note that Minitab reports the values of R^2 , R_{adj}^2 , C_p , and $S = \sqrt{MS_E}$ for each model.

12-6: Aspects of Multiple Regression Modeling

12-6.3 Selection of Variables and Model Building All Possible Regressions – Example 12-14

From Table 12-18 we see that the three-variable equation with $x_2 = \text{aroma}$, $x_4 = \text{flavor}$, and $x_5 = \text{oakiness}$ produces the minimum C_p equation, whereas the four-variable model, which adds $x_1 = \text{clarity}$ to the previous three regressors, results in maximum R_{adj}^2 (or minimum MS_E).

The three-variable model is

$$\hat{y} = 6.47 + 0.580x_2 + 1.20x_4 - 0.602x_5$$

and the four-variable model is

$$\hat{y} = 4.99 + 1.79x_1 + 0.530x_2 + 1.26x_4 - 0.659x_5$$

12-6: Aspects of Multiple Regression Modeling

12-6.3 Selection of Variables and Model Building All Possible Regressions – Example 12-14

Figure 12-12 A matrix of Scatter plots from Minitab for the Wine Quality Data.

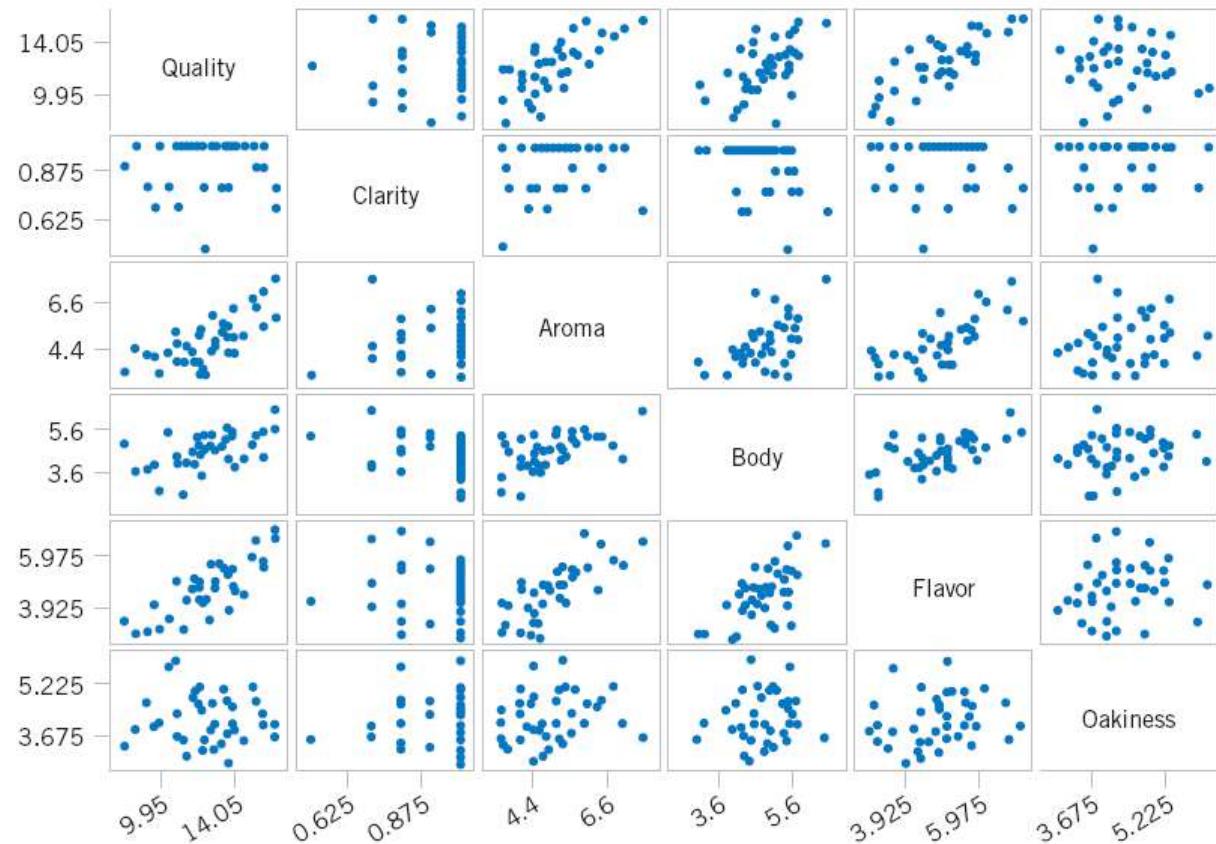


Figure 12-12 A matrix of scatter plots from Minitab for the wine quality data.

Table 12-18 Minitab All Possible Regressions Output for the Wine Quality Data

Best Subsets Regression: Quality versus Clarity, Aroma, . . .

Response is Quality

Vars	R-Sq	R-Sq (adj)	C-p	S	O
1	62.4	61.4	9.0	1.2712	C a
1	50.0	48.6	23.2	1.4658	I F k
1	30.1	28.2	46.0	1.7335	a A 1 i
2	66.1	64.2	6.8	1.2242	r r B a n
2	65.9	63.9	7.1	1.2288	i o o v e
2	63.3	61.2	10.0	1.2733	t m d o s
3	70.4	67.8	3.9	1.1613	y a y r s X
3	68.0	65.2	6.6	1.2068	X X X X
3	66.5	63.5	8.4	1.2357	X X X X X
4	71.5	68.0	4.7	1.1568	X X X X X
4	70.5	66.9	5.8	1.1769	X X X X X X
4	69.3	65.6	7.1	1.1996	X X X X X X
5	72.1	67.7	6.0	1.1625	X X X X X X X

12-6.3: Selection of Variables and Model Building - Stepwise Regression

Example 12-14

Table 12-19 Minitab Stepwise Regression Output for the Wine Quality Data

Stepwise Regression: Quality versus Clarity, Aroma, ...

Alpha-to-Enter: 0.15 Alpha-to-Remove: 0.15

Response is Quality on 5 predictors, with N = 38

Step	1	2	3
Constant	4.941	6.912	6.467
Flavor	1.57	1.64	1.20
T-Value	7.73	8.25	4.36
P-Value	0.000	0.000	0.000
Oakiness		-0.54	-0.60
T-Value		-1.95	-2.28
P-Value		0.059	0.029
Aroma			0.58
T-Value			2.21
P-Value			0.034
S	1.27	1.22	1.16
R-Sq	62.42	66.11	70.38
R-Sq(adj)	61.37	64.17	67.76
C-p	9.0	6.8	3.9

12-6.3: Selection of Variables and Model Building - Backward Regression

Table 12-20 Minitab Backward Elimination Output for the Wine Quality Data

Stepwise Regression: Quality versus Clarity, Aroma, ...

Backward elimination. Alpha-to-Remove: 0.1

Response is Quality on 5 predictors, with N = 38

Step	1	2	3
Constant	3.997	4.986	6.467
Clarity	2.3	1.8	
T-Value	1.35	1.12	
P-Value	0.187	0.269	
Aroma	0.48	0.53	0.58
T-Value	1.77	2.00	2.21
P-Value	0.086	0.054	0.034
Body	0.27		
T-Value	0.82		
P-Value	0.418		
Flavor	1.17	1.26	1.20
T-Value	3.84	4.52	4.36
P-Value	0.001	0.000	0.000
Oakiness	-0.68	-0.66	-0.60
T-Value	-2.52	-2.46	-2.28
P-Value	0.017	0.019	0.029
S	1.16	1.16	1.16
R-Sq	72.06	71.47	70.38
R-Sq(adj)	67.69	68.01	67.76
C-p	6.0	4.7	3.9

Example 12-14

12-6: Aspects of Multiple Regression Modeling

12-6.4 Multicollinearity

Variance Inflation Factor (VIF)

$$VIF(\beta_j) = \frac{1}{(1 - R_j^2)} \quad j = 1, 2, \dots, k \quad (12-51)$$

12-6: Aspects of Multiple Regression Modeling

12-6.4 Multicollinearity

The presence of multicollinearity can be detected in several ways. Two of the more easily understood of these are:

1. The **variance inflation factors**, defined in equation 12-50, are very useful measures of multicollinearity. The larger the variance inflation factor, the more severe the multicollinearity. Some authors have suggested that if any variance inflation factor exceeds 10, multicollinearity is a problem. Other authors consider this value too liberal and suggest that the variance inflation factors should not exceed 4 or 5. Minitab will calculate the variance inflation factors. Table 12-4 presents the Minitab multiple regression output for the wire bond pull strength data. Since both VIF_1 and VIF_2 are small, there is no problem with multicollinearity.
2. If the F -test for significance of regression is significant, but tests on the individual regression coefficients are not significant, multicollinearity may be present.

Important Terms & Concepts

All possible regressions	Model parameters & their interpretation in multiple regression
Analysis of variance test in multiple regression	Multicollinearity
Categorical variables	Multiple regression
Confidence intervals on the mean response	Outliers
Cp statistic	Polynomial regression model
Extra sum of squares method	Prediction interval on a future observation
Hidden extrapolation	PRESS statistic
Indicator variables	Residual analysis & model adequacy checking
Inference (test & intervals) on individual model parameters	Significance of regression
Influential observations	Stepwise regression & related methods
	Variance Inflation Factor (VIF)

Foundation of Data Science & Analytics

(FDSA)

Model Evaluation

Dr. Arun K. Timalsina

Materials Adaptation :

Classification: Basic Concepts

- Classification: Basic Concepts
- Decision Tree Induction
- Bayes Classification Methods
- Rule-Based Classification
- Model Evaluation and Selection
- Techniques to Improve Classification Accuracy:
Ensemble Methods
- Summary



Model Evaluation and Selection

- Evaluation metrics: How can we measure accuracy? Other metrics to consider?
- Use **validation test set** of class-labeled tuples instead of training set when assessing accuracy
- Methods for estimating a classifier's accuracy:
 - Holdout method, random subsampling
 - Cross-validation
 - Bootstrap
- Comparing classifiers:
 - Confidence intervals
 - Cost-benefit analysis and ROC Curves

Classifier Evaluation Metrics: Confusion Matrix

Confusion Matrix:

Actual class\Predicted class	C_1	$\neg C_1$
C_1	True Positives (TP)	False Negatives (FN)
$\neg C_1$	False Positives (FP)	True Negatives (TN)

Example of Confusion Matrix:

Actual class\Predicted class	buy_computer = yes	buy_computer = no	Total
buy_computer = yes	6954	46	7000
buy_computer = no	412	2588	3000
Total	7366	2634	10000

- Given m classes, an entry, $CM_{i,j}$ in a **confusion matrix** indicates # of tuples in class i that were labeled by the classifier as class j
- May have extra rows/columns to provide totals

Multi-class Problem : Confusion Matrix

	C1	C2	C3
C1	TP	Other	Other
C2	Other	TP	Other
C3	Other	Other	TP

Classifier Evaluation Metrics: Accuracy, Error Rate, Sensitivity and Specificity

A\P	C	-C	
C	TP	FN	P
-C	FP	TN	N
	P'	N'	All

- Classifier Accuracy, or *Recognition Rate*
- Percentage of test set tuples that are correctly classified
$$\text{Accuracy} = (\text{TP} + \text{TN})/\text{All}$$
- Error rate: $1 - \text{accuracy}$, or
$$\text{Error rate} = (\text{FP} + \text{FN})/\text{All}$$

Classifier Evaluation Metrics: Accuracy, Error Rate, Sensitivity and Specificity

A\P	C	$\neg C$	
C	TP	FN	P
$\neg C$	FP	TN	N
	P'	N'	All

- **Classifier Accuracy**, or recognition rate: percentage of test set tuples that are correctly classified

$$\text{Accuracy} = (\text{TP} + \text{TN})/\text{All}$$

- **Error rate**: $1 - \text{accuracy}$, or
$$\text{Error rate} = (\text{FP} + \text{FN})/\text{All}$$

- **Class Imbalance Problem**:
 - One class may be *rare*, e.g. fraud, or HIV-positive
 - Significant *majority of the negative class* and minority of the positive class
- **Sensitivity**: True Positive recognition rate
 - $\text{Sensitivity} = \text{TP}/\text{P}$
- **Specificity**: True Negative recognition rate
 - $\text{Specificity} = \text{TN}/\text{N}$

Classifier Evaluation Metrics: Precision and Recall, and F-measures

- **Precision:** exactness – what % of tuples that the classifier labeled as positive are actually positive

$$precision = \frac{TP}{TP + FP}$$

- **Recall:** completeness – what % of positive tuples did the classifier label as positive?

$$recall = \frac{TP}{TP + FN}$$

- Perfect score is 1.0
- Inverse relationship between precision & recall
- **F measure (F_1 or F-score):** harmonic mean of precision and recall,

$$F = \frac{2 \times precision \times recall}{precision + recall}$$

- F_β : weighted measure of precision and recall
 - assigns β times as much weight to recall as to precision

$$F_\beta = \frac{(1 + \beta^2) \times precision \times recall}{\beta^2 \times precision + recall}$$

Classifier Evaluation Metrics: Example

Actual Class\Predicted class	cancer = yes	cancer = no	Total	Recognition(%)
cancer = yes	90	210	300	
cancer = no	140	9560	9700	
Total	230	9770	10000	96.50 (accuracy)

Classifier Evaluation Metrics: Example

Actual Class\Predicted class	cancer = yes	cancer = no	Total	Recognition(%)
cancer = yes	90	210	300	
cancer = no	140	9560	9700	
Total	230	9770	10000	96.50 (<i>accuracy</i>)

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F = \frac{2 \times precision \times recall}{precision + recall}$$

Classifier Evaluation Metrics: Example

Actual Class\Predicted class	cancer = yes	cancer = no	Total	Recognition(%)
cancer = yes	90	210	300	
cancer = no	140	9560	9700	
Total	230	9770	10000	96.50 (<i>accuracy</i>)

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

- Precision = $90/230 = 39.13\%$

- Recall = $90/300 = 30.00\%$

Classifier Evaluation Metrics: Example

Actual Class\Predicted class	cancer = yes	cancer = no	Total	Recognition(%)
cancer = yes	90	210	300	
cancer = no	140	9560	9700	
Total	230	9770	10000	96.50 (<i>accuracy</i>)

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

- Precision = $90/230 = 39.13\%$

- Recall = $90/300 = 30.00\%$

$$F = \frac{2 \times precision \times recall}{precision + recall}$$

$$F = (2 * 39.13 * 30) / (39.13 + 30)$$

$$F = 33.96 \%$$

Evaluating Classifier Accuracy:

1. Holdout Method

■ Holdout method

- Given data is randomly partitioned into two independent sets
 - Training set (e.g., 2/3) for model construction
 - Test set (e.g., 1/3) for accuracy estimation
- Random sampling: a variation of holdout
 - Repeat holdout k times, accuracy = avg. of the accuracies obtained

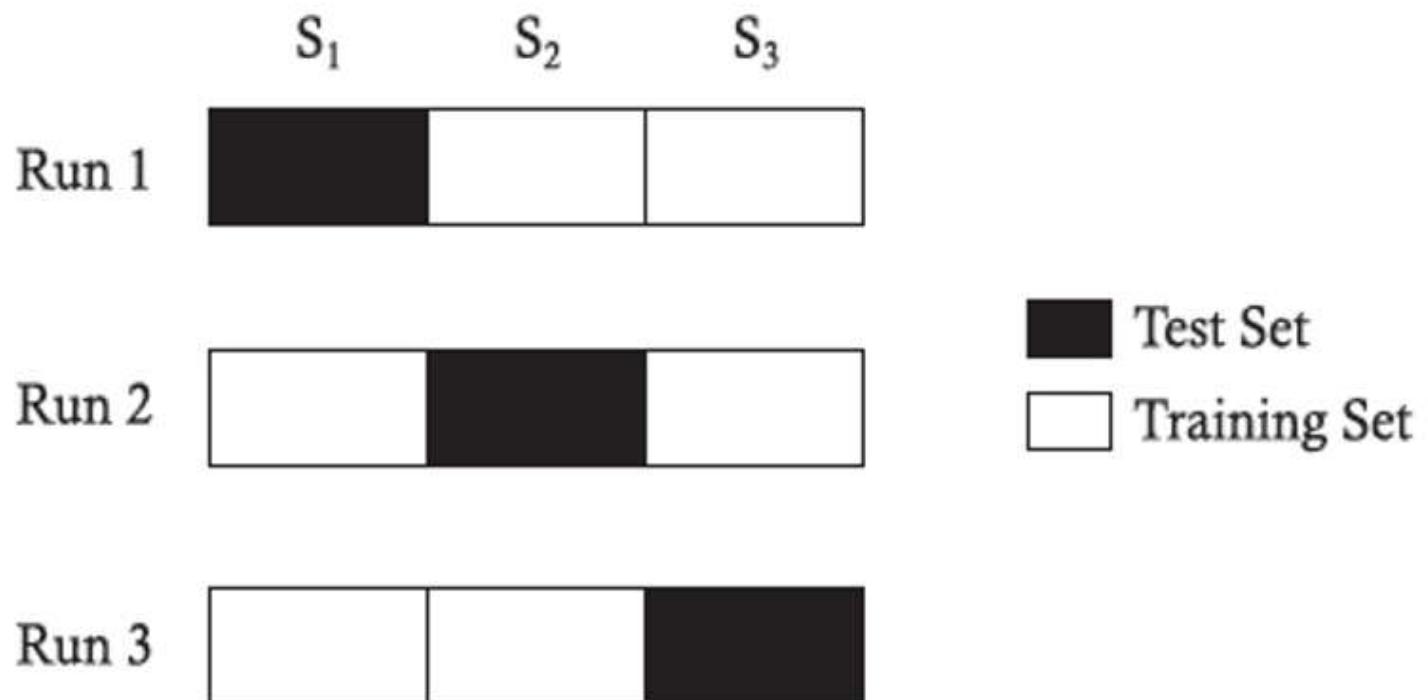
Evaluating Classifier Accuracy:

2. Cross-Validation Methods

- **Cross-validation** (k -fold, where $k = 10$ is most popular)
 - Randomly partition the data into k *mutually exclusive* subsets, each approximately equal size
 - At i -th iteration, use D_i as test set and others as training set
 - Leave-one-out: k folds where $k = \#$ of tuples, for small sized data
 - *Stratified cross-validation*: folds are stratified so that class dist. in each fold is approx. the same as that in the initial data

K-fold Cross-Validation

3-fold cross-validation



Evaluating Classifier Accuracy: 3. Bootstrap

■ Bootstrap

- Works well with small data sets
- Samples the given training tuples uniformly *with replacement*
 - i.e., each time a tuple is selected, it is equally likely to be selected again and re-added to the training set
- Repeat the sampling procedure k times, overall accuracy of the model:

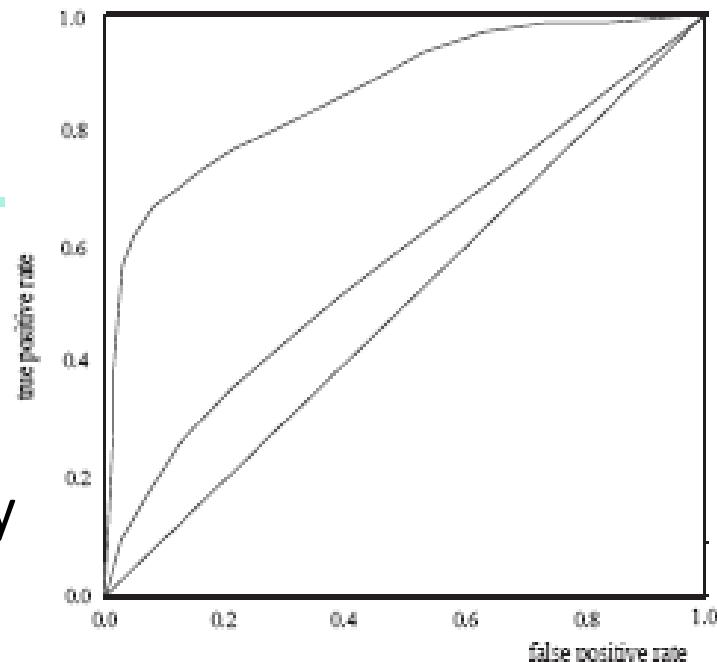
Evaluating Classifier Accuracy: 3. Bootstrap

- **Bootstrap** : Works well with small data sets
- Several bootstrap methods, and a common one is **.632 bootstrap**
 - A data set with d tuples is sampled d times, with replacement, resulting in a training set of d samples. The data tuples that did not make it into the training set end up forming the test set. About 63.2% of the original data end up in the bootstrap, and the remaining 36.8% form the test set (since $(1 - 1/d)^d \approx e^{-1} = 0.368$)
 - Repeating sampling procedure k times, overall accuracy of model:

$$Acc(M) = \frac{1}{k} \sum_{i=1}^k (0.632 \times Acc(M_i)_{test_set} + 0.368 \times Acc(M_i)_{train_set})$$

Model Selection: ROC Curves

- **ROC** (Receiver Operating Characteristics) curves: for visual comparison of classification models
- Originated from signal detection theory
- Shows the trade-off between the true positive rate and the false positive rate
- The area under the ROC curve is a measure of the accuracy of the model
- Rank the test tuples in decreasing order: the one that is most likely to belong to the positive class appears at the top of the list
- The closer to the diagonal line (i.e., the closer the area is to 0.5), the less accurate is the model



- Vertical axis represents the true positive rate
- Horizontal axis rep. the false positive rate
- The plot also shows a diagonal line
- A model with perfect accuracy will have an area of 1.0

Issues Affecting Model Selection

- **Accuracy**
 - classifier accuracy: predicting class label
- **Speed**
 - time to construct the model (training time)
 - time to use the model (classification/prediction time)
- **Robustness:** handling noise and missing values
- **Scalability:** efficiency in disk-resident databases
- **Interpretability**
 - understanding and insight provided by the model
- Other measures, e.g., goodness of rules, such as decision tree size or compactness of classification rules

Estimating Confidence Intervals: Classifier Models M_1 vs. M_2

- Suppose we have 2 classifiers, M_1 and M_2 , which one is better?
- Use 10-fold cross-validation to obtain $\overline{err}(M_1)$ and $\overline{err}(M_2)$
- These mean error rates are just *estimates* of error on the true population of *future* data cases
- What if the difference between the 2 error rates is just attributed to *chance*?
 - Use a **test of statistical significance**
 - Obtain **confidence limits** for our error estimates

Estimating Confidence Intervals: Null Hypothesis

- Perform 10-fold cross-validation
- Assume samples follow a **t distribution** with **$k-1$ degrees of freedom** (here, $k=10$)
- Use **t-test** (or **Student's t-test**)
- **Null Hypothesis:** M_1 & M_2 are the same
- If we can **reject** null hypothesis, then
 - we conclude that the difference between M_1 & M_2 is **statistically significant**
 - Choose model with lower error rate

Estimating Confidence Intervals: t-test

- If only 1 test set available: **pairwise comparison**
 - For i^{th} round of 10-fold cross-validation, the same cross partitioning is used to obtain $\text{err}(M_1)_i$ and $\text{err}(M_2)_i$
 - Average over 10 rounds to get $\overline{\text{err}}(M_1)$ and $\overline{\text{err}}(M_2)$
 - **t-test computes t-statistic with $k-1$ degrees of freedom:**
- $t = \frac{\overline{\text{err}}(M_1) - \overline{\text{err}}(M_2)}{\sqrt{\text{var}(M_1 - M_2)/k}}$ where
$$\text{var}(M_1 - M_2) = \frac{1}{k} \sum_{i=1}^k \left[\text{err}(M_1)_i - \text{err}(M_2)_i - (\overline{\text{err}}(M_1) - \overline{\text{err}}(M_2)) \right]^2$$
- If two test sets available: use **non-paired t-test**

where

$$\text{var}(M_1 - M_2) = \sqrt{\frac{\text{var}(M_1)}{k_1} + \frac{\text{var}(M_2)}{k_2}},$$

where k_1 & k_2 are # of cross-validation samples used for M_1 & M_2 , resp.

Estimating Confidence Intervals: Table for t-distribution

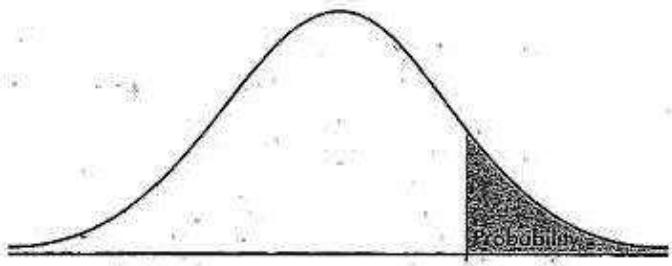


TABLE B: *t*-DISTRIBUTION CRITICAL VALUES

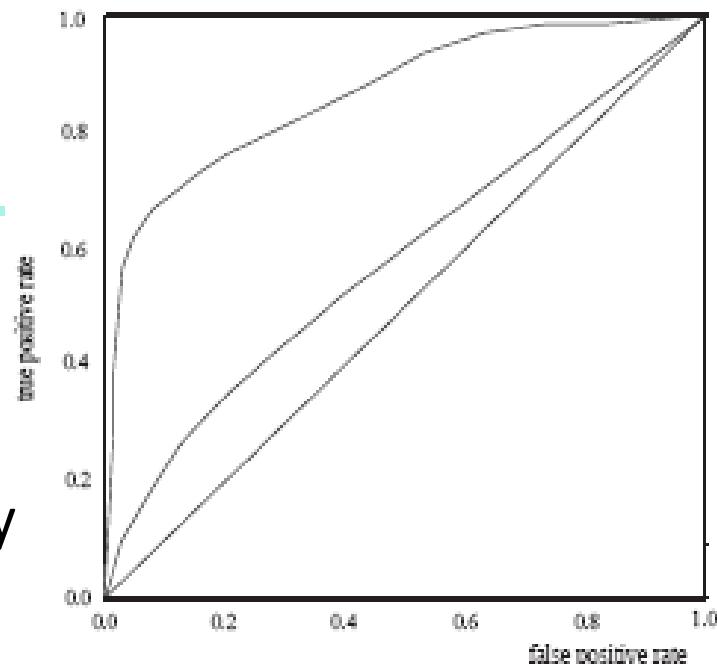
df	Tail probability <i>p</i>											
	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	.685	.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	.679	.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	.679	.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	.678	.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	.677	.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	.675	.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
∞	.674	.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level <i>C</i>											

Estimating Confidence Intervals: Statistical Significance

- Are M_1 & M_2 significantly different?
 - Compute t . Select *significance level* (e.g. $sig = 5\%$)
 - Consult table for t-distribution: Find *t value* corresponding to *k-1 degrees of freedom* (here, 9)
 - t-distribution is symmetric: typically upper % points of distribution shown → look up value for **confidence limit** $z=sig/2$ (here, 0.025)
 - If $t > z$ or $t < -z$, then t value lies in rejection region:
 - **Reject null hypothesis** that mean error rates of M_1 & M_2 are same
 - Conclude: statistically significant difference between M_1 & M_2
 - **Otherwise**, conclude that any difference is **chance**

Model Selection: ROC Curves

- **ROC** (Receiver Operating Characteristics) curves: for visual comparison of classification models
- Originated from signal detection theory
- Shows the trade-off between the true positive rate and the false positive rate
- The area under the ROC curve is a measure of the accuracy of the model
- Rank the test tuples in decreasing order: the one that is most likely to belong to the positive class appears at the top of the list
- The closer to the diagonal line (i.e., the closer the area is to 0.5), the less accurate is the model



- Vertical axis represents the true positive rate
- Horizontal axis rep. the false positive rate
- The plot also shows a diagonal line
- A model with perfect accuracy will have an area of 1.0

Issues Affecting Model Selection

- **Accuracy**
 - classifier accuracy: predicting class label
- **Speed**
 - time to construct the model (training time)
 - time to use the model (classification/prediction time)
- **Robustness:** handling noise and missing values
- **Scalability:** efficiency in disk-resident databases
- **Interpretability**
 - understanding and insight provided by the model
- Other measures, e.g., goodness of rules, such as decision tree size or compactness of classification rules

Classification of Class-Imbalanced Data Sets

- Class-imbalance problem: Rare positive example but numerous negative ones, e.g., medical diagnosis, fraud, oil-spill, fault, etc.
- Traditional methods assume a balanced distribution of classes and equal error costs: not suitable for class-imbalanced data
- Typical methods for imbalance data in 2-class classification:
 - **Oversampling**: re-sampling of data from positive class
 - **Under-sampling**: randomly eliminate tuples from negative class
 - **Threshold-moving**: moves the decision threshold, t , so that the rare class tuples are easier to classify, and hence, less chance of costly false negative errors
 - Ensemble techniques: Ensemble multiple classifiers introduced above
- Still difficult for class imbalance problem on multiclass tasks

Issues: Evaluating Classification Methods

- Accuracy
 - classifier accuracy: predicting class label
 - predictor accuracy: guessing value of predicted attributes
- Speed
 - time to construct the model (training time)
 - time to use the model (classification/prediction time)
- Robustness: handling noise and missing values
- Scalability: efficiency in disk-resident databases
- Interpretability
 - understanding and insight provided by the model
- Other measures, e.g., goodness of rules, such as decision tree size or compactness of classification rules

Predictor Error Measures

- Measure predictor accuracy: measure how far off the predicted value is from the actual known value
- **Loss function:** measures the error betw. y_i and the predicted value y'_i
 - Absolute error: $|y_i - y'_i|$
 - Squared error: $(y_i - y'_i)^2$
- Test error (generalization error): the average loss over the test set
 - Mean absolute error: $\frac{\sum_{i=1}^d |y_i - y'_i|}{d}$ Mean squared error: $\frac{\sum_{i=1}^d (y_i - y'_i)^2}{d}$
 - Relative absolute error: $\frac{\sum_{i=1}^d |y_i - y'_i|}{\sum_{i=1}^d |y_i - \bar{y}|}$ Relative squared error: $\frac{\sum_{i=1}^d (y_i - y'_i)^2}{\sum_{i=1}^d (y_i - \bar{y})^2}$

The mean squared-error exaggerates the presence of outliers

Popularly use (square) root mean-square error, similarly, root relative squared error

Cluster Validation

- For supervised classification we have a variety of measures to evaluate how good our model is
 - Accuracy, precision, recall
- For cluster analysis, the analogous question is how to evaluate the “goodness” of the resulting clusters?
- But “clusters are in the eye of the beholder”!
- Then why do we want to evaluate them?
 - To avoid finding patterns in noise
 - To compare clustering algorithms
 - To compare two sets of clusters
 - To compare two clusters

Internal Measures: Cohesion and Separation

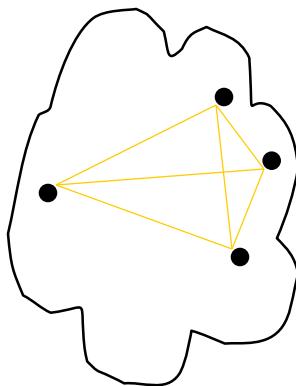
- **Cluster Cohesion:** Measures how closely related are objects in a cluster
 - Example: SSE
- **Cluster Separation:** Measure how distinct or well-separated a cluster is from other clusters
- Example: Squared Error
 - Cohesion is measured by the within cluster sum of squares (SSE)
$$SSE = WSS = \sum_i \sum_{x \in C_i} (x - m_i)^2$$
 - Separation is measured by the between cluster sum of squares

$$BSS = \sum |C_i| (m - m_i)^2$$

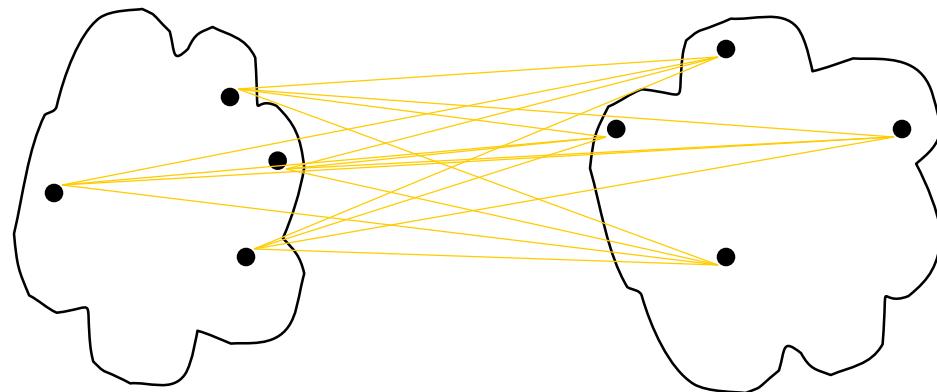
- Where $|C_i|$ is the size of cluster i

Internal Measures: Cohesion and Separation

- A proximity graph based approach can also be used for cohesion and separation.
 - Cluster cohesion is the sum of the weight of all links within a cluster.
 - Cluster separation is the sum of the weights between nodes in the cluster and nodes outside the cluster.



cohesion

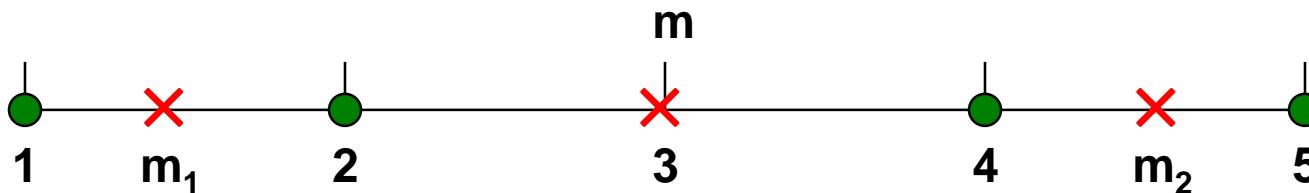


separation

Internal Measures: Cohesion and Separation

■ Example: SSE

- BSS + WSS = constant



K=1 cluster: $SSE = WSS = (1 - 3)^2 + (2 - 3)^2 + (4 - 3)^2 + (5 - 3)^2 = 10$

$$BSS = 4 \times (3 - 3)^2 = 0$$

$$Total = 10 + 0 = 10$$

K=2 clusters:

$$SSE = WSS = (1 - 1.5)^2 + (2 - 1.5)^2 + (4 - 4.5)^2 + (5 - 4.5)^2 = 1$$

$$BSS = 2 \times (3 - 1.5)^2 + 2 \times (4.5 - 3)^2 = 9$$

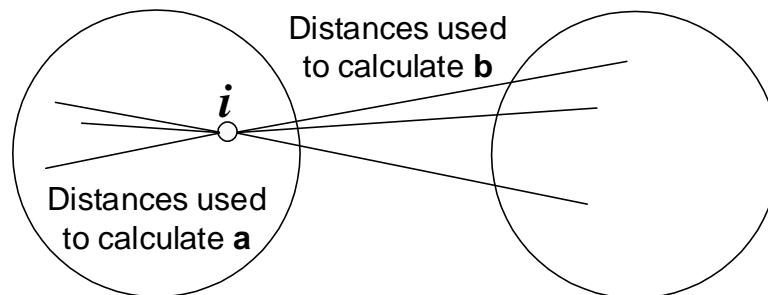
$$Total = 1 + 9 = 10$$

Internal Measures: Silhouette Coefficient

- Silhouette coefficient combines ideas of both cohesion and separation, but for individual points, as well as clusters and clusterings
- For an individual point, i
 - Calculate a = average distance of i to the points in its cluster
 - Calculate b = min (average distance of i to points in another cluster)
 - The silhouette coefficient for a point is then given by

$$s = (b - a) / \max(a,b)$$

- Typically between 0 and 1.
- The closer to 1 the better.



