DATA ANALYTICS

Unit 5: Advanced Techniques

Swati Pratap Jagdale
Department of Computer Science and Engineering

DATA ANALYTICS

Unit 5: Latent Semantic Analysis (LSA)

Swati Pratap Jagdale

Department of Computer Science and Engineering

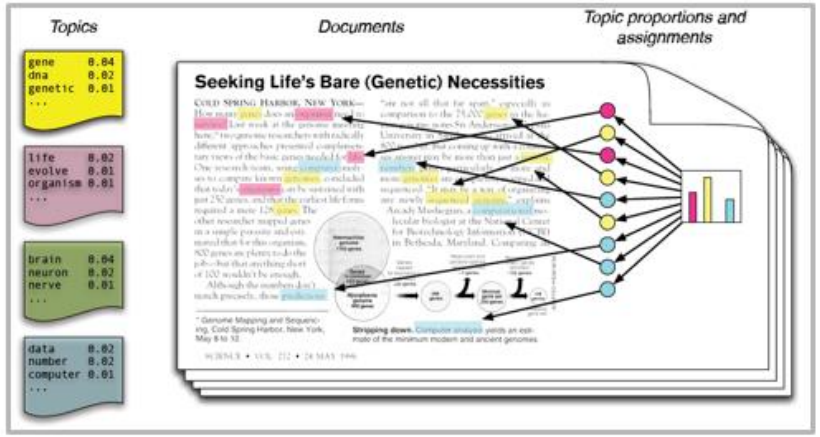
Latent Semantic Analysis (LSA)

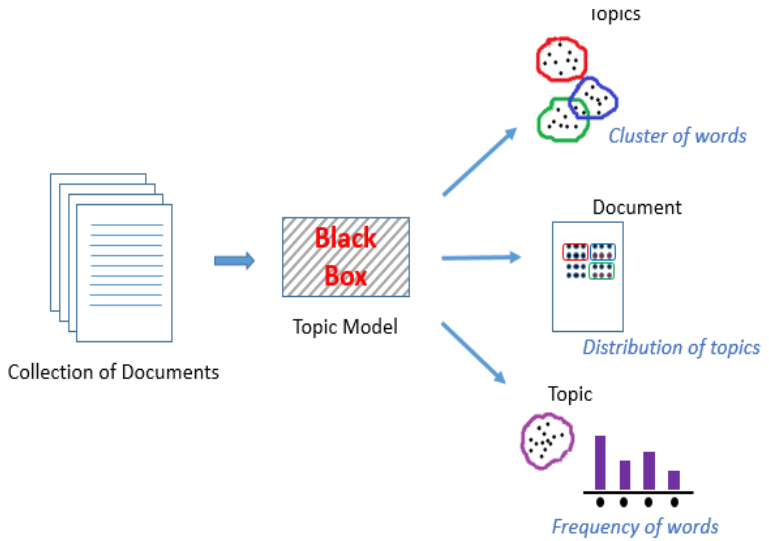
- Content-based recommendation systems: How do we extract keywords or 'topics' (latent patterns) in large documents (news articles, movie plots, book blurbs, relevant job descriptions, etc.) to create summaries or retrieve meaningful information?

- Latent Semantic Analysis, or LSA, is one of the foundation techniques in topic modeling.

- **What is a topic model?**
  An unsupervised technique to discover topics across various text documents.

  Every topic is defined by the proportion of different words it contains.





The problem and the vector space model

- Information Retrieval in the 1980s:
- Problem: Given a collection of documents: retrieve documents that are relevant to a given query match terms in documents to terms in query

- Solution approach: The vector space method
  - Create a term (rows) by document (columns) matrix, based on occurrence
  - Translate into vectors in a vector space; one vector for each document
  - Use cosine to measure distance between vectors (documents)
    small angle = large cosine = similar
    large angle = small cosine = dissimilar

- What can go wrong with this?

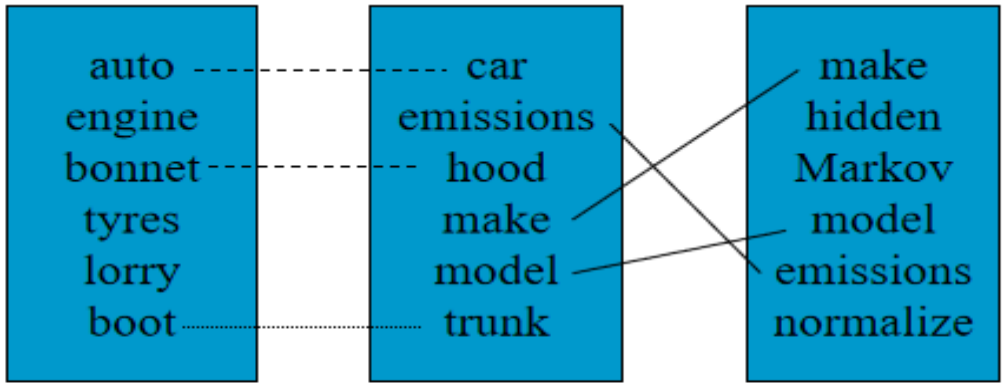
  - I liked his last **novel** quite a lot.
  - We would like to go for a **novel** marketing campaign.

In the first sentence, the word ‘novel’ refers to a book, and in the second sentence it means new or fresh.

Merely mapping words to documents will not suffice as a representation

Motivation for Latent Semantic Analysis (LSA)

- Vector Space Model (from Lillian Lee) has two problems; handling synonymy and polysemy



Synonymy

Will have small cosine

but are related

Not handling this leads to poor recall

Polysemy

Will have large cosine

but not truly related

Not handling this leads to poor precision

Latent Semantic Indexing was proposed to address these two problems to map 'concepts' better for effective Information Retrieval

Steps involved in Latent Semantic Analysis (LSA)

- Let's say we have m number of text documents with n number of total unique terms (words). We wish to extract k topics from all the text data in the documents. The number of topics, k, has to be specified by the user. Generate a document-term matrix of shape n x m
  An m x n term by document matrix (more generally term by context) tend to be sparse
- Convert matrix entries to weights, typically:
  - L(i,j) * G(i): local and global
  - a_ij -> log(freq(a_ij)) divided by entropy for row (-sum (p logp), over p: entries in the row) weight directly by estimated importance in passage
  - weight inversely by degree to which knowing that a word occurred, provides information about the passage it appeared in
- Rank-reduced Singular Value Decomposition (SVD) performed on matrix all but the k highest singular values are set to 0 produces k-dimensional approximation of the original matrix (in least-squares sense) this is the "semantic space"
- Compute similarities between entities in semantic space (usually with cosine)

Terms

Documents

	T1	T2	T3	...	Tn
D1	0.2	0.1	0.5	...	0.1
D2	0.1	0.3	0.4	...	0.3
D3	0.3	0.1	0.1	...	0.5
...	...	...	...	...	...
Dm	0.2	0.1	0.2	...	0.1

Singular Value Decomposition (SVD)

- SVD is basically a factorization of the matrix. Here, we reduce the number of rows (which means the number of words) while preserving the similarity structure among columns (which means paragraphs).
- Unique mathematical decomposition of a matrix into the product of three matrices: two with orthonormal columns and
  one with singular values on the diagonal
- A tool for dimension reduction
  - similarity measure based on co-occurrence
  - finds optimal projection into low-dimensional space

- Can be viewed as a method for rotating the axes in n-dimensional space, so that
  - the first axis runs along the direction of the largest variation among the documents
  - the second dimension runs along the direction with the second largest variation and so on
  - Generalized least-squares method

DATA ANALYTICS

Latent Semantic Analysis (LSA)

- **A Simple Example:** Technical Memo Titles

Topic: Human Computer Interaction (HCI)
c1: *Human* machine *interface* for ABC *computer* applications
c2: A *survey* of *user* opinion of *computer system response time*
c3: The *EPS user interface* management *system*
c4: *System* and *human system* engineering testing of *EPS*
c5: Relation of *user* perceived *response time* to error measurement

Topic: Graph theory (conceptually disjoint from HCI)
m1: The generation of random, binary, ordered *trees*
m2: The intersection *graph* of paths in *trees*
m3: *Graph minors* IV: Widths of *trees* and well-quasi-ordering
m4: *Graph minors*: A *survey*

Complete paper with detailed notes

Latent Semantic Analysis (LSA)

- Create a term-document matrix

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

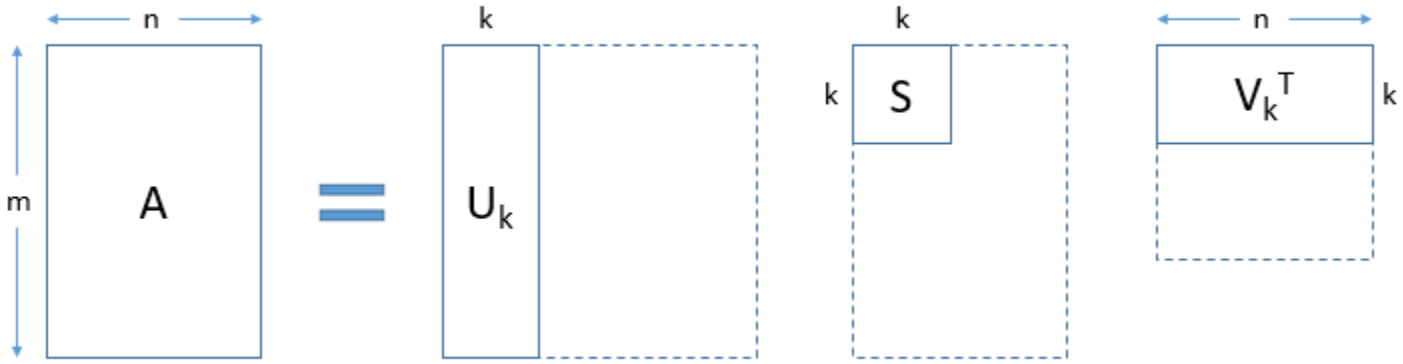
A word by context matrix, ***A***, formed from the titles of five articles about human-computer interaction and four about graph theory. Cell entries are the number of times that a word (rows) appeared in a title (columns) for words that appeared in at least two titles.

r (human,user) = -.38 r (human,minors) = -.29

Singular Value Decomposition

- The mxn term-document matrix is subject to singular value decomposition

$$A = USV^T$$



- Rank-reduced Singular Value Decomposition (SVD) performed on matrix, all but the k highest singular values are set to 0; this produces a k-dimensional approximation of the original matrix (in least-squares sense) this is the “semantic space”
- Compute similarities between entities in semantic space (usually with cosine)

Latent Semantic Analysis (LSA)

- Select k=2 rows of U

{U} =

0.22	-0.11	0.29	-0.41	-0.11	-0.34	0.52	-0.06	-0.41
0.20	-0.07	0.14	-0.55	0.28	0.50	-0.07	-0.01	-0.11
0.24	0.04	-0.16	-0.59	-0.11	-0.25	-0.30	0.06	0.49
0.40	0.06	-0.34	0.10	0.33	0.38	0.00	0.00	0.01
0.64	-0.17	0.36	0.33	-0.16	-0.21	-0.17	0.03	0.27
0.27	0.11	-0.43	0.07	0.08	-0.17	0.28	-0.02	-0.05
0.27	0.11	-0.43	0.07	0.08	-0.17	0.28	-0.02	-0.05
0.30	-0.14	0.33	0.19	0.11	0.27	0.03	-0.02	-0.17
0.21	0.27	-0.18	-0.03	-0.54	0.08	-0.47	-0.04	-0.58
0.01	0.49	0.23	0.03	0.59	-0.39	-0.29	0.25	-0.23
0.04	0.62	0.22	0.00	-0.07	0.11	0.16	-0.68	0.23
0.03	0.45	0.14	-0.01	-0.30	0.28	0.34	0.68	0.18

Latent Semantic Analysis (LSA)

- Select k=2 dimensions of Σ (the two largest Eigenvalues; denoted by the 2x2 submatrix S)

$\{\Sigma\} =$

$$\begin{matrix} 3.34 & & & & & & & & \\ & 2.54 & & & & & & & \\ & & 2.35 & & & & & & \\ & & & 1.64 & & & & & \\ & & & & 1.50 & & & & \\ & & & & & 1.31 & & & \\ & & & & & & 0.85 & & \\ & & & & & & & 0.56 & \\ & & & & & & & & 0.36 \end{matrix}$$

Latent Semantic Analysis (LSA)

- Select k=2 columns of V (or rows of V^T)

{V} =

0.20	0.61	0.46	0.54	0.28	0.00	0.01	0.02	0.08
-0.06	0.17	-0.13	-0.23	0.11	0.19	0.44	0.62	0.53
0.11	-0.50	0.21	0.57	-0.51	0.10	0.19	0.25	0.08
-0.95	-0.03	0.04	0.27	0.15	0.02	0.02	0.01	-0.03
0.05	-0.21	0.38	-0.21	0.33	0.39	0.35	0.15	-0.60
-0.08	-0.26	0.72	-0.37	0.03	-0.30	-0.21	0.00	0.36
0.18	-0.43	-0.24	0.26	0.67	-0.34	-0.15	0.25	0.04
-0.01	0.05	0.01	-0.02	-0.06	0.45	-0.76	0.45	-0.07
-0.06	0.24	0.02	-0.08	-0.26	-0.62	0.02	0.52	-0.45

Latent Semantic Analysis (LSA)

- Original term-document matrix

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

r (human.user) = -.38 r (human.minors) = -.29

After LSA

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	0.16	0.40	0.38	0.47	0.18	-0.05	-0.12	-0.16	-0.09
interface	0.14	0.37	0.33	0.40	0.16	-0.03	-0.07	-0.10	-0.04
computer	0.15	0.51	0.36	0.41	0.24	0.02	0.06	0.09	0.12
user	0.26	0.84	0.61	0.70	0.39	0.03	0.08	0.12	0.19
system	0.45	1.23	1.05	1.27	0.56	-0.07	-0.15	-0.21	-0.05
response	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
time	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
EPS	0.22	0.55	0.51	0.63	0.24	-0.07	-0.14	-0.20	-0.11
survey	0.10	0.53	0.23	0.21	0.27	0.14	0.31	0.44	0.42
trees	-0.06	0.23	-0.14	-0.27	0.14	0.24	0.55	0.77	0.66
graph	-0.06	0.34	-0.15	-0.30	0.20	0.31	0.69	0.98	0.85
minors	-0.04	0.25	-0.10	-0.21	0.15	0.22	0.50	0.71	0.62

r (human,user) = .94 r (human,minors) = -.83

Effect of SVD on the correlation matrix

LSA Titles example:
Correlations between titles in raw data

	c1	*c2*	*c3*	*c4*	*c5*	*m1*	*m2*	*m3*
c2	-0.19							
c3	0.00	0.00						
c4	0.00	0.00	0.47					
c5	-0.33	0.58	0.00	-0.31				
m1	-0.17	-0.30	-0.21	-0.16	-0.17			
m2	-0.26	-0.45	-0.32	-0.24	-0.26	0.67		
m3	-0.33	-0.58	-0.41	-0.31	-0.33	0.52	0.77	
m4	-0.33	-0.19	-0.41	-0.31	-0.33	-0.17	0.26	0.56

0.02
-0.30 0.44

Correlations in first-two dimension space post LSA

c2	0.91							
c3	1.00	0.91						
c4	1.00	0.88	1.00					
c5	0.85	0.99	0.85	0.81				
m1	-0.85	-0.56	-0.85	-0.88	-0.45			
m2	-0.85	-0.56	-0.85	-0.88	-0.44	1.00		
m3	-0.85	-0.56	-0.85	-0.88	-0.44	1.00	1.00	
m4	-0.81	-0.50	-0.81	-0.84	-0.37	1.00	1.00	1.00

0.92
-0.72 1.00

Some FAQ's on LSA

- **Why, and under what circumstances would reducing the dimensionality of representation be beneficial?**
  - When the original data are generated from a source of the same dimensionality and general structure as the reconstruction.
  - Suppose, for example, that speakers or writers generate paragraphs by choosing words from a k-dimensional space in such a way that words in the same paragraph tend to be selected from nearby locations. If listeners or readers try to infer the similarity of meaning from these data, they will do better if they reconstruct the full set of relations in the same number of dimensions as the source. Among other things, given the right analysis, this will allow the system to infer that two words from nearby locations in semantic space have similar meanings even though they are never used in the same passage, or that they have quite different meanings even though they often occur in the same utterances.