- Can calculate the average silhouette coefficient for a cluster or a clustering

External Measures of Cluster Validity:

Entropy and Purity

Table 5.9. K-means Clustering Results for LA Document Data Set

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Total	354	555	341	943	273	738	1.1450	0.7203

entropy For each cluster, the class distribution of the data is calculated first, i.e., for cluster j we compute p_{ij} , the ‘probability’ that a member of cluster j belongs to class i as follows: $p_{ij} = m_{ij}/m_j$, where m_j is the number of values in cluster j and m_{ij} is the number of values of class i in cluster j . Then using this class distribution, the entropy of each cluster j is calculated using the standard formula $e_j = \sum_{i=1}^L p_{ij} \log_2 p_{ij}$, where the L is the number of classes. The total entropy for a set of clusters is calculated as the sum of the entropies of each cluster weighted by the size of each cluster, i.e., $e = \sum_{i=1}^K \frac{m_i}{m} e_j$, where m_j is the size of cluster j , K is the number of clusters, and m is the total number of data points.

purity Using the terminology derived for entropy, the purity of cluster j , is given by $purity_j = \max p_{ij}$ and the overall purity of a clustering by $purity = \sum_{i=1}^K \frac{m_i}{m} purity_j$.

References (1)

- C. Apte and S. Weiss. **Data mining with decision trees and decision rules**. Future Generation Computer Systems, 13, 1997
- C. M. Bishop, **Neural Networks for Pattern Recognition**. Oxford University Press, 1995
- L. Breiman, J. Friedman, R. Olshen, and C. Stone. **Classification and Regression Trees**. Wadsworth International Group, 1984
- C. J. C. Burges. **A Tutorial on Support Vector Machines for Pattern Recognition**. *Data Mining and Knowledge Discovery*, 2(2): 121-168, 1998
- P. K. Chan and S. J. Stolfo. **Learning arbiter and combiner trees from partitioned data for scaling machine learning**. KDD'95
- H. Cheng, X. Yan, J. Han, and C.-W. Hsu, **Discriminative Frequent Pattern Analysis for Effective Classification**, ICDE'07
- H. Cheng, X. Yan, J. Han, and P. S. Yu, **Direct Discriminative Pattern Mining for Effective Classification**, ICDE'08
- W. Cohen. **Fast effective rule induction**. ICML'95
- G. Cong, K.-L. Tan, A. K. H. Tung, and X. Xu. **Mining top-k covering rule groups for gene expression data**. SIGMOD'05

References (2)

- A. J. Dobson. **An Introduction to Generalized Linear Models**. Chapman & Hall, 1990.
- G. Dong and J. Li. **Efficient mining of emerging patterns: Discovering trends and differences**. KDD'99.
- R. O. Duda, P. E. Hart, and D. G. Stork. **Pattern Classification**, 2ed. John Wiley, 2001
- U. M. Fayyad. **Branching on attribute values in decision tree generation**. AAAI'94.
- Y. Freund and R. E. Schapire. **A decision-theoretic generalization of on-line learning and an application to boosting**. J. Computer and System Sciences, 1997.
- J. Gehrke, R. Ramakrishnan, and V. Ganti. **Rainforest: A framework for fast decision tree construction of large datasets**. VLDB'98.
- J. Gehrke, V. Gant, R. Ramakrishnan, and W.-Y. Loh, **BOAT -- Optimistic Decision Tree Construction**. SIGMOD'99.
- T. Hastie, R. Tibshirani, and J. Friedman. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. Springer-Verlag, 2001.
- D. Heckerman, D. Geiger, and D. M. Chickering. **Learning Bayesian networks: The combination of knowledge and statistical data**. Machine Learning, 1995.
- W. Li, J. Han, and J. Pei, **CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules**, ICDM'01.

References (3)

- T.-S. Lim, W.-Y. Loh, and Y.-S. Shih. **A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms.** Machine Learning, 2000.
- J. Magidson. **The Chaid approach to segmentation modeling: Chi-squared automatic interaction detection.** In R. P. Bagozzi, editor, Advanced Methods of Marketing Research, Blackwell Business, 1994.
- M. Mehta, R. Agrawal, and J. Rissanen. **SLIQ : A fast scalable classifier for data mining.** EDBT'96.
- T. M. Mitchell. **Machine Learning.** McGraw Hill, 1997.
- S. K. Murthy, **Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey,** Data Mining and Knowledge Discovery 2(4): 345-389, 1998
- J. R. Quinlan. **Induction of decision trees.** *Machine Learning*, 1:81-106, 1986.
- J. R. Quinlan and R. M. Cameron-Jones. **FOIL: A midterm report.** ECML'93.
- J. R. Quinlan. **C4.5: Programs for Machine Learning.** Morgan Kaufmann, 1993.
- J. R. Quinlan. **Bagging, boosting, and c4.5.** AAAI'96.

References (4)

- R. Rastogi and K. Shim. **Public: A decision tree classifier that integrates building and pruning.** VLDB'98.
- J. Shafer, R. Agrawal, and M. Mehta. **SPRINT : A scalable parallel classifier for data mining.** VLDB'96.
- J. W. Shavlik and T. G. Dietterich. **Readings in Machine Learning.** Morgan Kaufmann, 1990.
- P. Tan, M. Steinbach, and V. Kumar. **Introduction to Data Mining.** Addison Wesley, 2005.
- S. M. Weiss and C. A. Kulikowski. **Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems.** Morgan Kaufman, 1991.
- S. M. Weiss and N. Indurkha. **Predictive Data Mining.** Morgan Kaufmann, 1997.
- I. H. Witten and E. Frank. **Data Mining: Practical Machine Learning Tools and Techniques**, 2ed. Morgan Kaufmann, 2005.
- X. Yin and J. Han. **CPAR: Classification based on predictive association rules.** SDM'03
- H. Yu, J. Yang, and J. Han. **Classifying large data sets using SVM with hierarchical clusters.** KDD'03.

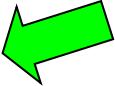
Foundation of Data Science and Analytics :

Association Mining (Frequent Pattern Analysis) - Basic Concepts

Dr. Arun K. Timalsina

Materials Adaptation :

Mining Frequent Patterns, Association and Correlations

- Basic Concepts 
- Frequent Itemset Mining Methods
- Which Patterns Are Interesting?—Pattern Evaluation Methods
- Summary

What Is Frequent Pattern Analysis?

- **Frequent pattern**: a pattern (a set of items, subsequences, substructures, etc.) that occurs frequently in a data set
- First proposed by Agrawal, Imielinski, and Swami [AIS93] in the context of **frequent itemsets** and **association rule mining**
- Motivation: Finding inherent regularities in data
 - What products were often purchased together?— Beer and diapers?!
 - What are the subsequent purchases after buying a PC?
 - What kinds of DNA are sensitive to this new drug?
 - Can we automatically classify web documents?
- Applications
 - Basket data analysis, cross-marketing, catalog design, sale campaign analysis, Web log (click stream) analysis, and DNA sequence analysis.

Why Is Freq. Pattern Mining Important?

- Freq. pattern: An intrinsic and important property of datasets
- Foundation for many essential data mining tasks
 - Association, correlation, and causality analysis
 - Sequential, structural (e.g., sub-graph) patterns
 - Pattern analysis in spatiotemporal, multimedia, time-series, and stream data
 - Classification: discriminative, frequent pattern analysis
 - Cluster analysis: frequent pattern-based clustering
 - Data warehousing: iceberg cube and cube-gradient
 - Semantic data compression: fascicles
 - Broad applications

Market Basket Example

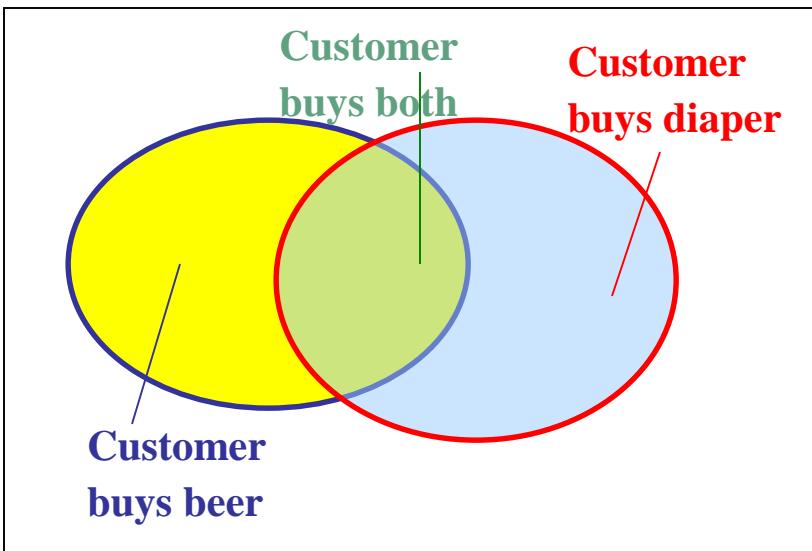


- Where should detergents be placed in the Store to maximize their sales?
- Are window cleaning products purchased when detergents and orange juice are bought together?
- Is soda typically purchased with bananas? Does the brand of soda make a difference?
- How are the demographics of the neighborhood affecting what customers are buying?

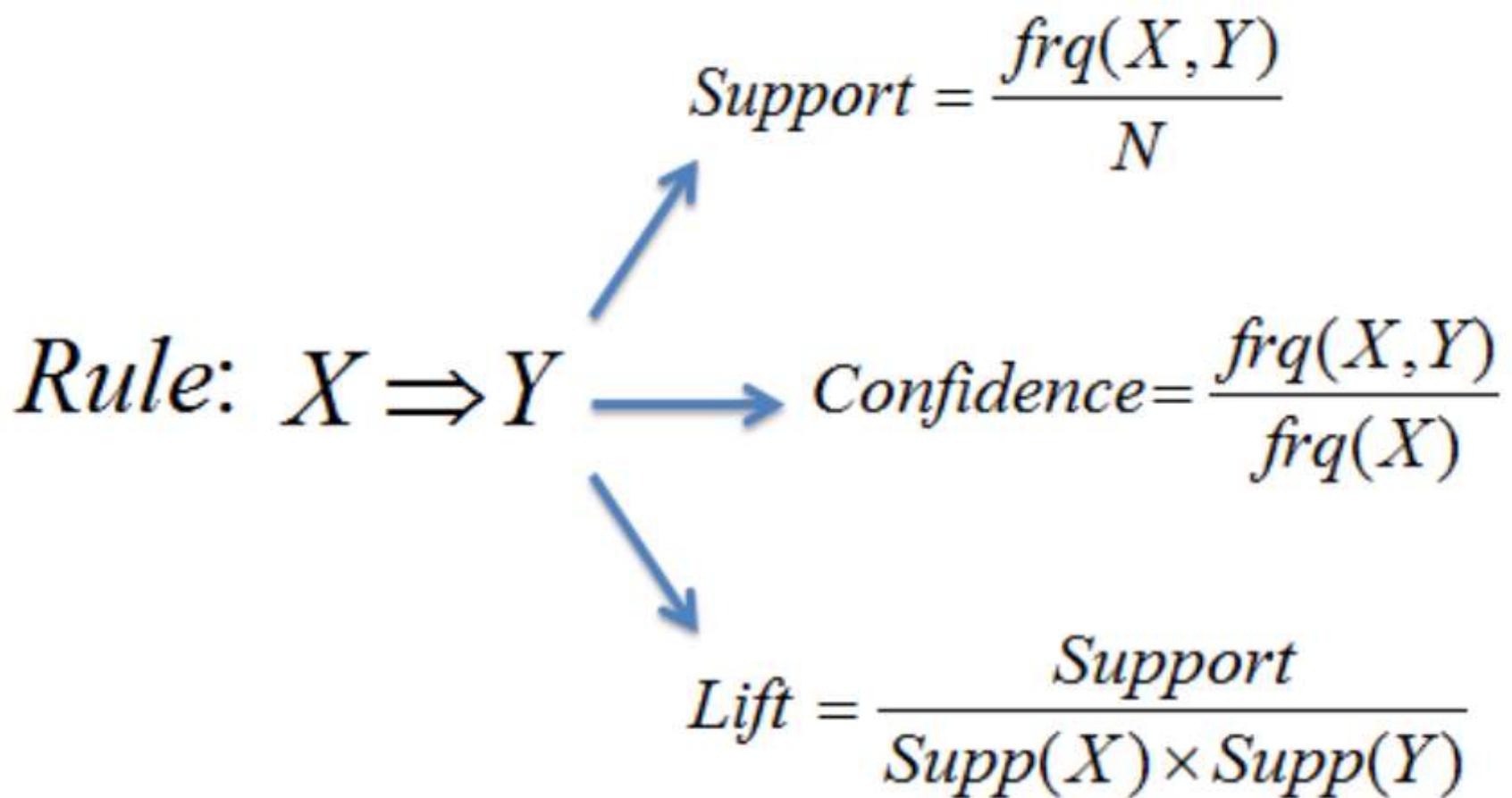
Image source: deepclimate.org

Basic Concepts: Frequent Patterns

Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk



- **itemset**: A set of one or more items
- **k-itemset** $X = \{x_1, \dots, x_k\}$
- **(absolute) support**, or, **support count** of X : Frequency or occurrence of an itemset X
- **(relative) support**, s , is the fraction of transactions that contains X (i.e., the **probability** that a transaction contains X)
- An itemset X is **frequent** if X 's support is no less than a **minsup** threshold



$$Support = \frac{frq(X, Y)}{N}$$

$$Rule: X \Rightarrow Y \longrightarrow Confidence = \frac{frq(X, Y)}{frq(X)}$$

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$



Transactions

$$Support = \frac{frq(X, Y)}{N}$$

Rule: $X \Rightarrow Y$

$$Confidence = \frac{frq(X, Y)}{frq(X)}$$

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

Transactions →



Calculated
values of
different
measures



Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$			
$A \Rightarrow C$			
$B \& C \Rightarrow D$			

Rule: $X \Rightarrow Y$

$$Support = \frac{frq(X, Y)}{N}$$

$$Confidence = \frac{frq(X, Y)}{frq(X)}$$

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

Transactions →



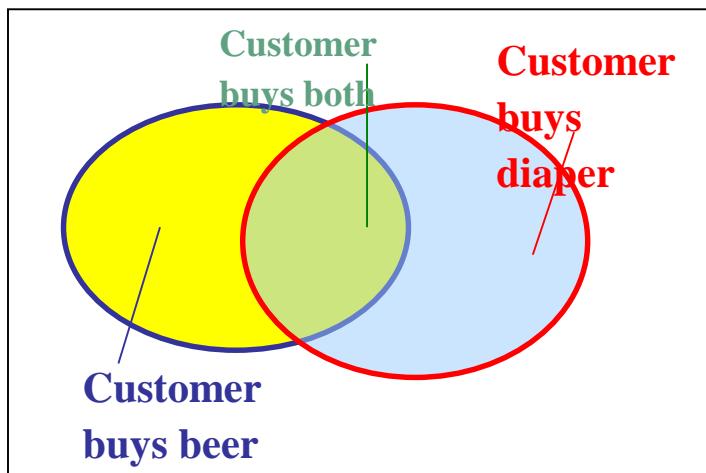
Calculated
values of
different
measures



Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \& C \Rightarrow D$	1/5	1/3	5/9

Basic Concepts: Association Rules

Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk



- Find all the rules $X \rightarrow Y$ with minimum support and confidence
 - **support**, s , probability that a transaction contains $X \cup Y$
 - **confidence**, c , conditional probability that a transaction having X also contains Y
- Let $\text{minsup} = 50\%$, $\text{minconf} = 50\%$
- Freq. Pat.: Beer:3, Nuts:3, Diaper:4, Eggs:3, {Beer, Diaper}:3
- Association rules: $X \rightarrow Y$ (s , c)
 - $\text{Beer} \rightarrow \text{Diaper}$ (60%, 100%)
 - $\text{Diaper} \rightarrow \text{Beer}$ (60%, 75%)

Closed Patterns and Max-Patterns

- A long pattern contains a combinatorial number of sub-patterns, e.g., $\{a_1, \dots, a_{100}\}$ contains $(_{100}^1) + (_{100}^2) + \dots + (_{100}^{100}) = 2^{100} - 1 = 1.27*10^{30}$ sub-patterns!
- Solution: Mine *closed patterns* and *max-patterns* instead
- An itemset X is **closed** if X is *frequent* and there exists *no super-pattern Y ⊃ X, with the same support as X* (proposed by Pasquier, et al. @ ICDT'99)
- An itemset X is a **max-pattern** if X is frequent and there exists no frequent super-pattern Y ⊃ X (proposed by Bayardo @ SIGMOD'98)
- Closed pattern is a lossless compression of freq. patterns
 - Reducing the # of patterns and rules

Closed Patterns and Max-Patterns

- Exercise. DB = { $\langle a_1, \dots, a_{100} \rangle$, $\langle a_1, \dots, a_{50} \rangle$ }
 - Min_sup = 1.
- What is the set of closed itemset?
 - $\langle a_1, \dots, a_{100} \rangle$: 1
 - $\langle a_1, \dots, a_{50} \rangle$: 2
- What is the set of max-pattern?
 - $\langle a_1, \dots, a_{100} \rangle$: 1
- What is the set of all patterns?
 - !!

Computational Complexity of Frequent Itemset Mining

- How many itemsets are potentially to be generated in the worst case?
 - The number of frequent itemsets to be generated is sensitive to the *minsup* threshold
 - When *minsup* is low, there exist potentially an exponential number of frequent itemsets
 - The worst case: If all rules exist then 2^N from powerset formula otherwise sum of combinatorial formula $C(N, R)$; approximation : N^R where N: # distinct items, and R: max length of transactions, M,N both large

Computational Complexity of Frequent Itemset Mining

- How many itemsets are potentially to be generated in the worst case?
 - The number of frequent itemsets to be generated is sensitive to the *minsup* threshold
 - When *minsup* is low, there exist potentially an exponential number of frequent itemsets
 - The worst case: M^N where M: # distinct items, and N: max length of transactions
- The worst case complexity vs. the expected probability
 - Ex. Suppose BhatBhateni has 10^4 kinds of products
 - The chance to pick up one product 10^{-4}
 - The chance to pick up a particular set of 10 products: $\sim 10^{-40}$
 - What is the chance this particular set of 10 products to be frequent 10^3 times in 10^9 transactions?

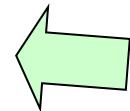
Mining Frequent Patterns, Association and Correlations: Basic Concepts and Methods

- Basic Concepts
- Frequent Itemset Mining Methods
- Which Patterns Are Interesting?—Pattern Evaluation Methods
- Summary



Scalable Frequent Itemset Mining Methods

- Apriori: A Candidate Generation-and-Test Approach
- Improving the Efficiency of Apriori
- FP-Growth: A Frequent Pattern-Growth Approach



The Downward Closure Property and Scalable Mining Methods

- The **downward closure** property of frequent patterns
 - Any subset of a frequent itemset must be frequent
 - If **{beer, diaper, nuts}** is frequent,
 - so is **{beer, diaper}**
 - : i.e., every transaction having {beer, diaper, nuts} also contains {beer, diaper}
- Scalable mining methods: Three major approaches
 - **Apriori** (Agrawal & Srikant@VLDB'94)
 - **Frequent pattern growth** (FPgrowth—Han,Pei&Yin @SIGMOD'00)
 - **Vertical data format approach** (Charm—Zaki & Hsiao @SDM'02)

Apriori: A Candidate Generation & Test Approach

- Apriori pruning principle: If there is **any** itemset which is infrequent, its superset should not be generated/tested! (Agrawal & Srikant @VLDB'94, Mannila, et al. @ KDD' 94)
- Method:
 - Initially, scan DB once to get frequent 1-itemset
 - Generate length $(k+1)$ **candidate** itemsets from length k **frequent** itemsets
 - **Test** the candidates against DB
 - Terminate when no frequent or candidate set can be generated

The Apriori Algorithm—An Example

Database TDB

Tid	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

1st scan $\rightarrow C_1$

Threshold : $Sup_{min} = 2$

Itemset	sup
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

The Apriori Algorithm—An Example

Database TDB

Tid	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

1st scan $\rightarrow C_1$

Threshold : $\text{Sup}_{\min} = 2$

Itemset	sup
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

The Apriori Algorithm—An Example

Database TDB

Tid	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

1st scan $\rightarrow C_1$

Threshold : $Sup_{min} = 2$

Itemset	sup
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

L_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{E}	3

The Apriori Algorithm—An Example

Database TDB

Tid	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

1st scan $\rightarrow C_1$

Threshold : $Sup_{min} = 2$

Itemset	sup
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

L_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{E}	3

The Apriori Algorithm—An Example

Database TDB $\text{Sup}_{\min} = 2$

Tid	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

C_1
1st scan

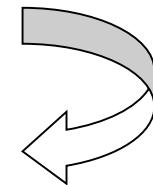
Itemset	sup
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

L_1

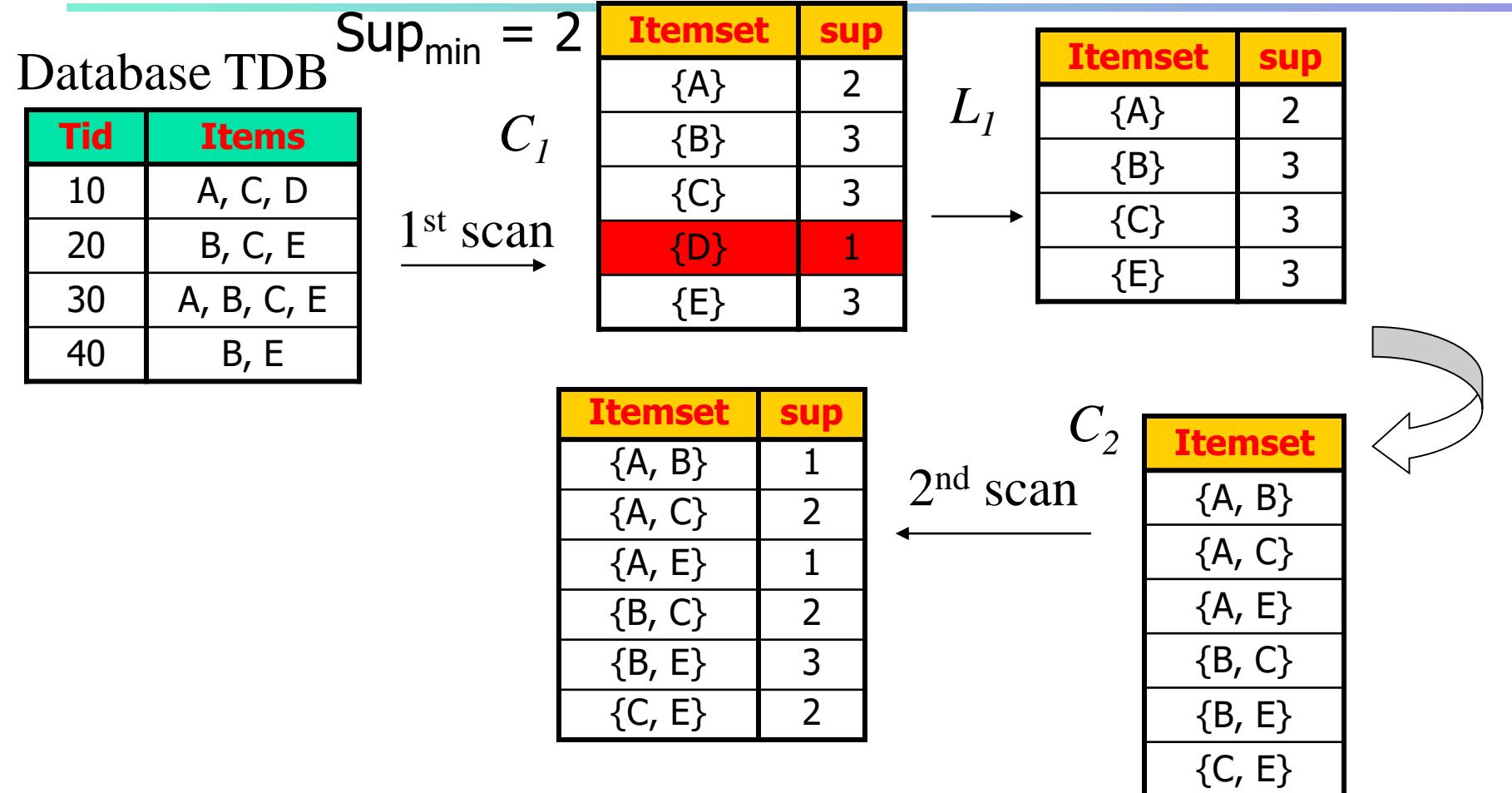
Itemset	sup
{A}	2
{B}	3
{C}	3
{E}	3

C_2

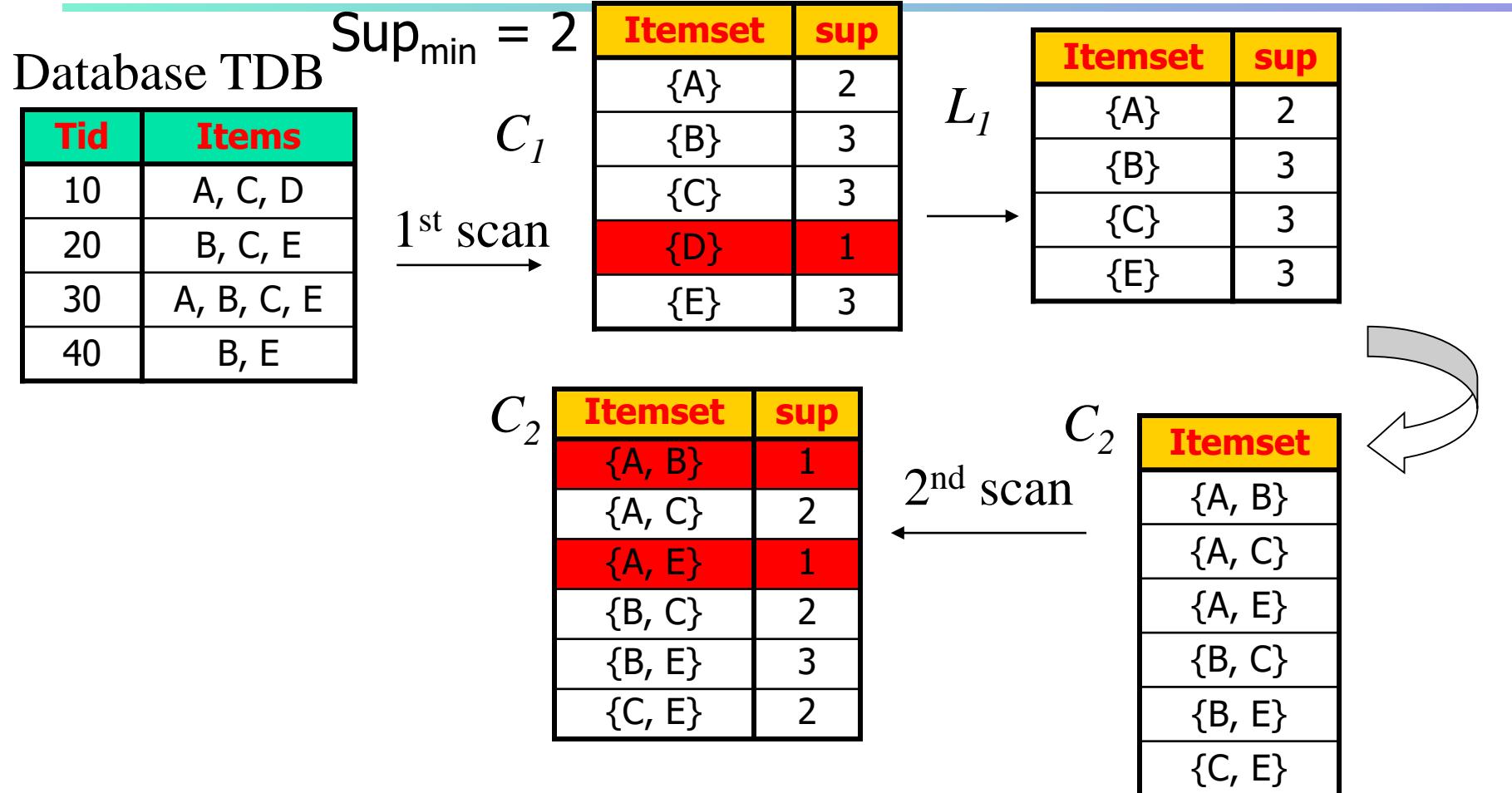
Itemset
{A, B}
{A, C}
{A, E}
{B, C}
{B, E}
{C, E}



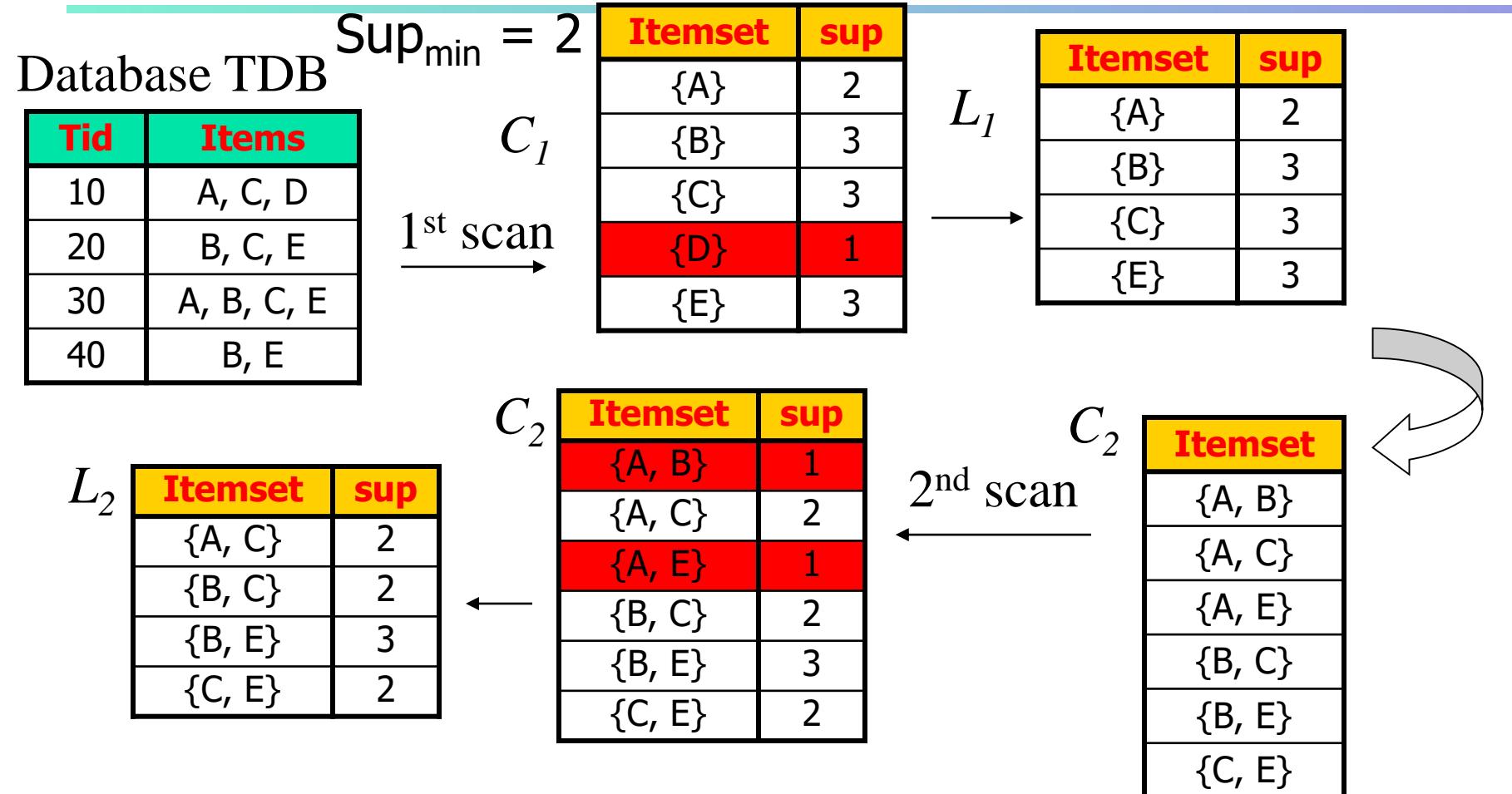
The Apriori Algorithm—An Example



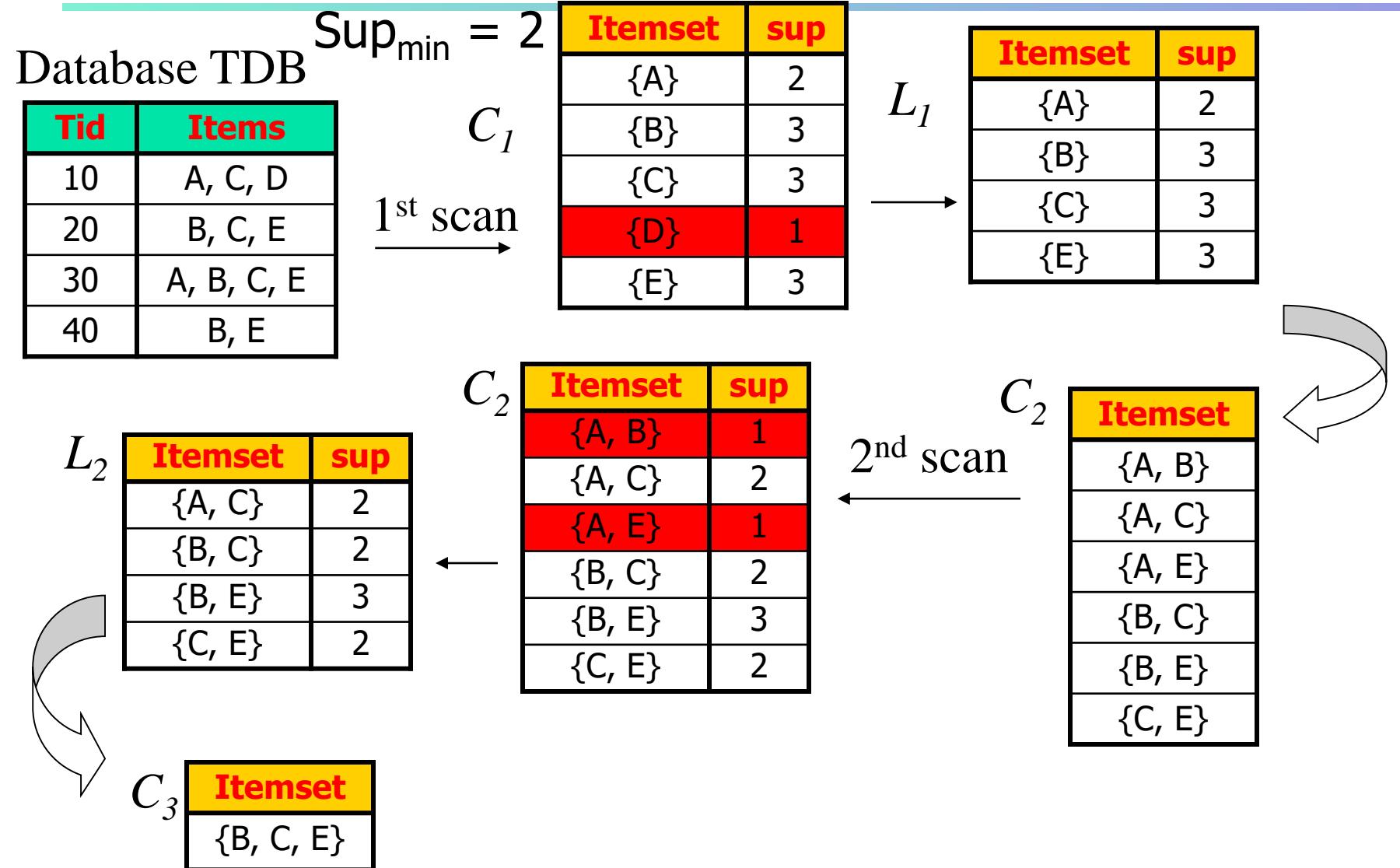
The Apriori Algorithm—An Example



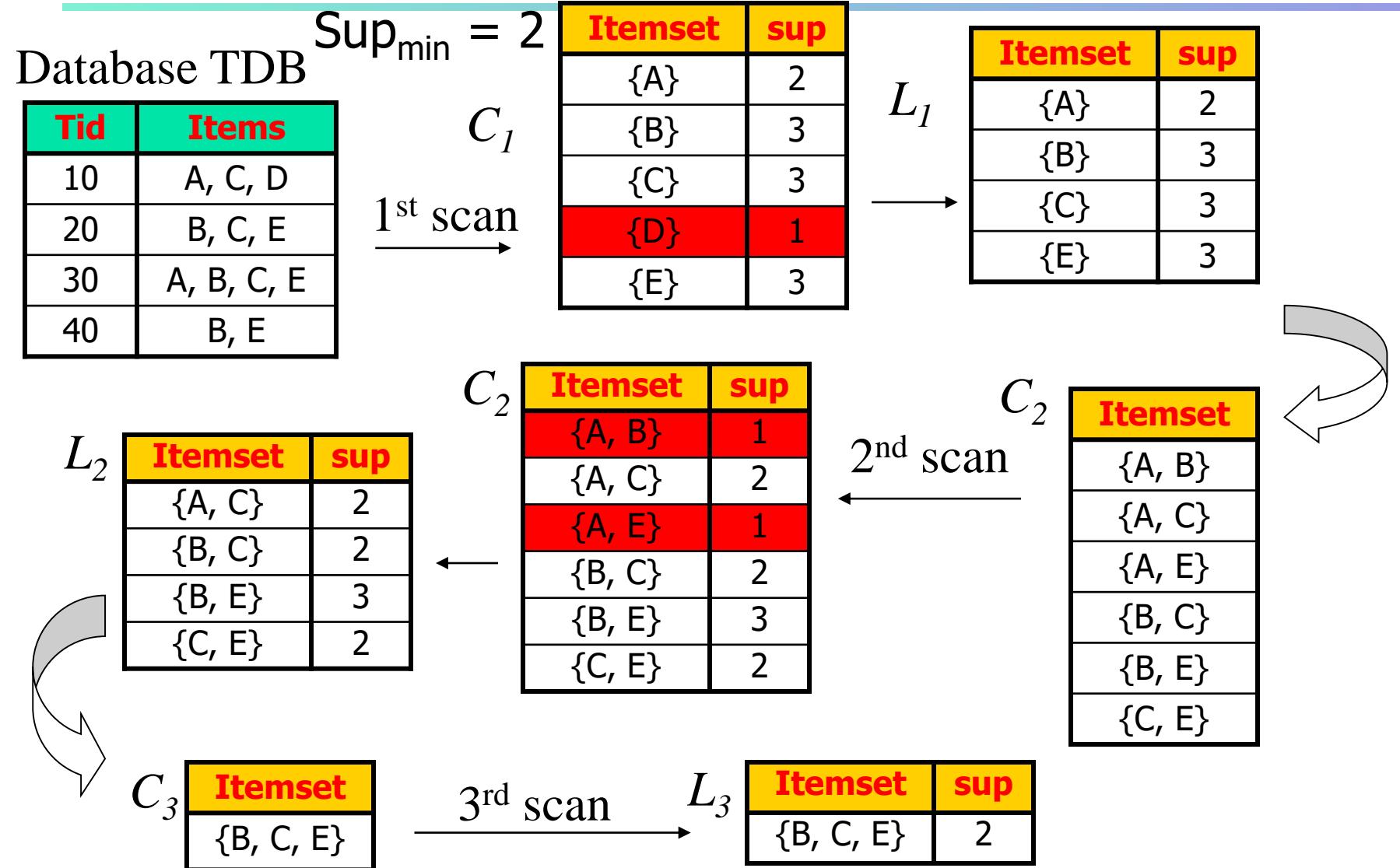
The Apriori Algorithm—An Example



The Apriori Algorithm—An Example



The Apriori Algorithm—An Example



The Apriori Algorithm (Pseudo-Code)

C_k : Candidate itemset of size k

L_k : frequent itemset of size k

$L_1 = \{\text{frequent items}\};$

for ($k = 1; L_k \neq \emptyset; k++$) **do begin**

C_{k+1} = candidates generated from L_k ;

for each transaction t in database do

 increment the count of all candidates in C_{k+1} that
 are contained in t

L_{k+1} = candidates in C_{k+1} with min_support

end

return $\cup_k L_k$

Implementation of Apriori

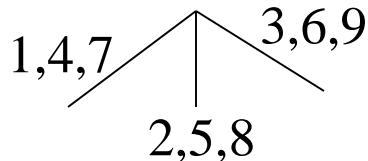
- How to generate candidates?
 - Step 1: self-joining L_k
 - Step 2: pruning
- Example of Candidate-generation
 - $L_3 = \{abc, abd, acd, ace, bcd\}$
 - Self-joining: $L_3 * L_3$
 - $abcd$ from abc and abd
 - $acde$ from acd and ace
 - Pruning:
 - $acde$ is removed because ade is not in L_3
 - $C_4 = \{abcd\}$

How to Count Supports of Candidates?

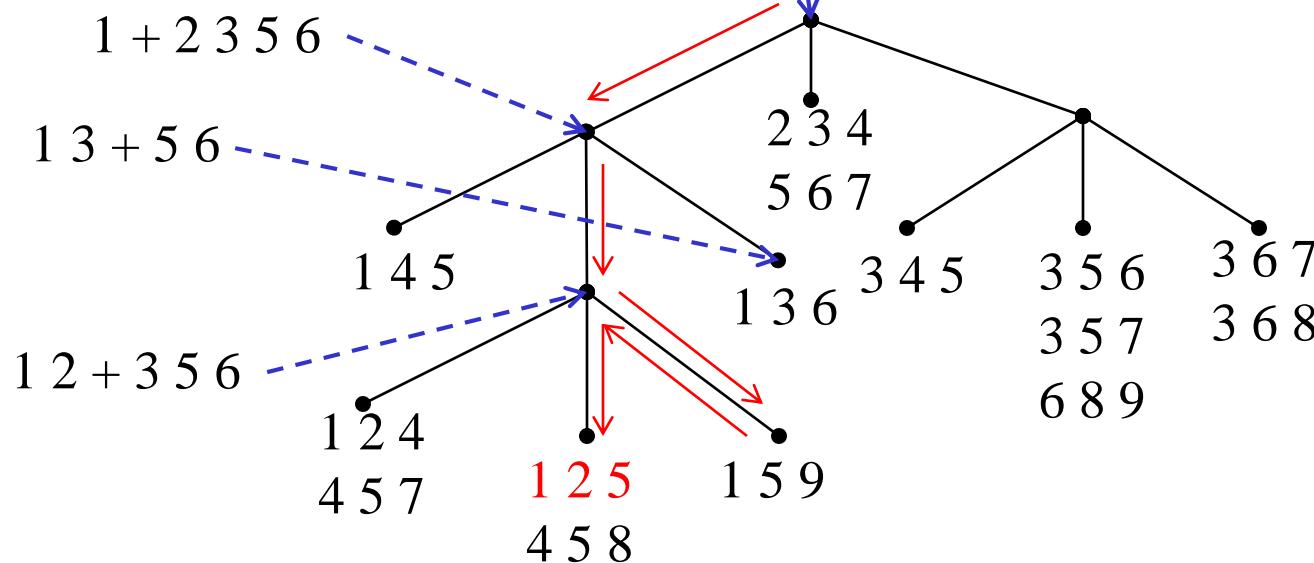
- Why counting supports of candidates a problem?
 - The total number of candidates can be very huge
 - One transaction may contain many candidates
- Method:
 - Candidate itemsets are stored in a *hash-tree*
 - *Leaf node* of hash-tree contains a list of itemsets and counts
 - *Interior node* contains a hash table
 - *Subset function*: finds all the candidates contained in a transaction

Counting Supports of Candidates Using Hash Tree

Subset function



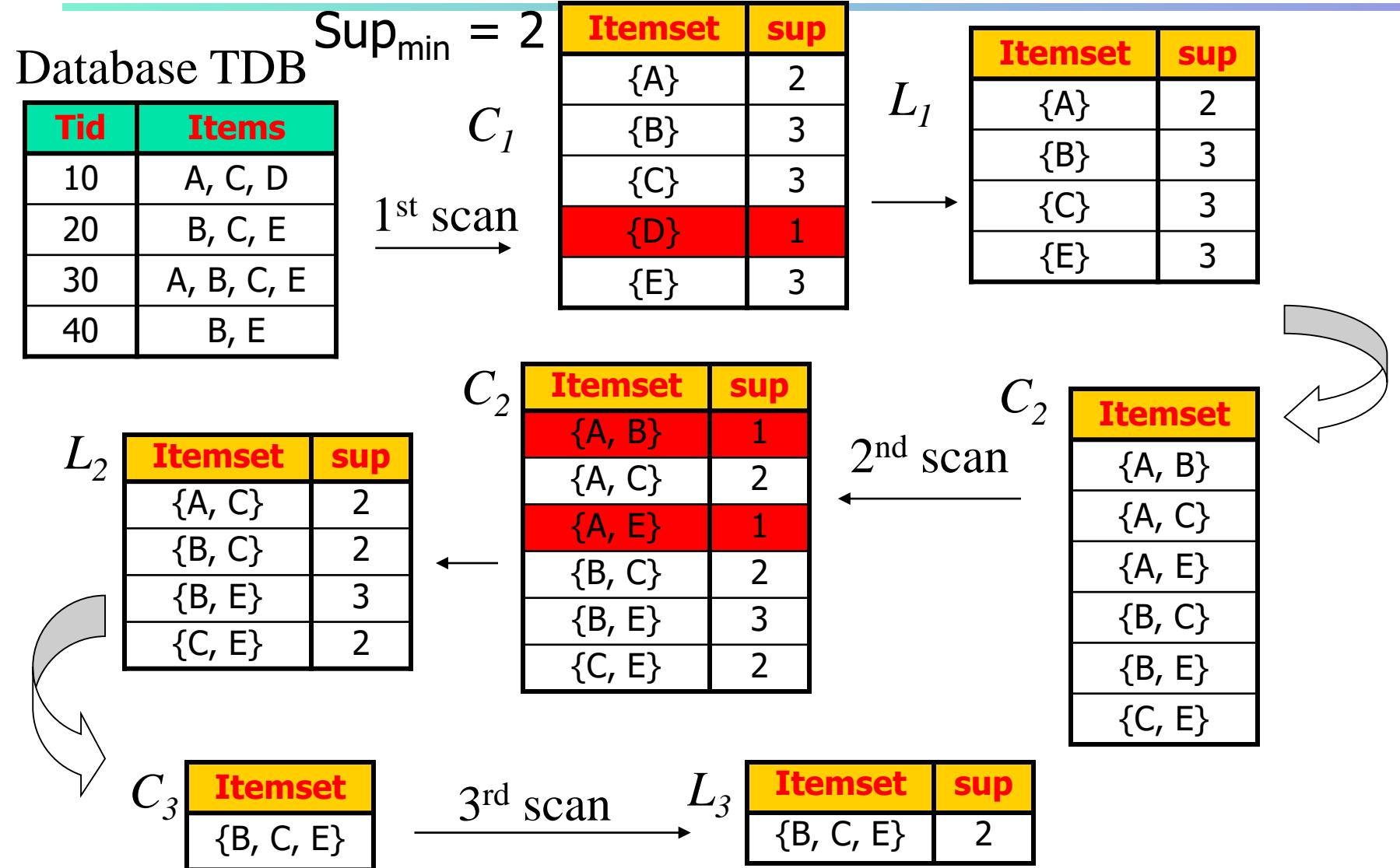
Transaction: 1 2 3 5 6



Candidate Generation: An SQL Implementation

- SQL Implementation of candidate generation
 - Suppose the items in L_{k-1} are listed in an order
 - Step 1: self-joining L_{k-1}
insert into C_k
select $p.item_1, p.item_2, \dots, p.item_{k-1}, q.item_{k-1}$
from $L_{k-1} p, L_{k-1} q$
where $p.item_1 = q.item_1, \dots, p.item_{k-2} = q.item_{k-2}, p.item_{k-1} < q.item_{k-1}$
 - Step 2: pruning
forall **itemsets** c in C_k do
 forall **($k-1$ -subsets** s of c)
 - if** (s is not in L_{k-1}) **then delete** c from C_k
- Use object-relational extensions like UDFs, BLOBs, and Table functions for efficient implementation [See: S. Sarawagi, S. Thomas, and R. Agrawal. Integrating association rule mining with relational database systems: Alternatives and implications. SIGMOD'98]

The Apriori Algorithm—An Example



2019 –Exam Question

Here is the sample transaction data records of famous grocery store, TTP (Taja Tarkari Pasal) on a particular morning of a day. Considering minimum support 40% and minimum confidence of 50%, answer the followings:

A	Aaloo
B	Bandaa
C	Chamsur
D	Dhaniya
G	Ghiraunla
K	Kaauli
M	Mula
P	Paalungo
R	RaayoSaag
T	Tamaatar

T1	K, A, T, D
T2	A,B,D
T3	C, P, M
T4	A,G,K,T
T5	R,M,A,B
T6	A,B,C,D,P
T7	K,M,A,D,T
T8	R,C,P,D
T9	B,R,D,K,T
T10	T,K,M,A

- (a) Using Apriori algorithm, identify the list of frequent items. (You need to show all the steps of calculations.)
(b) Find out all strong association rules of TTP transaction (i.e. $X \wedge Y \rightarrow Z$).

L_1

C_2

L_2

L_3

C_1

A	7	K	5
B	4	M	4
C	3	P	3
D	6	R	3
G	1	T	5

2019 –Exam Question

Here is the sample transaction data records of famous grocery store, TTP (Taja Tarkari Pasal) on a particular morning of a day. Considering minimum support 40% and minimum confidence of 50%, answer the followings:

A	Aaloo
B	Bandaa
C	Chamsur
D	Dhaniya
G	Ghiraunla
K	Kaauli
M	Mula
P	Paalungo
R	RaayoSaag
T	Tamaatar

T1	K, A, T, D
T2	A,B,D
T3	C, P, M
T4	A,G,K,T
T5	R,M,A,B
T6	A,B,C,D,P
T7	K,M,A,D,T
T8	R,C,P,D
T9	B,R,D,K,T
T10	T,K,M,A

A	7	K	5
B	4	M	4
C	3	P	3
D	6	R	3
G	1	T	5

- (a) Using Apriori algorithm, identify the list of frequent items. (You need to show all the steps of calculations.)
(b) Find out all strong association rules of TTP transaction (i.e. $X \wedge Y \rightarrow Z$).

L_1

C_2

L_2

L_3

A	7
B	4
D	6
K	5
M	4
T	5

2019 –Exam Question

Here is the sample transaction data records of famous grocery store, TTP (Taja Tarkari Pasal) on a particular morning of a day. Considering minimum support 40% and minimum confidence of 50%, answer the followings:

A	Aaloo
B	Bandaa
C	Chamsur
D	Dhaniya
G	Ghiraunla
K	Kaauli
M	Mula
P	Paalungo
R	RaayoSaag
T	Tamaatar

T1	K, A, T, D
T2	A,B,D
T3	C, P, M
T4	A,G,K,T
T5	R,M,A,B
T6	A,B,C,D,P
T7	K,M,A,D,T
T8	R,C,P,D
T9	B,R,D,K,T
T10	T,K,M,A

A	7	K	5
B	4	M	4
C	3	P	3
D	6	R	3
G	1	T	5

- (a) Using Apriori algorithm, identify the list of frequent items. (You need to show all the steps of calculations.)
- (b) Find out all strong association rules of TTP transaction (i.e. $X \wedge Y \rightarrow Z$).

L_1

A	7
B	4
D	6
K	5
M	4
T	5

C_2

L_2

AD	4
AK	4
AT	4
KT	5

L_3

2019 –Exam Question

Here is the sample transaction data records of famous grocery store, TTP (Taja Tarkari Pasal) on a particular morning of a day. Considering minimum support 40% and minimum confidence of 50%, answer the followings:

A	Aaloo
B	Bandaa
C	Chamsur
D	Dhaniya
G	Ghiraunla
K	Kaauli
M	Mula
P	Paalungo
R	RaayoSaag
T	Tamaatar

T1	K, A, T, D
T2	A,B,D
T3	C, P, M
T4	A,G,K,T
T5	R,M,A,B
T6	A,B,C,D,P
T7	K,M,A,D,T
T8	R,C,P,D
T9	B,R,D,K,T
T10	T,K,M,A

A	7	K	5
B	4	M	4
C	3	P	3
D	6	R	3
G	1	T	5

- (a) Using Apriori algorithm, identify the list of frequent items. (You need to show all the steps of calculations.)
- (b) Find out all strong association rules of TTP transaction (i.e. $X \wedge Y \rightarrow Z$).

L_1

A	7
B	4
D	6
K	5
M	4
T	5

C_2

L_2

AD	4
AK	4
AT	4
KT	5

L_3

AKT	4

2019 –Exam Question

Here is the sample transaction data records of famous grocery store, TTP (Taja Tarkari Pasal) on a particular morning of a day. Considering minimum support 40% and minimum confidence of 50%, answer the followings:

A	Aaloo
B	Bandaa
C	Chamsur
D	Dhaniya
G	Ghiraunla
K	Kaauli
M	Mula
P	Paalungo
R	RaayoSaag
T	Tamaatar

T1	K, A, T, D
T2	A,B,D
T3	C, P, M
T4	A,G,K,T
T5	R,M,A,B
T6	A,B,C,D,P
T7	K,M,A,D,T
T8	R,C,P,D
T9	B,R,D,K,T
T10	T,K,M,A

- (a) Using Apriori algorithm, identify the list of frequent items. (You need to show all the steps of calculations.)
- (b) Find out all strong association rules of TTP transaction (i.e. $X \wedge Y \rightarrow Z$).

L_3

AKT	4

$A \wedge K \rightarrow T$

$A \wedge T \rightarrow K$

$K \wedge T \rightarrow A$

2019 –Exam Question

Here is the sample transaction data records of famous grocery store, TTP (Taja Tarkari Pasal) on a particular morning of a day. Considering minimum support 40% and minimum confidence of 50%, answer the followings:

A	Aaloo
B	Bandaa
C	Chamsur
D	Dhaniya
G	Ghiraunla
K	Kaauli
M	Mula
P	Paalungo
R	RaayoSaag
T	Tamaatar

T1	K, A, T, D
T2	A,B,D
T3	C, P, M
T4	A,G,K,T
T5	R,M,A,B
T6	A,B,C,D,P
T7	K,M,A,D,T
T8	R,C,P,D
T9	B,R,D,K,T
T10	T,K,M,A

- (a) Using Apriori algorithm, identify the list of frequent items. (You need to show all the steps of calculations.)
(b) Find out all strong association rules of TTP transaction (i.e. $X \wedge Y \rightarrow Z$).

L_3

$$A \wedge K \rightarrow T : s(A \wedge K \wedge T) ; s(A \wedge K \wedge T) / s(A \wedge K)$$

AKT	4
-----	---

$$S(A \wedge K \wedge T) = 4$$

$$S(A \wedge K) = 4$$

C

2019 –Exam Question

Here is the sample transaction data records of famous grocery store, TTP (Taja Tarkari Pasal) on a particular morning of a day. Considering minimum support 40% and minimum confidence of 50%, answer the followings:

A	Aaloo
B	Bandaa
C	Chamsur
D	Dhaniya
G	Ghiraunla
K	Kaauli
M	Mula
P	Paalungo
R	RaayoSaag
T	Tamaatar

T1	K, A, T, D
T2	A,B,D
T3	C, P, M
T4	A,G,K,T
T5	R,M,A,B
T6	A,B,C,D,P
T7	K,M,A,D,T
T8	R,C,P,D
T9	B,R,D,K,T
T10	T,K,M,A

- (a) Using Apriori algorithm, identify the list of frequent items. (You need to show all the steps of calculations.)
- (b) Find out all strong association rules of TTP transaction (i.e. $X \wedge Y \rightarrow Z$).

L_3

AKT	4

$$A \wedge K \rightarrow T \quad (s, c) = (0.4, 1)$$

$$A \wedge T \rightarrow K \quad (s, c) = (0.4, 1)$$

$$K \wedge T \rightarrow A \quad (s, c) = (0.4, 0.8)$$

2019 –Exam Question

Here is the sample transaction data records of famous grocery store, TTP (Taja Tarkari Pasal) on a particular morning of a day. Considering minimum support 40% and minimum confidence of 50%, answer the followings:

A	Aaloo
B	Bandaa
C	Chamsur
D	Dhaniya
G	Ghiraunla
K	Kaauli
M	Mula
P	Paalungo
R	RaayoSaag
T	Tamaatar

T1	K, A, T, D
T2	A,B,D
T3	C, P, M
T4	A,G,K,T
T5	R,M,A,B
T6	A,B,C,D,P
T7	K,M,A,D,T
T8	R,C,P,D
T9	B,R,D,K,T
T10	T,K,M,A

- (a) Using Apriori algorithm, identify the list of frequent items. (You need to show all the steps of calculations.)
- (b) Find out all strong association rules of TTP transaction (i.e. $X \wedge Y \rightarrow Z$).

L_3

AKT	4

$$A \wedge K \rightarrow T \quad (s, c) = (0.4, 1)$$

$$A \wedge T \rightarrow K \quad (s, c) = (0.4, 1)$$

$$K \wedge T \rightarrow A \quad (s, c) = (0.4, 0.8)$$

$$A \wedge K \rightarrow T$$

$$S = p(A \wedge K \wedge T) = 4/10 = 0.4$$

$$C = p(A \wedge K \wedge T) / p(A \wedge K) = 0.4/0.4 = 1$$

2019 –Exam Question

Here is the sample transaction data records of famous grocery store, TTP (Taja Tarkari Pasal) on a particular morning of a day. Considering minimum support 40% and minimum confidence of 50%, answer the followings:

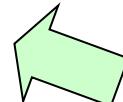
A	Aaloo
B	Bandaa
C	Chamsur
D	Dhaniya
G	Ghiraunla
K	Kaauli
M	Mula
P	Paalungo
R	RaayoSaag
T	Tamaatar

T1	K, A, T, D
T2	A,B,D
T3	C, P, M
T4	A,G,K,T
T5	R,M,A,B
T6	A,B,C,D,P
T7	K,M,A,D,T
T8	R,C,P,D
T9	B,R,D,K,T
T10	T,K,M,A

- (a) Using Apriori algorithm, identify the list of frequent items. (You need to show all the steps of calculations.)
- (b) Find out all strong association rules of TTP transaction (i.e. $X \wedge Y \rightarrow Z$).
- c) If the support threshold is decreased to 30% and confidence threshold is decreased to 40%, will there be any association rules to be added?

Scalable Frequent Itemset Mining Methods

- Apriori: A Candidate Generation-and-Test Approach
- Improving the Efficiency of Apriori
- FP-Growth: A Frequent Pattern-Growth Approach



Further Improvement of the Apriori Method

- Major computational challenges
 - Multiple scans of transaction database
 - Huge number of candidates
 - Tedium workload of support counting for candidates
- Improving Apriori: general ideas
 - Reduce passes of transaction database scans
 - Shrink number of candidates
 - Facilitate support counting of candidates

Partition: Scan Database Only Twice

- Any itemset that is potentially frequent in DB must be frequent in at least one of the partitions of DB
 - Scan 1: partition database and find local frequent patterns
 - Scan 2: consolidate global frequent patterns
- A. Savasere, E. Omiecinski and S. Navathe, *VLDB'95*

$$\boxed{\text{ }} + \boxed{\text{ }} + \dots + \boxed{\text{ }} = \text{DB}$$
$$\text{sup}_1(i) < \sigma \text{DB}_1 \quad \text{sup}_2(i) < \sigma \text{DB}_2 \quad \quad \quad \text{sup}_k(i) < \sigma \text{DB}_k \quad \text{sup}(i) < \sigma \text{DB}$$

DHP: Reduce the Number of Candidates

- A k -itemset whose corresponding hashing bucket count is below the threshold cannot be frequent

- Candidates: a, b, c, d, e
- Hash entries
 - {ab, ad, ae}
 - {bd, be, de}

count	itemsets
35	{ab, ad, ae}
88	{bd, be, de}
.	.
.	.
.	.
102	{yz, qs, wt}

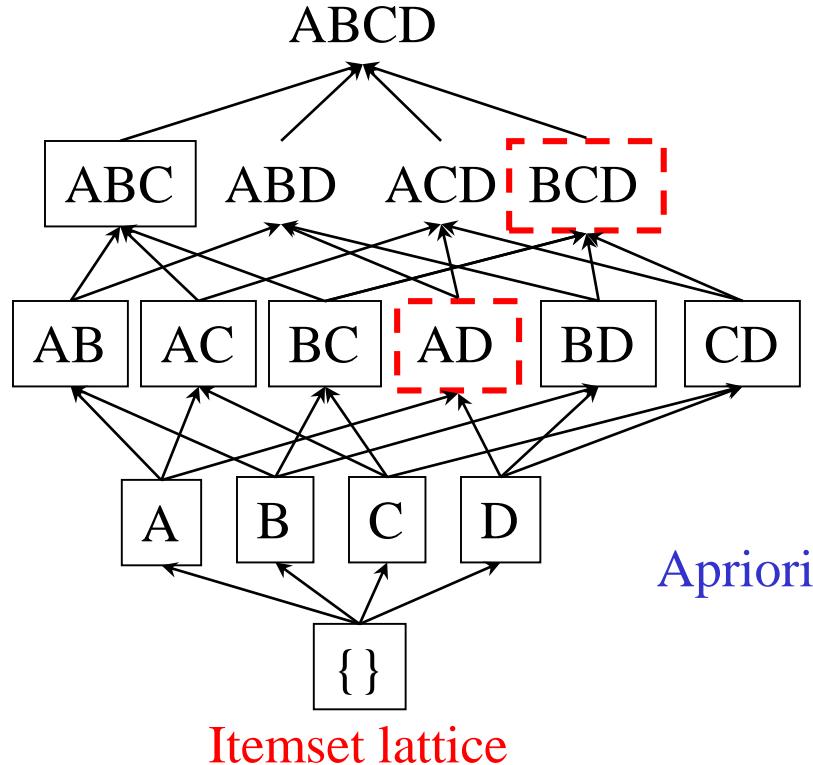
Hash Table

- Frequent 1-itemset: a, b, d, e
- ab is not a candidate 2-itemset if the sum of count of {ab, ad, ae} is below support threshold
- J. Park, M. Chen, and P. Yu. An effective hash-based algorithm for mining association rules. *SIGMOD'95*

Sampling for Frequent Patterns

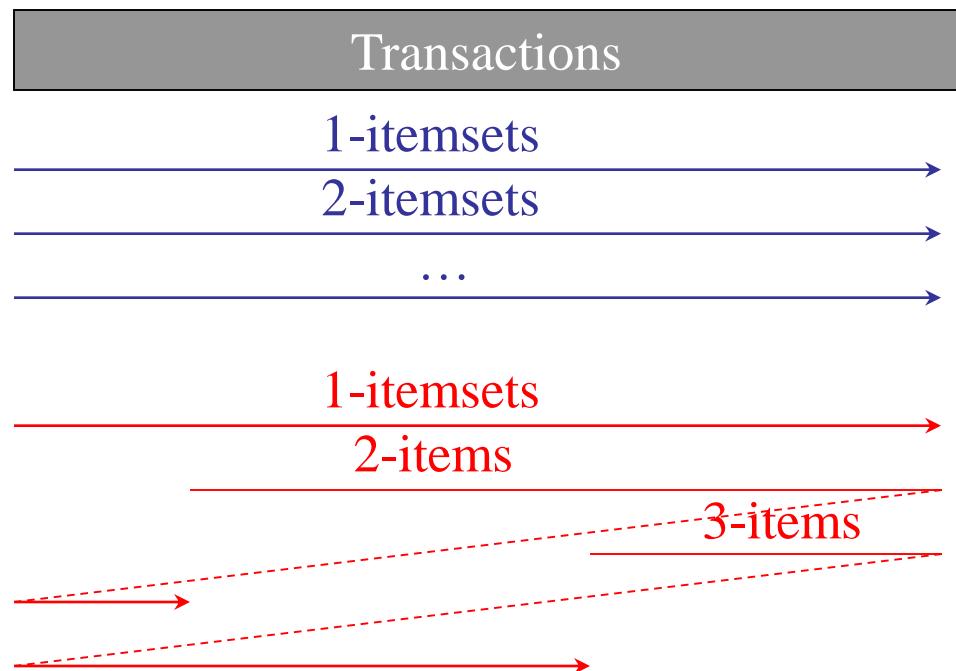
- Select a sample of original database, mine frequent patterns within sample using Apriori
- Scan database once to verify frequent itemsets found in sample, only *borders* of closure of frequent patterns are checked : Ex.: check *abcd* instead of *ab, ac, ..., etc.*
- Scan database again to find missed frequent patterns
- H. Toivonen. Sampling large databases for association rules. In *VLDB'96*

DIC: Reduce Number of Scans



S. Brin R. Motwani, J. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. *SIGMOD'97*

- Once both A & D are determined frequent, counting of AD begins
- Once all length-2 subsets of BCD are determined frequent, the counting of BCD begins



Pattern-Growth Approach: Mining Frequent Patterns Without Candidate Generation

- Bottlenecks of the Apriori approach
 - Breadth-first (i.e., level-wise) search
 - Candidate generation and test
 - Often generates a huge number of candidates
- The FP-Growth Approach (J. Han, J. Pei, and Y. Yin, SIGMOD' 00)
 - Depth-first search
 - Avoid explicit candidate generation
- Major philosophy: Grow long patterns from short ones using local frequent items only
 - “abc” is a frequent pattern
 - Get all transactions having “abc”, i.e., project DB on abc: DB|abc
 - “d” is a local frequent item in DB|abc → abcd is a frequent pattern

Construct FP-tree from a Transaction Database

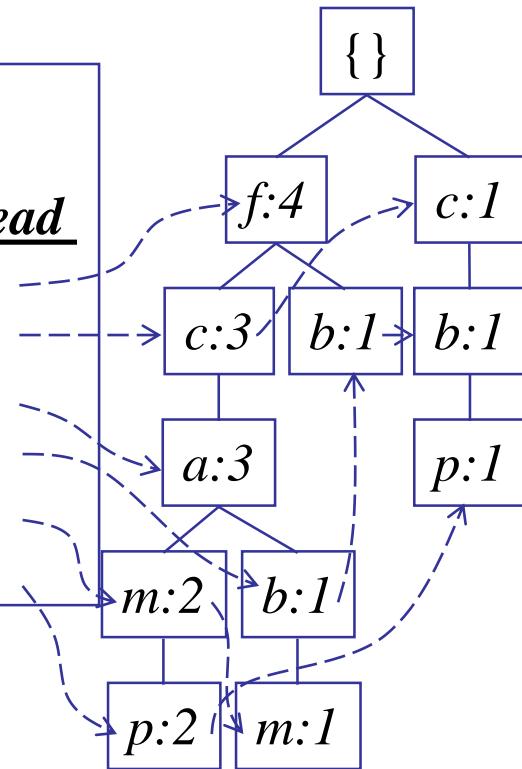
<i>TID</i>	<i>Items bought</i>	<i>(ordered) frequent items</i>
100	{f, a, c, d, g, i, m, p}	{f, c, a, m, p}
200	{a, b, c, f, l, m, o}	{f, c, a, b, m}
300	{b, f, h, j, o, w}	{f, b}
400	{b, c, k, s, p}	{c, b, p}
500	{a, f, c, e, l, p, m, n}	{f, c, a, m, p}

min_support = 3

1. Scan DB once, find frequent 1-itemset (single item pattern)
2. Sort frequent items in frequency descending order, f-list
3. Scan DB again, construct FP-tree

Header Table

<i>Item frequency head</i>	
f	4
c	4
a	3
b	3
m	3
p	3



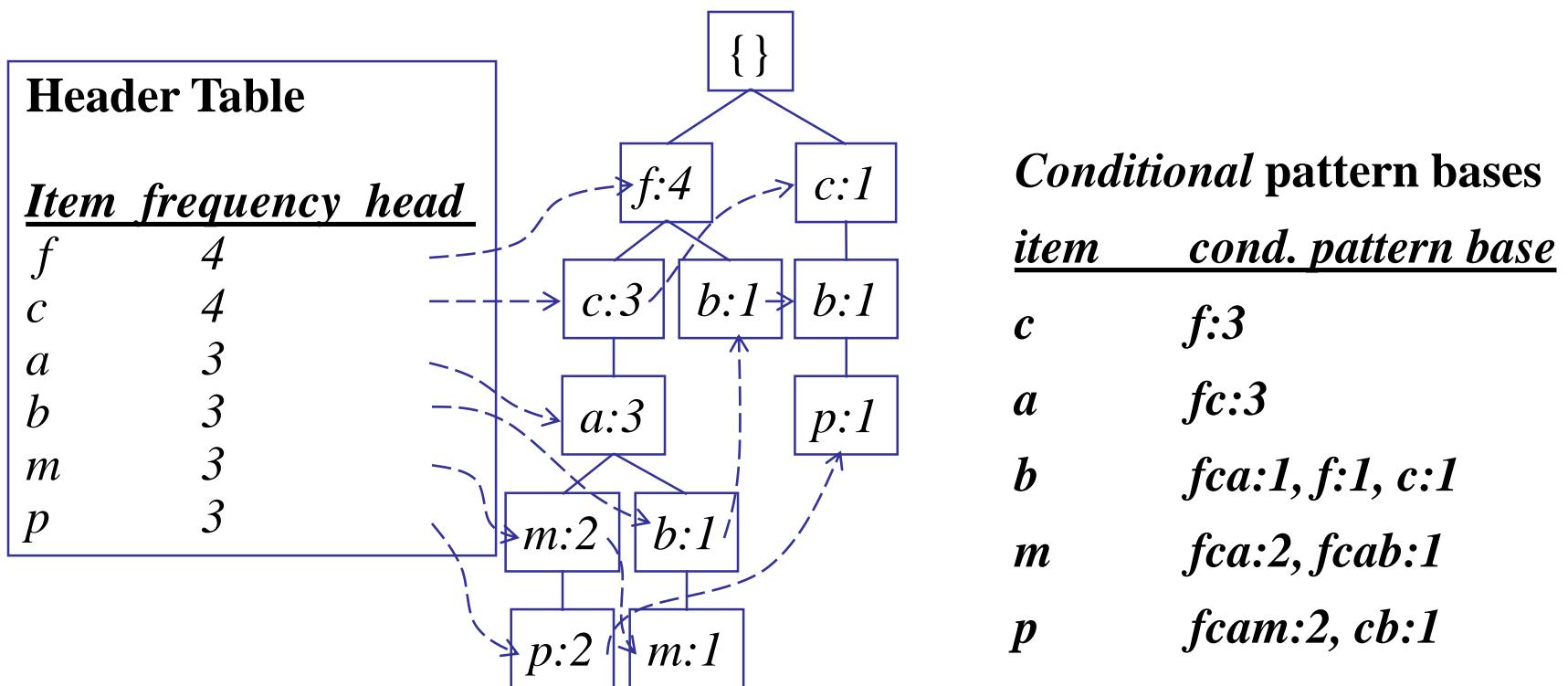
$$\text{F-list} = \text{f-c-a-b-m-p}$$

Partition Patterns and Databases

- Frequent patterns can be partitioned into subsets according to f-list
 - F-list = f-c-a-b-m-p
 - Patterns containing p
 - Patterns having m but no p
 - ...
 - Patterns having c but no a nor b, m, p
 - Pattern f
- Completeness and non-redundancy