- **How is k, the number of dimensions to be retained in LSA, selected?**
  - Empirically
  - Some external criterion of validity is sought, such as the performance on a synonym test or prediction of the missing words in passages if some portion are deleted in forming the initial matrix.

Pros and Cons of LSA

Pros:

- LSA is fast and easy to implement.
- It gives decent results, much better than a plain vector space model.

Cons:

- Since it is a linear model, it might not do well on datasets with non-linear dependencies.
- LSA assumes a Gaussian distribution of the terms in the documents, which may not be true for all problems.
- LSA involves SVD, which is computationally intensive and hard to update as new data comes up.

Additional References

https://www.analyticsvidhya.com/blog/2018/10/stepwise-guide-topic-modeling-latent-semantic-analysis/

https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-python/

Paper on LSA with a detailed explanation of the example:
http://lsa.colorado.edu/papers/dp1.LSAintro.pdf

DATA ANALYTICS

Unit 5: Sparse data processing, sparse PCA

Swati Pratap Jagdale

Department of Computer Science and Engineering

Sparse Data Processing

- Sparse data is when there are more 0's than there are entries in the data matrix (or flatfile or record, etc.)

User	Movie	Rating
A	Parasite	5
A	Joker	4
A	Avengers: Endgame	4
B	Parasite	2
B	Spotlight	4
B	The Great Beauty	3
C	Avengers: Endgame	5
D	There will be blood	4
E	Avengers: Endgame	4

Dense matrix

Users					Movies						Target
A	B	C	D	E	Parasite	Joker	Avengers	Spotlight	The Great Beauty	There will be blood	Rating
1	0	0	0	0	1	0	0	0	0	0	5
1	0	0	0	0	0	1	0	0	0	0	4
1	0	0	0	0	0	0	1	0	0	0	4
0	1	0	0	0	1	0	0	0	1	0	2
0	1	0	0	0	0	0	0	1	0	0	4
0	1	0	0	0	0	0	0	0	1	0	3
0	0	1	0	0	0	0	1	0	0	0	5
0	0	0	1	0	0	0	0	0	0	1	4
0	0	0	0	1	0	0	1	0	0	0	4

Sparse matrix

- What is the problem?
  - Space complexity: very large term-document matrices or matrices showing links between websites or users need to be stored in memory for processing
  - Time complexity: it takes needlessly long to perform operations on sparse matrices (given most of the data is 0 and need not be processed!)

https://machinelearningmastery.com/sparse-matrices-for-machine-learning/
https://towardsdatascience.com/working-with-sparse-data-sets-in-pandas-and-sklearn-d26c1cfbe067

Sparse Data Processing

Some workarounds

- Ignore zero values; only nonzero values can be stored and processed
- Use a different representation:

  - Dictionary of Keys. A dictionary is used where a row and column index is mapped to a value.

  - List of Lists. Each row of the matrix is stored as a list, with each sublist containing the column index and the value.

  - Coordinate List. A list of tuples is stored with each tuple containing the row index, column index, and the value.

  - Compressed Sparse Row (CSR). The sparse matrix is represented using three one-dimensional arrays for the non-zero values, the extents of the rows, and the column indexes.

  - Compressed Sparse Column. The same as the Compressed Sparse Row method except the column indices are compressed and read first before the row indices.

- Use dimensionality reduction techniques such as sparse PCA

https://machinelearningmastery.com/sparse-matrices-for-machine-learning/
https://towardsdatascience.com/working-with-sparse-data-sets-in-pandas-and-sklearn-d26c1cfbe067

Sparse PCA

- **Sparse principal component analysis (sparse PCA)** is a specialised technique used in statistical analysis and, in particular, in the analysis of multivariate data sets.
- It extends the classic method of principal component analysis (PCA) for the reduction of dimensionality of data by introducing sparsity structures to the input variables.
- A particular disadvantage of ordinary PCA is that the principal components are usually linear combinations of all input variables.
- Sparse PCA overcomes this disadvantage by finding linear combinations that contain just a few input variables.
- Contemporary datasets often have the number of input variables comparable with or even much larger than the number of samples.

Sparse PCA

Mathematical Formulation

- Given a data matrix X n rows (each row is an independent sample) with p columns (attributes)
- One assumes each column of X has mean zero, otherwise one can subtract column-wise mean from each element of X.
- Let $\Sigma = \frac{1}{n-1} X^\top X$ be the empirical covariance matrix of X, which has

dimensions p x p. Given integer k, 1<=k<=p, the sparse PCA problem can be formulated as maximizing the variance along a direction represented by vector v, while constraining its cardinality.

$$\begin{aligned} \max \quad & v^T \Sigma v \\ \text{subject to} \quad & \|v\|_2 = 1 \\ & \|v\|_0 \leq k. \end{aligned}$$

- The constraints specify that *v* is a unit vector and $||v||_0$ represents the L_0 norm of v, defined as the number of its non-zero components (<=k).
- k is much smaller than p; the result is the k-sparse largest eigenvalue.
- If one takes k=p, the problem reduces to the ordinary PCA, and the optimal value becomes the largest eigenvalue of covariance matrix Σ.

Sparse PCA

- After finding the optimal solution v, one deflates Σ to obtain a new matrix.

$$\Sigma_1 = \Sigma - (v^T \Sigma v) v v^T,$$

- Iterate this process to obtain further principal components.
- However, unlike PCA, sparse PCA cannot guarantee that different principal components are orthogonal. In order to achieve orthogonality, additional constraints must be enforced.

- The following equivalent definition is in matrix form.
- Let, V be a p×p symmetric matrix, one can rewrite the sparse PCA problem as:

$$\begin{aligned} \max \quad & Tr(\Sigma V) \\ \text{subject to} \quad & Tr(V) = 1 \\ & \|V\|_0 \leq k^2 \\ & Rank(V) = 1, V \succeq 0. \end{aligned}$$

- Tr is the matrix trace, and $||V||_0$ represents the non-zero elements in matrix V. The last line specifies that V has matrix rank one and is positive semidefinite. The last line means that one has **V= vv^T**

$$\begin{aligned} \max \quad & Tr(\Sigma V) \\ \text{subject to} \quad & Tr(V) = 1 \\ & \mathbf{1}^T |V| \mathbf{1} \leq k \\ & V \succeq 0. \end{aligned}$$

Applications of Sparse PCA

Financial Data Analysis

- Suppose ordinary PCA is applied to a dataset where each input variable represents a different asset, it may generate principal components that are weighted combination of all the assets
- In contrast, sparse PCA would produce principal components that are weighted combination of only a few input assets, so one can easily interpret its meaning.
- Furthermore, if one uses a trading strategy based on these principal components, fewer assets imply less transaction costs.

Biology

- Consider a dataset where each input variable corresponds to a specific gene. Sparse PCA can produce a principal component that involves only a few genes, so researchers can focus on these specific genes for further analysis.

High-dimensional Hypothesis Testing

- Contemporary datasets often have the number of input variables (p) comparable with or even much larger than the number of samples (n)
- It has been shown that if p/n does not converge to zero, the classical PCA is not consistent. But sparse PCA can retain consistency even if p>>n

References

https://www.analyticsvidhya.com/blog/2014/01/logistic-regression-rare-event/

https://en.wikipedia.org/wiki/Sparse_PCA#Financial_Data_Analysis

DATA ANALYTICS

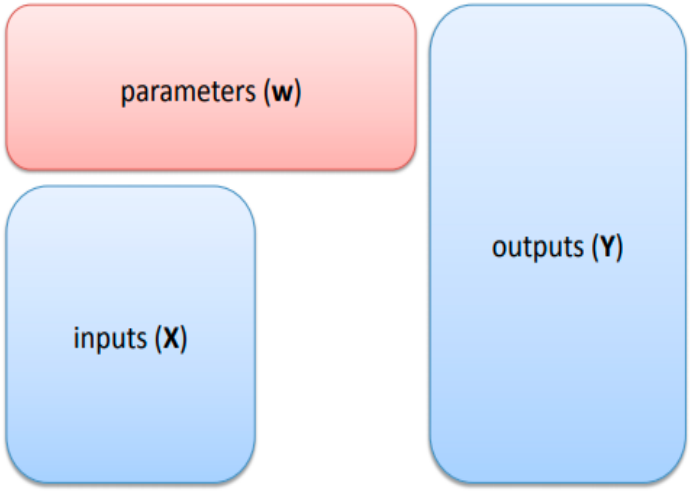
Unit 5: Concept of hidden variables

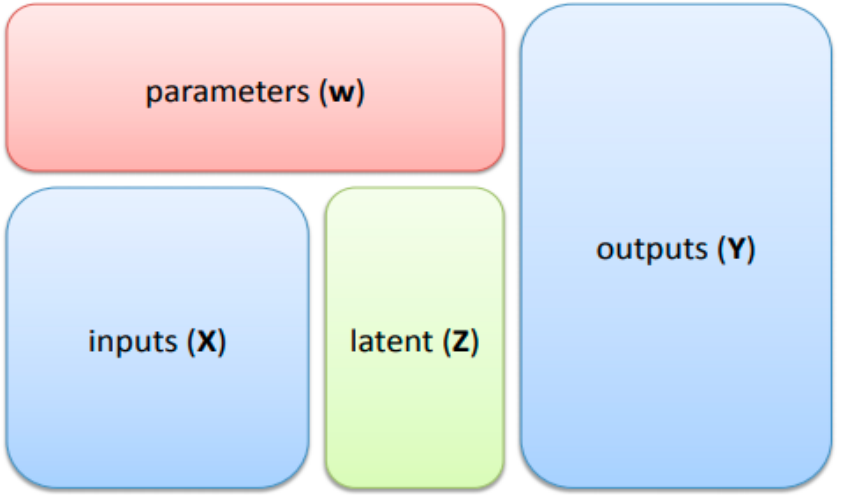
Swati Pratap Jagdale

Department of Computer Science and Engineering

Hidden Variables

- Random variables in supervised learning



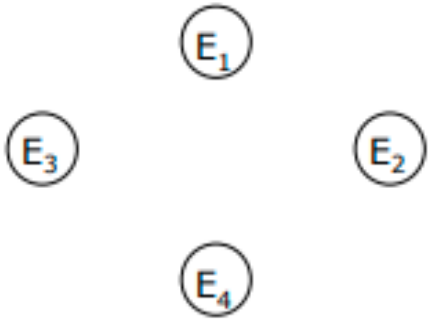


- 'Hidden' (or latent) variable is one that we never 'see'
- Not even in training
- Sometimes we believe they are real
  and sometimes they approximate reality (as it happens in Physics)

- 'Learning' or 'decoding' are both understood as inference problems

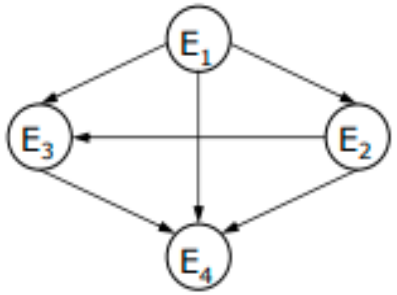
Learning With Hidden Variables

- Consider a situation in which you can observe a whole bunch of different evidence variables, E_1 through E_n. Maybe they're all the different symptoms that a patient might have. Or maybe they represent different movies and whether someone likes them.



Learning With Hidden Variables

- If those variables are all conditionally dependent on one another, then we'd need a highly connected graph that's capable of representing the entire joint distribution between the variables.
- Because the last node has n-1 parents, it will take on the order of 2^n parameters to specify the conditional probability tables in this network.



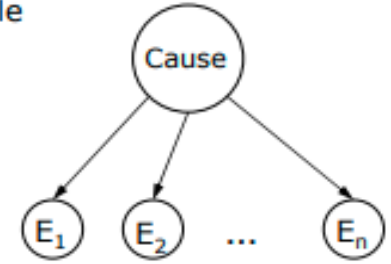
Without the cause, all the evidence is dependent on each other

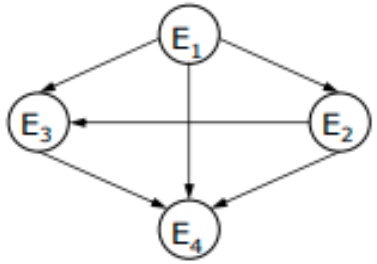
$O(2^n)$ parameters

Learning With Hidden Variables

- But, in some cases, we can get a considerably simpler model by introducing an additional “cause” node.
- It might represent the underlying disease state that was causing the patients’ symptoms or some division of people into those who like westerns and those who like comedies.