Find Patterns Having P From P-conditional Database

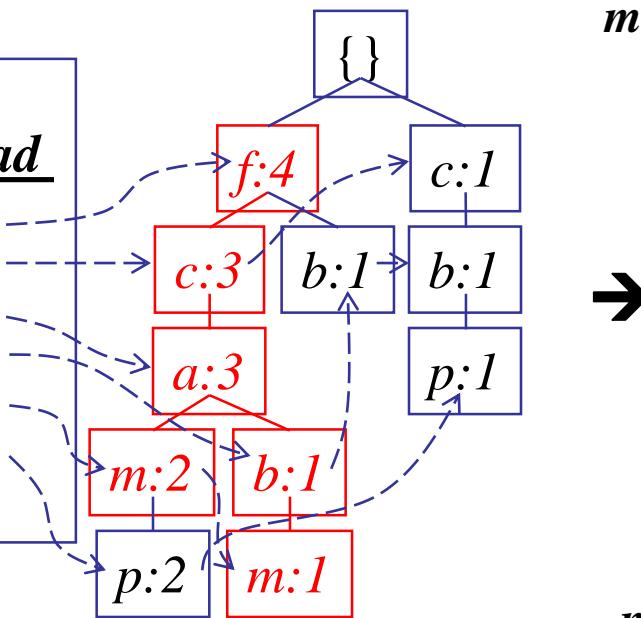
- Starting at the frequent item header table in the FP-tree
- Traverse the FP-tree by following the link of each frequent item p
- Accumulate all of *transformed prefix paths* of item p to form p 's conditional pattern base



From Conditional Pattern-bases to Conditional FP-trees

- For each pattern-base
 - Accumulate the count for each item in the base
 - Construct the FP-tree for frequent items of the pattern base

Header Table	
	<i>Item frequency head</i>
f	4
c	4
a	3
b	3
m	3
p	3



m-conditional pattern base:
 $fca:2, fcab:1$

All frequent patterns relate to *m*

→

- {}
- |
- $f:3 \rightarrow fm, cm, am,$
- |
- $c:3 \rightarrow fcm, fam, cam,$
- |
- $a:3 \rightarrow fcam,$

m-conditional FP-tree

Benefits of the FP-tree Structure

- Completeness
 - Preserve complete information for frequent pattern mining
 - Never break a long pattern of any transaction
- Compactness
 - Reduce irrelevant info—infrequent items are gone
 - Items in frequency descending order: the more frequently occurring, the more likely to be shared
 - Never be larger than the original database (not count node-links and the *count* field)

The Frequent Pattern Growth Mining Method

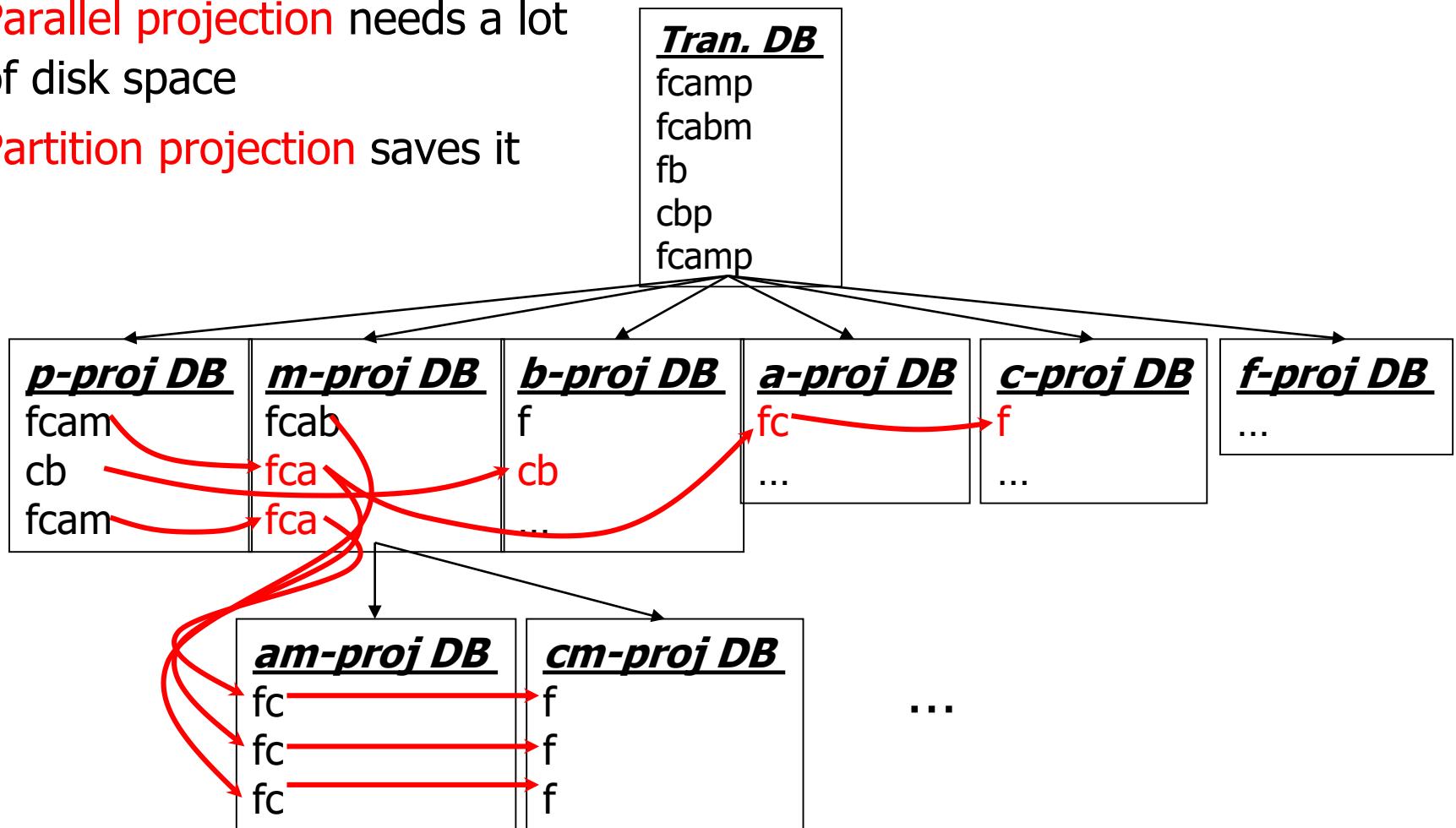
- Idea: Frequent pattern growth
 - Recursively grow frequent patterns by pattern and database partition
- Method
 - For each frequent item, construct its conditional pattern-base, and then its conditional FP-tree
 - Repeat the process on each newly created conditional FP-tree
 - Until the resulting FP-tree is empty, or it contains only one path—single path will generate all the combinations of its sub-paths, each of which is a frequent pattern

Scaling FP-growth by Database Projection

- What about if FP-tree cannot fit in memory?
 - DB projection
- First partition a database into a set of projected DBs
- Then construct and mine FP-tree for each projected DB
- Parallel projection vs. partition projection techniques
 - Parallel projection
 - Project the DB in parallel for each frequent item
 - Parallel projection is space costly
 - All the partitions can be processed in parallel
 - Partition projection
 - Partition the DB based on the ordered frequent items
 - Passing unprocessed parts to subsequent partitions

Partition-Based Projection

- Parallel projection needs a lot of disk space
- Partition projection saves it



Advantages of the Pattern Growth Approach

- Divide-and-conquer:
 - Decompose both the mining task and DB according to the frequent patterns obtained so far
 - Lead to focused search of smaller databases
- Other factors
 - No candidate generation, no candidate test
 - Compressed database: FP-tree structure
 - No repeated scan of entire database
 - Basic ops: counting local freq items and building sub FP-tree, no pattern search and matching
- A good open-source implementation and refinement of FP-Growth
 - FP-Growth+ (Grahne and J. Zhu, FIMI'03)

Further Improvements of Mining Methods

- AFOPT (Liu, et al. @ KDD'03)
 - A “push-right” method for mining condensed frequent pattern (CFP) tree
- Carpenter (Pan, et al. @ KDD'03)
 - Mine data sets with small rows but numerous columns
 - Construct a row-enumeration tree for efficient mining
- FPgrowth+ (Grahne and Zhu, FIMI'03)
 - Efficiently Using Prefix-Trees in Mining Frequent Itemsets, Proc. ICDM'03 Int. Workshop on Frequent Itemset Mining Implementations (FIMI'03), Melbourne, FL, Nov. 2003
- TD-Close (Liu, et al, SDM'06)

Extension of Pattern Growth Mining Methodology

- Mining closed frequent itemsets and max-patterns
 - CLOSET (DMKD'00), FPclose, and FPMax (Grahne & Zhu, Fimi'03)
- Mining sequential patterns
 - PrefixSpan (ICDE'01), CloSpan (SDM'03), BIDE (ICDE'04)
- Mining graph patterns
 - gSpan (ICDM'02), CloseGraph (KDD'03)
- Constraint-based mining of frequent patterns
 - Convertible constraints (ICDE'01), gPrune (PAKDD'03)
- Computing iceberg data cubes with complex measures
 - H-tree, H-cubing, and Star-cubing (SIGMOD'01, VLDB'03)
- Pattern-growth-based Clustering
 - MaPle (Pei, et al., ICDM'03)
- Pattern-Growth-Based Classification
 - Mining frequent and discriminative patterns (Cheng, et al, ICDE'07)

Foundation of Data Science and Analytics

Association Analysis: Advance Concepts

Material Adaptation:

Introduction to Data Mining, By Tan, Steinbach, Karpatne, Kumar

Continuous and Categorical Attributes

How to apply association analysis to non-asymmetric binary variables?

Gender	...	Age	Annual Income	No of hours spent online per week	No of email accounts	Privacy Concern
Female	...	26	90K	20	4	Yes
Male	...	51	135K	10	2	No
Male	...	29	80K	10	3	Yes
Female	...	45	120K	15	3	Yes
Female	...	31	95K	20	5	Yes
Male	...	25	55K	25	5	Yes
Male	...	37	100K	10	1	No
Male	...	41	65K	8	2	No
Female	...	26	85K	12	1	No
...

Example of Association Rule:

$$\{\text{Gender}=\text{Male}, \text{Age} \in [21,30]\} \rightarrow \{\text{No of hours online} \geq 10\}$$

Handling Categorical Attributes

- Example: Internet Usage Data

Gender	Level of Education	State	Computer at Home	Online Auction	Chat Online	Online Banking	Privacy Concerns
Female	Graduate	Illinois	Yes	Yes	Daily	Yes	Yes
Male	College	California	No	No	Never	No	No
Male	Graduate	Michigan	Yes	Yes	Monthly	Yes	Yes
Female	College	Virginia	No	Yes	Never	Yes	Yes
Female	Graduate	California	Yes	No	Never	No	Yes
Male	College	Minnesota	Yes	Yes	Weekly	Yes	Yes
Male	College	Alaska	Yes	Yes	Daily	Yes	No
Male	High School	Oregon	Yes	No	Never	No	No
Female	Graduate	Texas	No	No	Monthly	No	No
...

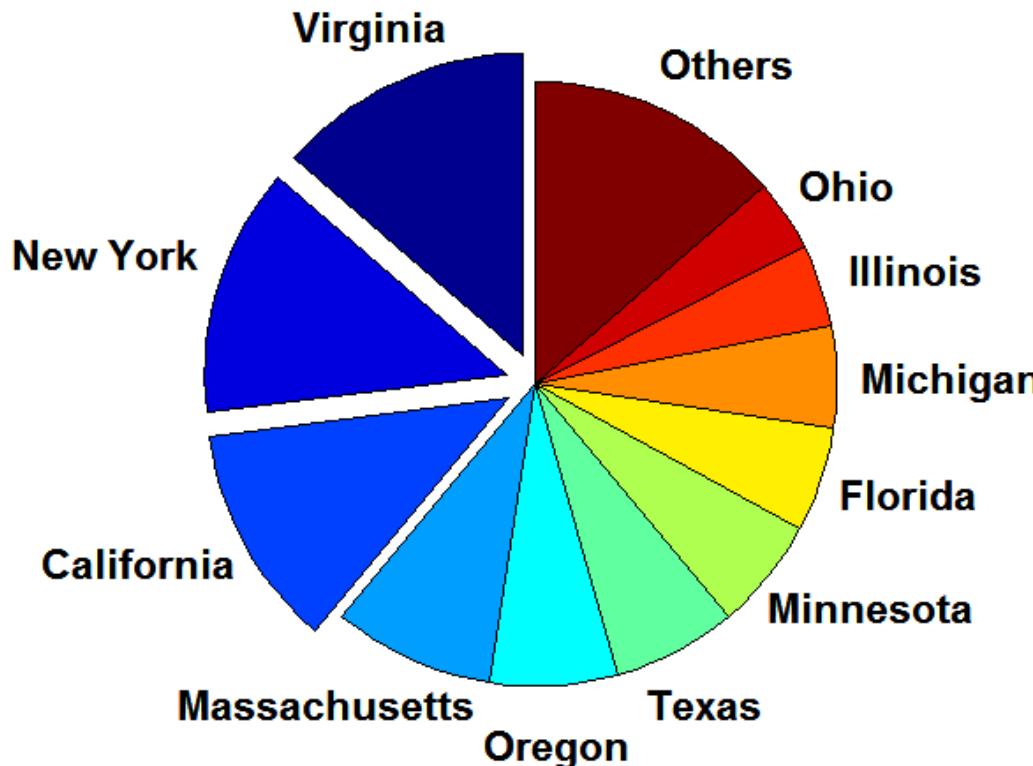
{Level of Education=Graduate, Online Banking=Yes}
→ {Privacy Concerns = Yes}

Handling Categorical Attributes

- Introduce a new “item” for each distinct attribute-value pair

Handling Categorical Attributes

- Some attributes can have many possible values
 - Many of their attribute values have very low support
 - ◆ Potential solution: Aggregate the low-support attribute values



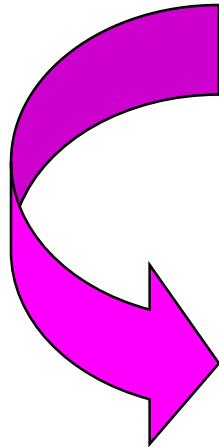
Handling Categorical Attributes

- Distribution of attribute values can be highly skewed
 - Ex.: 85% of survey participants own a computer at home
 - ◆ Most records have Computer at home = Yes
 - ◆ Computation becomes expensive; many frequent itemsets involving the binary item (Computer at home = Yes)
 - ◆ Potential solution:
 - discard the highly frequent items
 - Use alternative measures such as h-confidence
- Computational Complexity
 - Binarizing the data increases the number of items
 - But the width of the “transactions” remain the same as the number of original (non-binarized) attributes
 - Produce more frequent itemsets but maximum size of frequent itemset is limited to the number of original attributes

Handling Continuous Attributes

- Different methods:
 - Discretization-based
 - (Other methods; NOT Discussed)
- Different kinds of rules can be produced:
 - $\{\text{Age} \in [21,30], \text{No of hours online} \in [10,20]\}$
→ {Chat Online = Yes}
 - $\{\text{Age} \in [15,30], \text{Covid-Positive} = \text{Yes}\}$
→ Full_recovery

Discretization-based Methods

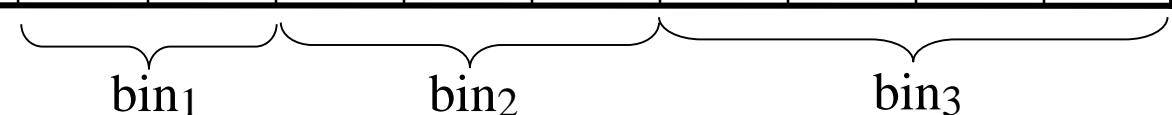


Gender	...	Age	Annual Income	No of hours spent online per week	No of email accounts	Privacy Concern
Female	...	26	90K	20	4	Yes
Male	...	51	135K	10	2	No
Male	...	29	80K	10	3	Yes
Female	...	45	120K	15	3	Yes
Female	...	31	95K	20	5	Yes
Male	...	25	55K	25	5	Yes
Male	...	37	100K	10	1	No
Male	...	41	65K	8	2	No
Female	...	26	85K	12	1	No
...

Discretization-based Methods

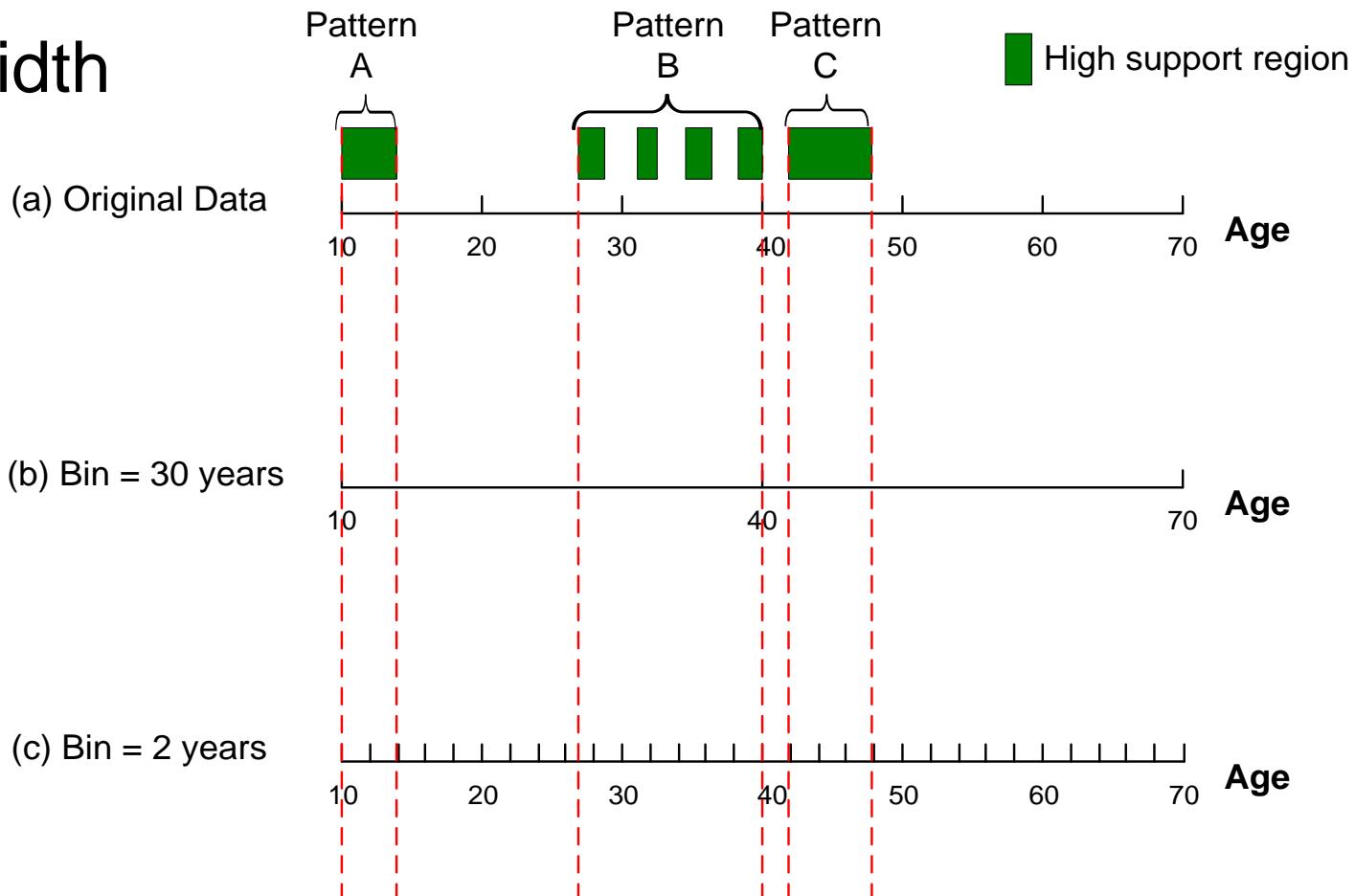
- Unsupervised:
 - Equal-width binning <1 2 3> <4 5 6> <7 8 9>
 - Equal-depth binning <1 2> <3 4 5 6 7> <8 9>
 - Cluster-based
- Supervised discretization

	Continuous attribute, v								
	1	2	3	4	5	6	7	8	9
Chat Online = Yes	0	0	20	10	20	0	0	0	0
Chat Online = No	150	100	0	0	0	100	100	150	100


bin1 bin2 bin3

Discretization Issues

- Interval width



Pattern A: Age $\in [10, 15)$ \rightarrow Chat Online = Never

Pattern B: Age $\in [26, 41)$ \rightarrow Chat Online = Never

Pattern C: Age $\in [42, 48)$ \rightarrow Online Banking = Yes

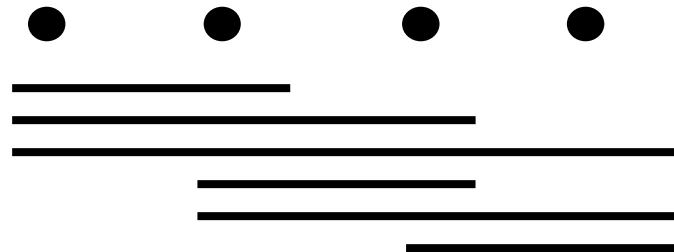
Discretization Issues

- Interval too wide (e.g., Bin size= 30)
 - May merge several disparate patterns
 - ◆ Patterns A and B are merged together
 - May lose some of the interesting patterns
 - ◆ Pattern C may not have enough confidence
- Interval too narrow (e.g., Bin size = 2)
 - Pattern A is broken up into two smaller patterns
 - ◆ Can recover the pattern by merging adjacent subpatterns
 - Pattern B is broken up into smaller patterns
 - ◆ Cannot recover the pattern by merging adjacent subpatterns
 - Some windows may not meet support threshold

Discretization: all possible intervals

Number of intervals = k

Total number of Adjacent intervals = $k(k-1)/2$



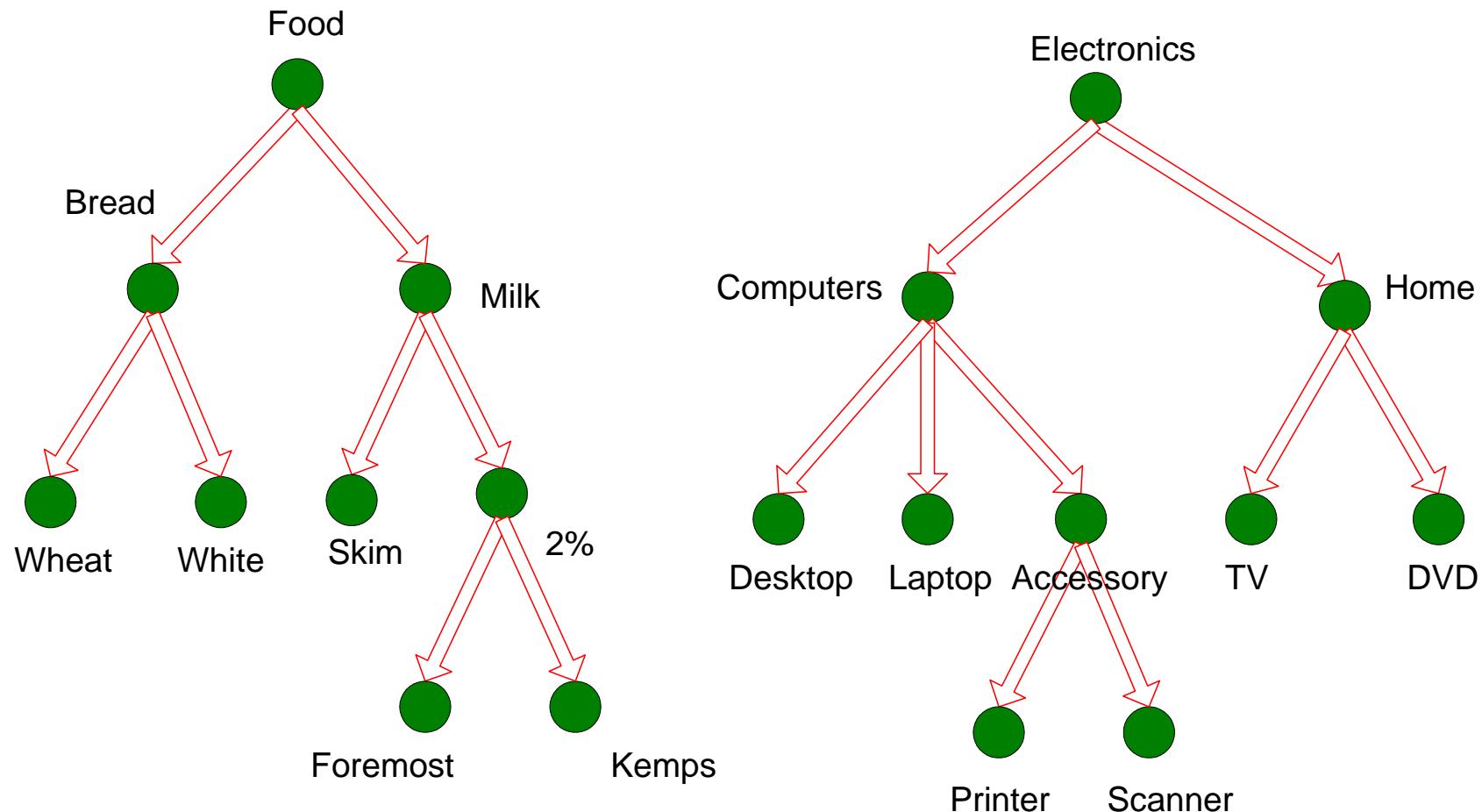
- Execution time

- If the range is partitioned into k intervals, there are $O(k^2)$ new items
- If an interval $[a,b)$ is frequent, then all intervals that subsume $[a,b)$ must also be frequent
 - ◆ E.g.: if $\{\text{Age } \in [21,25), \text{ Chat Online=Yes}\}$ is frequent, then $\{\text{Age } \in [10,50), \text{ Chat Online=Yes}\}$ is also frequent
- Improve efficiency:
 - ◆ Use maximum support to avoid intervals that are too wide

Multi-level Association Rules

- Why should we incorporate concept hierarchy?
 - Rules at lower levels may not have enough support to appear in any frequent itemsets
 - Rules at lower levels of the hierarchy are overly specific
 - ◆ e.g., following rules are indicative of association between milk and bread
 - skim milk → white bread,
 - 2% milk → wheat bread,
 - skim milk → wheat bread, etc.
 - Rules at higher level of hierarchy may be too generic
 - ◆ e.g., electronics → food

Concept Hierarchies



Multi-level Association Rules

- How do support and confidence vary as we traverse the concept hierarchy?
 - If $\sigma(X_1 \cup Y_1) \geq \text{minsup}$,
and X is parent of X_1 , Y is parent of Y_1
then $\sigma(X \cup Y_1) \geq \text{minsup}$, $\sigma(X_1 \cup Y) \geq \text{minsup}$
 $\sigma(X \cup Y) \geq \text{minsup}$
 - If $\text{conf}(X_1 \Rightarrow Y_1) \geq \text{minconf}$,
then $\text{conf}(X_1 \Rightarrow Y) \geq \text{minconf}$

Multi-level Association Rules

- Approach 1:
 - Extend current association rule formulation by augmenting each transaction with higher level items

Original Transaction: {skim milk, wheat bread}

Augmented Transaction:

{skim milk, wheat bread, milk, bread, food}

- Issues:
 - Items that reside at higher levels have much higher support counts
 - ◆ if support threshold is low, too many frequent patterns involving items from the higher levels
 - Increased dimensionality of the data

Multi-level Association Rules

- Approach 2:
 - Generate frequent patterns at highest level first
 - Then, generate frequent patterns at the next highest level, and so on
- Issues:
 - I/O requirements will increase dramatically because we need to perform more passes over the data
 - May miss some potentially interesting cross-level association patterns

Dimensionality Reduction (PCA ..)

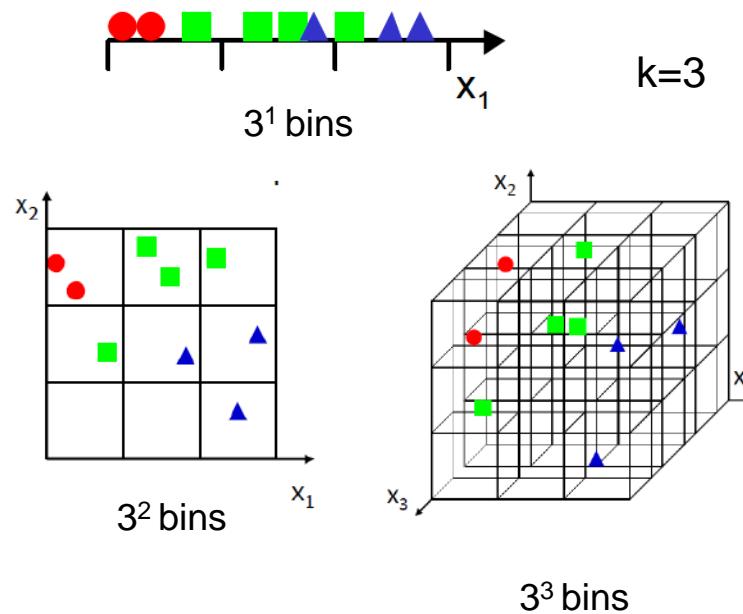
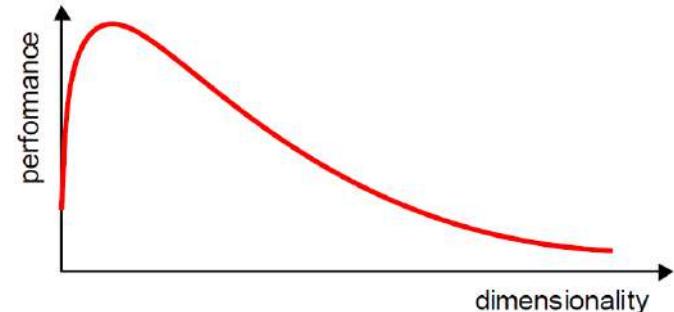
Working with High-Dimensional Data

- High-dimensional data
 - Many applications: text documents, DNA micro-array data
 - Major challenges:
 - Many irrelevant dimensions may mask class and clusters
 - Measure becomes meaningless—e.g. distance → equi-distance
 - Issues like Clusters may exist only in some subspaces
- Methods
 - Feature transformation: only effective if most dimensions are relevant
 - PCA & SVD useful only when features are highly correlated/redundant
 - Feature selection: wrapper or filter approaches
 - useful to find a subspace where the data have nice clusters
 - Subspace-clustering: find clusters in all the possible subspaces
 - CLIQUE, ProClus, and frequent pattern-based clustering

Curse of Dimensionality

- Increasing the number of features will not always improve classification accuracy.
- In practice, the inclusion of more features might actually lead to **worse** performance.
- The number of training examples required increases **exponentially** with dimensionality d (i.e., k^d).

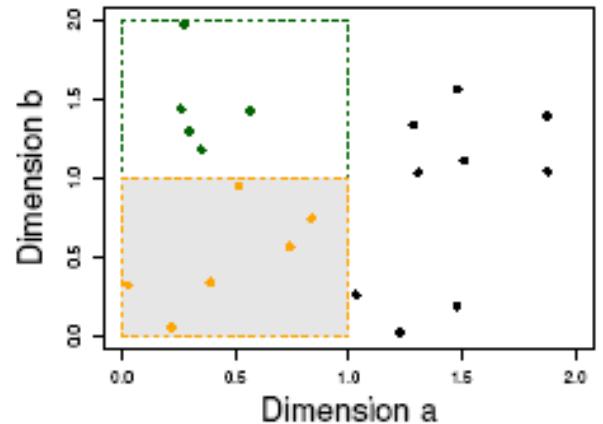
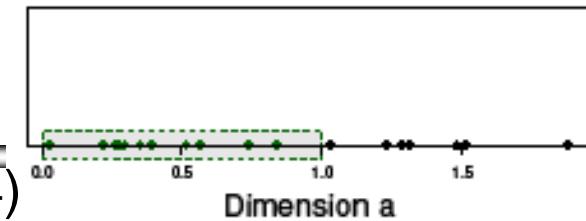
k : number of bins per feature



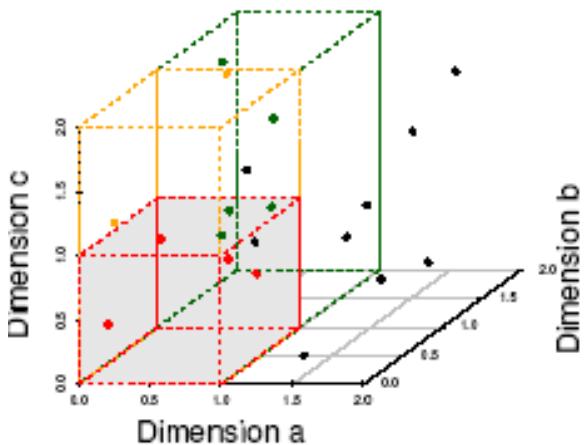
Curse of Dimensionality

(graphs adapted from Parsons et al. KDD Explorations 2004)

- Data in only one dimension is relatively packed
- Adding a dimension “stretch” the points across that dimension, making them further apart
- Adding more dimensions will make the points further apart—high dimensional data is extremely sparse
- Distance measure becomes meaningless—due to equi-distance



(b) 6 Objects in One Unit Bin

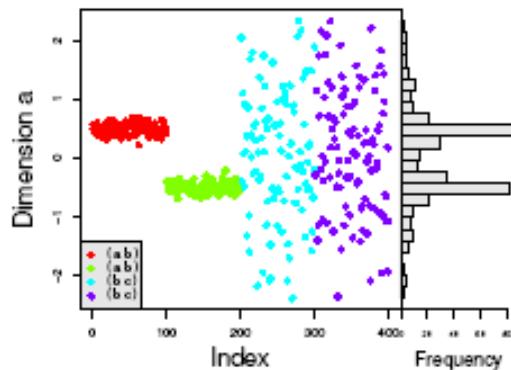
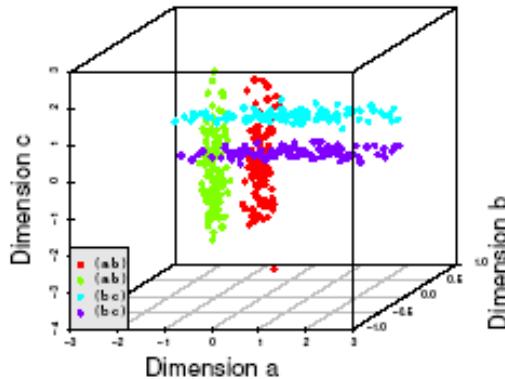


(c) 4 Objects in One Unit Bin

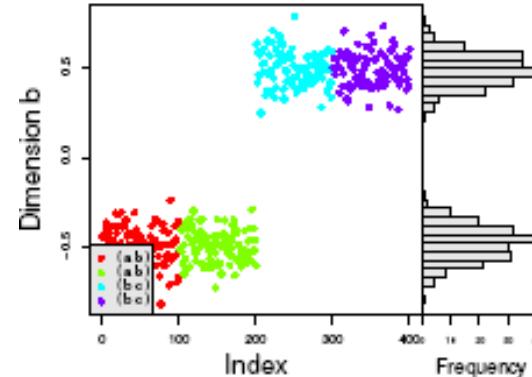
Why Subspace Clustering?