Cause is unobservable



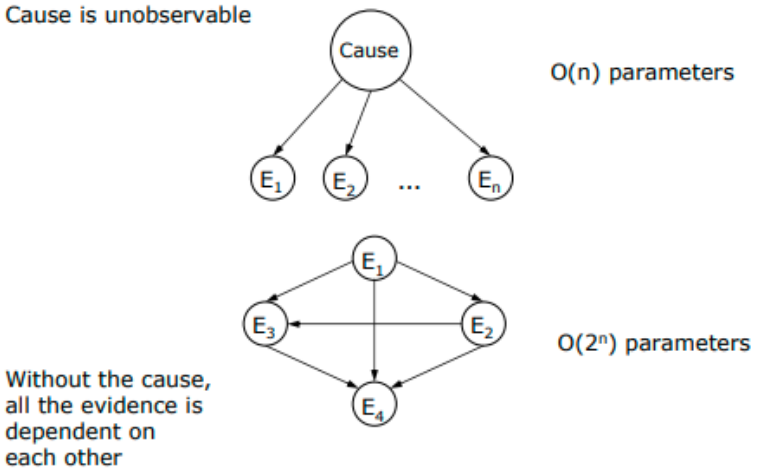


$O(2^n)$ parameters

Without the cause, all the evidence is dependent on each other

Learning With Hidden Variables

- In the simpler model, the evidence variables are conditionally independent given the causes. That means that it would only require on the order of n parameters to describe all the CPTs in the network, because at each node, we just need a table of size 2 (if the cause is binary; or k if the cause can take on k values), and one (or k-1) parameter to specify the probability of the cause.



- So, what if you think there's a hidden cause? How can you learn a network with unobservable variables?

Simpson's Paradox

- Edward Hugh Simpson, a statistician and former cryptanalyst at Bletchley Park, described the statistical phenomenon - Simpson's paradox

- The art of data science is seeing beyond the data — using and developing methods and tools to get an idea of what that hidden reality looks like.

- Simpson's paradox showcases the importance of skepticism and interpreting data with respect to the real world, and also the dangers of oversimplifying a more complex truth by trying to see the whole story from a single data-viewpoint.

Simpson's Paradox

- ***Simpson's Paradox:***

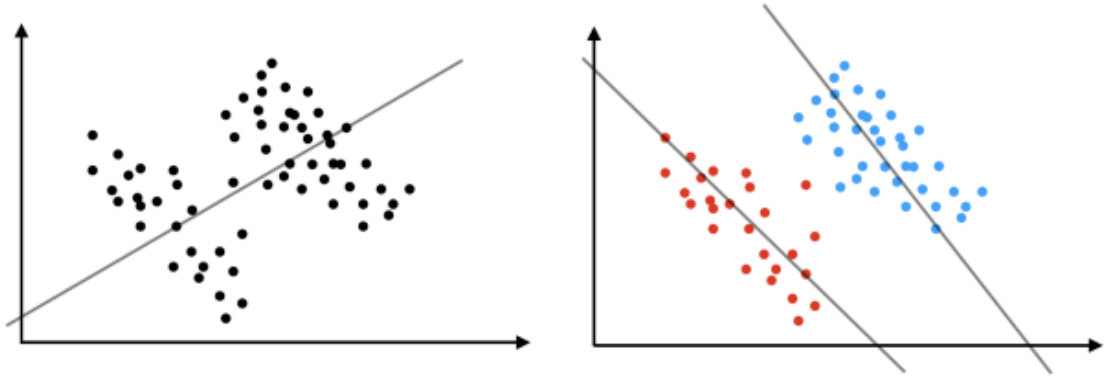
A trend or result that is present when data is put into groups that reverses or disappears when the data is combined.

Example: UC Berkley's suspected gender-bias.
At the beginning of the academic year in 1973, UC Berkeley's graduate school had admitted roughly 44% of their male applicants and 35% of their female applicants.
Was there a discrimination against female applicants?

When the data was studied department-wise, it was observed that:

- there was a statistically significant gender bias **in favor of women** for 4 out of the 6 departments, and no significant gender bias in the remaining 2
- It is discovered that **women tended to apply to departments that admitted a smaller percentage of applicants overall**, and that this hidden variable affected the marginal values for the percentage of accepted applicants in such a way as to reverse the trend that existed in the data as a whole

Simpson's Paradox



A visual example: the overall trend reverses when data is grouped by some colour-represented category.

Simpson's Paradox

A simple example in business:

- Suppose the soft drinks industry is trying to choose between two new flavors they have produced. We could sample public opinion on the two flavors

Flavour	Sample Size	# Liked Flavour
Sinful Strawberry	1000	800
Passionate Peach	1000	750

- 80% of people enjoyed 'Sinful Strawberry' whereas only 75% of people enjoyed 'Passionate Peach'. So 'Sinful Strawberry' is more likely to be the preferred flavor.

Simpson's Paradox

- Some other information while conducting the survey, such as the sex of the person sampling the drink. What happens if we split our data up by sex?

- 84.4% of men and 40% of women liked 'Sinful Strawberry' whereas 85.7% of men and 50% of women liked 'Passionate Peach'

Flavour	# Men	# Liked Flavour (Men)	# Women	# Liked Flavour (Women)
Sinful Strawberry	900	760	100	40
Passionate Peach	700	600	300	150

Simpson's Paradox

- *According to our sample data, generally people prefer 'Sinful Strawberry', but both men and women separately prefer 'Passionate Peach'.*
- *This is an example of Simpson's Paradox!*

Flavour	# Men	# Liked Flavour (Men)	# Women	# Liked Flavour (Women)
Sinful Strawberry	900	760	100	40
Passionate Peach	700	600	300	150

Simpson's Paradox

Lurking variables (Hidden Variables)

- Simpson's paradox arises when there are hidden variables that split data into multiple separate distributions.
- Such a hidden variable is aptly referred to as a **lurking variable**, and they can often be difficult to identify.

Consider the lurking variable (sex) and a little bit of probability theory:-

- P(Liked Strawberry) = P(Liked Strawberry | Man)P(Man) + P(Liked Strawberry | Woman)P(Woman)

- 800/1000 = (760/900)×(900/1000) + (40/100)×(100/1000)

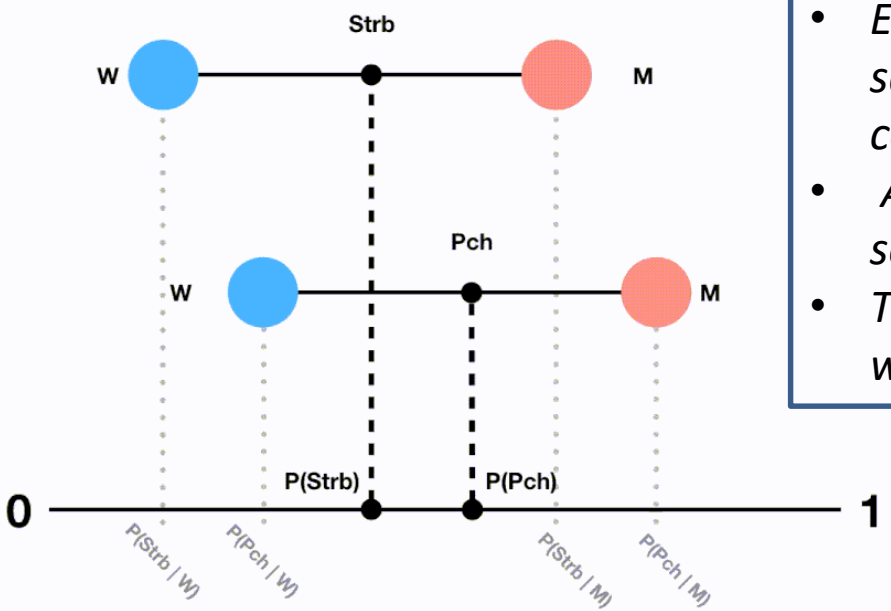
- P(Liked Peach) = P(Liked Peach | Man)P(Man) + P(Liked Peach | Woman)P(Woman)

- 750/1000 = (600/700)×(700/1000) + (150/300)×(300/1000)

Simpson's Paradox

- **Lurking variables (Hidden Variables)**

- We can think of the marginal probabilities of sex (P(Man) and P(Woman)) as weights that, in the case of 'Sinful Strawberry', cause the total probability to be significantly shifted towards the male opinion.



- *Each coloured circle represents either the men or women that sampled each flavour, the position of the centre of each circle corresponds to that group's probability of liking the flavour.*
- *As the circles grow (i.e. sample proportions change) we can see how the marginal probability of liking the flavour changes.*
- *The marginal distributions shift and switch as samples become weighted with respect to the lurking variable (sex).*

Approaches to Infer Hidden Variables

From other variables in the model i.e., from observations or evidence

- Viterbi algorithm for decoding the transition of hidden states
- “Learning” of Hidden Markov Models (or expectation maximization (EM))

These are popular approaches to infer the transition probabilities between hidden states or the effect of hidden variables on a system

Think about this:
Given a model, how does one postulate the presence of a hidden (or latent) variable?

(This is where an understanding of the problem domain comes in…)

References

http://www.cs.cmu.edu/~nasmith/psnlp/lecture5.pdf

https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-825-techniques-in-artificial-intelligence-sma-5504-fall-2002/lecture-notes/Lecture18FinalPart1.pdf

https://towardsdatascience.com/simpsons-paradox-and-interpreting-data-6a0443516765

THANK YOU

Swati Pratap Jagdale
Department of Computer Science
swatigambhire@pes.edu

DATA ANALYTICS

Unit 5: Advanced Techniques

Swati Pratap Jagdale
Department of Computer Science and Engineering

DATA ANALYTICS

Unit 5: Confounding Variables

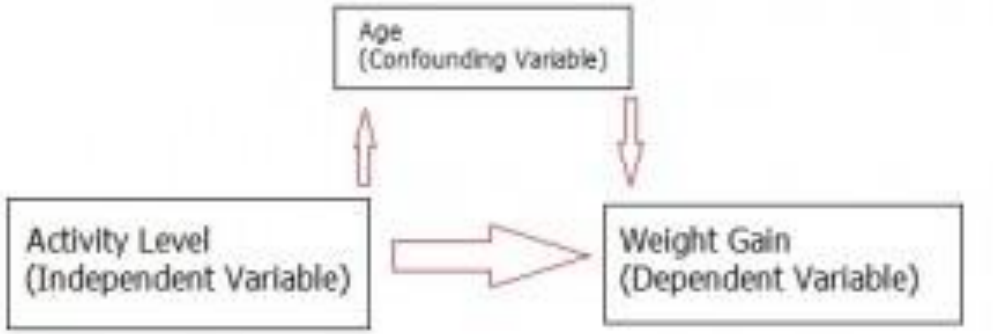
Swati Pratap Jagdale

Department of Computer Science and Engineering

Confounding Variables

- **What is a Confounding Variable?**

- A confounding variable is an “extra” variable that you didn’t account for. They can ruin an experiment and give you useless results. They can suggest there is correlation when in fact there isn’t.
- They can even introduce bias. That’s why it is important to know what one is, and how to avoid getting them into your experiment in the first place.



A confounding variable can have a hidden effect on your experiment’s outcome.

Confounding Variables

- In an experiment, the independent variable typically has an effect on your dependent variable.
- For example, if you are researching whether lack of exercise leads to weight gain, then
  lack of exercise -- independent variable
  weight gain -- dependent variable.
- Confounding variables are any other variable (like age) that also has an effect on your dependent variable (weight gain). They are like extra independent variables that are having a hidden effect on your dependent variables. A confounding variable can also be related to the independent variable (with increase in age, there may be a decrease in the duration or intensity of exercise).

- Confounding variables can cause two major problems:
  - Increase variance
  - Introduce bias.

Confounding Variables

Example:

- You test 200 volunteers (100 men and 100 women). You find that lack of exercise leads to weight gain.
- One problem with your experiment is that is lacks any control variables. For example, the use of placebos, or random assignment to groups.
- So you really can't say for sure whether lack of exercise leads to weight gain. One confounding variable is how much people eat. It's also possible that men eat more than women; this could also make sex a confounding variable.

- A poor study design like this could lead to bias.
- For example, if all of the women in the study were middle-aged, and all of the men were aged 16, age would have a direct effect on weight gain. That makes age a confounding variable.

Confounding Bias

Confounding Bias

- Bias is usually a result of errors in data collection or measurement.
- However, one definition of bias is "***…the tendency of a statistic to overestimate or underestimate a parameter***", so in this sense, confounding is a type of bias.

- Confounding bias is the result of having confounding variables in your model. It has a direction, depending on if it over- or underestimates the effects of your model:
- Positive confounding is when the observed association is biased away from the null. In other words, it overestimates the effect.
- Negative confounding is when the observed association is biased toward the null. In other words, it underestimates the effect.

Confounding Bias

How to Reduce Confounding Variables?

- Make sure you identify all of the possible confounding variables in your study.
- Make a list of everything you can think of and one by one, consider whether those listed items might influence the outcome of your study. Usually, someone has done a similar study before you. So check the academic databases for ideas about what to include on your list.

- Once you have figured out the variables, techniques to reduce the effect of those confounding variables:
  - Bias can be eliminated with random samples.
  - Introduce control variables to control for confounding variables. For example, you could control for age by only measuring 30 year olds.
  - Within subjects designs test the same subjects each time. Anything could happen to the test subject in the “between” period so this doesn’t make for perfect immunity from confounding variables.
  - Counterbalancing can be used if you have paired designs. In counterbalancing, half of the group is measured under condition 1 and half is measured under condition 2.

Terminology and identifying a confounding variable

Synonyms for Confounding Variables and Omitted Variable Bias

- Confounding variables or confounders, and lurking variables.

- A confounding variable is closely related to both the independent and dependent variables in a study. An independent variable represents the suppose *cause*, while the dependent variable is the supposed *effect*.
- A confounding variable is a third variable that influences both the independent and dependent variables. Failing to account for confounding variables can cause you to wrongly estimate the relationship between your independent and dependent variables.

- How do we identify a confounding variable?
  1. There must be three or more variables in the study
     two variables => one is the cause the other is the effect)
  2. The variable we suspect is a confounding variable, changes systematically with at least one of the variables we are measuring (either independent or dependent)
  3. Identify extraneous variables that relate to subjects (age, gender, etc.), the environment in which the study is conducted (weather, location, etc.) and to the two variables we are explicitly measuring to test for systematic changes to identify a confounding variable.

http://dissertation.laerd.com/extraneous-and-confounding-variables-p3.php

Hidden variable vs confounding variable

A hidden variable could be connecting two variables that are spuriously correlated (temperature/ season that connects the two spuriously correlated variables: ice-cream sales and number of robberies)

A confounding variable is related to two variables that are not spuriously correlated (hypothyroidism -> causes increase in weight due to lower metabolic rate; but lower metabolic rate implies lower energy levels; this lowers the ability of one to exercise which in turn leads to increase in weight)

Lack of exercise -> increase in weight;
the two are connected not spuriously

Confounding Bias

Omitting confounding variables from your regression model can bias the coefficient estimates.

- When you are assessing the effects of the independent variables in the regression output, this bias can produce the following problems:
  - Overestimate the strength of an effect.
  - Underestimate the strength of an effect.
  - Change the sign of an effect.
  - Mask an effect that actually exist

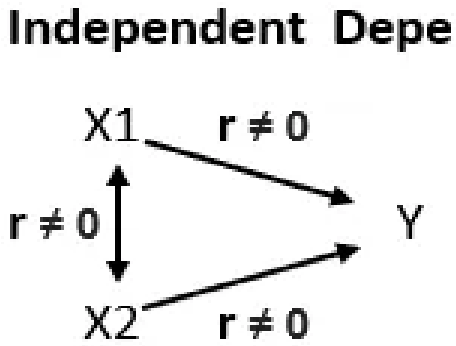
What Conditions Cause Omitted Variable Bias?

- How does this bias occur? How can variables you leave out of the model affect the variables that you include in the model?

- For omitted variable bias to occur, the following two conditions must exist:
- The omitted variable must correlate with the dependent variable.
- The omitted variable must correlate with at least one independent variable that is in the regression model.

Confounding Bias

- There must be non-zero correlations (r) on all three sides of the triangle.
- This correlation structure causes confounding variables that are not in the model to bias the estimates that appear in your regression results. For example, removing either X variable will bias the other X variable.



- The amount of bias depends on the strength of these correlations.
- Strong correlations produce greater bias.
- If the relationships are weak, the bias might not be severe.
- And, if the omitted variable is not correlated with another independent variable at all, excluding it does not produce bias.

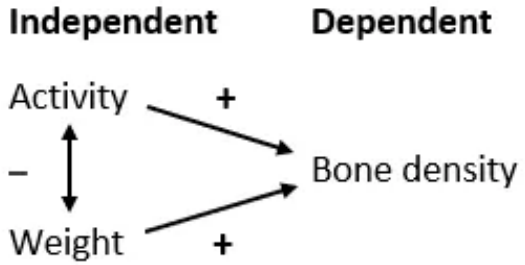
Confounding Bias

- **Example of How Confounding Variables Can Produce Bias**

Example:

- In a biomechanics lab, One study assessed the effects of physical activity on bone density.
- They measured various characteristics including the subjects' activity levels, their weights, and bone densities among many others.
- Theories about how our bodies build bone suggest that there should be a positive correlation between activity level and bone density. In other words, higher activity produces greater bone density.
- Simple regression analysis to determine whether there is a relationship between activity and bone density... there was no relationship at all!

- They included activity level as the only independent variable, but it turns out there is another variable that correlates with both activity and bone density—the subject's weight.

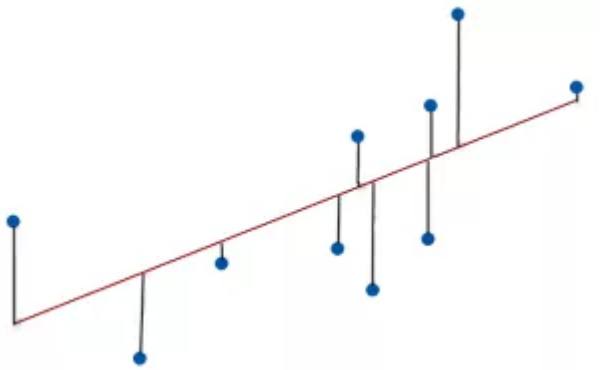


The diagram shows the signs of the correlations between the variables.

Confounding Bias

- Correlations, Residuals, and OLS Assumptions

- When you satisfy the ordinary least squares (OLS) assumptions, the Gauss-Markov theorem states that your estimates will be unbiased and have minimum variance.



Residuals = Observed value – Fitted value

Omitted variable bias occurs because the residuals violate one of the assumptions.

Confounding Bias

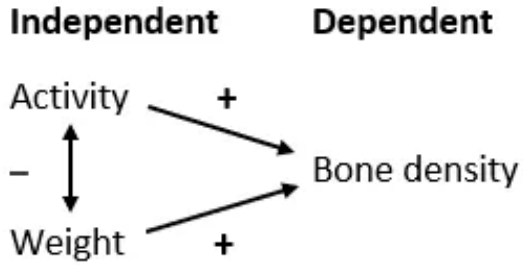
- Consider, regression model with two significant independent variables, X1 and X2. These independent variables correlate with each other and the dependent variable—which are the requirements for omitted variable bias.

- Now, imagine that we take variable X2 out of the model. It is the confounding variable. Here's what happens:

- The model fits the data less well because we have removed a significant explanatory variable. Consequently, the gap between the observed values and the fitted values increases. These gaps are the residuals.
- The degree to which each residual increases depends on the relationship between X2 and the dependent variable. Consequently, the residuals correlate with X2.
- X1 correlates with X2, and X2 correlates with the residuals. Ergo, variable X1 correlates with the residuals.
- This condition violates the ordinary least squares assumption that independent variables in the model do not correlate with the residuals. Violations of this assumption produce biased estimates.

Confounding Bias

	Included and Omitted: Negative Correlation	Included and Omitted: Positive Correlation
Included and Dependent: Negative Correlation	Positive bias: coefficient is overestimated.	Negative bias: coefficient is underestimated.
Included and Dependent: Positive Correlation	Negative bias: coefficient is underestimated.	Positive bias: coefficient is overestimated.



The table summarizes these relationships and the direction of bias.

Included (activity) and omitted (weight) are negatively correlated.
The included variable (weight) and the dependent variable (bone density) have a positive relationship, implies the result has a negative bias.

References

https://www.statisticshowto.com/experimental-design/confounding-variable/

https://statisticsbyjim.com/regression/confounding-variables-bias/

THANK YOU

Swati Pratap Jagdale
Department of Computer Science
swatigambhire@pes.edu

DATA ANALYTICS

Unit 5:Introduction to Stochastic models and Markov processes (first order)

Bharathi R

Department of Computer Science and Engineering

Introduction Stochastic Process

- Stochastic models are powerful tools which can be used for solving problems which are dynamic in nature, that is, the values of the random variables change with time.

- Stochastic process is defined as a collection of random variables
  $\{X_n, n \geq 0\}$ indexed by time (however, index can be other than time).

- The value (cash flow) that the random variable X_n can take is called the **state of the stochastic process at time n**.

- The set of all possible values the random variable can take is called the **state space**.

1.Poisson Process: Examples

Generally we would like to count the number of events that occur over a period of time. Following are few examples of counting process:

1. Retail stores would like to predict footfall (number of customers visiting the store) over a period
of time.
2. Call centres would like to predict the number of calls they receive over a period of time.
3. Number of customer arrivals at banks, airports, restaurants, and any service centres.
4. Demand for spare parts of capital equipment caused due to failure of parts over a period of time.
5. Number of insurance claims received at an insurance company.

1.Poisson Process

Homogeneous Poisson Process (HPP) is a stochastic counting process $N(t)$ with the following properties:

$N(0) = 0$, that is the number of events by time t = 0 is zero.

$N(t)$ has independent increments. That is if $t_0 < t_1 < t_2 < \ldots < < t_n$, then $N(t_1) - N(t_0)$, $N(t_2) - N(t_1)$, …, $N(t_n) - N(t_{n-1})$ are independent.

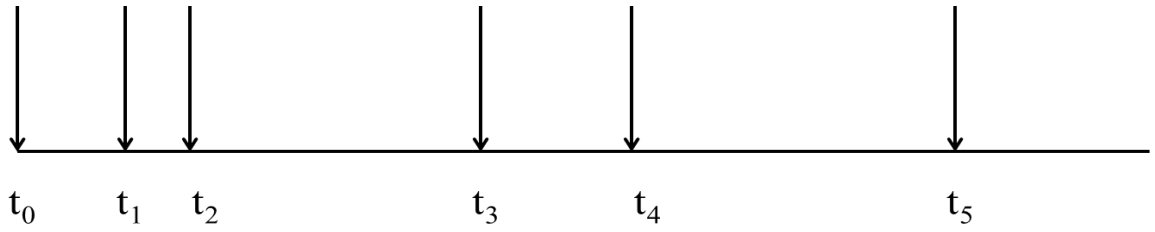


Figure above shows a Poisson process of events in which $(t_1 - t_0)$, $(t_2 - t_1)$, $(t_3 - t_2)$ are time between events.

1. Poisson Process

The number of events by time t, N(t), follows a Poisson distribution, that is

$$P[N(t)=n]=\frac{e^{-\lambda t}\times(\lambda t)^{n}}{n!}$$

Cumulative distribution of number of events by time t in a Poisson process is given by

$$P[N(t)\le n]=\sum_{i=0}^{n}P[N(t)=i]=\sum_{i=0}^{n}\frac{e^{-\lambda t}\times(\lambda t)^{i}}{i!}$$

The mean, $E[N(t)]$, and variance, $\mathrm{Var}[N(t)]$, of a Poisson process $N(t)$ are given by

$E[N(t)] = \lambda t$
$\mathrm{Var}[N(t)] = \lambda t$

In the case of Poisson process, the time between events follows an exponential distribution with parameter λ, that is the time between events have a density function $f(t) = \lambda e^{-\lambda t}$ and cumulative distribution function $F(t) = 1 - e^{-\lambda t}$.

DATA ANALYTICS

Poisson Process: Example

Johny Sparewala (JS) is a supplier of aircraft flight control system spares based out of Mumbai, India. The demand for hydraulic pumps used in the flight control system follows a Poisson process. Sample data (50 cases) on time between demands (measured in number of days) for hydraulic pumps are shown in Table 16.1

TABLE 16.1 Time between demands (in days) for hydraulic pumps

104	90	45	32	12	6	30	23	58	118
80	12	216	71	29	188	15	88	88	94
63	125	108	42	77	65	18	25	30	16
92	114	151	10	26	182	175	189	14	11
83	418	21	19	73	31	175	14	226	8

Example Continued

(a) Calculate the expected number of demand for hydraulic pump spares for next two years.

(b) Johny Sparewala would like to ensure that the demand for spares over next two years is met in at least 90% of the cases from the spares stocked (called fill rate) since lead time to manufacture a part is more than 2 years. Calculate the inventory of spares that would give at least 90% fill rate.

Solution

(a) To calculate the expected number of demand for spares for two years, we have to estimate the parameter λ of the Poisson distribution. The maximum likelihood estimate of λ is given by

$$\hat{\lambda} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} X_i} = 0.0125$$

where X_i is the time between failure of i^{th} case and $\frac{1}{n}\sum_{i=1} X_i$ is the mean time between failure.

The expected number of demand for spares, $E[N(t)]$, for 2 years (2×365 days) is given by

$$E[N(t)] = E[N(2\times 365)] = \hat{\lambda} \times t = 0.0125 \times 2 \times 365 = 9.125$$

Solution

(b) To ensure that the demand for spares is met 90% of the time, we have to calculate smallest k such that

$$\sum_{i=0}^{k} \frac{e^{-\hat{\lambda}t} \times (\hat{\lambda}t)^i}{i!} \geq 0.90$$

Table 16.2 shows density and cumulative distribution function values of Poisson process for different values of k.

TABLE 16.2 Poisson density and distribution function for different values of k

k	Poisson Density	Cumulative	k	Poisson Density	Cumulative
0	0.0001	0.0001	11	0.0996	0.7907
1	0.0010	0.0011	12	0.0758	0.8665
2	0.0045	0.0056	13	0.0532	0.9197
3	0.0138	0.0194	14	0.0347	0.9543
4	0.0315	0.0509	15	0.0211	0.9754
5	0.0574	0.1083	16	0.0120	0.9875
6	0.0873	0.1956	17	0.0065	0.9939
7	0.1138	0.3095	18	0.0033	0.9972
8	0.1298	0.4393	19	0.0016	0.9988
9	0.1316	0.5709	20	0.0007	0.9995
10	0.1201	0.6911	21	0.0003	0.9998

Solution

Smallest value of k for which the cumulative probability is greater than 0.90 is 13. That is, JS should stock 13 spares to ensure that they meet demand for spares in 90% of the cases over a two-year period. The probability density function of Poisson distribution with mean 9.125 is shown in Figure 16.2.

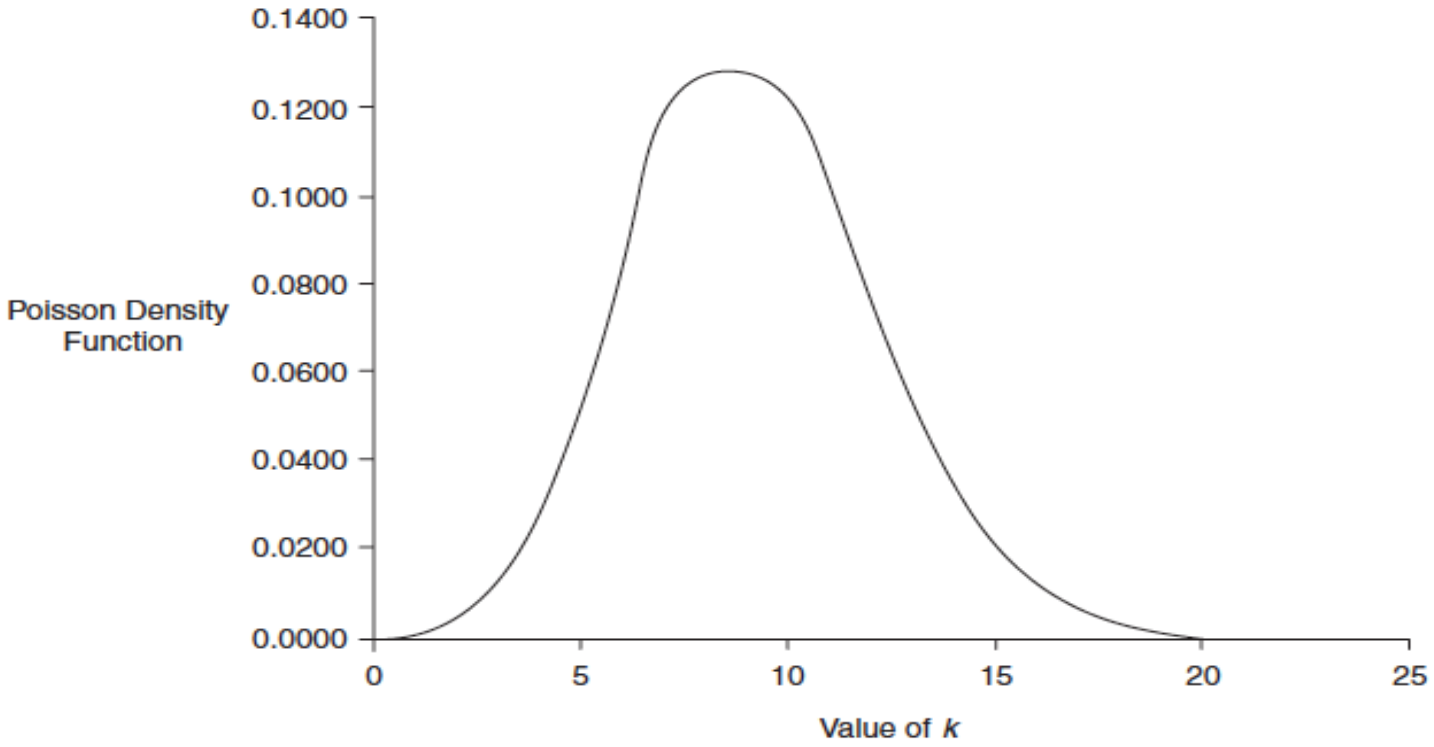


FIGURE 16.2 Poisson process density function.