(adapted from Parsons et al. SIGKDD Explorations 2004)

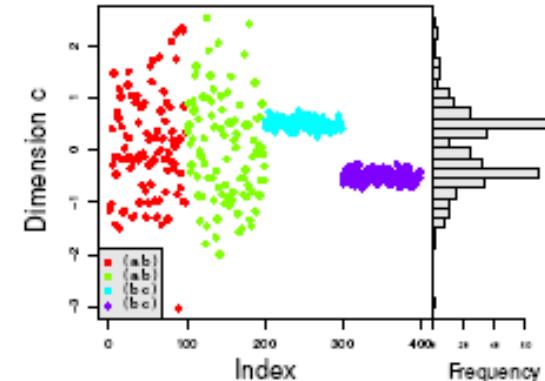
- Clusters may exist only in some subspaces
- Subspace-clustering: find clusters in all the subspaces



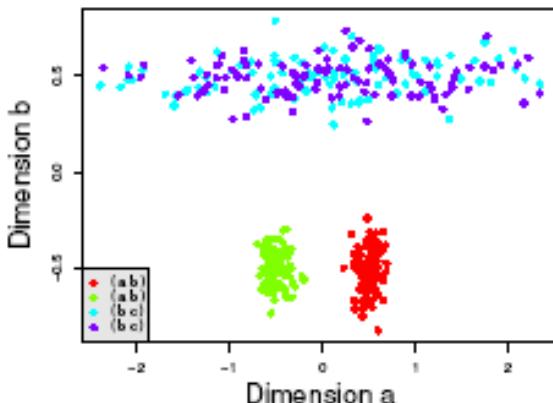
(a) Dimension a



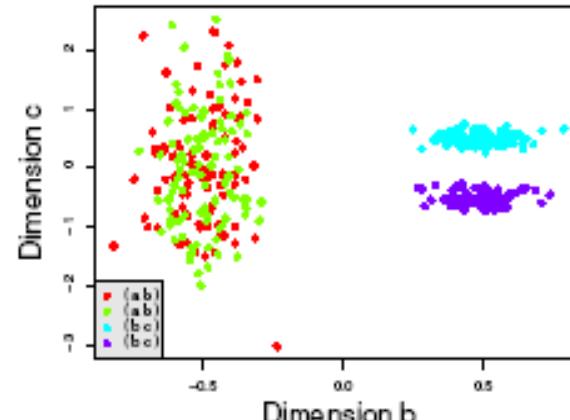
(b) Dimension b



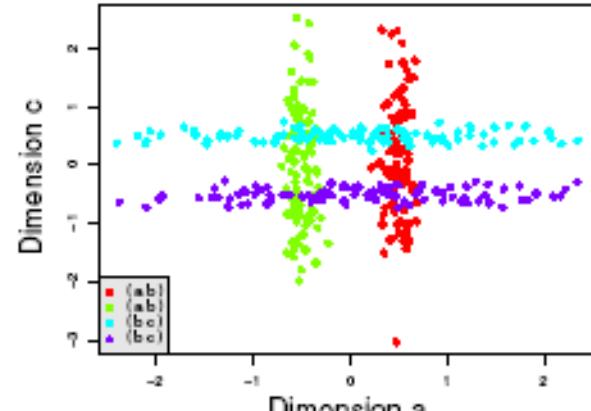
(c) Dimension c



(a) Dims a & b



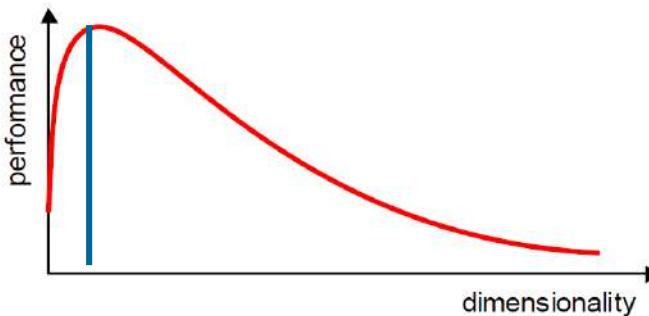
(b) Dims b & c



(c) Dims a & c

Dimensionality Reduction

- What is the objective?
 - Choose an optimum set of features of lower dimensionality to **improve** classification accuracy.



- Different methods can be used to reduce dimensionality:
 - Feature extraction
 - Feature selection

Dimensionality Reduction (cont'd)

Feature extraction: finds a set of **new** features (i.e., through some mapping **f()**) from the **existing** features.

Feature selection: chooses a subset of the **original** features.

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ \vdots \\ x_N \end{bmatrix} \xrightarrow{f(\mathbf{x})} \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ \vdots \\ y_K \end{bmatrix}$$

The mapping $f()$ could be **linear** or **non-linear**

K << N

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ \vdots \\ x_N \end{bmatrix} \rightarrow \mathbf{y} = \begin{bmatrix} x_{i_1} \\ x_{i_2} \\ \vdots \\ \vdots \\ \vdots \\ x_{i_K} \end{bmatrix}$$

K << N

Feature Extraction

- **Linear** combinations are particularly attractive because they are simpler to compute and analytically tractable.
- Given $\mathbf{x} \in \mathbb{R}^N$, find an **K x N** matrix **T** such that:

$$\mathbf{y} = \mathbf{T}\mathbf{x} \in \mathbb{R}^K \text{ where } K \ll N$$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ x_N \end{bmatrix} \xrightarrow{\mathbf{T}} \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \cdot \\ \cdot \\ y_K \end{bmatrix}$$

This is a **projection** from the N-dimensional space to a K-dimensional space.

Feature Extraction (cont'd)

- From a mathematical point of view, finding an **optimum** mapping $\mathbf{y}=f(\mathbf{x})$ is equivalent to optimizing an **objective** criterion.
- Different methods use different objective criteria, e.g.,
 - **Minimize Information Loss**: represent the data as accurately as possible in the lower-dimensional space.
 - **Maximize Discriminatory Information**: enhance the class-discriminatory information in the lower-dimensional space.

Feature Extraction (cont'd)

- Popular **linear** feature extraction methods:
 - Principal Components Analysis (PCA): Seeks a projection that **preserves** as much **information** in the data as possible.
 - Linear Discriminant Analysis (LDA): Seeks a projection that **best discriminates** the data.
- Many other methods:
 - Making features as independent as possible (**Independent Component Analysis or ICA**).
 - Retaining interesting directions (**Projection Pursuit**).
 - Embedding to lower dimensional manifolds (**Isomap, Locally Linear Embedding or LLE**).

Vector Representation

- A vector $\mathbf{x} \in \mathbb{R}^n$ can be represented by **n** components:
- Assuming the standard base $\langle v_1, v_2, \dots, v_N \rangle$ (i.e., unit vectors in each dimension), x_i can be obtained by **projecting** \mathbf{x} along the direction of v_i :
- \mathbf{x} can be “**reconstructed**” from its projections as follows:
- Since the basis vectors are the same for all $\mathbf{x} \in \mathbb{R}^n$ (standard basis), we typically represent them as a **n**-component vector.

$$\mathbf{x} : \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_N \end{bmatrix}$$

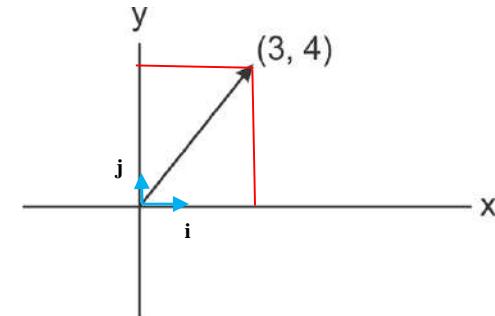
$$x_i = \frac{\mathbf{x}^T v_i}{v_i^T v_i} = \mathbf{x}^T v_i$$

$$\mathbf{x} = \sum_{i=1}^N x_i v_i = x_1 v_1 + x_2 v_2 + \dots + x_N v_N$$

Vector Representation (cont'd)

- Example assuming n=2:

$$\mathbf{x} : \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$$



- Assuming the standard base $\langle v_1=i, v_2=j \rangle$, x_i can be obtained by projecting x along the direction of v_i :

$$x_1 = \mathbf{x}^T i = [3 \quad 4] \begin{bmatrix} 1 \\ 0 \end{bmatrix} = 3$$

$$x_2 = \mathbf{x}^T j = [3 \quad 4] \begin{bmatrix} 0 \\ 1 \end{bmatrix} = 4$$

- \mathbf{x} can be “reconstructed” from its projections as follows:

$$\mathbf{x} = 3i + 4j$$

Principal Component Analysis (PCA)

- If $\mathbf{x} \in \mathbb{R}^N$, then it can be written as a linear combination of an **orthonormal** set of N basis vectors $\langle \mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_N \rangle$ in \mathbb{R}^N (e.g., using the standard base):

$$\mathbf{v}_i^T \mathbf{v}_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{x} = \sum_{i=1}^N x_i \mathbf{v}_i = x_1 \mathbf{v}_1 + x_2 \mathbf{v}_2 + \dots + x_N \mathbf{v}_N$$

where $x_i = \frac{\mathbf{x}^T \mathbf{v}_i}{\mathbf{v}_i^T \mathbf{v}_i} = \mathbf{x}^T \mathbf{v}_i$

$$\mathbf{x}: \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}$$

- PCA seeks to **approximate** \mathbf{x} in a **subspace** of \mathbb{R}^N using a **new** set of $K < N$ basis vectors $\langle \mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_K \rangle$ in \mathbb{R}^N :

$$\hat{\mathbf{x}} = \sum_{i=1}^K y_i \mathbf{u}_i = y_1 \mathbf{u}_1 + y_2 \mathbf{u}_2 + \dots + y_K \mathbf{u}_K$$

(reconstruction)

where $y_i = \frac{\mathbf{x}^T \mathbf{u}_i}{\mathbf{u}_i^T \mathbf{u}_i} = \mathbf{x}^T \mathbf{u}_i$

$$\hat{\mathbf{x}}: \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_K \end{bmatrix}$$

such that $\|\mathbf{x} - \hat{\mathbf{x}}\|$ is **minimized!**
(i.e., minimize information loss)

Principal Component Analysis (PCA)

- The “**optimal**” set of basis vectors $\langle u_1, u_2, \dots, u_K \rangle$ can be found as follows (we will see why):

(1) Find the **eigenvectors** u_i of the **covariance** matrix of the (training) data Σ_x

$$\Sigma_x u_i = \lambda_i u_i$$

(2) Choose the K “**largest**” eigenvectors u_i (i.e., corresponding to the K “**largest**” eigenvalues λ_i)

$\langle u_1, u_2, \dots, u_K \rangle$ correspond to the “optimal” basis!

We refer to the “**largest**” eigenvectors u_i as **principal components**.

PCA - Steps

- Suppose we are given $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M$ ($N \times 1$) vectors
- N: # of features

Step 1: compute sample mean

$$\bar{\mathbf{x}} = \frac{1}{M} \sum_{i=1}^M \mathbf{x}_i$$

M: # data

Step 2: subtract sample mean (i.e., center data at zero)

$$\Phi_i = \mathbf{x}_i - \bar{\mathbf{x}}$$

Step 3: compute the sample covariance matrix Σ_x

$$\Sigma_x = \frac{1}{M} \sum_{i=1}^M (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T = \frac{1}{M} \sum_{i=1}^M \Phi_i \Phi_i^T = \frac{1}{M} A A^T$$

where $A = [\Phi_1 \ \Phi_2 \ \dots \ \Phi_M]$
i.e., the columns of A are the Φ_i
($N \times M$ matrix)

PCA - Steps

Step 4: compute the eigenvalues/eigenvectors of Σ_x

$$\Sigma_x u_i = \lambda_i u_i$$

where we assume $\lambda_1 > \lambda_2 > \dots > \lambda_N$

Note : most software packages return the eigenvalues (and corresponding eigenvectors) in **decreasing** order – if not, you can explicitly put them in this order)

Since Σ_x is symmetric, $\langle u_1, u_2, \dots, u_N \rangle$ form an **orthogonal basis** in R^N and we can represent **any** $x \in R^N$ as:

$$x - \bar{x} = \sum_{i=1}^N y_i u_i = y_1 u_1 + y_2 u_2 + \dots + y_N u_N$$

$$y_i = \frac{(x - \bar{x})^T u_i}{u_i^T u_i} = (x - \bar{x})^T u_i \quad \text{if } \|u_i\| = 1$$

i.e., this is
just a “**change**”
of basis!

$$x - \bar{x} : \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \rightarrow \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

Note : most software packages **normalize** u_i to unit length to simplify calculations; if not, you can explicitly normalize them)

PCA - Steps

Step 5: dimensionality reduction step – approximate \mathbf{x} using only the **first K eigenvectors** ($K \ll N$) (i.e., corresponding to the **K largest eigenvalues** where K is a **parameter**):

$$\mathbf{x} - \bar{\mathbf{x}} = \sum_{i=1}^N y_i \mathbf{u}_i = y_1 \mathbf{u}_1 + y_2 \mathbf{u}_2 + \dots + y_N \mathbf{u}_N$$

 approximate \mathbf{x} by $\hat{\mathbf{x}}$
using first K eigenvectors only

$$\hat{\mathbf{x}} - \bar{\mathbf{x}} = \sum_{i=1}^K y_i \mathbf{u}_i = y_1 \mathbf{u}_1 + y_2 \mathbf{u}_2 + \dots + y_K \mathbf{u}_K$$

(reconstruction)

$$\mathbf{x} - \bar{\mathbf{x}} : \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ \vdots \\ x_N \end{bmatrix} \rightarrow \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ \vdots \\ y_N \end{bmatrix} \rightarrow \hat{\mathbf{x}} - \bar{\mathbf{x}} : \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ \vdots \\ y_K \end{bmatrix}$$

note that if $K=N$, then 
(i.e., zero reconstruction error)

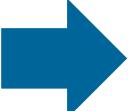
What is the Linear Transformation implied by PCA?

- The linear transformation $\mathbf{y} = \mathbf{T}\mathbf{x}$ which performs the dimensionality reduction in PCA is:

$$\hat{\mathbf{x}} - \bar{\mathbf{x}} = \sum_{i=1}^K y_i u_i = y_1 u_1 + y_2 u_2 + \dots + y_K u_K$$

$$(\hat{\mathbf{x}} - \bar{\mathbf{x}}) = U \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_K \end{bmatrix} \quad \text{where } U = [u_1 \ u_2 \ \dots \ u_K] \quad N \times K \text{ matrix}$$

i.e., the **columns** of U are the the first K eigenvectors of Σ_x


$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_K \end{bmatrix} = U^T (\hat{\mathbf{x}} - \bar{\mathbf{x}}) \quad \mathbf{T} = \mathbf{U}^T \quad K \times N \text{ matrix}$$

i.e., the **rows** of T are the first K eigenvectors of Σ_x

What is the form of Σ_y ?

$$\Sigma_x = \frac{1}{M} \sum_{i=1}^M (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T = \frac{1}{M} \sum_{i=1}^M \Phi_i \Phi_i^T$$

Using diagonalization:

$$\Sigma_x = P \Lambda P^T$$

The columns of P are the
eigenvectors of Σ_x

The diagonal elements of
 Λ are the **eigenvalues** of Σ_x
or the **variances**

$$\mathbf{y}_i = U^T (\mathbf{x}_i - \bar{\mathbf{x}}) = P^T \Phi_i$$

$$\Sigma_y = \frac{1}{M} \sum_{i=1}^M (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^T = \frac{1}{M} \sum_{i=1}^M (\mathbf{y}_i)(\mathbf{y}_i)^T = \frac{1}{M} \sum_{i=1}^M (P^T \Phi_i)(P^T \Phi_i)^T =$$

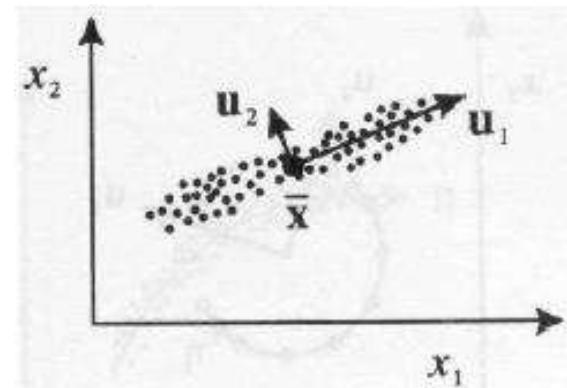
$$\frac{1}{M} \sum_{i=1}^M (P^T \Phi_i)(\Phi_i^T P) = P^T \left(\frac{1}{M} \sum_{i=1}^M \Phi_i \Phi_i^T \right) P = P^T \Sigma_x P = P^T (P \Lambda P^T) P = \Lambda$$

$$\Sigma_y = \Lambda$$

PCA de-correlates the data!
Preserves original variances!

Interpretation of PCA

- PCA chooses the **eigenvectors** of the covariance matrix corresponding to the **largest** eigenvalues.
- The **eigenvalues** correspond to the **variance** of the data along the eigenvector directions.
- Therefore, PCA projects the data along the directions where the data varies **most**.
- PCA preserves as much **information** in the data by preserving as much **variance** in the data.



u_1 : direction of **max** variance
 u_2 : orthogonal to u_1

Example

- Compute the PCA of the following dataset:

(1,2),(3,3),(3,5),(5,4),(5,6),(6,5),(8,7),(9,8)

- Compute the sample covariance matrix is:

$$\hat{\Sigma} = \frac{1}{n} \sum_{k=1}^n (\mathbf{x}_k - \hat{\mu})(\mathbf{x}_k - \hat{\mu})^t$$

$$\Sigma_x = \begin{bmatrix} 6.25 & 4.25 \\ 4.25 & 3.5 \end{bmatrix}$$

- The eigenvalues can be computed by finding the roots of the characteristic polynomial:

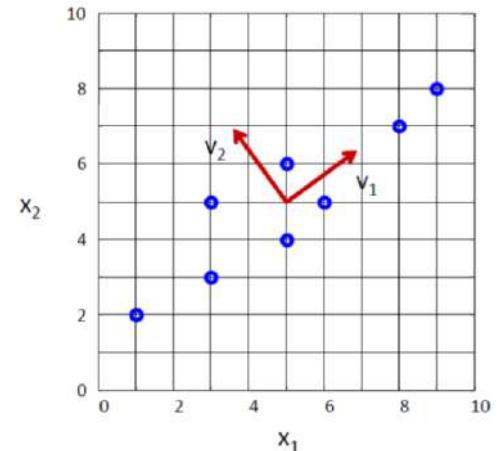
$$\begin{aligned}\Sigma_x v &= \lambda v \Rightarrow |\Sigma_x - \lambda I| = 0 \\ &\Rightarrow \begin{vmatrix} 6.25 - \lambda & 4.25 \\ 4.25 & 3.5 - \lambda \end{vmatrix} = 0 \\ &\Rightarrow \lambda_1 = 9.34; \lambda_2 = 0.41\end{aligned}$$

Example (cont'd)

- The eigenvectors are the solutions of the systems:

$$\sum_{\mathbf{x}} \mathbf{u}_i = \lambda_i \mathbf{u}_i$$

$$\begin{bmatrix} 6.25 & 4.25 \\ 4.25 & 3.5 \end{bmatrix} \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = \begin{bmatrix} \lambda_1 v_{11} \\ \lambda_1 v_{12} \end{bmatrix} \Rightarrow \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = \begin{bmatrix} 0.81 \\ 0.59 \end{bmatrix}$$
$$\begin{bmatrix} 6.25 & 4.25 \\ 4.25 & 3.5 \end{bmatrix} \begin{bmatrix} v_{21} \\ v_{22} \end{bmatrix} = \begin{bmatrix} \lambda_2 v_{21} \\ \lambda_2 v_{22} \end{bmatrix} \Rightarrow \begin{bmatrix} v_{21} \\ v_{22} \end{bmatrix} = \begin{bmatrix} -0.59 \\ 0.81 \end{bmatrix}$$



Note: if \mathbf{u}_i is a solution, then $c\mathbf{u}_i$ is also a solution where $c \neq 0$.

Eigenvectors can be normalized to unit-length using:

$$\hat{\mathbf{v}}_i = \frac{\mathbf{v}_i}{\| \mathbf{v}_i \|}$$

How do we choose K ?

- K is typically chosen based on how much **information (variance)** we want to preserve:

Choose the **smallest**
 K that satisfies
the following
inequality:

$$\frac{\sum_{i=1}^K \lambda_i}{\sum_{i=1}^N \lambda_i} > T \quad \text{where } T \text{ is a threshold (e.g., 0.9)}$$

- If $T=0.9$, for example, we “**preserve**” 90% of the information (variance) in the data.
- If $K=N$, then we “**preserve**” 100% of the information in the data (i.e., just a “**change**” of basis and $\hat{\mathbf{x}} = \mathbf{x}$)

Approximation Error

- The approximation error (or reconstruction error) can be computed by:

$$\| \mathbf{x} - \hat{\mathbf{x}} \|$$

where $\hat{\mathbf{x}} = \sum_{i=1}^K y_i u_i + \bar{\mathbf{x}} = y_1 u_1 + y_2 u_2 + \dots + y_K u_K + \bar{\mathbf{x}}$
(reconstruction)

- It can also be shown that the approximation error can be computed as follows:

$$\| \mathbf{x} - \hat{\mathbf{x}} \| = \frac{1}{2} \sum_{i=K+1}^N \lambda_i$$

Data Normalization

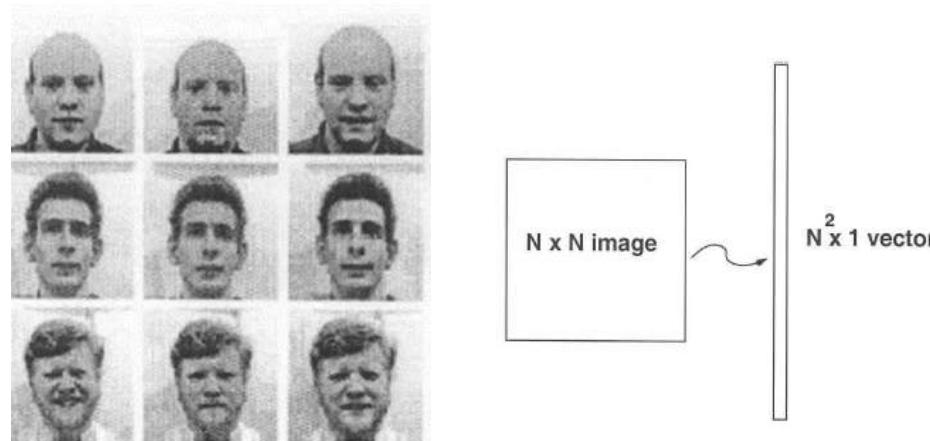
- The principal components are dependent on the ***units*** used to measure the original variables as well as on the ***range*** of values they assume.
- Data should **always** be normalized prior to using PCA.
- A common normalization method is to transform all the data to have **zero mean** and **unit standard deviation**:

$$\frac{x_i - \mu}{\sigma}$$

where μ and σ are the mean and standard deviation of the i -th feature x_i

Application to Images

- The goal is to represent images in a space of lower dimensionality using PCA.
 - Useful for various applications, e.g., face recognition, image compression, etc.
- Given **M** images of size **N x N**, first represent each image as a 1D vector (i.e., by stacking the rows together).
 - Note that for **face recognition**, faces must be **centered** and of the same **size**.



MATRIX FACTORIZATION

&

RECOMMENDER SYSTEMS

Outline

- **Recommender Systems**
 - Content Filtering
 - Collaborative Filtering
 - CF: Neighborhood Methods
 - CF: Latent Factor Methods
- **Matrix Factorization**
 - User / item vectors
 - Prediction model
 - Training by SGD
- **Extra: Matrix Multiplication in ML**
 - Matrix Factorization
 - Linear Regression
 - PCA
 - (Autoencoders)
 - K-means

Recommender Systems

A Common Challenge:

- Assume you're a company selling **items** of some sort: movies, songs, products, etc.
- Company collects millions of **ratings** from **users** of their **items**
- To maximize profit / user happiness, you want to **recommend** items that users are likely to want

Recommender Systems

The screenshot shows the Amazon homepage with a large banner at the top for 'NEW & INTERESTING FINDS ON AMAZON' and a 'EXPLORE' button. Below the banner, there's a search bar with the text 'All' and a magnifying glass icon. To the right of the search bar are promotional banners for 'CYBER MONDAY' and 'DEALS WEEK'. The navigation bar includes links for 'Departments', 'Browsing History', 'Matt's Amazon.com', 'Cyber Monday', 'Gift Cards & Registry', 'Sell', 'Help', 'Hello, Matt Your Account', 'Prime', 'Lists', and a shopping cart icon with the number '1'.

Matt's Amazon

You could be seeing useful stuff here!
Sign in to get your order status, balances and rewards.

Sign In

Recommended for you, Matt

Buy It Again in Grocery
14 ITEMS

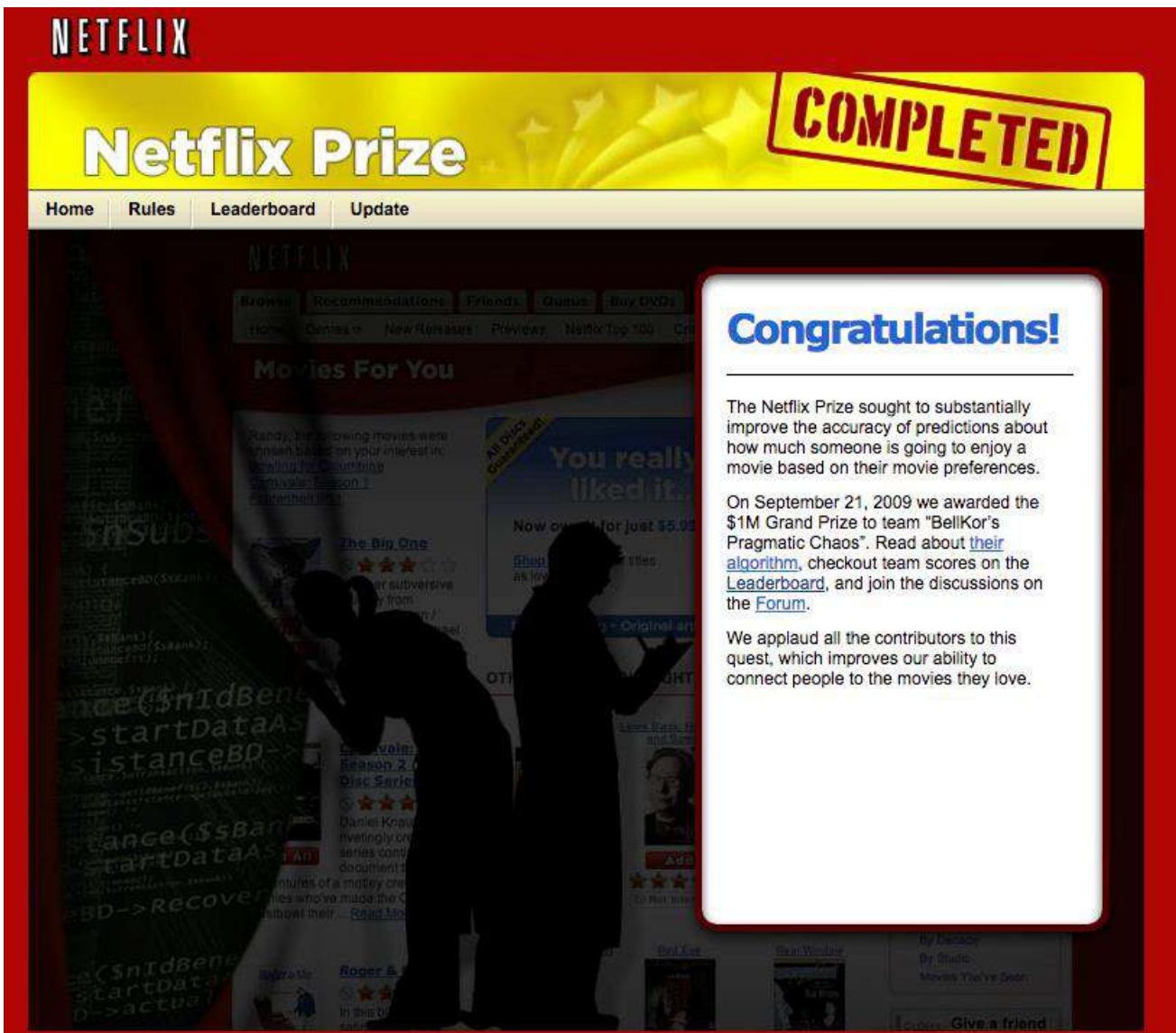
Buy It Again in Pets
6 ITEMS

Buy It Again in Baby Products
5 ITEMS

Engineering Books
86 ITEMS

PROBABILISTIC GRAPHICAL MODELS
PRINCIPLES AND TECHNIQUES
DAPHNE KOLLER AND JIRI FRIEDMAN

Recommender Systems



Recommender Systems

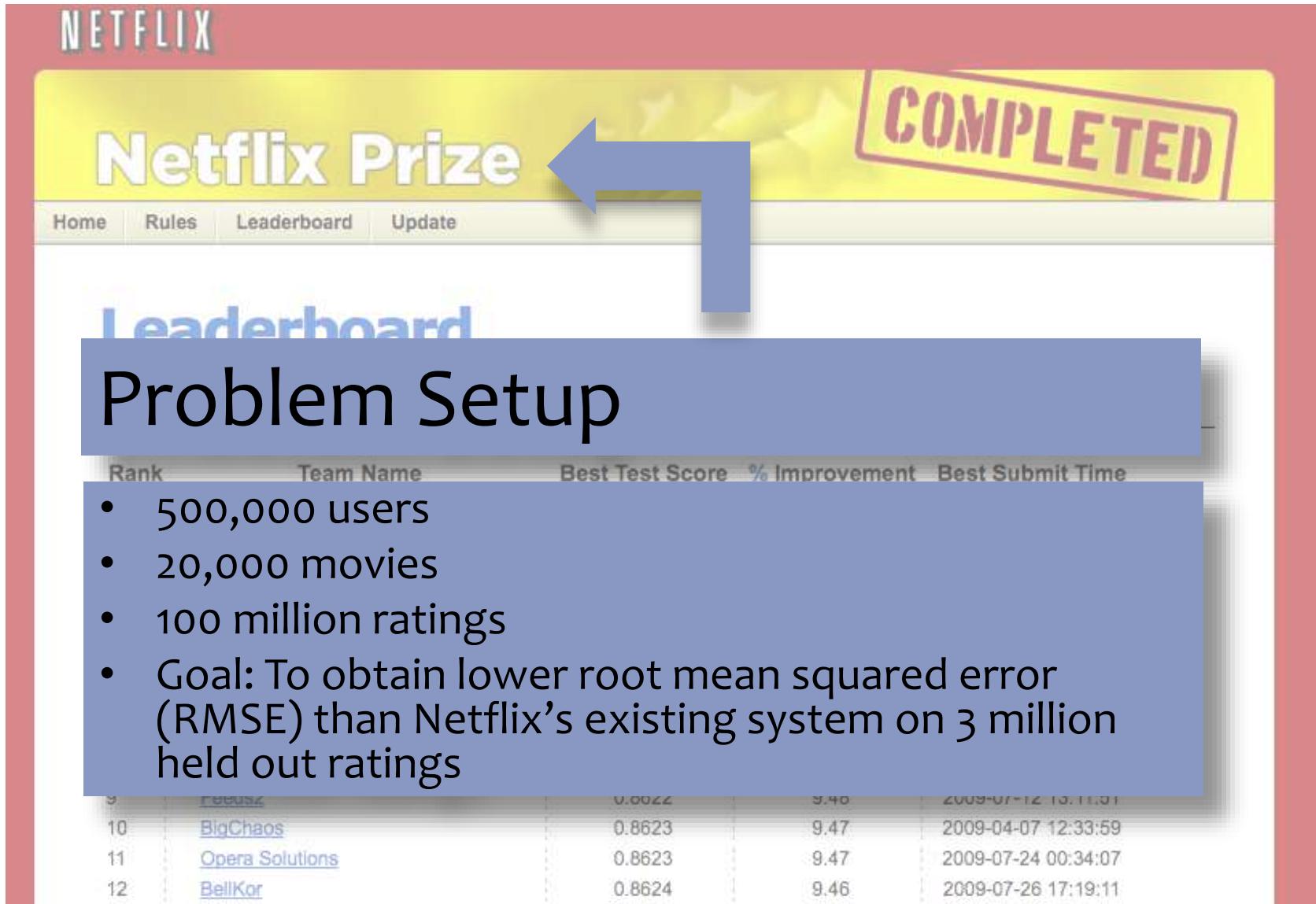
The image shows a screenshot of the Netflix Prize website. On the left, there's a blurred background of the 'Movies For You' section of the Netflix homepage. On the right, a prominent yellow banner at the top says 'CONGRATULATIONS!' in large blue letters. Below this, a large blue header reads 'Congratulations!'. Underneath, there's a paragraph of text followed by another paragraph. At the bottom, there are links for 'FAQ' and 'Forum'.

The Netflix Prize sought to substantially improve the accuracy of predictions about how much someone is going to enjoy a movie based on their movie preferences.

On September 21, 2009 we awarded the \$1M Grand Prize to team "BellKor's Pragmatic Chaos". Read about [their algorithm](#), checkout team scores on the [Leaderboard](#), and join the discussions on the [Forum](#).

We applaud all the contributors to this quest, which improves our ability to connect people to the movies they love.

Recommender Systems



The image shows a screenshot of the Netflix Prize website. At the top, there's a yellow banner with the text "Netflix Prize" and a large red "COMPLETED" stamp. Below the banner is a navigation bar with links for "Home", "Rules", "Leaderboard", and "Update". A blue arrow points from the word "Leaderboard" in the navigation bar down to the "Leaderboard" section of the page. The "Leaderboard" section has a blue header and displays a table of results. The columns in the table are "Rank", "Team Name", "Best Test Score", "% Improvement", and "Best Submit Time". The table lists four teams: "Freesurk" (rank 9), "BigChaos" (rank 10), "Opera Solutions" (rank 11), and "BellKor" (rank 12). All teams have a best test score of 0.8622 or 0.8623, a percentage improvement of 9.40 or 9.47, and a best submit time between July 12, 2009, and July 26, 2009.