2. Compound Poisson Process

Compound Poisson process is a stochastic process $X(t)$ where the arrival of events follows a Poisson process and each arrival is associated with another independent and identically distributed random variable Y_i.

Compound Poisson process $X(t)$ is a continuous-time stochastic process defined as

$$X(t) = \sum_{k=1}^{N(t)} Y_k$$

where $N(t)$ is a Poisson process with mean λt and Y_i are independent and identically distributed random variables with mean $E(Y_i)$ and variance $\text{Var}(Y_i)$.

The mean and variance of the compound Poisson process $X(t)$ are given by (Ross, 2010)

$$E[X(t)] = \mu_{X(t)} = \lambda t \times E(Y_i) \tag{16.6}$$

$$\text{Var}[X(t)] = \sigma^2_{X(t)} = \lambda t \times E(Y_i^2) = \lambda t \times (VAR(Y_i) + [E(Y_i)]^2) \tag{16.7}$$

For large t, we can show that the compound Poisson process follows an approximate normal distribution with mean $\mu_{X(t)}$ and standard deviation $\sigma_{X(t)}$.

Compound Poisson Process: Example

Customers arrive at an average rate of 12 per hour to withdraw money from an ATM and the arrivals follow a Poisson process. The money withdrawn are independent and identically distributed with mean and variance INR 4200 and 2,50,000, respectively. If the ATM has INR 6,00,000 cash, what is the probability that it will run out of cash in 10 hours?

Solution

Solution:

The mean and standard deviation of the compound Poisson process $X(t)$ can be calculated as described below:

Mean of compound Poisson process is

$$\mu_{X(t)} = \lambda t \times E(Y_i) = 12 \times 10 \times 4200 = 5,04,000$$

Variance of compound Poisson process is

$$\sigma^2_{X(t)} = \lambda t \times (\text{Var}(Y_i) + [E(Y_i)]^2) = 12 \times 10 \times (250000 + 4200^2) = 21468 \times 10^5$$

Standard deviation of compound Poisson process is

$$\sigma_{X(t)} = \sqrt{\sigma^2_{X(t)}} = \sqrt{21468 \times 10^5} = 46333.57$$

Probability that the cash withdrawal will exceed INR 6,00,000 is given by

$$P(X(t) \geq 6,00,000) = P\left(Z \geq \frac{6,00,000 - 504000}{46333.57}\right) = P(Z \geq 2.0719) = 0.0191$$

That is, there is approximately 2% chance that the ATM will run out of cash in 10 hours.

3. Markov Chains

The condition $P\left[X_{n+1}=j \mid X_0=i_0, X_1=i_1,\ldots, X_n=i\right]=P\left[X_{n+1}=j \mid X_n=i\right]$ is called Markov property named after the Russian mathematician *A A Markov*.

If the state space *S* is discrete then the stochastic process

$$\{X_n,\ n = 0, 1, 2, \ldots\}$$

that satisfies the condition is called a Markov chain

One Step Transition Probabilities of Markov Chains

Let $\{X_n,\ n = 0, 1, 2, \ldots\}$ be a Markov chain with state space *S*. Then the conditional probability

$$P\left[X_{n+1}=j \mid X_0=i_0, X_1=i_1,\ldots, X_n=i\right]=P\left[X_{n+1}=j \mid X_n=i\right]=P_{ij}$$

is called the one-step transition probability. P_{ij} gives conditional probability of moving from state *i* to stage *j* in one period.

One-Step Transition Probabilities of Markov Chain

P_{ij} gives conditional probability of moving from state i to stage j in one period.
One-step transition probabilities between all states in the state space are expressed in the form of one-step transition probability matrix as shown below

$$\mathbf{P} = \mathbf{P_{ij}} =$$

	1	2	…	n
1	P_{11}	P_{12}	…	P_{1n}
2	P_{21}	P_{22}	…	P_{2n}
…	…	…	…	…
n	P_{n1}	P_{n2}	…	P_{nn}

m-step transition probability

An *m*-step transition probability in a Markov chain is given by

$$P_{ij}^{(m)} = P(X_{n+m} = j \mid X_n = i)$$

The m-step transition probability $P_{ij}^{(m)}$ can be written as

$$P_{ij}^{(m)} = \sum_{r=1}^{n} P_{ir}^{k} \times P_{rj}^{(m-k)}, \quad 0 < k < m$$

Estimation of One-Step Transition Probabilities of Markov Chain

Transition probabilities of a Markov chain are estimated using maximum likelihood estimate (MLE) from the transition data (Anderson and Goodman, 1957).

The MLE estimate of the transition probability P_{ij} (probability of moving from state i to state j in one step) is given by

$$\hat{P}_{ij} = \frac{N_{ij}}{\sum_{k=1}^{m} N_{ik}}$$

where N_{ij} is number of cases in which $X_n = i$ (state at time n is i) and $X_{n+1} = j$ (state at time $n + 1$ is j).

Hypothesis Tests for Markov Chain: Anderson Goodman Test

The null and alternative hypotheses to check whether the sequence of random variables follows a Markov chain is stated below

H_0: The sequences of transitions (X_1, X_2, …, X_n) are independent (zero-order Markov chain)

H_A: The sequences of transitions (X_1, X_2, …, X_n) are dependent (first-order Markov chain)

The corresponding test statistic is

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \left(\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \right)$$

where

O_{ij} = Observed number of transitions from state i to state j in one period.

E_{ij} = Expected number of transitions from state i to state j assuming independence.

DATA ANALYTICS

Testing Time Homogeneity of Transition Matrices: Likelihood Ratio Test

Anderson and Goodman (1957) suggested a likelihood ratio test for checking whether the transition probability matrices are time homogeneous. The null and alternative hypotheses of the likelihood ratio tests are

$$H_0: P_{ij}(t) = P_{ij},\ t = 1, 2, 3, 4, \text{and } 5$$
$$H_A: P_{ij}(t) \neq P_{ij},\ t = 1, 2, 3, 4, \text{and } 5$$

The test statistic is a likelihood test ratio statistic and is given by (Anderson and Goodman, 1957):

$$\lambda = \prod_t \prod_{i,j} \left[\frac{\hat{P}_{ij}}{\hat{P}_{ij}(t)} \right]^{n_{ij}(t)}$$

Summary

1.Most problems in analytics are dynamic in nature and thus require collection of random variables to model the problem.

2. Stochastic process is a collection of random variables usually indexed by time t and used while modelling problems that are not independent and identically distributed.

3. Poisson process is a counting process that is used in decision-making scenarios such as capacity planning and spare parts demand forecasting. Compound Poisson process can be used to study problems such as cash replenishments at ATMs, total insurance claims, etc.

4. Markov chain is one of the most powerful models in analytics with applications across industry sectors. Google's PageRank algorithm is based on Markov chain.

5. Asset availability, market share, customer retention probability, and customer lifetime value are few applications of Markov chain in analytics.

References

Text Book:

Chapter 16.1 -16.4.4 “Business Analytics,

The Science of Data-Driven Decision Making”, by U. Dinesh Kumar, Wiley 2017

DATA ANALYTICS

Unit 5: Advanced Techniques

Bharathi R

Department of Computer Science and Engineering

DATA ANALYTICS

Unit 5: Markov chains in Predictive Analytics

Bharathi R

Department of Computer Science and Engineering

Markov Chains in Predictive Analytics

One of the primary applications of Markov chain is predicting the values of Xn in the future.
For example, assume that the initial distribution of customers in 4 states is

$P_I = (450, 225, 175, 150)$.

Assume that the one-step transition matrix P is as

$$\mathbf{P} = \begin{pmatrix} & 1 & 2 & 3 & 4 \\ 1 & 0.8189 & 0.0882 & 0.0472 & 0.0457 \\ 2 & 0.1128 & 0.7180 & 0.0902 & 0.0789 \\ 3 & 0.2077 & 0.0984 & 0.6011 & 0.0929 \\ 4 & 0.0663 & 0.0964 & 0.0964 & 0.7410 \end{pmatrix}$$

Using Chapman–Kolmogorov relationship [Eq. (16.12)], we can show that the distribution of customers after n periods is given by $\mathbf{P}_I \times \mathbf{P}^n$, where $\mathbf{P}_I$ is the initial distribution of customers across various states and $\mathbf{P}$ is the one-step transition matrix. For example, the distribution of customers after 4 weeks is $\mathbf{P}_I \times \mathbf{P}^4$. That is

$$\begin{pmatrix} 450 & 225 & 175 & 150 \end{pmatrix} \times \begin{pmatrix} 0.8189 & 0.0882 & 0.0472 & 0.0457 \\ 0.1128 & 0.7180 & 0.0902 & 0.0789 \\ 0.2077 & 0.0849 & 0.6011 & 0.0929 \\ 0.0663 & 0.0964 & 0.0964 & 0.7410 \end{pmatrix}^4 = \begin{pmatrix} 417.84 & 243.06 & 150.69 & 188.41 \end{pmatrix}$$

So after 4 periods, the distribution of customers will be

State 1: 417.84; State 2: 243.06; State 3: 150.69; and State 4: 188.41

Stationary Distribution in a Markov Chain

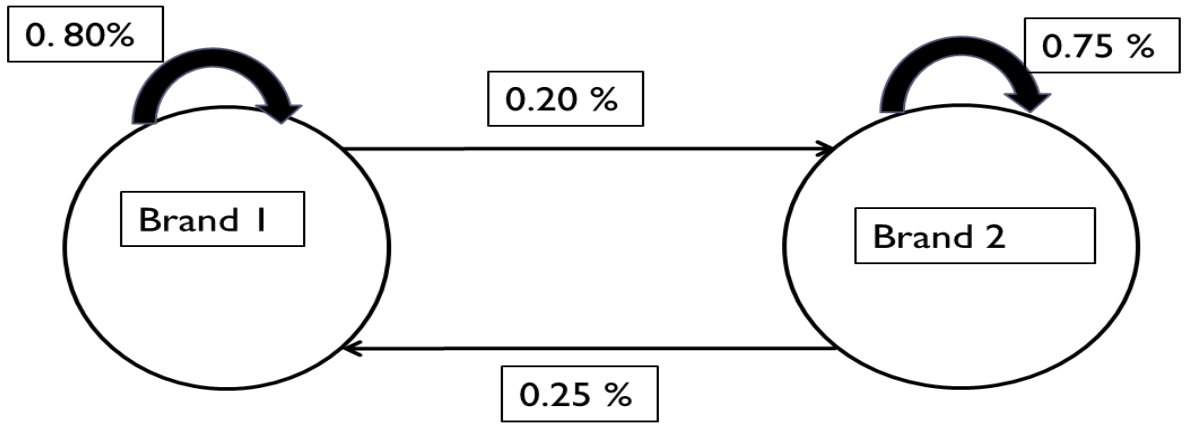
Consider brand switching between two brands (B1 and B2) and let the initial market share be as shown in the following vector:

$P_I = (0.2\ 0.8)$

Transition probability

	Brand 1	Brand 2
Brand 1	0.80	0.20
Brand 2	0.25	0.75

State transition diagram between brands.



DATA ANALYTICS

Stationary Distribution in a Markov Chain

In Table given, both rows of the matrix $\mathbf{P^n}$ converge to 0.555556 and 0.444444 as the value of n increases.

The market share of brands 1 and 2 converges to 0.555556 and 0.444444, respectively.

The values (0.555556, 0.444444) are the stationary probability distribution of the Markov chain or equilibrium probabilities.

The values can be interpreted as long-run market shares of the brands.

TABLE 16.7 Shows the values of P^n and the market share of brands after *n* periods (P_iP^n)

		Brand 1	Brand 2	Market Share *n* Periods		
P^1		0.2	0.8			
		Brand 1	Brand 2		Brand 1	Brand 2
P	Brand 1	0.8	0.2	1 (P_iP^1)	0.36	0.64
	Brand 2	0.25	0.75			
P^2	Brand 1	0.69	0.31	2 (P_iP^2)	0.448	0.552
	Brand 2	0.3875	0.6125			
P^4	Brand 1	0.596225	0.403775	4 (P_iP^4)	0.52302	0.47698
	Brand 2	0.504719	0.495281			
P^8	Brand 1	0.559277	0.440723	8 (P_iP^8)	0.552578	0.447422
	Brand 2	0.550904	0.449096			
P^{16}	Brand 1	0.555587	0.444413	16 (P_iP^{18})	0.555531	0.444469
	Brand 2	0.555517	0.444483			
P^{32}	Brand 1	0.555556	0.444444	32 (P_iP^{32})	0.555556	0.444444
	Brand 2	0.555556	0.444444			
P^{64}	Brand 1	0.555556	0.444444	64 (P_iP^{64})	0.555556	0.444444
	Brand 2	0.555556	0.444444			

Stationary Distribution in a Markov Chain

Let $\pi = (\pi_1, \pi_2, \ldots, \pi_m)$ be the stationary distribution. Then it satisfies the following system of equations:

$$\pi_j = \sum_{k=1}^{m} \pi_k P_{kj} \tag{16.20}$$

$$\sum_{k=1}^{m} \pi_k = 1 \tag{16.21}$$

The system of equations in Eq. (16.20) can be written as

$$\pi = \pi \mathrm{P} \tag{16.22}$$

Stationary Distribution in a Markov Chain

The stationary distribution equation for the matrix in Transition Table is given by

$$(\pi_1 \quad \pi_2) = (\pi_1 \quad \pi_2)\begin{pmatrix} 0.80 & 0.20 \\ 0.25 & 0.75 \end{pmatrix}$$

That is

$$\pi_1 = 0.80\pi_1 + 0.25\pi_2$$
$$\pi_2 = 0.20\pi_1 + 0.75\pi_2$$

Since π_1 and π_2 are probabilities, we have

$$\pi_1 + \pi_2 = 1$$

$$0.20\pi_1 - 0.25\pi_2 = 0$$
$$\pi_1 + \pi_2 = 1$$

Solving the above system of equations, we get $\pi_1 = 0.555556$ and $\pi_2 = 0.444444$. That is, in the long run, the markets shares of brand 1 and brand 2 will converge to 0.555556 and 0.444444, respectively. The stationary distribution will be independent of the initial probability distribution P_I.

DATA ANALYTICS

Regular Matrix

A matrix P is called a regular matrix, when for some *n*, all entries of P^n will be greater than zero, that is for some *n* $P_{ij}^{n} > 0$

Consider the matrix:

$$\mathbf{P} = \begin{pmatrix} 0.2 & 0 & 0.8 \\ 0.5 & 0 & 0.5 \\ 0.3 & 0.7 & 0 \end{pmatrix}$$

Then

$$\mathbf{P^2} = \begin{pmatrix} 0.28 & 0.56 & 0.16 \\ 0.25 & 0.35 & 0.4 \\ 0.41 & 0 & 0.59 \end{pmatrix} \text{ and } \mathbf{P^3} = \begin{pmatrix} 0.384 & 0.112 & 0.504 \\ 0.345 & 0.280 & 0.375 \\ 0.259 & 0.413 & 0.328 \end{pmatrix}$$

Note that although the matrix P has zero entries ($P_{12} = P_{22} = P_{33} = 0$), in $\mathbf{P^3}$ all entries are greater than zero, thus matrix **P** is a **regular matrix**. A regular matrix will have stationary distribution and satisfy the system of equations as shown below,

$$\pi_j = \sum_{k=1}^{m} \pi_k P_{kj} \qquad \sum_{k=1}^{m} \pi_k = 1$$

Example

The number of flights cancelled by an airline daily is modelled using a Markov chain. The states of the chain and the description of states are given in Table 16.8. The revenue loss (in millions of rupees) due to cancellation of flights in various states is given in Table 16.9. The transition probability matrix between states is shown in Table 16.10.

(a) If there are no cancellations initially, what is the probability that there will be at least one cancellation after 2 days?

(b) Calculate the steady-state expected loss due to cancellation of flights.

TABLE 16.8 States representing cancellation of flights

State	Description
0	No cancellations
1	One cancellation
2	Two cancellations
3	More than 2 cancellations

TABLE 16.9 Revenue loss due to cancellations

State	0	1	2	3
Loss	0	4.5	10.0	16.0

TABLE 16.10 State transition matrix between flight cancellations

	0	1	2	3
0	0.45	0.30	0.20	0.05
1	0.15	0.60	0.15	0.10
2	0.10	0.30	0.40	0.20
3	0	0.10	0.70	0.20

Example

Solution:

(a) If there are no cancellations initially, then the initial state vector is $P_I = [1\ 0\ 0\ 0]$. The probability distribution after two days is

$$P_I P^2 = \begin{pmatrix}1 & 0 & 0 & 0\end{pmatrix}\begin{pmatrix}0.45 & 0.30 & 0.20 & 0.05\\ 0.15 & 0.60 & 0.15 & 0.10\\ 0.10 & 0.30 & 0.40 & 0.20\\ 0 & 0.10 & 0.70 & 0.20\end{pmatrix}^2 = \begin{pmatrix}0.2675 & 0.38 & 0.25 & 0.1025\end{pmatrix}$$

Probability that there will be at least one cancellation after 2 days $= 0.38 + 0.25 + 0.1025$
$= 0.7325$.

Example

(b) To calculate the steady-state expected loss, we have to calculate the steady-state distribution. The steady-state distribution will satisfy the following system of equations:

$$\pi_0 = 0.45\pi_0 + 0.15\pi_1 + 0.10\pi_2$$
$$\pi_1 = 0.30\pi_0 + 0.60\pi_1 + 0.30\pi_2 + 0.10\pi_3$$
$$\pi_2 = 0.20\pi_0 + 0.15\pi_1 + 0.40\pi_2 + 0.70\pi_3$$
$$\pi_3 = 0.05\pi_0 + 0.10\pi_1 + 0.20\pi_2 + 0.20\pi_3$$
$$\pi_0 + \pi_1 + \pi_2 + \pi_3 = 1$$

TABLE 16.10 State transition matrix between flight cancellations

	0	1	2	3
0	0.45	0.30	0.20	0.05
1	0.15	0.60	0.15	0.10
2	0.10	0.30	0.40	0.20
3	0	0.10	0.70	0.20

Solving the above system of equations we get

$$(\pi_0 \quad \pi_1 \quad \pi_2 \quad \pi_3) = (0.163 \quad 0.390 \quad 0.311 \quad 0.137)$$

The steady-state expected loss is $\sum_{i=0}^{3} \pi_i \times L_i$, where L_i is the expected revenue loss in state i (Table 16.9). Hence

$$\sum_{i=0}^{3} \pi_i \times L_i = 0.163 \times 0 + 0.390 \times 4.5 + 0.311 \times 10 + 0.137 \times 16 = 7.05$$

DATA ANALYTICS

Classification of States in a Markov Chain

Not all Markov chains will have stationary probability distribution.
To derive the necessary and sufficient conditions for existence of stationary distribution of a Markov chain, we have to understand different classes of states that exist in a Markov chain.

The classes of states are

- Accessible state
- Communicating state
- Recurrent and Transient state
- Positive recurrent and Null-Recurent state
- Periodic and Aperiodic state

Accessible State

A state *j* is accessible (or reachable) from state *i* if there exists a *n* such that

$$P_{ij}^{n} > 0$$

That is, there exists a path from state *i* to state *j*.

Communicating States

Two states i and j are communicating states when there exists n and m such that

$$P_{ij}^{n} > 0 \text{ and } P_{ji}^{m} > 0.$$

That is, state j can be reached (accessible) from state i and
similarly state i can be reached from state j.

A Markov chain is called irreducible if all states of the chain communicate with each other.

Recurrent and Transient States

A state i of a Markov chain is called a recurrent state when

$$\sum_{n=1}^{\infty} P_{ii}^{n} = \infty$$

That is, if the state i is recurrent then the Markov chain will visit state i infinite number of times in the long run.

If state i is recurrent and states i and j are communicating states, then state j is also a recurrent state.

A state k of a Markov chain is called a transient state when

$$\sum_{n=1}^{\infty} P_{kk}^{n} < \infty$$

That is, state k is called a transient state when $\sum_{n=1}^{\infty} P_{kk}^{n} < \infty$ is finite.

This means it is possible that the Markov chain may not return to state k in the long run.

First Passage Time and Mean Recurrence Time

First passage time is the probability that the Markov chain will enter state *i* exactly after *n* steps for the first time after leaving state *i*, that is

$$f_{ii}^{n} = P\left[X_n = i, X_k \neq i, k = 1,2,\ldots,n-1 \mid X_0 = i\right]$$

Mean recurrence time is the average time taken to return to state *i* after leaving state *i*. Mean recurrence time μ_{ii} is given by

$$\mu_{ii} = \sum_{n=1}^{\infty} n \times f_{ii}^{n}$$

If the mean recurrence time is finite (μ_{ii} is finite), then the recurrent state is called a **positive recurrent state** and if it is infinite then it is called **null-recurrent state**.

Periodic and Aperiodic State

Periodic state is a special case of recurrent state in which $d(i)$ is the greatest common divisor of n such that

$$P_{ii}^{n} > 0$$

If $d(i) = 1$, it is called aperiodic state and if $d(i) \geq 2$, then it is called a periodic state.

In the matrix shown in Table, for state 1

$$P_{11}^{3} = 1, P_{11}^{6} = 1, P_{11}^{9} = 1, \text{ and } P_{11}^{n} = 0$$

when n is not a multiple of 3. That is, the greatest common divisor is 3 which means that the periodicity is 3.

TABLE 16.11 Transition matrix

		1	2	3
	1	0	1	0
$P=$	2	0	0	1
	3	1	0	0

Ergodic Markov Chain

A state *i* of a Markov chain is ergodic when it is positive recurrent and aperiodic. Markov chain in which all states are positive recurrent and aperiodic is called an ergodic Markov chain.

For an ergodic Markov chain, a stationary distribution exists that satisfies the system of equations

$$\pi_j = \sum_{k=1}^{m} \pi_k P_{kj}$$

$$\sum_{k=1}^{m} \pi_k = 1$$

Limiting Probability

In a Markov chain, the limiting probability is given by:

$$\lim_{n \to \infty} p_{ij}^{n}$$

The main difference between limiting probability and stationary distribution is that, stationary distribution when exists is unique. Whereas limiting probability may not be unique.

DATA ANALYTICS

Unit 5: Markov chains contd.

Bharathi R

Department of Computer Science and Engineering

Markov Chains with Absorbing States

A state *i* of a Markov chain is called an absorbing state when $P_{ii} = 1$, that is if the system enters this state, it will remain in the same state. Many real-life problems such as bad debt (non-performing assets), bankruptcy, customer churn, employee attrition and so on can be modelled using absorbing state Markov chain.

Absorbing state Markov chain is a Markov chain in which there is at least one state k such that $P_{kk} = 1$. Non-absorbing states in an absorbing state Markov chain are transient states.

While using absorbing state Markov chains in analytics problem solving, we would like to learn the following from an absorbing state Markov chain:

1. The probability of eventual absorption to a specific absorbing state (when there are more than one absorbing states) from various transient states of the Markov chain.

2. The expected time to absorption from a transient state to absorbing states.

The above questions are answered using canonical form of the transition matrix.

Canonical Form of the Transition Matrix of an Absorbing State Markov Chain

The rows of the transition probability matrix of an absorbing state Markov chain can be rearranged such that the top rows are assigned for absorbing states followed by transient states (the idea here is to group the absorbing state and non-absorbing states).

Let A and T be the set of absorbing and transient states, respectively, in the Markov chain.

Then the transition probability matrix can be arranged such that

P =		A	T
	A	I	0
	T	R	Q

The Matrix P

The matrix P is divided into 4 matrices I, 0, R, and Q, where

- Matrix **I** is the identity matrix. It corresponds to transition within absorbing states.

- Matrix **0** is a matrix in which all elements are zero. Here the elements correspond to transition between an absorbing state and transient states.

- Matrix **R** represents the probability of absorption from a transient state to an absorbing state.

- Matrix **Q** represents the transition between transient states.

Expected time to absorption

To calculate the eventual probability of absorption, we would like to calculate the long-run (limiting probability) value of R in the above matrix. When we multiply the canonical form of the matrix, we get

$$\mathbf{P^n} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \left(\sum_{k=0}^{n-1} \mathbf{Q}^k\right)\mathbf{R} & \mathbf{Q^n} \end{pmatrix}$$

For large n, the matrix $\left(\sum_{k=0}^{n-1} \mathbf{Q}^k\right)\mathbf{R}$ will give the probability of eventual absorption to an absorbing state.

As $n \to \infty$, we can show that $\sum_{k=0}^{n-1} \mathbf{Q}^k = \mathbf{F} = (\mathbf{I}-\mathbf{Q})^{-1}$. The matrix F is called the fundamental matrix and the matrix FR is the probability of eventual absorption into an absorbing state from a transient state. The expected time to absorption is given by

Expected time to absorption = Fc

where c is a unit vector. That is, the row sum of the fundamental matrix gives the expected duration for absorption (that is, expected time it takes to reach an absorbing state from a transient state).

Example 16.5

Airwaves India (AI) is a mobile phone service provider based in Allahabad, India that provides several value-added services such as mobile data, video conferencing, etc. The market is highly competitive and AI faces high churn rate among its customers. The customers of AI are categorized into different states as listed below:

STATE 1 : Customer churn that generated no revenue/profit
STATE 2 : Customer churn that generated INR 200 profit per month on average (customer uses the service only for incoming calls and data)
STATE 3 :Customer state that generated INR 300 profit per month on average
STATE 4 :Customer state that generated INR 400 profit per month on average
STATE 5 :Customer state that generated INR 600 profit per month on average
STATE 6 :Customer state that generated INR 800 profit per month on average

Example 16.5

The transition probability values between different states are shown in Table 16.12.

TABLE 16.12 Transition probability matrix (based on monthly data)

	1	2	3	4	5	6
1	1	0	0	0	0	0
2	0	1	0	0	0	0
3	0.05	0.05	0.90	0	0	0
4	0.10	0.05	0	0.80	0.05	0
5	0.20	0.10	0	0.05	0.60	0.05
6	0.10	0.20	0	0	0	0.70

(a) If a customer is in state 6, calculate the probability of eventual absorption in state 2?
(b) Calculate the expected value of time taken to absorption if the current state is 4.

Solution: 16.5

(a) To calculate the probability of absorption of a customer in state 6 to state 2, we have to calculate **FR**.

TABLE 16.12 Transition probability matrix (based on monthly data)

	1	2	3	4	5	6
1	1	0	0	0	0	0
2	0	1	0	0	0	0
3	0.05	0.05	0.90	0	0	0
4	0.10	0.05	0	0.80	0.05	0
5	0.20	0.10	0	0.05	0.60	0.05
6	0.10	0.20	0	0	0	0.70

The matrix Q is given by

$$Q = \begin{bmatrix} 0.9 & 0 & 0 & 0 \\ 0 & 0.8 & 0.05 & 0 \\ 0 & 0.05 & 0.6 & 0.05 \\ 0 & 0 & 0 & 0.7 \end{bmatrix}$$

$$I - Q = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 0.9 & 0 & 0 & 0 \\ 0 & 0.8 & 0.05 & 0 \\ 0 & 0.05 & 0.6 & 0.05 \\ 0 & 0 & 0 & 0.7 \end{bmatrix} = \begin{bmatrix} 0.1 & 0 & 0 & 0 \\ 0 & 0.2 & -0.05 & 0 \\ 0 & -0.05 & 0.4 & -0.05 \\ 0 & 0 & 0 & 0.3 \end{bmatrix}$$

$$F = (1 - Q)^{-1} = \begin{bmatrix} 10 & 0 & 0 & 0 \\ 0 & 5.1613 & 0.6452 & 0.1075 \\ 0 & 0.6452 & 2.5806 & 0.4301 \\ 0 & 0 & 0 & 3.3333 \end{bmatrix}$$

Solution: 16.5

Probability of absorption FR is given by

$$FR = \begin{bmatrix} 10 & 0 & 0 & 0 \\ 0 & 5.1613 & 0.6452 & 0.1075 \\ 0 & 0.6452 & 2.5806 & 0.4301 \\ 0 & 0 & 0 & 3.3333 \end{bmatrix} \times \begin{bmatrix} 0.05 & 0.05 \\ 0.1 & 0.05 \\ 0.2 & 0.1 \\ 0.1 & 0.2 \end{bmatrix} = \begin{bmatrix} 0.5 & 0.5 \\ 0.6559 & 0.3441 \\ 0.6237 & 0.3763 \\ 0.3333 & 0.6667 \end{bmatrix}$$

That is, if the current customer state is 6, the probability of absorption into churn state 2 is 0.6667.