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
9	Freesurk	0.8622	9.40	2009-07-12 13:11:01
10	BigChaos	0.8623	9.47	2009-04-07 12:33:59
11	Opera Solutions	0.8623	9.47	2009-07-24 00:34:07
12	BellKor	0.8624	9.46	2009-07-26 17:19:11

Recommender Systems

The image shows a screenshot of the Netflix Prize Leaderboard page. At the top, there is a yellow banner with the text "Netflix Prize" and a large red stamp that says "COMPLETED". Below the banner, there is a navigation bar with links for "Home", "Rules", "Leaderboard", and "Update". The main title "Leaderboard" is displayed in a large blue font. Below the title, there is a message "Showing Test Score. [Click here to show quiz score](#)". The table below lists the top 12 teams, their test scores, improvement percentages, and submission times. The winning team, "BellKor's Pragmatic Chaos", is highlighted in a blue row.

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries !	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09
8	Dace	0.8612	9.59	2009-07-24 17:18:43
9	Feeds2	0.8622	9.48	2009-07-12 13:11:51
10	BigChaos	0.8623	9.47	2009-04-07 12:33:59
11	Opera Solutions	0.8623	9.47	2009-07-24 00:34:07
12	BellKor	0.8624	9.46	2009-07-26 17:19:11

Recommender Systems

- **Setup:**
 - **Items:**
movies, songs, products, etc.
(often many thousands)
 - **Users:**
watchers, listeners, purchasers, etc.
(often many millions)
 - **Feedback:**
5-star ratings, not-clicking ‘next’,
purchases, etc.
- **Key Assumptions:**
 - Can represent ratings numerically
as a user/item matrix
 - Users only rate a small number of
items (the matrix is sparse)

	Doctor Strange	Star Trek: Beyond	Zootopia
Alice	1		5
Bob	3	4	
Charlie	3	5	2

Recommender Systems

The screenshot shows the Netflix Prize Leaderboard page. At the top, the Netflix logo is visible, followed by the text "Netflix Prize" and a large red stamp that says "COMPLETED". Below this, there is a navigation bar with links for "Home", "Rules", "Leaderboard", and "Update". The main section is titled "Leaderboard" in large blue letters. Below it, a message says "Showing Test Score. [Click here to show quiz score](#)". The table below lists the top 12 teams, their scores, improvement percentages, and submission times.

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries !	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09
8	Dace	0.8612	9.59	2009-07-24 17:18:43
9	Feeds2	0.8622	9.48	2009-07-12 13:11:51
10	BigChaos	0.8623	9.47	2009-04-07 12:33:59
11	Opera Solutions	0.8623	9.47	2009-07-24 00:34:07
12	BellKor	0.8624	9.46	2009-07-26 17:19:11

Two Types of Recommender Systems

Content Filtering

- Example: [Pandora.com](#) music recommendations (Music Genome Project)
- **Con:** Assumes access to **side information** about items (e.g. properties of a song)
- **Pro:** Got a **new item** to add? No problem, just be sure to include the side information

Collaborative Filtering

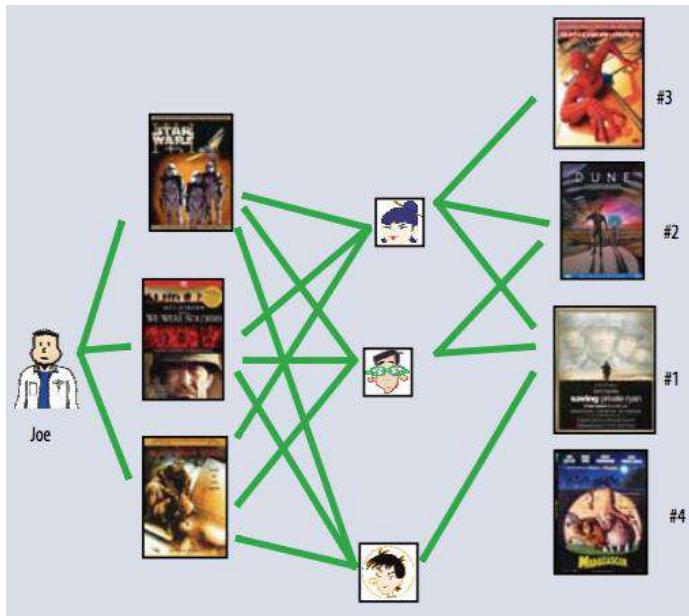
- Example: [Netflix](#) movie recommendations
- **Pro:** Does not assume access to **side information** about items (e.g. does not need to know about movie genres)
- **Con:** Does not work on **new items** that have no ratings

Collaborative Filtering

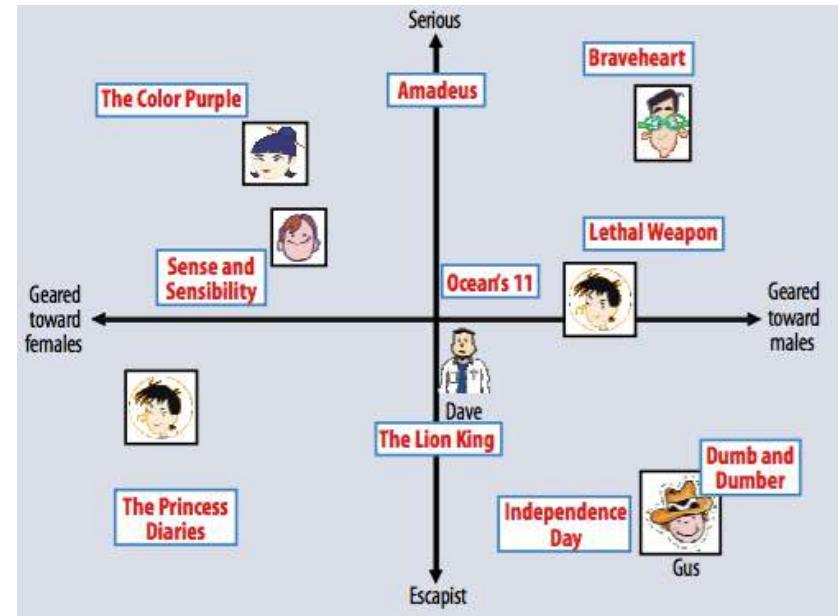
- **Everyday Examples of Collaborative Filtering...**
 - Bestseller lists
 - Top 40 music lists
 - The “recent returns” shelf at the library
 - Unmarked but well-used paths thru the woods
 - The printer room at work
 - “Read any good books lately?”
 - ...
- **Common insight:** personal tastes are correlated
 - If Alice and Bob both like X and Alice likes Y then Bob is more likely to like Y
 - especially (perhaps) if Bob knows Alice

Two Types of Collaborative Filtering

1. Neighborhood Methods

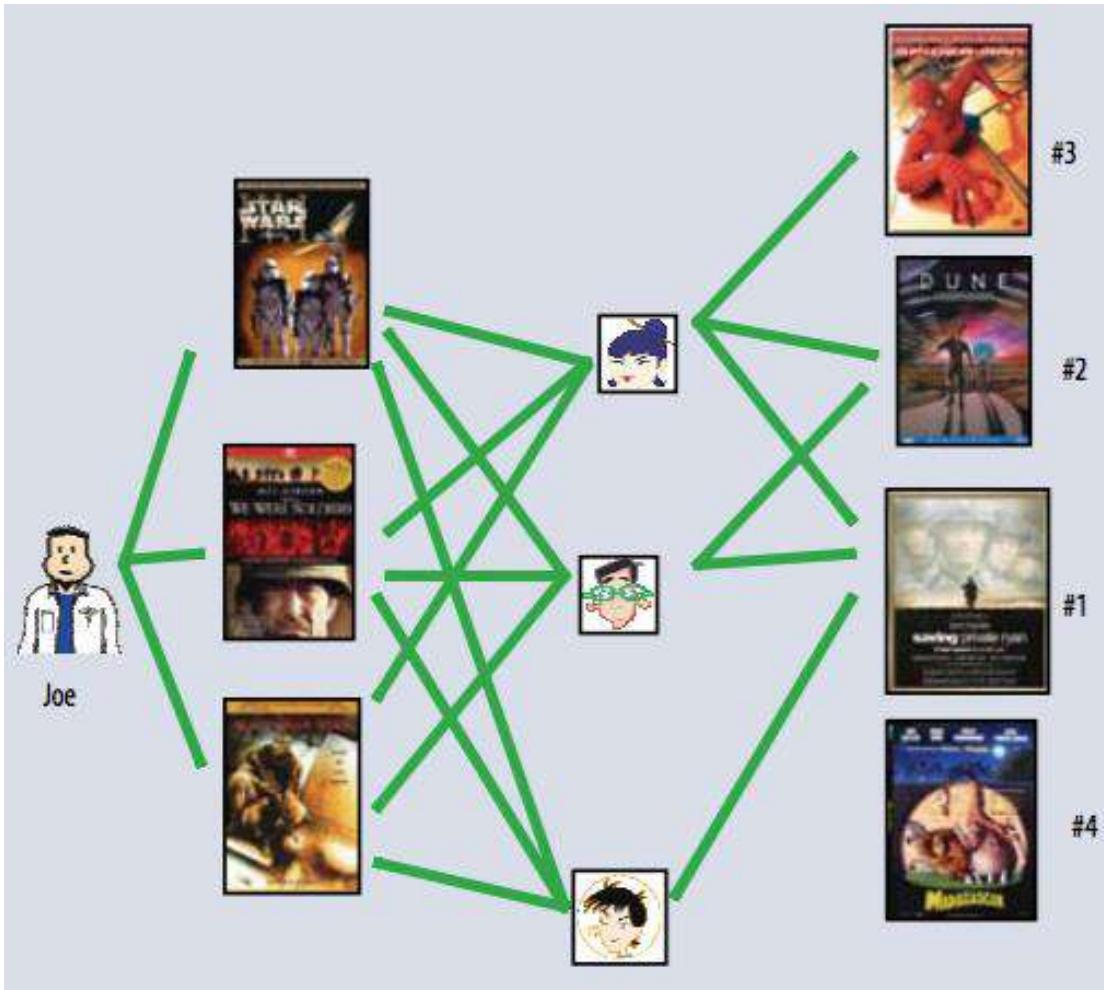


2. Latent Factor Methods



Two Types of Collaborative Filtering

1. Neighborhood Methods



In the figure, assume that a green line indicates the movie was **watched**

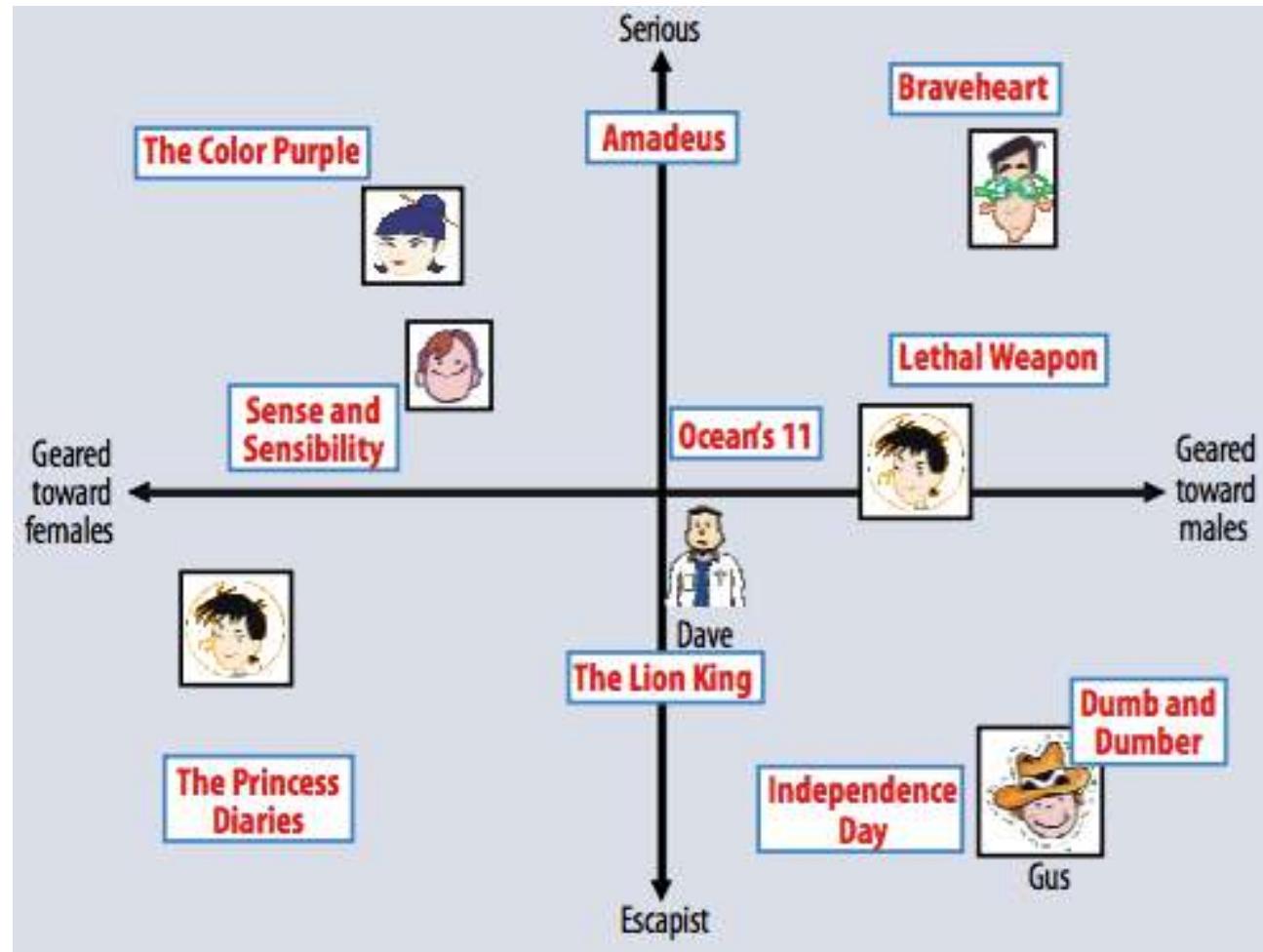
Algorithm:

1. **Find neighbors** based on similarity of movie preferences
2. **Recommend** movies that those neighbors watched

Two Types of Collaborative Filtering

2. Latent Factor Methods

- Assume that both movies and users live in some **low-dimensional space** describing their properties
- **Recommend** a movie based on its **proximity** to the user in the latent space



MATRIX FACTORIZATION

Matrix Factorization (with matrices)

- User vectors:

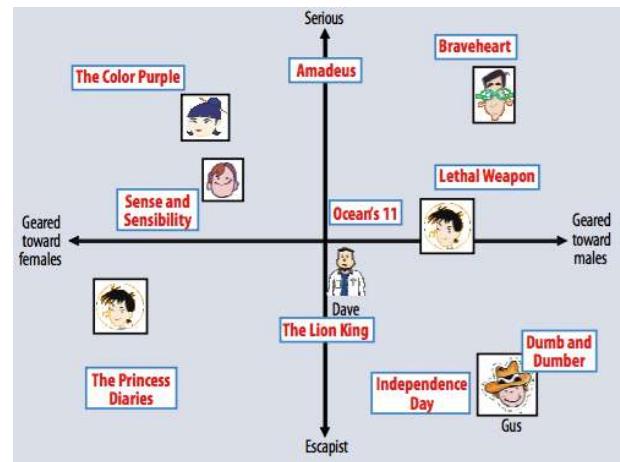
$$(W_{u*})^T \in \mathbb{R}^r$$

- Item vectors:

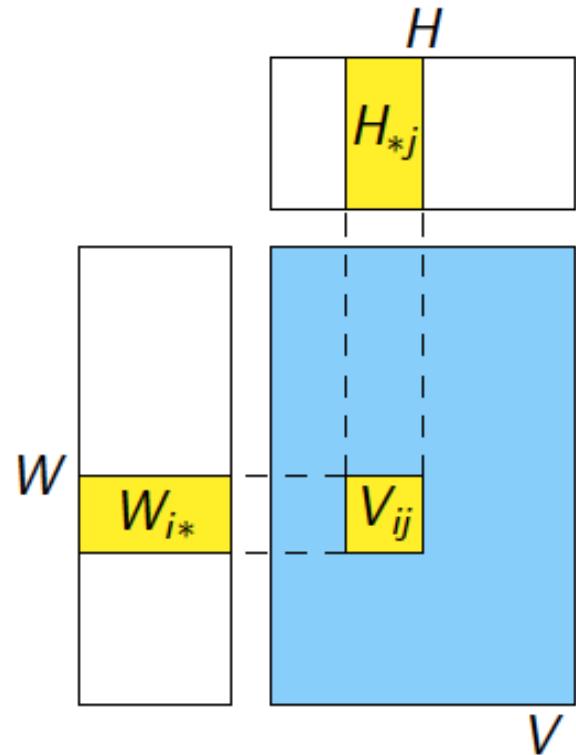
$$H_{*i} \in \mathbb{R}^r$$

- Rating prediction:

$$\begin{aligned} V_{ui} &= W_{u*} H_{*i} \\ &= [WH]_{ui} \end{aligned}$$



Figures from Koren et al. (2009)



Figures from Gemulla et al. (2011)₁₇

Matrix Factorization (with vectors)

- User vectors:

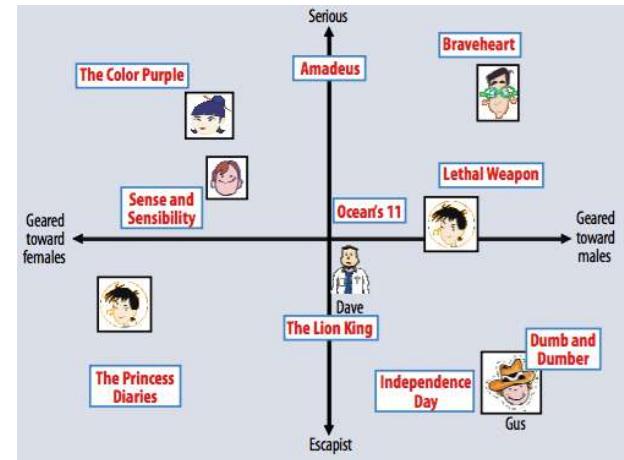
$$\mathbf{w}_u \in \mathbb{R}^r$$

- Item vectors:

$$\mathbf{h}_i \in \mathbb{R}^r$$

- Rating prediction:

$$v_{ui} = \mathbf{w}_u^T \mathbf{h}_i$$



Figures from Koren et al. (2009)

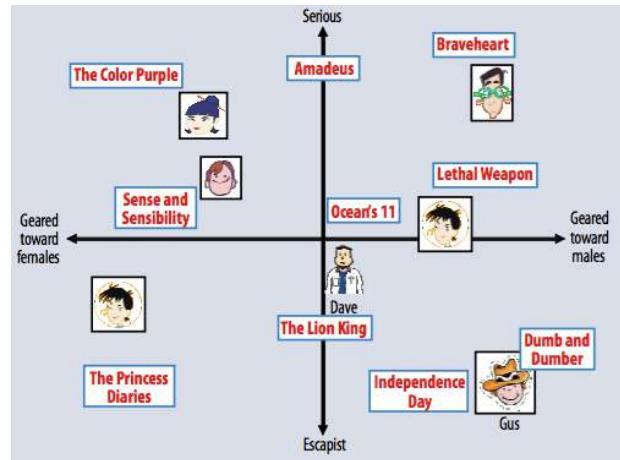
Matrix Factorization (with vectors)

- Set of non-zero entries:

$$\mathcal{Z} = \{(u, i) : v_{ui} \neq 0\}$$

- Objective:

$$\operatorname{argmin}_{\mathbf{w}, \mathbf{h}} \sum_{(u, i) \in \mathcal{Z}} (v_{ui} - \mathbf{w}_u^T \mathbf{h}_i)^2$$

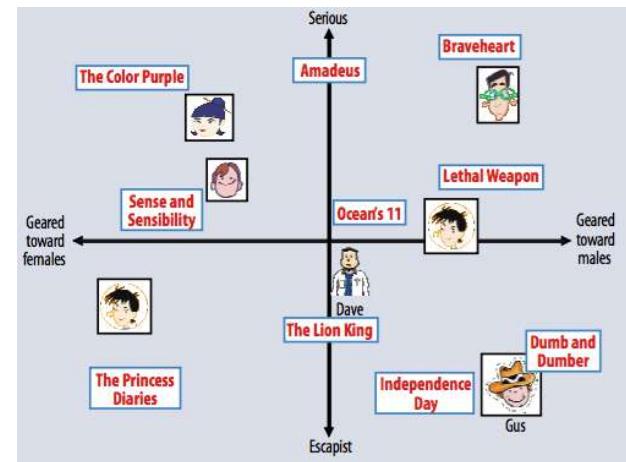


Figures from Koren et al. (2009)

Matrix Factorization (with vectors)

- Regularized Objective:

$$\operatorname{argmin}_{\mathbf{w}, \mathbf{h}} \sum_{(u,i) \in \mathcal{Z}} (v_{ui} - \mathbf{w}_u^T \mathbf{h}_i)^2 + \lambda \left(\sum_i \|\mathbf{w}_i\|^2 + \sum_u \|\mathbf{h}_u\|^2 \right)$$



Figures from Koren et al. (2009)

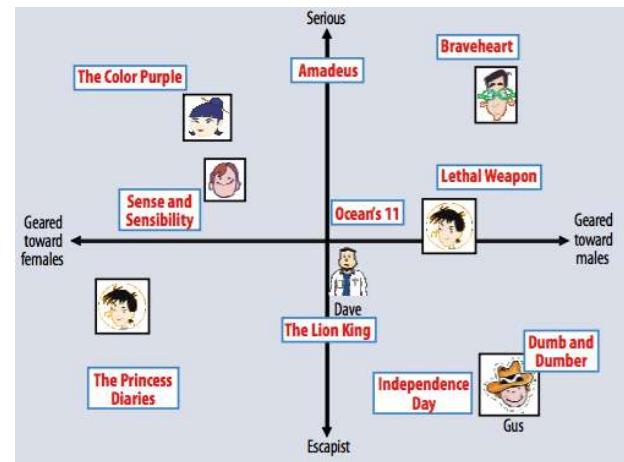
Matrix Factorization (with vectors)

- Regularized Objective:

$$\begin{aligned} \operatorname{argmin}_{\mathbf{w}, \mathbf{h}} \sum_{(u,i) \in \mathcal{Z}} & (v_{ui} - \mathbf{w}_u^T \mathbf{h}_i)^2 \\ & + \lambda \left(\sum_i \|\mathbf{w}_i\|^2 + \sum_u \|\mathbf{h}_u\|^2 \right) \end{aligned}$$

- Stochastic Gradient Descent (SGD) update for random (u,i):

$$\begin{aligned} e_{ui} &\leftarrow v_{ui} - \mathbf{w}_u^T \mathbf{h}_i \\ \mathbf{w}_u &\leftarrow \mathbf{w}_u + \gamma (e_{ui} \mathbf{h}_i - \lambda \mathbf{w}_u) \\ \mathbf{h}_i &\leftarrow \mathbf{h}_i + \gamma (e_{ui} \mathbf{w}_u - \lambda \mathbf{h}_i) \end{aligned}$$



Figures from Koren et al. (2009)

Matrix Factorization (with matrices)

- User vectors:

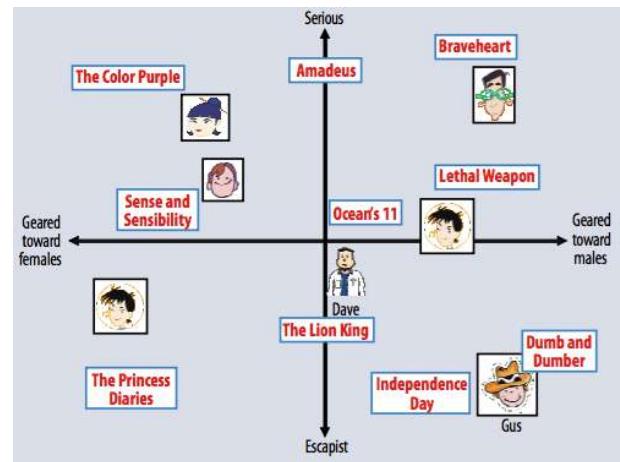
$$(W_{u*})^T \in \mathbb{R}^r$$

- Item vectors:

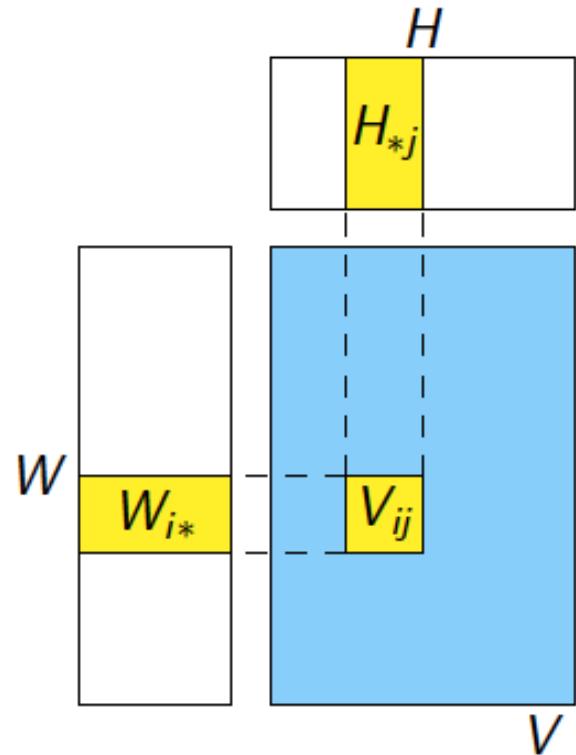
$$H_{*i} \in \mathbb{R}^r$$

- Rating prediction:

$$\begin{aligned} V_{ui} &= W_{u*} H_{*i} \\ &= [WH]_{ui} \end{aligned}$$



Figures from Koren et al. (2009)



Figures from Gemulla et al. (2011)₂₂

Matrix Factorization (with matrices)

- SGD (Stochastic Gradient Descent)

require that the loss can be written as

$$L = \sum_{(i,j) \in Z} l(\mathbf{V}_{ij}, \mathbf{W}_{i*}, \mathbf{H}_{*j})$$

Algorithm 1 SGD for Matrix Factorization

Require: A training set Z , initial values \mathbf{W}_0 and \mathbf{H}_0
while not converged **do** {step}

Select a training point $(i, j) \in Z$ uniformly at random.

$$\mathbf{W}'_{i*} \leftarrow \mathbf{W}_{i*} - \epsilon_n N \frac{\partial}{\partial \mathbf{W}_{i*}} l(\mathbf{V}_{ij}, \mathbf{W}_{i*}, \mathbf{H}_{*j})$$

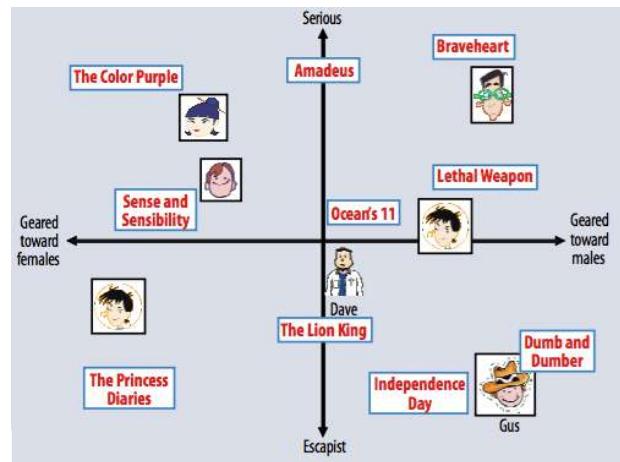
$$\mathbf{H}_{*j} \leftarrow \mathbf{H}_{*j} - \epsilon_n N \frac{\partial}{\partial \mathbf{H}_{*j}} l(\mathbf{V}_{ij}, \mathbf{W}_{i*}, \mathbf{H}_{*j})$$

$$\mathbf{W}_{i*} \leftarrow \mathbf{W}'_{i*}$$

end while

step size

Figure from Gemulla et al. (2011)



Figures from Koren et al. (2009)

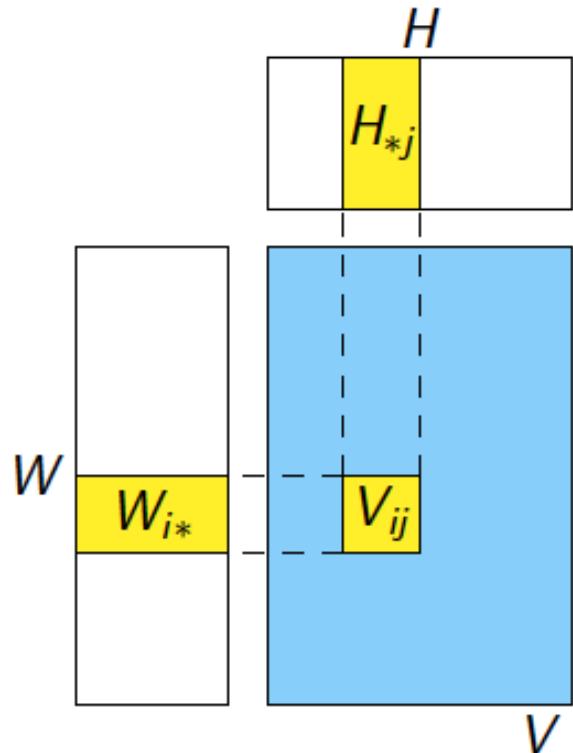


Figure from Gemulla et al. (2011)₂₃

Matrix Factorization

Example Factors

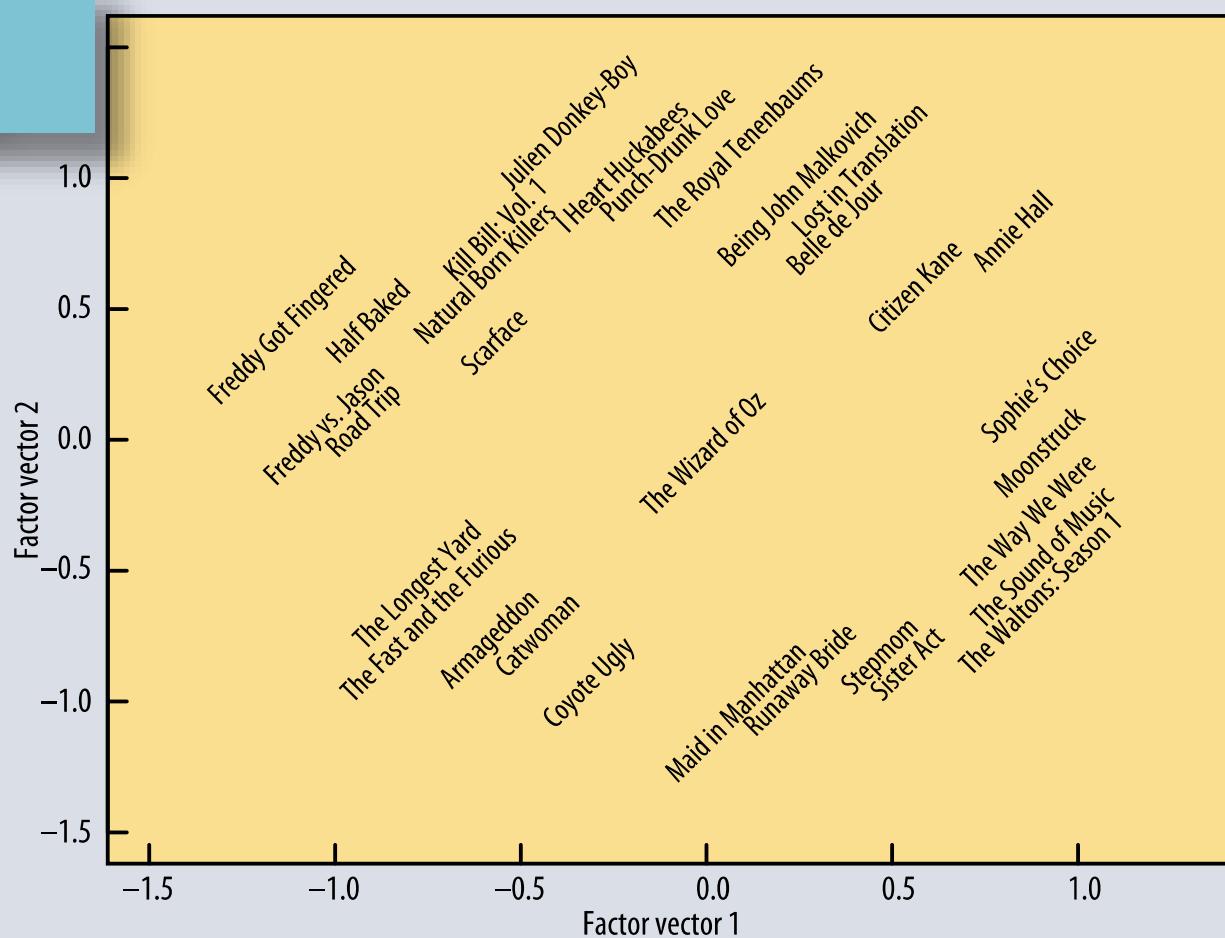
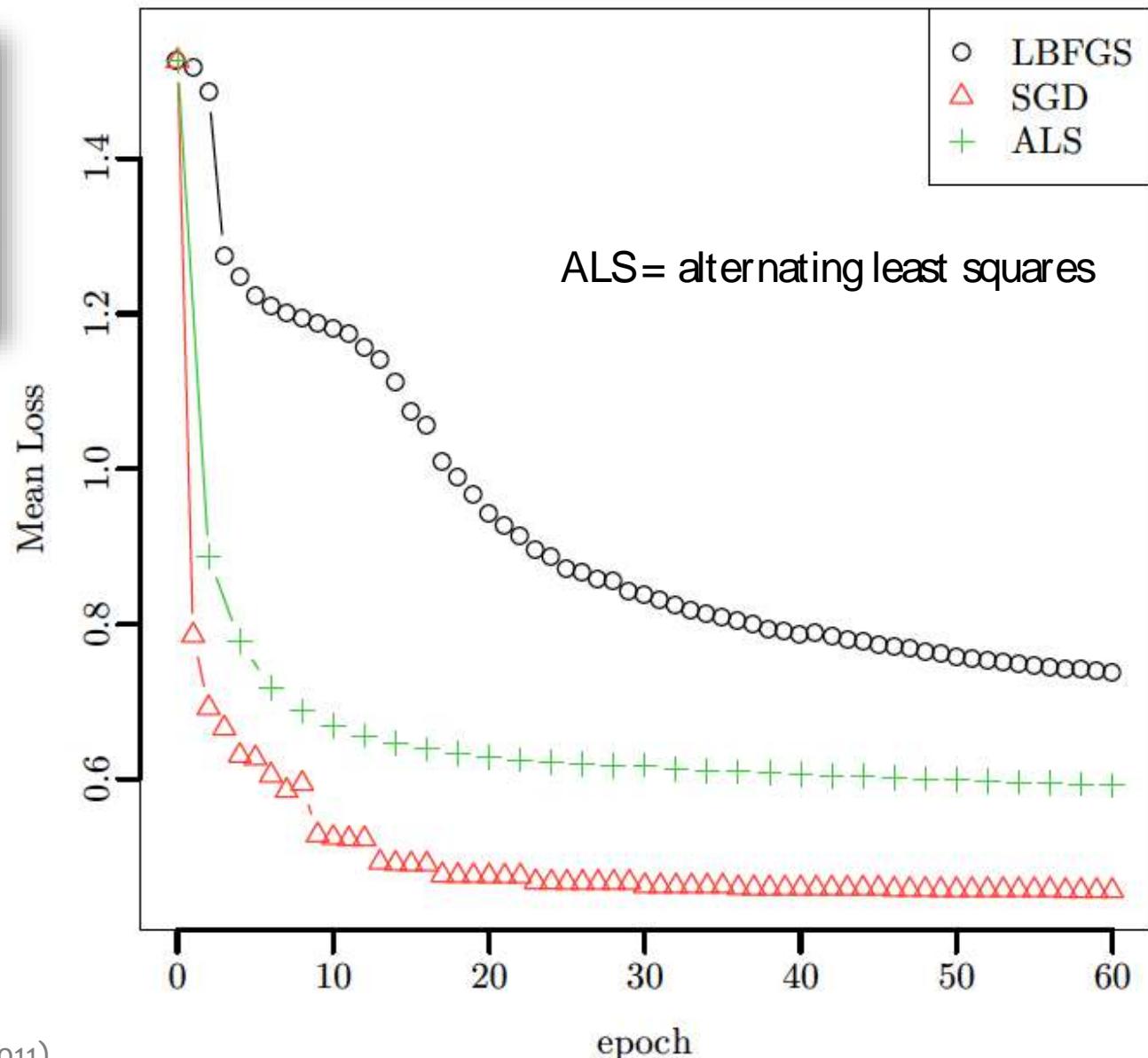