(b) Expected value of time to absorption is given by Fc:

$$Fc = \begin{bmatrix} 10 & 0 & 0 & 0 \\ 0 & 5.1613 & 0.6452 & 0.1075 \\ 0 & 0.6452 & 2.5806 & 0.4301 \\ 0 & 0 & 0 & 3.3333 \end{bmatrix} \times \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 10 \\ 5.91 \\ 3.65 \\ 3.33 \end{bmatrix}$$

Expected value of time to absorption when the current state is 4 is 5.91 months.

References

Text Book:

"Business Analytics, The Science of Data-Driven Decision Making", U. Dinesh Kumar, Wiley 2017

Markov chains contd (absorbing states, expected duration to reach a state) [ch 16.4.5 - ch 16.8]

THANK YOU

Prof Bharathi R

Professor, Department of Computer Science

rbharathi@pes.edu

DATA ANALYTICS

Unit 5: Advanced Techniques

Bharathi R

Associate Professor, Department of Computer Science and Engineering

DATA ANALYTICS

Unit 5: Markov chains contd.

Bharathi R

Department of Computer Science and Engineering

Expected Duration to Reach a State from Other States

The expected duration to reach a state *j* from state *i* can be calculated by solving the following system of equations:
Then, $E_{i,j}$ satisfies the following system of equations:

$$E_{i,j} = 1 + \sum_k P_{i,k} E_{k,j} \quad \forall\, i,\ i \neq j$$

$$\mathbf{E_{j,j} = 0}$$

Example 16.6

The transition probability matrix calculated based on monthly data is shown in Table 16.13. Calculate the expected duration (in months) for the process to reach state 7 from state 4.
The percentage of non-performing assets at a bank is classified into the following seven states:

State	State Description
1	NPA is less than 1%
2	NPA is between 1% and 2%
3	NPA is between 2% and 3%
4	NPA is between 3% and 4%
5	NPA is between 4% and 5%
6	NPA is between 5% and 6%
7	NPA greater than 6%

TABLE 16.13 Transition probability matrix between NPA states

	1	2	3	4	5	6	7
1	0.95	0.05	0	0	0	0	0
2	0.10	0.85	0.05	0	0	0	0
3	0	0.10	0.80	0.10	0	0	0
4	0	0	0.15	0.70	0.15	0	0
5	0	0	0	0.15	0.65	0.20	0
6	0	0	0	0	0.20	0.60	0.20
7	0	0	0	0	0	0.10	0.90

Solution for Example 16.6

Let $E_{4,7}$ be the expected number of duration for the process to reach state 7 from state 4. Then it satisfies the following system of equations:

$$E_{4,7} = 1 + 0.15E_{3,7} + 0.70E_{4,7} + 0.15E_{5,7}$$
$$E_{3,7} = 1 + 0.10E_{2,7} + 0.80E_{3,7} + 0.10E_{4,7}$$
$$E_{5,7} = 1 + 0.15E_{4,7} + 0.65E_{5,7} + 0.20E_{6,7}$$
$$E_{6,7} = 1 + 0.20E_{5,7} + 0.60E_{6,7} + 0.20E_{7,7}$$
$$E_{2,7} = 1 + 0.10E_{1,7} + 0.85E_{2,7} + 0.05E_{3,7}$$
$$E_{1,7} = 1 + 0.95E_{1,7} + 0.05E_{2,7}$$
$$E_{7,7} = 0$$

TABLE 16.13 Transition probability matrix between NPA states

	1	2	3	4	5	6	7
1	0.95	0.05	0	0	0	0	0
2	0.10	0.85	0.05	0	0	0	0
3	0	0.10	0.80	0.10	0	0	0
4	0	0	0.15	0.70	0.15	0	0
5	0	0	0	0.15	0.65	0.20	0
6	0	0	0	0	0.20	0.60	0.20
7	0	0	0	0	0	0.10	0.90

Solving the system of equations we get $E_{4,7}$ = 206.6667.
That is, it takes approximately 207 months on average for the process to reach state 7 from state 4.

Calculation of Retention Probability and Customer Lifetime Value using Markov Chains

CLV is the net present value (NPV) of the future margin generated from its customers or customer segments. CLV is calculated usually at a customer segment level.

Ching *et al.* (2004) showed that the steady-state retention probability can be calculated using

$$R_t = \sum_{i=1}^{n} \frac{\pi_i}{\left(\sum_{j=1}^{n} \pi_j\right)} (1 - P_{i0}) = 1 - \frac{\pi_0 (1 - P_{00})}{1 - \pi_0}$$

where R_t is the steady-state retention probability.

DATA ANALYTICS

Calculation of Retention Probability and Customer Lifetime Value using Markov Chains

The customer lifetime value for *N* periods is given by (Pfeifer and Carraway, 2000):

$$CLV = \sum_{t=0}^{N} \frac{\mathbf{P_I} \times \mathbf{P^t R}}{(1+i)^t}$$

where

P_I is the initial distribution of customers in different states,

P is the transition probability matrix,

R is the reward vector (margin generated in each customer segments)

i is the interest rate

Note: discount rate d = 1/(1+i) is the discount factor

Example :16.7

The customers of Dubai Data Services (DDS) are classified into five categories as shown in Table 16.14 along with transition probability matrix. State 0 represents non-customers and the remaining states are different customer segments created based on the revenue generated. The average margin generated in different states is shown in Table 16.15 along with initial distribution of customers in millions.

Calculate the steady-state retention probability and CLV for 6 periods (N = 5) using a discount factor of d = 0.95.

TABLE 16.14 Customer states of DDS and transition matrix

	0	1	2	3	4
0	0.80	0.10	0.10	0	0
1	0.10	0.60	0.20	0.10	0
2	0.15	0.05	0.75	0.05	0
3	0.20	0	0.10	0.60	0.10
4	0.30	0	0	0.05	0.65

TABLE 16.15 Margin generated in different states

State	0	1	2	3	4
Average Margin	0	120	300	450	620
Customers (in millions)	55.8	6.5	4.1	2.3	1.6

Solution

The stationary distribution equations are

$$\pi_0 = 0.8\pi_0 + 0.10\pi_1 + 0.15\pi_2 + 0.20\pi_3 + 0.30\pi_4$$
$$\pi_1 = 0.1\pi_0 + 0.60\pi_1 + 0.05\pi_2$$
$$\pi_2 = 0.1\pi_0 + 0.2\pi_1 + 0.75\pi_2 + 0.10\pi_3$$
$$\pi_3 = 0.10\pi_1 + 0.05\pi_2 + 0.60\pi_3 + 0.05\pi_4$$
$$\pi_0 + \pi_1 + \pi_2 + \pi_3 + \pi_4 = 1$$

Solving the above system of equations, we get $\pi_0 = 0.4287$. The steady-state retention probability R_t is

$$R_t = 1 - \frac{\pi_0(1 - P_{00})}{1 - \pi_0} = 1 - \frac{0.4287 \times (1 - 0.80)}{1 - 0.4287} = 0.85$$

Customer lifetime value for $N = 5$ is

$$CLV = \sum_{t=0}^{5} \frac{\mathbf{P_I} \times \mathbf{P^t R}}{(1+i)^t}$$

where

$$P_I = \begin{pmatrix} 55.8 & 6.5 & 4.1 & 2.3 & 1.6 \end{pmatrix}$$

Reward vector $$\mathbf{R} = \begin{pmatrix} 0 \\ 120 \\ 300 \\ 450 \\ 620 \end{pmatrix}$$

TABLE 16.15 Margin generated in different states

State	0	1	2	3	4
Average Margin	0	120	300	450	620
Customers (in millions)	55.8	6.5	4.1	2.3	1.6

Substituting the values in CLV equation, we get CLV = 40181.59.

References

Text Book:

"Business Analytics, The Science of Data-Driven Decision Making", U. Dinesh Kumar, Wiley 2017

Markov chains contd (absorbing states, expected duration to reach a state) [ch 16.4.5 - ch 16.8]

DATA ANALYTICS

Unit 5: AB Testing

Bharathi R

Department of Computer Science and Engineering

1. What is A/B Testing?

- **A/B testing** (also known as **bucket testing** or **split-run testing**) is a user experience research methodology.
- A/B tests consist of a randomized experiment with two variants, A and B.
- It includes application of statistical hypothesis testing or "two-sample hypothesis testing" as used in the field of statistics.
- A/B testing is a way to compare two versions of a single variable, typically by testing a subject's response to variant A against variant B, and determining which of the two variants is more effective.

Outline

1. What is A/B Testing?
2. A/B Testing Examples in Popular Industries
3. Why Should You A/B Test?
4. How to Perform an A/B Test?
5. What Can You A/B Test?
6. 9 Mistakes to Avoid While A/B Testing
7. 6 Challenges of A/B Testing
8. How To Make an A/B Testing Calendar?
9. Summary

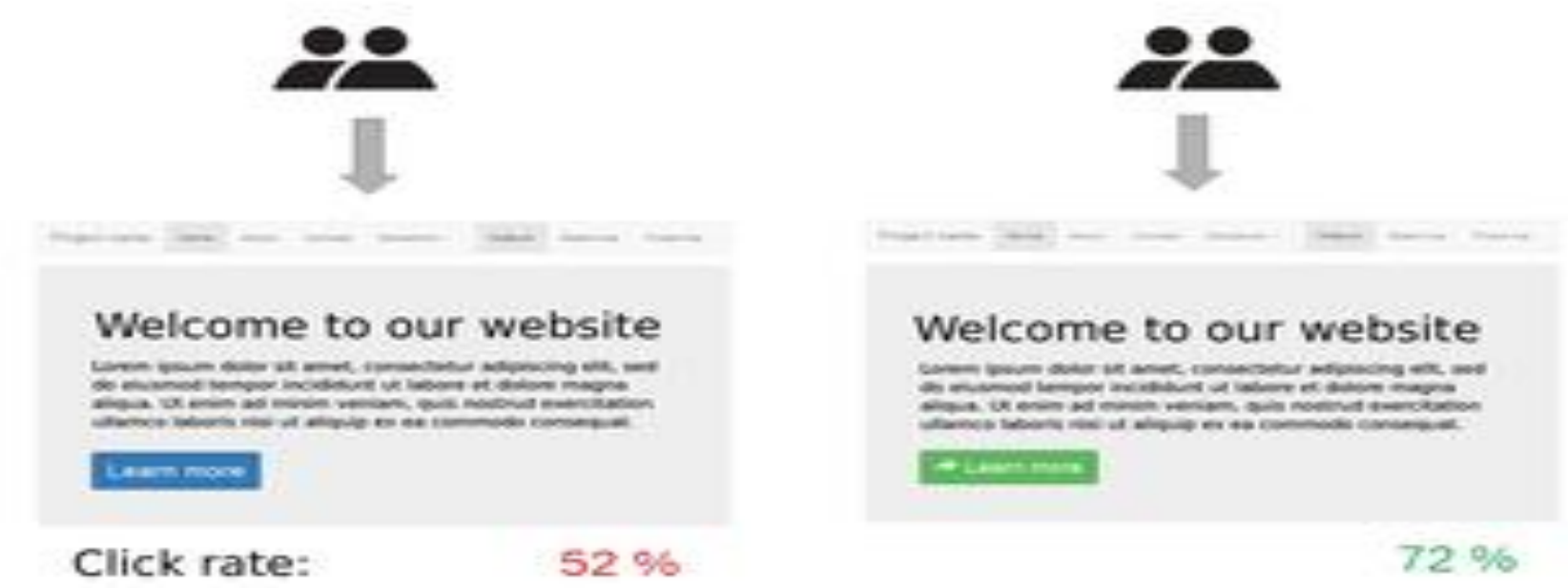
1. What is A/B Testing?

- A/B test is the shorthand for a simple controlled experiment.
- As the name implies, two versions (A and B) of a single variable are compared, which are identical except for one variation that might affect a user's behavior.
- A/B tests are widely considered the simplest form of controlled experiment.
- However, by adding more variants to the test, this becomes more complex.

DATA ANALYTICS

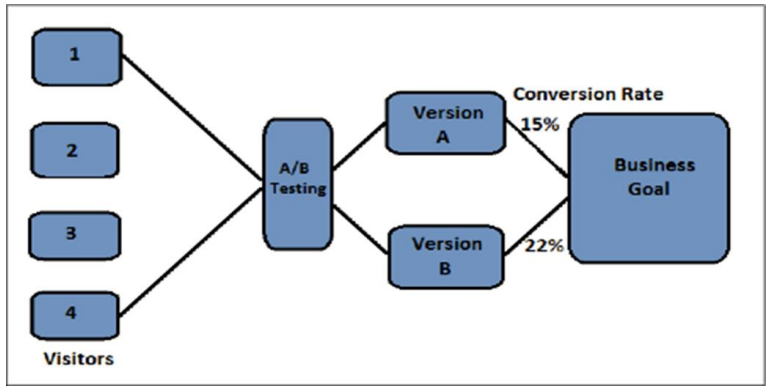
History

Example of A/B testing on a website. By randomly serving visitors two versions of a website that differ only in the design of a single button element, the relative efficacy of the two designs can be measured.



Conversion Rate

Let us assume that there is a web page and all the traffic is directed to this page. Now as a part of A/B Testing, you have made some minor changes like headlines, numbering, etc. on the same page and half of its traffic is directed to the modified version of this web page. Now you have version A and version B of the same web page and you can monitor the visitor's actions using statistics and analysis to determine the version that yields a higher conversion rate.



A conversion rate is defined as the instance, when any visitor on your website performs a desired action.

A/B Testing enables you to determine the best online marketing strategy for your business.