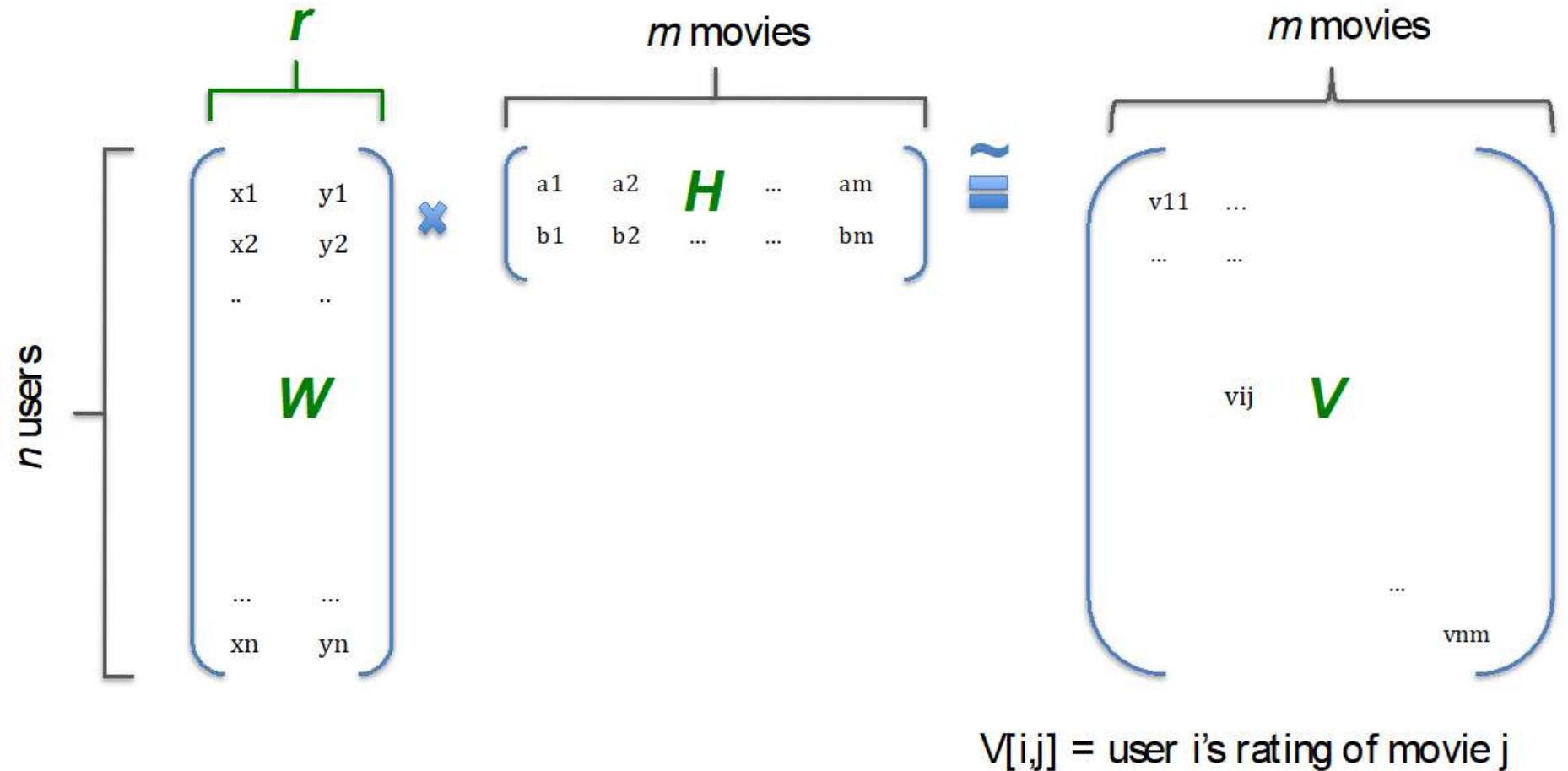
Figure 3. The first two vectors from a matrix decomposition of the Netflix Prize data. Selected movies are placed at the appropriate spot based on their factor vectors in two dimensions. The plot reveals distinct genres, including clusters of movies with strong female leads, fraternity humor, and quirky independent films.

Comparison of Optimization Algorithms

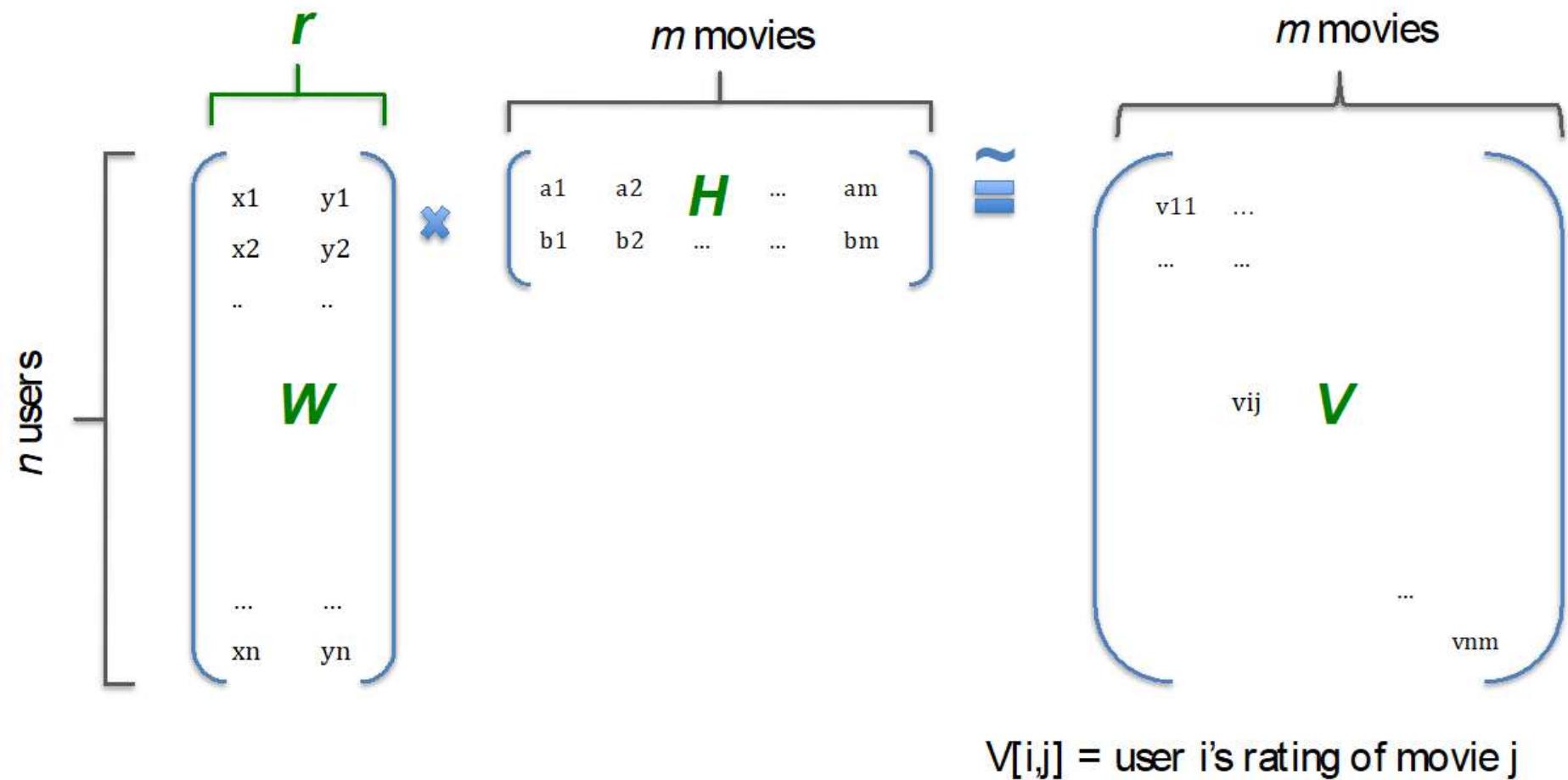
Matrix Factorization



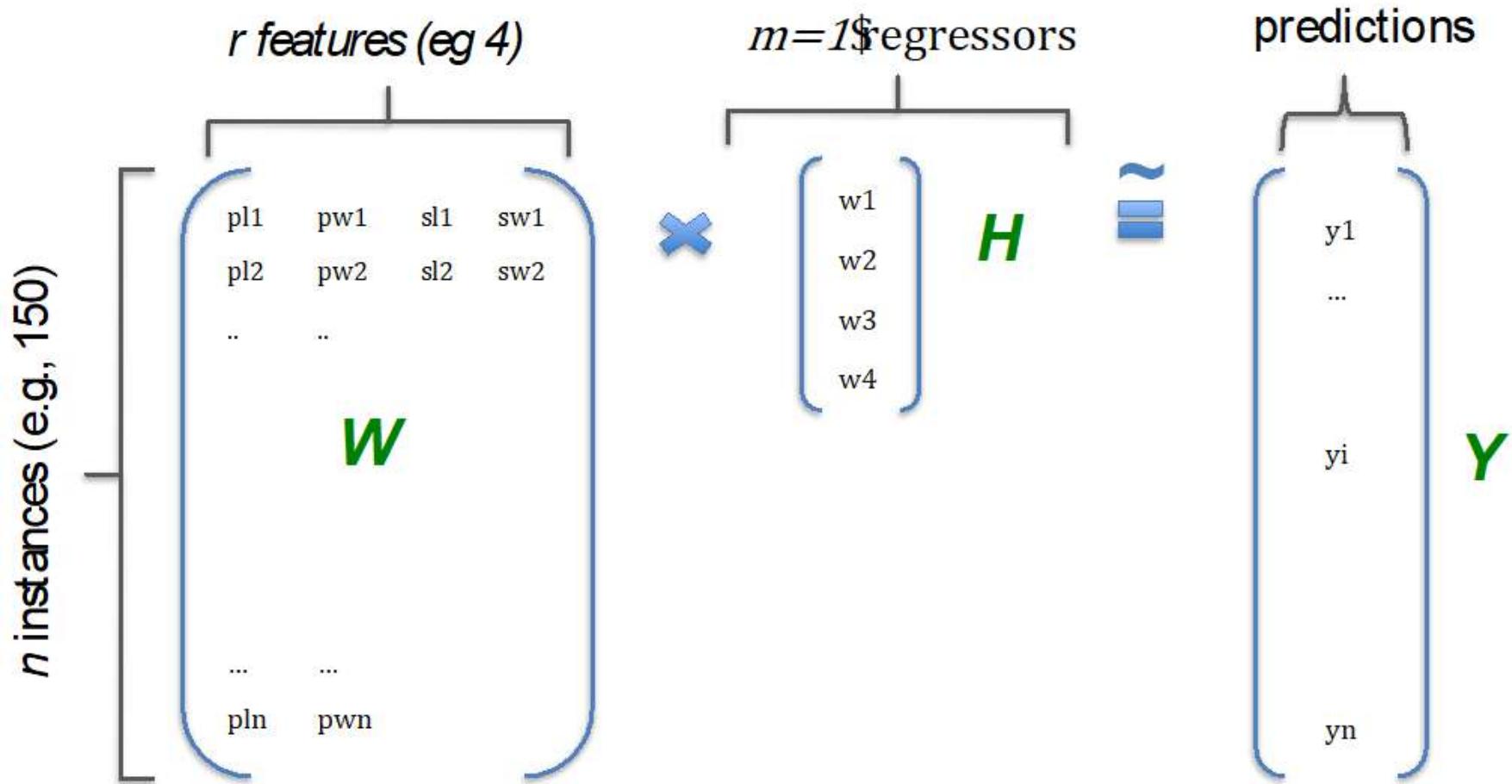
MATRIX MULTIPLICATION IN MACHINE LEARNING



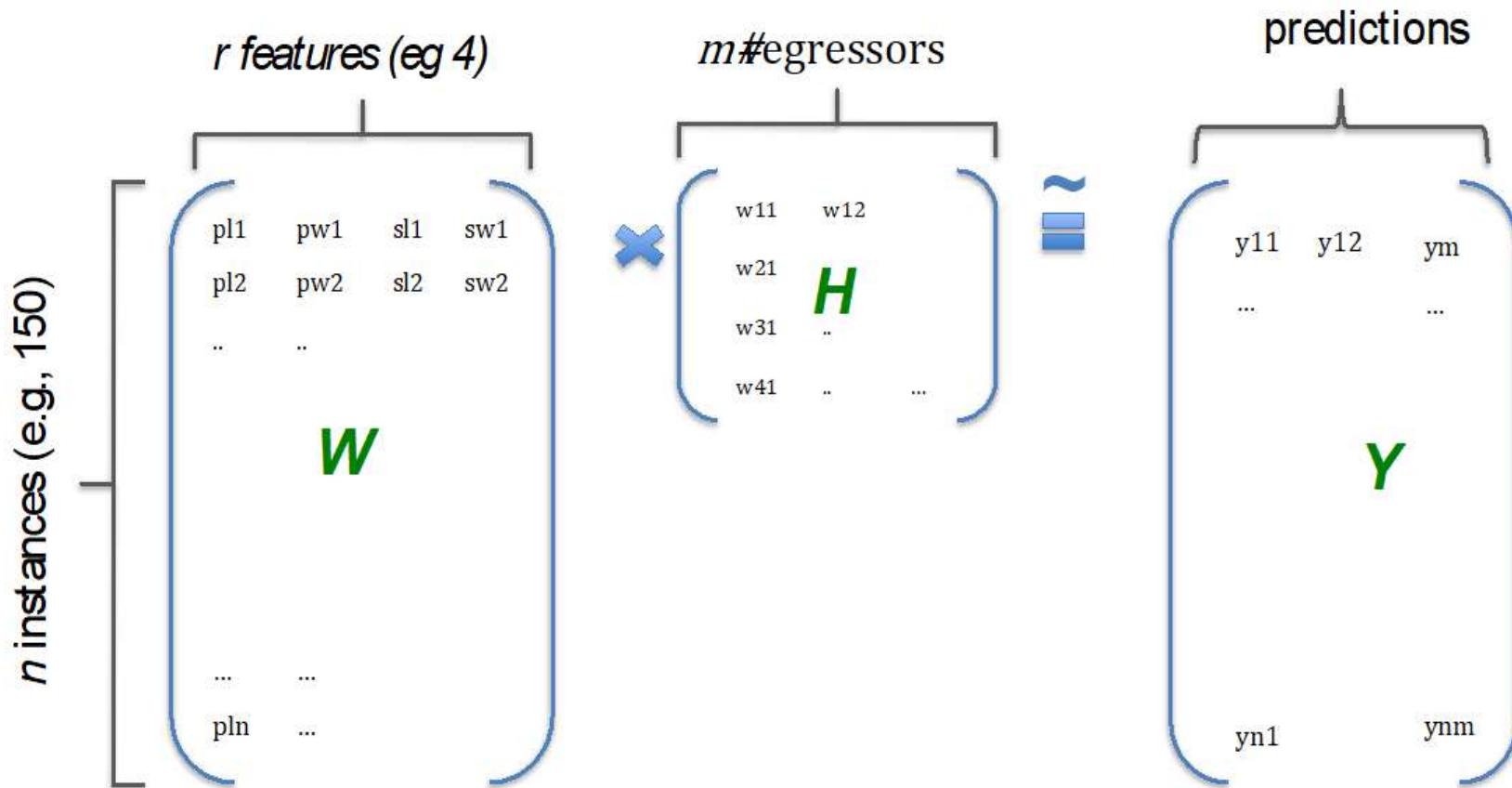
Recovering Latent Factors



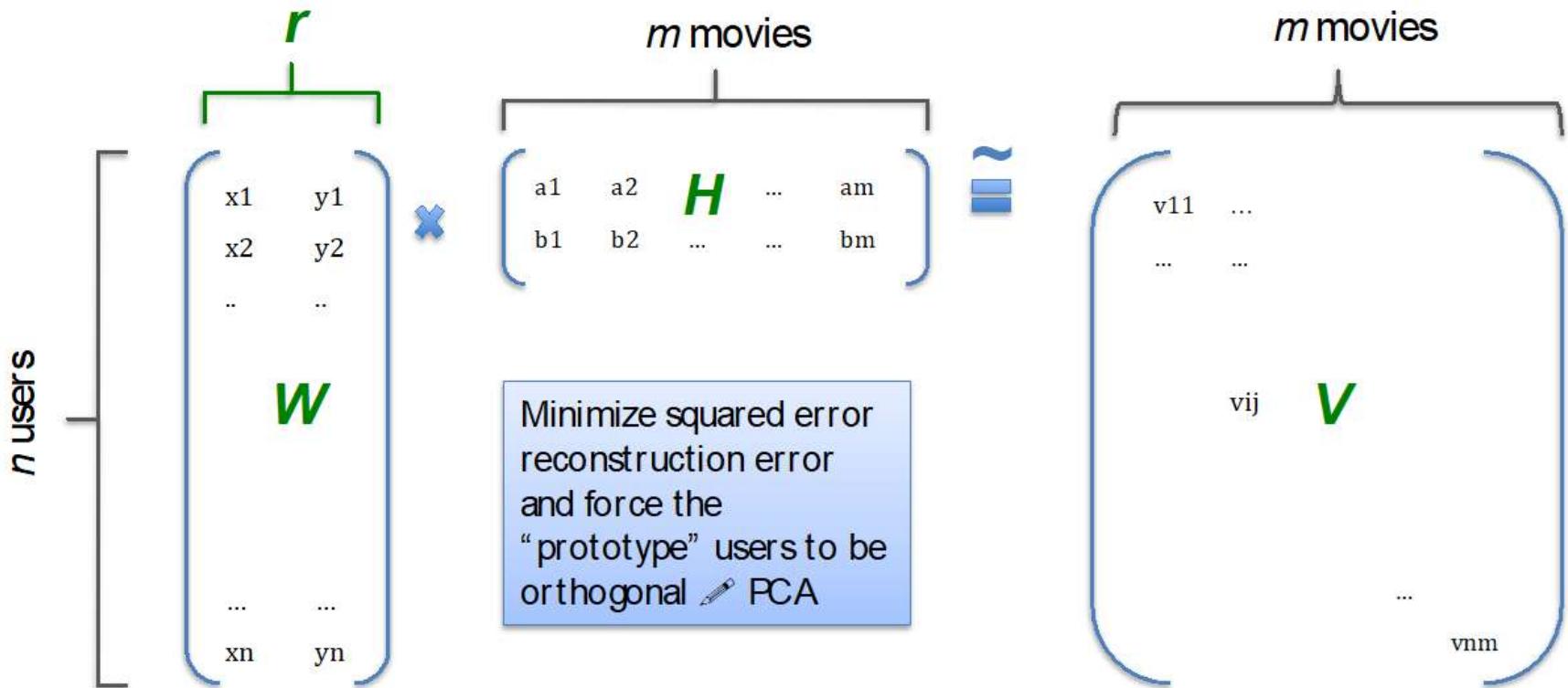
Is like Regression



Many Output at Once



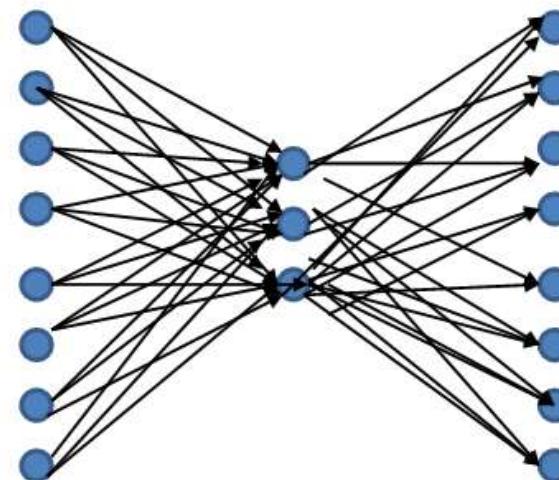
Similar like PCA



Auto-encoder and Non-Linear PCA

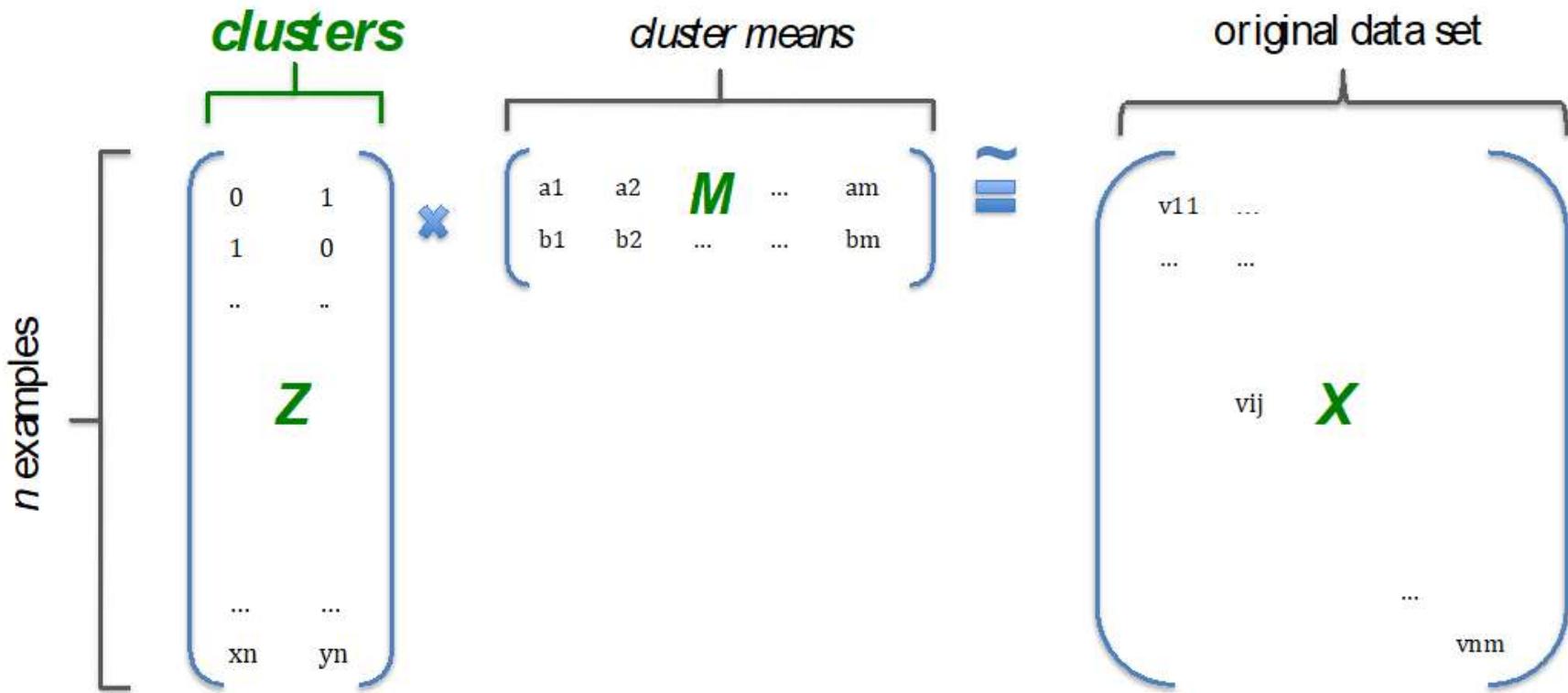
- Assume we would like to learn the following (trivial?) output function:
- Using the following network:
- With *linear* hidden units, how do the weights match up to W and H ?

Input	Output
00000001	00000001
00000010	00000010
00000101	00000100
00001000	00001000
00010000	00010000
00100000	00100000
01000000	01000000
10000000	10000000

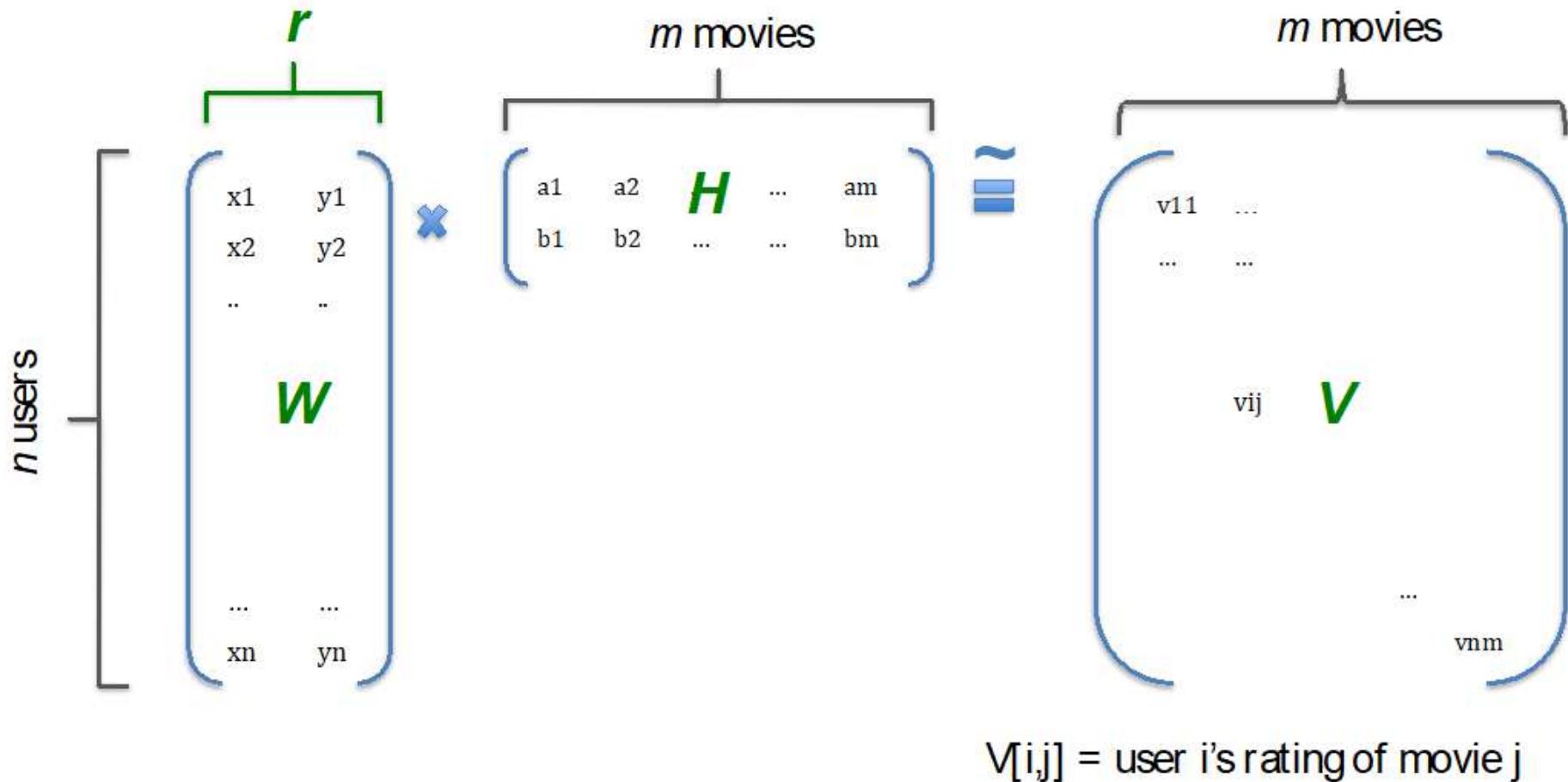


Like K-Means Clustering

indicators for r



Recovering Latent Factors in Matrix



Summary

- Recommender systems solve many **real-world (*large-scale) problems**
- Collaborative filtering by Matrix Factorization (MF) is an **efficient and effective** approach
- MF is just another example of a **common recipe**:
 1. define a model
 2. define an objective function
 3. optimize with SGD

Doing Good Data Science

The hard thing about being an ethical data scientist isn't understanding ethics. It's the junction between ethical ideas and practice. It's doing good data science.

There has been a lot of healthy discussion about data ethics lately. We want to be clear: that discussion is good, and necessary. But it's also not the biggest problem we face. We already have good standards for data ethics. The [ACM's code of ethics](#), which dates back to 1993, and is currently being updated, is clear, concise, and surprisingly forward-thinking; 25 years later, it's a great start for anyone thinking about ethics. The [American Statistical Association](#) has a good set of ethical guidelines for working with data. So, we're not working in a vacuum.

And we believe that most people want to be fair. Data scientists and software developers don't want to harm the people using their products. There are exceptions, of course; we call them criminals and con artists. [Defining "fairness" is difficult](#), and perhaps impossible, given the many crosscutting layers of "fairness" that we might be concerned with. But we don't have to solve that problem in advance, and it's not going to be solved in a simple statement of ethical principles, anyway.

The problem we face is different: how do we put ethical principles into practice? We're not talking about an abstract commitment to being fair. Ethical principles are worse than useless if we don't allow them to change our practice, if they don't have any effect on what we do day-to-day. For data scientists, whether you're doing classical data analysis or leading-edge AI, that's a big challenge. We need to understand how to build the software systems that implement fairness. That's what we mean by doing good data science.

Any code of data ethics will tell you that you shouldn't collect data from experimental subjects without informed consent. But that code won't tell you how to implement "informed consent." Informed consent is easy when you're interviewing a few dozen people in person for a psychology experiment. Informed consent means something different when someone clicks an item in an online catalog (hello, Amazon), and ads for that item start following them around *ad infinitum*. Do you use a pop-up to ask for permission to use their choice in targeted advertising? How many customers would you lose if you did so? Informed consent means something yet again when you're asking someone to fill out a profile for a social site, and you might (or might not) use that data for any number of experimental purposes. Do you pop up a consent form in impenetrable legalese that basically says "we will use your data, but we don't know for what"? Do you phrase this agreement as an opt-out, and hide it somewhere on the site where nobody will find it?

That's the sort of question we need to answer. And we need to find ways to share best practices. After the ethical principle, we have to think about the implementation of the ethical

principle. That isn't easy; it encompasses everything from user experience design to data management. How do we design the user experience so that our concern for fairness and ethics doesn't make an application unuseable? Bad as it might be to show users a pop-up with thousands of words of legalese, laboriously guiding users through careful and lengthy explanations isn't likely to meet with approval, either. How do we manage any sensitive data that we acquire? It's easy to say that applications shouldn't collect data about race, gender, disabilities, or other protected classes. But if you don't gather that data, you will have trouble testing whether your applications are fair to minorities. Machine learning has proven to be very good at figuring its own proxies for race and other classes. Your application wouldn't be the first system that was unfair despite the best intentions of its developers. Do you keep the data you need to test for fairness in a separate database, with separate access controls?

To put ethical principles into practice, we need space to be ethical. We need the ability to have conversations about what ethics means, what it will cost, and what solutions to implement. As technologists, we frequently share best practices at conferences, write blog posts, and develop open source technologies---but we rarely discuss problems such as how to obtain informed consent.

There are several facets to this space that we need to think about.

Foremost, we need corporate cultures in which discussions about fairness, about the proper use of data, and about the harm that can be done by inappropriate use of data can be considered. In turn, this means that we can't rush products out the door without thinking about how they're used. We can't allow "internet time" to mean ignoring the consequences. Computer security has shown us the consequences of ignoring the consequences: many companies that have never taken the time to implement good security practices and safeguards are now paying with damage to their reputations and their finances. We need to do the same when thinking about issues like fairness, accountability, and unintended consequences.

We particularly need to think about the unintended consequences of our use of data. It will never be possible to predict all the unintended consequences; we're only human, and our ability to foresee the future is limited. But plenty of unintended consequences could easily have been foreseen: for example, Facebook's "Year in Review" that [reminded people of deaths and other painful events](#). Moving fast and breaking things is unacceptable if we don't think about the things we are likely to break. And we need the space to do that thinking: space in project schedules, and space to tell management that a product needs to be rethought.

We also need space to stop the production line when something goes wrong. This idea goes back to Toyota's [Kanban](#): any assembly line worker can [stop the line](#) if they see something going wrong. The line doesn't restart until the problem is fixed. Workers don't have to fear

consequences from management for stopping the line; they are trusted, and expected to behave responsibly. What would it mean if we could do this with product features? If anyone at Facebook could have said "wait, we're getting complaints about Year in Review" and pulled it out of production until someone could investigate what was happening?

It's easy to imagine the screams from management. But it's not hard to imagine a Toyota-style "stop button" working. After all, Facebook is the poster child for continuous deployment, and they've often talked about how new employees push changes to production on their first day. Why not let employees pull features out of production? Where are the tools for instantaneous undeployment? They certainly exist; continuous deployment doesn't make sense if you can't roll back changes that didn't work. Yes, Facebook is a big, complicated company, with a big complicated product. So is Toyota. It worked for them.

The issue lurking behind all of these concerns is, of course, corporate culture. Corporate environments can be hostile to anything other than short-term profitability. That's a consequence of poor court decisions and economic doctrine, particularly in the US. But that inevitably leads us to the biggest issue: how to move the needle on corporate culture. Susan Etlinger has suggested that, in a time when public distrust and disenchantment is running high, [ethics is a good investment](#). Upper-level management is only starting to see this; changes to corporate culture won't happen quickly.

Users want to engage with companies and organizations they can trust not to take unfair advantage of them. Users want to deal with companies that will treat them and their data responsibly, not just as potential profit or engagement to be maximized. Those companies will be the ones that create space for ethics within their organizations. We, the data scientists, data engineers, AI and ML developers, and other data professionals, have to demand change. We can't leave it to people that "do" ethics. We can't expect management to hire trained ethicists and assign them to our teams. We need to live ethical values, not just talk about them. We need to think carefully about the consequences of our work. We must create space for ethics within our organizations. Cultural change may take time, but it will happen--if we are that change. That's what it means to do good data science.

Of Oaths and Checklists

"Oaths? We don't need no stinkin' oaths." (With apologies to Humphrey Bogart in *Treasure of the Sierra Madre*.)

Over the past year, there has been a great discussion of data ethics, motivated in part by discomfort over "fake news," targeted advertising, algorithmic bias, and the effect that data products have on individuals and on society. Concern about data ethics is hardly new; the [ACM](#), [IEEE](#), and the [American Statistical Association](#) all have ethical codes that address data. But the

intensity with which we've discussed ethics shows that something significant is happening: data science is coming of age and realizing its responsibilities. A better world won't come about simply because we use data; data has its dark underside.

The recent discussion frequently veers into a discussion of [data oaths](#), looking back to the ancient [Hippocratic Oath](#) for doctors. Much as we appreciate the work and the thought that goes into oaths, we are skeptical about their value. Oaths have several problems:

- They're one-shots. You take the oath once (if at all), and that's it. There's no reason to keep it in the front of your consciousness. You don't recite it each morning. Or evaluate regularly whether you're living up to the ideals.
- Oaths are a set of very general and broad principles. Discussions of the Hippocratic Oath begin with the phrase "First, do no harm," words that don't actually appear in the oath. But what does "do no harm" mean? For centuries doctors did very little but harm (many people died because doctors didn't believe they needed to wash their hands). The doctors just didn't know they were doing harm. Nice idea, but short on the execution. And data science (like medicine) is all about execution.
- Oaths can actually give cover to people and organizations who are doing unethical work. It's easy to think "we can't be unethical, because we endorsed this oath." It's not enough to say "don't be evil." You have to not be evil.
- Oaths do very little to connect theories and principles to practice. It is one thing to say "researchers must obtain informed consent"; it's an entirely different thing to get informed consent at internet scale. Or to teach users what "informed consent" means.

We are not suggesting that the principles embodied in oaths aren't important, just that they don't get us to the endpoint we want. They don't connect our ideas about what's good or just to the practices that create goodness and justice. We can talk a lot about the importance of being fair and unbiased without knowing about how to be fair and unbiased. At this point, the oath actually becomes dangerous: it becomes a tool to convince yourself that you're one of the good guys, that you're doing the right thing, when you really don't know.

Oaths are good at creating discussion---and, in the past year, they have created quite a lot of discussion. The discussion has been tremendously helpful in making people aware of issues like algorithmic fairness. The discussion has helped software developers and data scientists to understand that their work isn't value-neutral, that their work has real impact, both good and bad, on real people. And there has been a vigorous debate about what self-government means for data scientists, and what guiding principles would last longer than a few years. But we need to take the next step, and connect these ideas to practice. How will we do that?

In 2009, Atul Gawande wrote *The Checklist Manifesto* (Macmillan), a short book on how not to make big mistakes. He writes a lot about his practice as a surgeon. In a hospital, everyone knows what to do. Everyone knows that you're supposed to scrub down before the surgery. Everyone knows that you're not supposed to amputate the wrong leg. Everyone knows that you're not supposed to leave sponges and other equipment in patients when you close the incision.

But mistakes are made, particularly when people are in stressful environments. The surgeon operates on the wrong leg; the sponge is left behind; and so on. Gawande found that, simply by creating checklists for basic things you shouldn't forget, these mistakes could be eliminated almost completely. Yes, there were some doctors who found the idea of checklists insultingly simple; they were the ones who continued making mistakes.

Unlike oaths, checklists connect principle to practice. Everyone knows to scrub down before the operation. That's the principle. But if you have to check a box on a form after you've done it, you're not likely to forget. That's the practice. And checklists aren't one-shots. A checklist isn't something you read once at some initiation ceremony; a checklist is something you work through with every procedure.

What would a checklist for data science and machine learning look like? The [UK Government's Data Ethics Framework](#) and [Data Ethics Workbook](#) is one approach. They isolate seven principles, and link to detailed discussions of each. The workbook asks a number of open-ended questions to probe your compliance with these principles. Our criticism is that their process imposes a lot of overhead. While anyone going through their entire process will certainly have thought carefully about ethical issues, in practice, asking developers to fill out a workbook with substantive answers to 46 questions is an effective way to ensure that ethical thought doesn't happen.

We believe that checklists are built around simple, "have we done this?" questions---and they are effective because they are simple. They don't leave much room to wiggle. Either you've analyzed how a project can be abused, or you haven't. You've built a mechanism for gathering consent, or you haven't. Granted, it's still possible to take shortcuts: your analysis might be inadequate and your consent mechanism might be flawed, but you've at least gone on record for saying that you've done it.

Feel free to use and modify this checklist in your projects. It covers most of the bases that we've seen discussed in various data oaths. Go over the checklist when starting a project so the developers know what's needed and aren't surprised by a new set of requirements at the last minute. Then work through it whenever you release software. Go through it, and actually check off all the boxes before your product hits the public.

Here's a checklist for people who are working on data projects:

- Have we listed how this technology can be attacked or abused?
- Have we tested our training data to ensure it is fair and representative?
- Have we studied and understood possible sources of bias in our data?
- Does our team reflect diversity of opinions, backgrounds, and kinds of thought?
- What kind of user consent do we need to collect to use the data?
- Do we have a mechanism for gathering consent from users?
- Have we explained clearly what users are consenting to?
- Do we have a mechanism for redress if people are harmed by the results?
- Can we shut down this software in production if it is behaving badly?
- Have we tested for fairness with respect to different user groups?
- Have we tested for disparate error rates among different user groups?
- Do we test and monitor for model drift to ensure our software remains fair over time?
- Do we have a plan to protect and secure user data?

Oaths and codes of conduct have their value. The value of an oath isn't the pledge itself, but the process you go through in developing the oath. People who work with data are now having discussions that would never have taken place a decade ago. But discussions don't get the hard work done, and we need to get down to the hard work. We don't want to talk about how to use data ethically; we want to use data ethically. It's hypocritical to talk about ethics, but never do anything about it. We want to put our principles into practice. And that's what checklists will help us do.

The Five Cs

What does it take to build a good data product or service? Not just a product or service that's useful, or one that's commercially viable, but one that uses data ethically and responsibly.

We often talk about a product's technology or its user experience, but we rarely talk about how to build a data product in a responsible way that puts the user in the center of the conversation. Those products are badly needed. News that people "don't trust" the data products they use---or that use them---is common. While Facebook has received the most coverage, lack of trust isn't limited to a single platform. Lack of trust extends to nearly every consumer internet

company, to large traditional retailers, and to data collectors and brokers in industry and government.

Users lose trust because they feel abused by malicious ads; they feel abused by fake and misleading content, and they feel abused by "act first, and apologize profusely later" cultures at many of the major online companies. And users ought to feel abused by many abuses they don't even know about. Why was their insurance claim denied? Why weren't they approved for that loan? Were those decisions made by a system that was trained on biased data? The slogan goes, "Move fast and break things." But what if society is broken?

Data collection is a big business. Data is valuable: "the new oil," as the [Economist proclaimed](#). We've known that for some time. But the public provides the data under the assumption that we, the public, benefit from it. We also assume that data is collected and stored responsibly, and those who supply the data won't be harmed. Essentially it's a model of trust. But how do you restore trust once it's been broken? It's no use pretending that you're trustworthy when your actions have proven that you aren't. The only way to get trust back is to be trustworthy, and regaining that trust once you've lost it takes time.

There's no simple way to regain users' trust, but we'd like to suggest a "golden rule" for data as a starting point: "treat others' data as you would have others treat your own data." However, implementing a golden rule in the actual research and development process is challenging---just as it's hard to get from short, pithy oaths and pledges to actual practice.

What does it mean to treat others' data as you would treat your own? How many data scientists have actually thought about how their own data might be used and abused? And once you know how you'd like to see your data (and others' data) respected, how do you implement those ideas? The golden rule isn't enough by itself. We need guidelines to force discussions with the application development teams, application users, and those who might be harmed by the collection and use of data.

Five framing guidelines help us think about building data products. We call them the five Cs: consent, clarity, consistency, control (and transparency), and consequences (and harm). They're a framework for implementing the golden rule for data. Let's look at them one at a time.

Consent

You can't establish trust between the people who are providing data and the people who are using it without agreement about what data is being collected and how that data will be used. Agreement starts with obtaining consent to collect and use data. Unfortunately, the agreements between a service's users (people whose data is collected) and the service itself (which uses the data in many ways) are binary (meaning that you either accept or decline) and lack clarity. In business, when contracts are being negotiated between two parties, there are multiple

iterations (redlines) before the contract is settled. But when a user is agreeing to a contract with a data service, they either accept the terms or they don't get access. It's nonnegotiable.

For example, when you check into a hospital you are required to sign a form that gives them the right to use your data. Generally, there's no way to say that your data can be used for some purposes but not others. When you sign up for a loyalty card at your local pharmacy, you're agreeing that they can use your data in unspecified ways. Those ways certainly include targeted advertising (often phrased as "special offers"), but may also include selling your data (with or without anonymization) to other parties. And what happens to your data when one company buys another and uses data in ways that you didn't expect?

Data is frequently collected, used, and sold without consent. This includes organizations like Acxiom, Equifax, Experian, and Transunion, that collect data to assess financial risk, but many common brands also collect data without consent. In Europe, [Google collected data](#) from cameras mounted on cars to develop new mapping products. [AT&T and Comcast](#) both used cable set top boxes to collect data about their users, and [Samsung](#) collected voice recordings from TVs that respond to voice commands. There are many, many more examples of nonconsensual data collection. At every step of building a data product, it is essential to ask whether appropriate and necessary consent has been provided.

Clarity

Clarity is closely related to consent. You can't really consent to anything unless you're told clearly what you're consenting to. Users must have clarity about what data they are providing, what is going to be done with the data, and any downstream consequences of how their data is used. All too often, explanations of what data is collected or being sold are buried in lengthy legal documents that are rarely read carefully, if at all. Observant readers of Eventbrite's user agreement [recently discovered that listing an event gave the company the right to send a video team, and exclusive copyright to the recordings](#). And the only way to opt out was by writing to the company. The backlash was swift once people realized the potential impact, and Eventbrite removed the language.

Facebook users who played Cambridge Analytica's "This Is Your Digital Life" game may have understood that they were giving up their data; after all, they were answering questions, and those answers certainly went somewhere. But did they understand how that data might be used? Or that they were giving access to their friends' data behind the scenes? That's buried deep in Facebook's privacy settings.

Even when it seems obvious that their data is in a public forum, users frequently don't understand how that data could be used. Most Twitter users know that their public tweets are, in fact, public; but [many don't understand that their tweets can be collected and used for](#)

[research](#), or even that they are [for sale](#). This isn't to say that such usage is unethical; but as [Casey Fiesler](#) points out, the need isn't just to get consent, but to inform users what they're consenting to. That's clarity.

It really doesn't matter which service you use; you rarely get a simple explanation of what the service is doing with your data, and what consequences their actions might have. Unfortunately, the process of consent is often used to obfuscate the details and implications of what users may be agreeing to. And once data has escaped, there is no recourse. You can't take it back. Even if an organization is willing to delete the data, it's very difficult to prove that it has been deleted.

There are some notable exceptions: people like John Wilbanks are [working](#) to develop models that help users to understand the implications of their choices. Wilbanks' work helps people understand what happens when they provide [sensitive medical and health data to a service](#).

Consistency and Trust

Trust requires consistency over time. You can't trust someone who is unpredictable. They may have the best intentions, but they may not honor those intentions when you need them to. Or they may interpret their intentions in a strange and unpredictable way. And once broken, rebuilding trust may take a long time. Restoring trust requires a prolonged period of consistent behavior.

Consistency, and therefore trust, can be broken either explicitly or implicitly. An organization that exposes user data can do so intentionally or unintentionally. In the past years, we've seen many security incidents in which customer data was stolen: Yahoo!, Target, Anthem, local hospitals, government data, data brokers like Experian, and the list grows longer each day. Failing to safeguard customer data breaks trust---and safeguarding data means nothing if not consistency over time.

We've also seen frustration, anger, and surprise when users don't realize what they've agreed to. When Cambridge Analytica used Facebook's data to target vulnerable customers with highly specific advertisements, Facebook initially claimed that this was not a data breach. And while Facebook was technically correct, in that data was not stolen by an intruder, the public's perception was clearly different. This was a breach of trust, if not a breach of Facebook's perimeter. Facebook didn't consistently enforce its agreement with its customers. When the news broke, Facebook became unpredictable because most of its users had no idea what it would or wouldn't do. They didn't understand their user agreements, they didn't understand their complex privacy settings, and they didn't understand how Facebook would interpret those settings.

Control and Transparency

Once you have given your data to a service, you must be able to understand what is happening to your data. Can you control how the service uses your data? For example, Facebook asks for (but doesn't require) your political views, religious views, and gender preference. What happens if you change your mind about the data you've provided? If you decide you're rather keep your political affiliation quiet, do you know whether Facebook actually deletes that information? Do you know whether Facebook continues to use that information in ad placement?

All too often, users have no effective control over how their data is used. They are given all-or-nothing choices, or a convoluted set of options that make controlling access overwhelming and confusing. It's often impossible to reduce the amount of data collected, or to have data deleted later.

A major part of the shift in data privacy rights is moving to give users greater control of their data. For example, Europe's [General Data Protection Regulation](#) (GDPR) requires users' data to be provided to them at their request and removed from the system if they so desire.

Consequences

Data products are designed to add value for a particular user or system. As these products increase in sophistication, and have broader societal implications, it is essential to ask whether the data that is being collected could cause harm to an individual or a group. We continue to hear about unforeseen consequences and the "unknown unknowns" about using data and combining data sets. Risks can never be eliminated completely. However, many unforeseen consequences and unknown unknowns could be foreseen and known, if only people had tried. All too often, unknown unknowns are unknown because we don't want to know.

Due to potential issues around the use of data, laws and policies have been put in place to protect specific groups: for example, the [Children's Online Privacy Protection Act](#) (COPPA) protects children and their data. Likewise, there are laws to protect specific sensitive data sets: for example, the [Genetic Information Nondiscrimination Act](#) (GINA) was established in 2008 in response to rising fears that genetic testing could be used against a person or their family. Unfortunately, policy doesn't keep up with technology advances; neither of these laws have been updated. Given how rapidly technology is being adopted by society, the Obama administration realized that the pace of the regulatory process couldn't keep up. As a result, it created the roles of the US chief technology officer and chief data scientist. The Obama administration also established more than 40 chief data officers and scientists across the federal government. The result has been to make sure the regulatory process fosters innovation while ensuring the question of potential of harm is asked regularly and often.

Even philanthropic approaches can have unintended and harmful consequences. When, in 2006, AOL released anonymized search data to researchers, it proved possible to "de-

anonymize" the data and identify specific users. In 2018, [Strava opened up their data](#) to allow users to discover new places to run or bike. Strava didn't realize that members of the US military were using GPS-enabled wearables, and their activity exposed the locations of bases and patrol routes in Iraq and Afghanistan. Exposure became apparent after the product was released to the public, and people exploring the data started talking about their concerns.

While Strava and AOL triggered a chain of unforeseen consequences by releasing their data, it's important to understand that their data had the potential to be dangerous even if it wasn't released publicly. Collecting data that may seem innocuous and combining it with other data sets has real-world implications. Combining data sets frequently gives results that are much more powerful and dangerous than anything you might get from either data set on its own. For example, data about running routes could be combined with data from smart locks, telling thieves when a house or apartment was unoccupied, and for how long. The data could be stolen by an attacker, and the company wouldn't even recognize the damage.

It's easy to argue that Strava shouldn't have produced this product, or that AOL shouldn't have released their search data, but that ignores the data's potential for good. In both cases, well-intentioned data scientists were looking to help others. The problem is that they didn't think through the consequences and the potential risks.

It is possible to provide data for research without unintended side-effects. For example, the US Internal Revenue Service (IRS), in collaboration with researchers, opened a similar data set in a [tightly controlled manner](#) to help understand economic inequality. There were no negative repercussions or [major policy implications](#). Similarly the Department of Transportation releases data about [traffic fatalities](#). The [UK Biobank](#) (one of the largest collections of genomic data) has a sophisticated approach to opening up different levels of data. Other companies have successfully [opened up data for the public benefit](#), including [LinkedIn's Economic Graph project](#) and [Google Books' ngram viewer](#).

Many data sets that could provide tremendous benefits remain locked up on servers. Medical data that is fragmented across multiple institutions limits the pace of [research](#). And the data held on traffic from ride-sharing and GPS/mapping companies could transform approaches for traffic safety and congestion. But opening up that data to researchers requires careful planning.

Implementing the Five Cs

Data can improve our lives in many ways, from the mundane to the amazing. Good movie recommendations aren't a bad thing; if we could consolidate medical data from patients around the world, we could make some significant progress on treating diseases like cancer. But we won't get either better movie recommendations or better cancer treatments if we can't ensure

that the five Cs are implemented effectively. We won't get either if we can't treat others' data as carefully as we'd treat our own.

Over the past decade, the software industry has put significant effort into improving user experience (UX). Much of this investment has been in user-centric approaches to building products and services that depend on the data the collective user base provides. All this work has produced results: using software is, on the whole, easier and more enjoyable.

Unfortunately, these teams have either intentionally or unintentionally limited their efforts to providing users with immediate gratification or the ability to accomplish near-term goals.

"Growth hacking" focuses on getting people to sign up for services through viral mechanisms. We've seen few product teams that try to develop a user experience that balances immediate experience with long-term values.

In short, product teams haven't considered the impacts of the five Cs. For example, how should an application inform users about how their data will be used, and get their consent? That part of user experience can't be swept under the rug. And it can't mean making it easy for users to give consent, and difficult to say "no." It's all part of the total user experience. Users need to understand what they are consenting to and what effects that consent might have; if they don't, the designer's job isn't done.

Responsibility for the five Cs can't be limited to the designers. It's the responsibility of the entire team. The data scientists need to approach the problem asking "what if" scenarios that get to all of the five C's. The same is true for the product managers, business leaders, sales, marketing, and also executives.

The five Cs need to be part of every organization's culture. Product and design reviews should go over the five Cs regularly. They should consider developing a checklist before releasing a product to the public. All too often, we think of data products as minimal viable products (MVPs: prototypes to test whether the product has value to users). While that's a constructive approach for developing and testing new ideas, even MVPs must address the five Cs. The same is true for well-established products. New techniques may have been developed that could result in harm in unforeseen ways. In short, it's about taking responsibility for the products that are built. The five Cs are a mechanism to foster dialogue to ensure the products "do no harm."

Data's Day of Reckoning

Our lives are bathed in data: from recommendations about whom to "follow" or "friend" to data-driven autonomous vehicles. But in the past few years, it has become clear that the products and technologies we have created have been weaponized and used against us. Although we've benefited from the use of data in countless ways, it has also created a tension between individual privacy, public good, and corporate profits. Cathy O'Neil's *Weapons of Math*

Destruction (Broadway Books) and Virginia Eubanks' *Automating Inequality* (Macmillan) document the many ways that data has been used to harm the broader population.

Data science, machine learning, artificial intelligence, and related technologies are now facing a day of reckoning. It is time for us to take responsibility for our creations. What does it mean to take responsibility for building, maintaining, and managing data, technologies, and services? Responsibility is inevitably tangled with the complex incentives that surround the creation of any product. These incentives have been front and center in the conversations around the roles that social networks have played in the 2016 US elections, recruitment of terrorists, and online harassment. It has become very clear that the incentives of the organizations that build and own data products haven't aligned with the good of the people using those products.

These issues aren't new to the consumer internet. Other fields have had their days of reckoning. In medicine, widely publicized abuses include the [Tuskegee syphilis experiment](#), the case of [Henrietta Lacks](#) (whose cells were used for cancer research without her permission and without compensation), and human experiments performed during World War II by Nazis. The physics community had to grapple with the implications of the atomic bomb. Chemists and biologists have had to address the use of their research for chemical and biological weapons. Other engineering disciplines have realized that shoddy work has an impact on people's lives; [it's hard to ignore bridge collapses](#). As a result, professional societies were formed to maintain and enforce codes of conduct; government regulatory processes have established standards and penalties for work that is detrimental to society.

Ethics and Security Training

In many fields, ethics is an essential part of professional education. This isn't true in computer science, data science, artificial intelligence, or any related field. While courses on ethics exist at many schools, the ideas taught in ethics classes often aren't connected to existing projects or course work. Students may study ethical principles, but they don't learn how to implement those principles in their projects. As a result, they are ill-prepared for the challenges of the real world. They're not trained to think about ethical issues and how they affect design choices. They don't know how to have discussions about projects or technologies that may cause real-world harm.

Software security and ethics frequently go hand in hand, and our current practices for teaching security provide an example of what not to do. Security is usually taught as an elective, isolated from other classes about software development. For example, a class on databases may never discuss [SQL injection attacks](#). SQL injection would be addressed in classes on security, but not in the required database course. When a student submits a project in a database course, its vulnerability to hostile attack doesn't affect the grade; an automated grading system won't even

test it for vulnerabilities. Furthermore, a database course might not discuss architectural decisions that limit damage if an attacker gains access---for example, storing data elements such as names and Social Security numbers in different databases.

Teaching security in an elective is better than not teaching it at all; but the best way to produce programmers who really understand security is to incorporate it into assignments and grading within the core curriculum, in addition to teaching it in electives. The core curriculum ensures that everyone can recognize and deal with basic security problems; the electives can go into greater depth, and tie together issues from different disciplines ranging from physical security to cryptography. Security as an afterthought doesn't work in product development; why do we expect it to work in education? There is no industry in which security lapses haven't led to stolen data, affecting millions of individuals. Poor security practices have led to [serious vulnerabilities](#) in many consumer devices, from smart locks to smart light bulbs.

Ethics faces the same problem. Data ethics is taught at [many colleges and universities](#), but it's isolated from the rest of the curriculum. Courses in ethics help students think seriously about issues, but can't address questions like getting informed consent in the context of a real-world application. The White House report "[Preparing for the Future of Artificial Intelligence](#)" highlights the need for training in both ethics and security:

Ethical training for AI practitioners and students is a necessary part of the solution. Ideally, every student learning AI, computer science, or data science would be exposed to curriculum and discussion on related ethics and security topics. However, ethics alone is not sufficient. Ethics can help practitioners understand their responsibilities to all stakeholders, but ethical training should be augmented with technical tools and methods for putting good intentions into practice by doing the technical work needed to prevent unacceptable outcomes.

Ethics and security must be at the heart of the curriculum, not only as electives, or even isolated requirements. They must be integrated into every course at colleges, universities, online courses, and programming boot camps. They can't remain abstract, but need to be coupled with "technical tools and methods for putting good intentions into practice." And training can't stop upon graduation. Employers need to host regular forums and offer refresher courses to keep people up-to-date on the latest challenges and perspectives.

Developing Guiding Principles

The problem with ethical principles is that it's easy to forget about them when you're rushing: when you're trying to get a project finished on a tight, perhaps unrealistic, schedule. When the clock is ticking away toward a deadline, it's all too easy to forget everything you learned in class--even if that class connected ethics with solutions to real-world problems.

Checklists are a proven way to solve this problem. A checklist, as described by Atul Gawande in *The Checklist Manifesto* (Metropolitan Books), becomes part of the ritual. It's a short set of questions that you ask at the start of the project, and at every stage as you move toward release. You don't go to the next stage until you've answered all the questions affirmatively. Checklists have been shown to reduce mistakes in surgery; they're used very heavily by airline pilots, especially in emergencies; and they can help data professionals to not forget ethical issues, even when they are under pressure to deliver.

In Chapter 2, we proposed a checklist for developers working on data-driven applications. Feel free to use and to modify this checklist to fit your situation and use it in your projects. Our checklist doesn't reflect all the issues that you should be considering, and certainly doesn't reflect all the applications that people are currently developing, let alone in the future. If you add to it, though, try to keep the additions short; that's why checklists work.

The [Fairness, Accountability, and Transparency in Machine Learning](#) group (FAT/ML) advocates a similar approach. Their [Principles for Accountable Algorithms and a Social Impact Statement for Algorithms](#) suggests assessing the social impact statement of a project at least three times during its life: during design, pre-launch, and post-launch. Working through a social impact statement requires developers to think about the ethical consequences of their projects and address any problems that turn up. In a similar vein, the [Community Principles on Ethical Data Practices](#), which arose out of the [Data for Good Exchange \(D4GX\)](#), provides a set of values and principles that have been gathered through community discussion. They're a great start for any group that wants to create its own checklist. And Cathy O'Neil has proposed [auditing](#) machine learning algorithms for fairness.

Building Ethics into a Data-Driven Culture

Individual responsibility isn't sufficient. Ethics needs to be part of an organization's culture. We've seen many organizations recognize the value of developing a data-driven culture; we need to ensure ethics and security become part of that culture, too.

Security is gradually becoming a part of corporate culture: the [professional, financial, legal, and reputational consequences](#) of being a victim are too large to ignore. Organizations are experimenting with bug-bounty programs, sharing threats with each other, and collaborating with government agencies. Security teams are no longer simply corporate naysayers; they're charged with preventing serious damage to an organization's reputation and to finances.

Integrating ethics into corporate culture has been more challenging. A single team member may object to an approach, but it's easy for an individual to be overruled, and if there's no support for ethical thinking within the organization, that's likely to be where it ends. Ethical thinking is

important with or without corporate support, but it's more likely to make a difference when ethical action is a corporate value. Here are some ideas for building ethics into culture:

An individual needs to be empowered to stop the process before damage is done.

Toyota and [W. Edwards Deming](#) pioneered the use of the [andon cord](#) to improve quality and efficiency. Anyone who saw a problem could pull the cord, which would halt the production line. Senior managers as well as production line operators would then discuss the issue, make improvements, and restart the process.

Any member of a data team should be able to pull a virtual "andon cord," stopping production, whenever they see an issue. The product or feature stays offline until the team has a resolution. This way, an iterative process can be developed that avoids glossing over issues.

Anyone should be able to escalate issues for remediation without fear of retaliation.

There needs to be an escalation process for team members who don't feel their voice has been heard. The US Department of State has a [dissent channel](#) where any diplomat can make sure the Secretary of State hears their concerns. In health care, a path to escalate legal and ethical issues is required by law. For health care plans in the US, there is a compliance officer who reports directly to the board of directors.

Data-driven organizations need a similar model that allows people to escalate issues without the fear of reprisal. An escalation process could be implemented in several forms. For example, companies could work with an organization such as the Electronic Frontier Foundation (EFF) to develop a program that accepts and investigates whistleblower reports. The problem would be kept from public scrutiny unless specific criteria are violated. A similar approach could be implemented under an existing or new agency (e.g., a Consumer Data Protection Agency).

An ethical challenge should be part of the hiring process.

When hiring, companies frequently assess whether a candidate will be a "cultural fit." Interviewers ask questions that help them understand whether a candidate will work well with other team members. However, interviewers rarely ask questions about the candidate's ethical values.

Rather than asking a question with a right/wrong answer, we've found that it's best to pose a problem that lets us see how the candidate thinks about ethical and security choices. Here's a question we have used:

Assume we have a large set of demographic data. We're trying to evaluate individuals and we're not supposed to use race as an input. However, you discover a proxy for race with the other variables. What would you do?

This kind of question can start a dialogue about how to use the proxy variable. What effects does it have on people using the product? Are we making recommendations, or deciding whether to provide services? Are we implementing a legal requirement, or providing guidance about compliance? Discussing the question and possible answers will reveal the candidate's values.

Product reviews must ask questions about the product's impact.

Environmental impact statements predict the impact of construction projects on the public. We've already mentioned FAT/ML's proposed Social Impact Statements as an example of what might be done for data. In the social sciences and the biomedical industry, Institutional Review Boards (IRBs) assess the possible consequences of experiments before they're performed.

While both environmental impact statements and IRBs present problems for data products, data teams need to evaluate the impact of choices they make. Teams need to think about the consequences of their actions before releasing products. We believe that using a checklist is the best approach for ensuring good outcomes.

Teams must reflect diversity of thought, experiences, race, and background.

All too often, we hear about products that are culturally insensitive or overtly racist. One [notorious example](#) is an automated passport control system that doesn't let an individual proceed until a good digital image is captured. People of Asian ancestry reported that the system kept asking them to open their eyes, even though their eyes were open. Many cringe-worthy examples are well documented; they can often be traced to a lack of data or a lack of insight into the diversity of the population that will be impacted.

While there's no general solution to these problems of cultural sensitivity, diversity and inclusion are a tremendous help. Team members should be from the populations that will be impacted. They'll see issues well before anyone else. External peer reviews can help to reveal ethical issues that your team can't see. When you're deeply involved with a project, it can be hard to recognize problems that are obvious to outsiders.

Corporations must make their own principles clear.

Google's "Don't be evil" has always been a cute, but vague, maxim. Their recent statement, [Artificial Intelligence at Google: Our Principles](#), is more specific. In a similar vein, the face recognition startup Kairos has said that they [won't do business with law enforcement companies](#). Kairos' CEO writes that "the use of commercial face recognition in law enforcement or government surveillance of any kind is wrong."

However, it's important to realize that advocating for corporate ethical principles has consequences. Significant internal protest, and the [resignation of several developers](#) in protest

over Google's defense contracts, were needed to get their AI principles in place. Kairos is probably leaving a lot of money on the table. It's also important to realize that organizations frequently point to their ethical principles to divert attention from unethical projects.

Over the past few years, we've heard a lot about software startups that begin with a "minimal viable product," and adhere to Facebook's slogan, "move fast and break things." Is that incompatible with the approach we've just described? This is a false choice. Going fast doesn't mean breaking things. It is possible to build quickly and responsibly.

The lean/agile methodology used in many startups is a good way to expose ethical issues before they become problems. Developers start with a very simple product ("[the simplest thing that could possibly work](#)," according to Ward Cunningham's seminal phrase), demo it to users, get feedback, develop the next version, and repeat. The process continues for as many iterations as needed to get a product that's satisfactory. If a diverse group of users tests the product, the product development loop is likely to flush out systematic problems with bias and cultural insensitivity. The key is testing the product on a truly diverse group of users, not just a group that mirrors the expected customer base or the developers' backgrounds.

Regulation

In some industries, ethical standards have been imposed by law and regulation. The [Nuremberg Code](#) was developed in response to Nazi atrocities. It focuses on individual consent to participation in an experiment. After the [Tuskegee syphilis experiments](#) became public knowledge, the code was put into law in the [1974 National Research Act](#) and the [1975 Declaration of Helsinki](#). This push to codify ethical guidelines established the role of the institutional review board (IRB), and was adopted widely in the US via the [Common Rule](#).

In other industries, other regulatory bodies enforce ethical standards. These include the US Federal Trade Commission (FTC), which oversees commerce; the Nuclear Regulatory Commission (NRC), which oversees nuclear power plants; the Federal Food and Drug Administration (FDA), which oversees the safety of pharmaceuticals; and, most recently, the Consumer Finance Protection Bureau (CFPB), which oversees bankers and lenders on behalf of consumers.

The European Union's [General Data Protection Regulation](#) (GDPR) takes an aggressive approach to regulating data use and establishing a uniform data policy. In June 2018, California passed a [digital privacy law](#) similar to GDPR, despite the reservations of many online companies. One challenge of developing a policy framework is that the policy development process nearly always lags the pace of innovation, and isn't agile enough to keep policy iterative. By the time a policy has been formulated and approved, it almost always lags behind technology; but it's impossible for policy makers to iterate quickly enough to catch up with the newest technology.

Another problem is that the committees that make policy often lack experts with the necessary technical background. That can be good; technologists are too easily influenced by "[what technology wants](#)." But policies created by people who are technologically uninformed are frequently out of touch with reality: look at the [debate over back doors to encryption protocols](#).

Some have argued that organizations using data should adopt the Institutional Review Board (IRB) model from the biomedical industry. Unfortunately, while there are many positive aspects of the IRB, this isn't a viable approach. IRBs are complex, and they can't be agile; it's very difficult for IRBs to adapt to new ideas and technologies. It's why the Obama administration pushed for nearly eight years to update the Common Rule's models for consent to be consistent with digital technologies and to enable data mining.

Building Our Future

For some time, we've been aware of the ethical problems that arise from the use and abuse of data. Public outcry over Facebook will die down eventually, but the problems won't. We're looking at a future in which most vehicles are autonomous; we will be talking to robots with voices and speech patterns that are indistinguishable from humans; and where devices are listening to all our conversations, ready to make helpful suggestions about everything from restaurants and recipes to medical procedures. The results could be wonderful--or they could be a nightmarish dystopia.

It's data's day of reckoning. The shape of the future will depend a lot on what we do in the next few years. We need to incorporate ethics into all aspects of technical education and corporate culture; we need to give people the freedom to stop production if necessary, and to escalate concerns if they're not addressed; we need to incorporate diversity and ethics into hiring decisions; and we may need to consider regulation to protect the interests of individual users, and society as a whole.

Above all, talk about ethics! In "[It's time for data ethics conversations at the dinner table](#)," [Natalie Evans Harris](#) and others write that "we need to be having difficult conversations about our individual and collective responsibility to handle data ethically." This is the best single thing you can do to further data ethics: talk about it in meetings, at lunch, and even at dinner. Signing a data oath, or agreeing to a code of conduct, does little if you don't live and breathe ethics. Once you are living and breathing it, you will start to think differently about the code you write, the models you build, and the applications you create. The only way to create an ethical culture is to live it. The change won't take place magically, nor will it be easy--but it's necessary.

We can build a future we want to live in, or we can build a nightmare. The choice is up to us.

Case Studies

To help us think seriously about data ethics, we need case studies that we can discuss, argue about, and come to terms with as we engage with the real world. Good case studies give us the opportunity to think through problems before facing them in real life. And case studies show us that ethical problems aren't simple. They are multifaceted, and frequently there's no single right answer. And they help us to recognize there are few situations that don't raise ethical questions.

Princeton's [Center for Information Technology Policy](#) and [Center for Human Values](#) have created four anonymized [case studies](#) to promote the discussion of ethics. (More are in the pipeline, and may be available by the time you read this.) The first of these studies, [Automated Healthcare App](#), discusses a smartphone app designed to help adult onset diabetes patients. It raises issues like paternalism, consent, and even language choices. Is it OK to "nudge" patients toward more healthy behaviors? What about automatically moderating the users' discussion groups to emphasize scientifically accurate information? And how do you deal with minorities who don't respond to treatment as well? Could the problem be the language itself that is used to discuss treatment?

The next case study, [Dynamic Sound Identification](#), covers an application that can identify voices, raising issues about privacy, language, and even gender. How far should developers go in identifying potential harm that can be caused by an application? What are acceptable error rates for an application that can potentially do harm? How can a voice application handle people with different accents or dialects? And what responsibility do developers have when a small experimental tool is bought by a large corporation that wants to commercialize it?

The [Optimizing Schools](#) case study deals with the problem of finding at-risk children in school systems. Privacy and language are again an issue; it also raises the issue of how decisions to use data are made. Who makes those decisions, and who needs to be informed about them? What are the consequences when people find out how their data has been used? And how do you interpret the results of an experiment? Under what conditions can you say that a data experiment has really yielded improved educational results?

The final case study, [Law Enforcement Chatbots](#), raises issues about the trade-off between liberty and security, entrapment, openness and accountability, and compliance with international law.

None of these issues are simple, and there are few (if any) "right answers." For example, it's easy to react against perceived paternalism in a medical application, but the purpose of such an application is to encourage patients to comply with their treatment program. It's easy to object to monitoring students in a public school, but students are minors, and schools by nature handle a lot of private personal data. Where is the boundary between what is, and isn't, acceptable? What's important isn't getting to the correct answer on any issue, but to make sure

the issue is discussed and understood, and that we know what trade-offs we are making. What is important is that we get practice in discussing ethical issues and put that practice to work in our jobs. That's what these case studies give us.