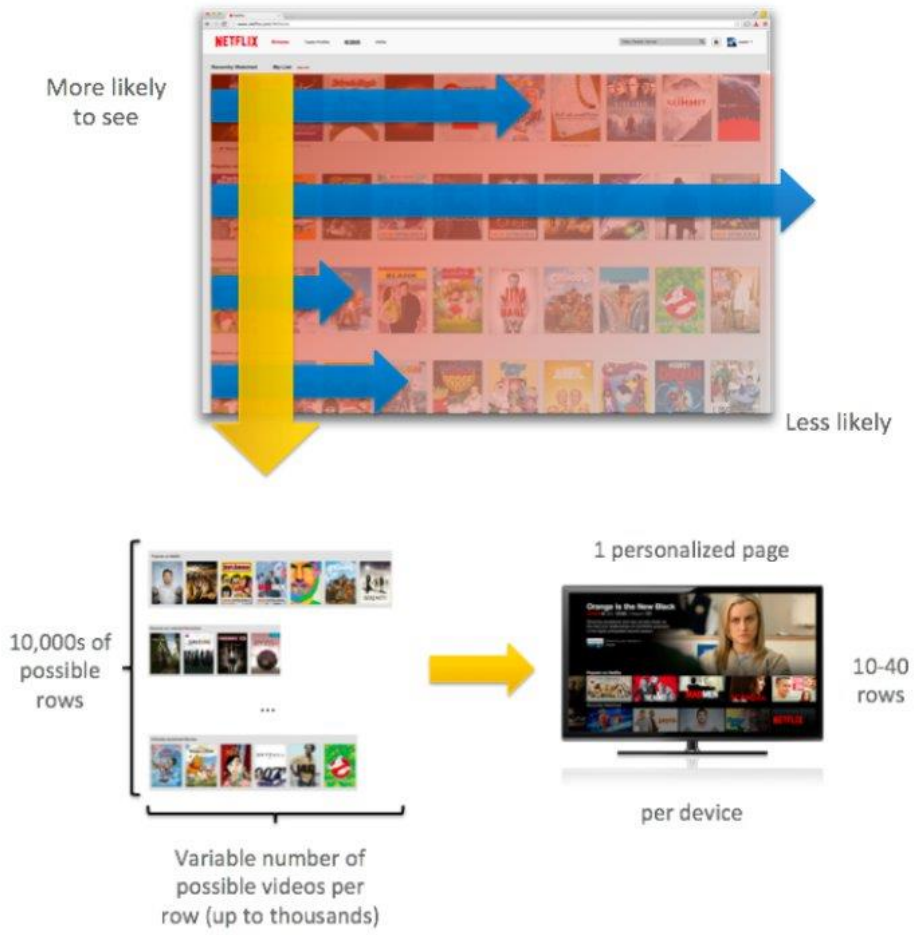
Take a look at the illustration. It shows that version A yields a conversion rate of 15% and version B yields a conversion rate of 22%.

2. A/B Testing Examples in Popular Industries

1. A/B testing in Media & Publishing Industry
2. A/B Testing in eCommerce Industry
3. A/B Testing in Travel Industry
4. A/B Testing in B2B/SaaS Industry

2. 1. A/B testing in Media & Publishing Industry

- Netflix uses personalization extensively for its homepage.
- Based on each user's profile, Netflix personalizes the homepage to provide the best user experience to each user.
- They decide how many rows go on the homepage and which shows/movies go into the rows based on the users streaming history and preferences.



2.2 A/B Testing in eCommerce Industry

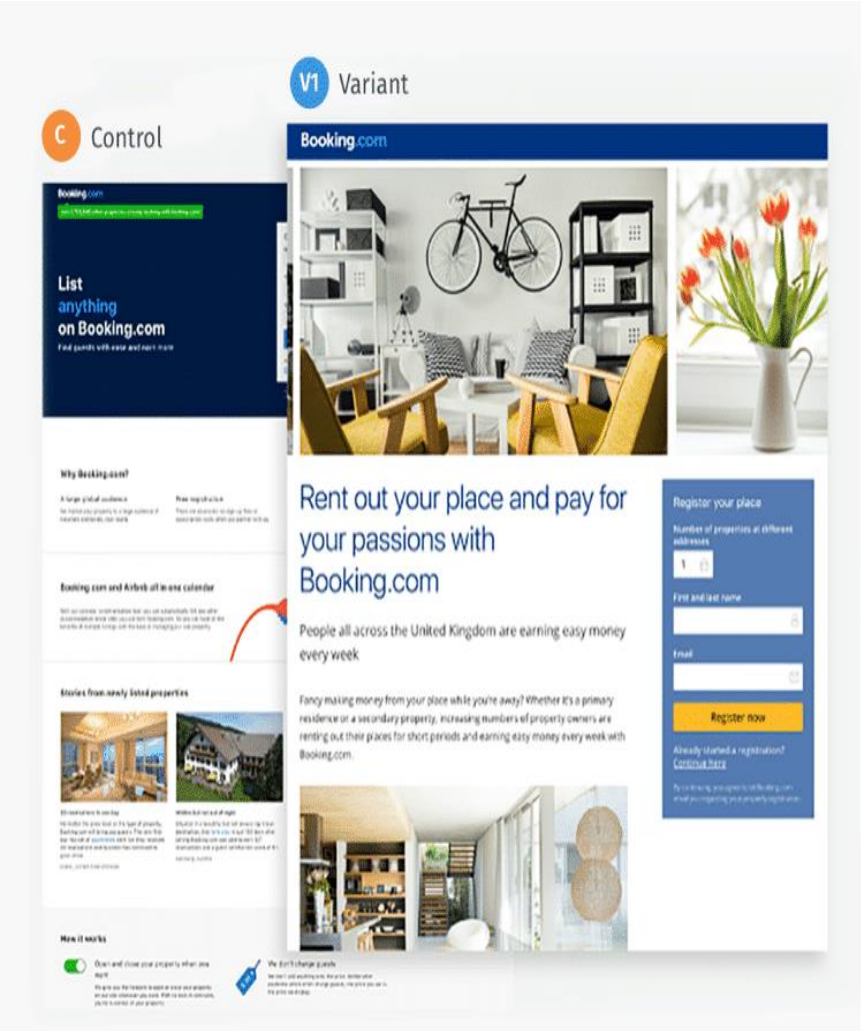
Amazon is at the forefront in conversion optimization partly due to the scale they operate at and partly due to their immense dedication to providing the best customer experience.

Amongst the many revolutionary practices they brought to the eCommerce industry, the most prolific one has been their '1-Click Ordering'.

This change had such a huge business impact that Amazon got it patented (now expired) in 1999. In fact, in 2000, even Apple bought a license for the same to be used in their online store.

2. 3. A/B Testing in Travel Industry

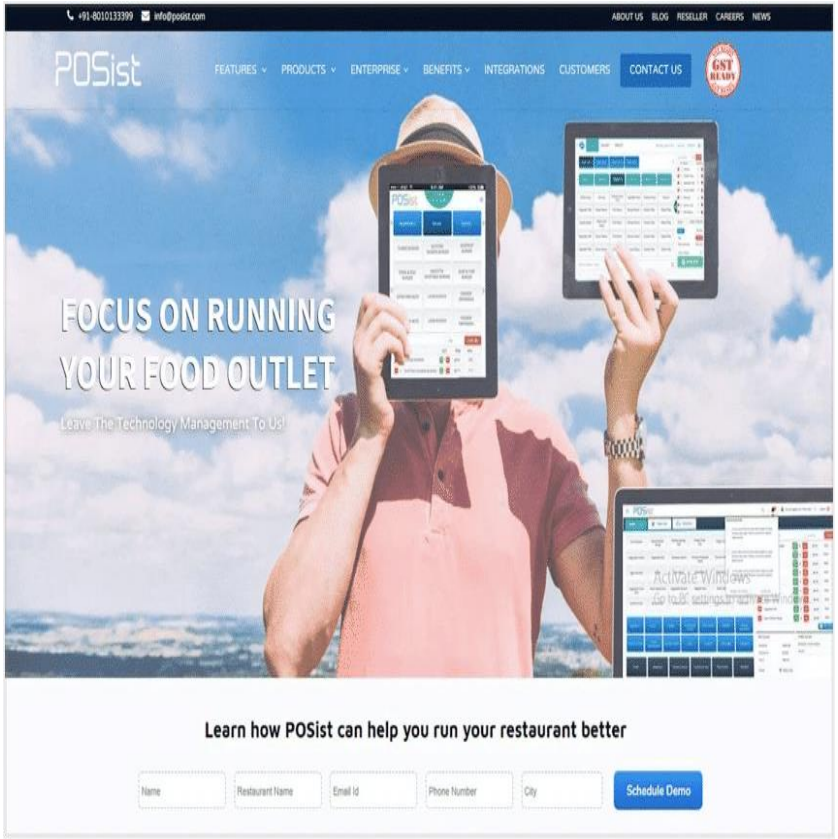
- In the travel industry, Booking.com easily surpasses all other eCommerce businesses when it comes to using A/B testing for their optimization needs.
- They test like it's nobody's business.
- From the day of its inception, Booking.com has treated A/B testing as the treadmill that introduces a flywheel effect for revenue.



2.4 A/B Testing in B2B/SaaS Industry

Generate high-quality leads for your sales team, increase the number of free trial requests, attract your target buyers, and perform other such actions by testing and polishing important elements of your demand generation engine.

POSist, a leading SaaS-based restaurant management platform with more than 5,000 customers at over 100 locations across six countries, wanted to increase their demo requests.



C Control

3.Why Should You A/B Test?

If B2B businesses today are unhappy with all the unqualified leads they get per month, eCommerce stores, on the other hand, are struggling with a high cart abandonment rate.

Meanwhile, media and publishing houses are also dealing with low viewer engagement. These core conversion metrics are affected by some common problems like leaks in the conversion funnel, drop-offs on the payment page, etc.

Let's see why you should do A/B testing to deal with all these problems:

How do you Perform an A/B Test?

A/B testing offers a very systematic way of finding out what works and what doesn't work in any given marketing campaign.
Most marketing efforts are geared toward driving more traffic.
But, as traffic acquisition becomes more difficult and expensive, it becomes paramount to offer the best experience to your users who come to your website.
This will help them achieve their goals and allow them to convert in the fastest and most efficient manner possible.
A/B testing in marketing allows you to make the most out of your existing traffic.
Broadly, it includes the following steps:

Step 1: Research
Step 2: Observe and Formulate Hypothesis
Step 3: Create Variations
Step 4: Run Test
- **Split URL Testing**
- **Multivariate Testing (MVT)**
- **Multipage Testing**

Step 5: Result Analysis and Deployment

Step 1: Research

Before building an A/B testing plan, one needs to conduct thorough research on how the website is currently performing.

Collect data on everything related to how many users are coming onto the site, which pages drive the most traffic, what are the various conversion goals of different pages etc.

The A/B testing tools used here can include quantitative website analytics tools such as Google Analytics, Omniture, Mixpanel, etc., which can help you figure out your most visited pages, pages with most time spent or pages with the highest bounce rate.

For example, to start by shortlisting pages which have the highest revenue potential or the highest daily traffic.

Step 2: Observe and Formulate Hypothesis

- Get closer to business goals by logging research observations and creating data-backed hypotheses aimed at increasing conversions.

.

- The qualitative and quantitative research tools can only help with gathering visitor behavior data.

- Analyze and make sense of that data.

- The best way to utilize every bit of data collated is to analyze it, to make keen observations on them, and then to draw website as well as user insights to formulate data-backed hypotheses.

- Once is hypothesis ready, test it against various parameters like how much confidence you have of it winning, its impact on macro goals, and how easy it is to set up and so on.

Step 3: Create Variations

- The next step in the testing program should be to create a variation based on your hypothesis, and A/B test it against the existing version (control).
- A variation is another version of the current version with changes that you want to test. Test multiple variations against the control to see which one works best.
- Create a variation based on the hypothesis of what might work from a UX perspective.
- For example, enough people not filling forms? Does the form have too many fields? Does it ask for personal information? Maybe try a variation with a shorter form or another variation by omitting fields that ask for personal information.

Step 4: Run Test

Before we get to this step, let's first explore how many kinds of testing methods are there and when to use which method.

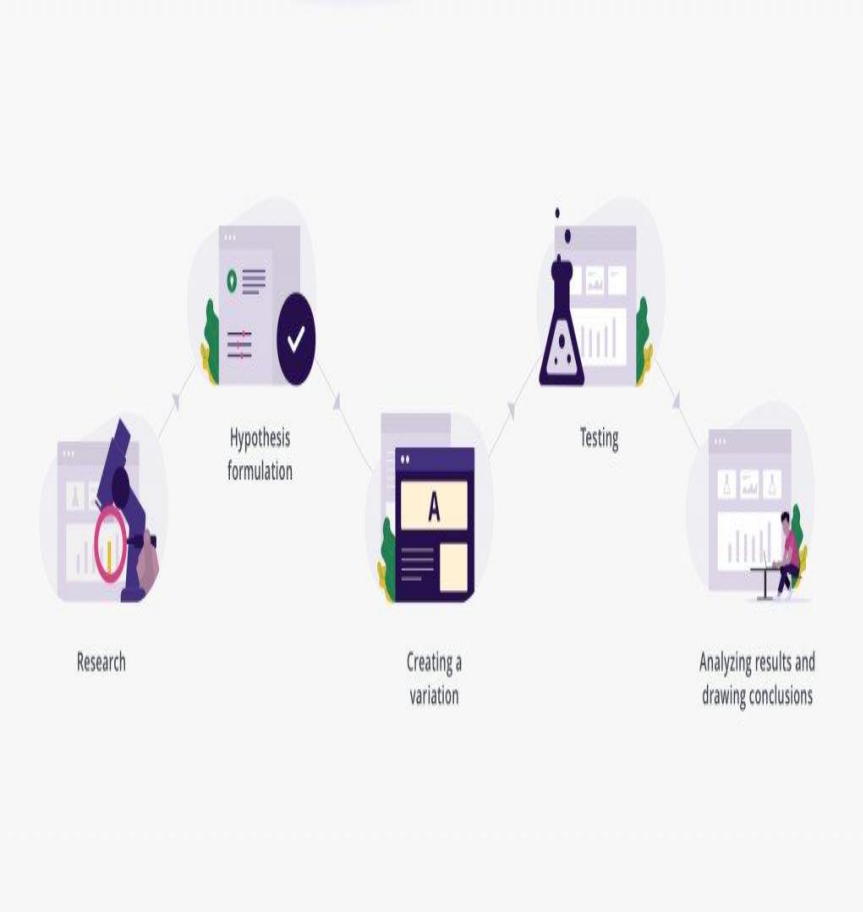
- A/B Testing,
- Multivariate Testing,
- Split URL Testing, and
- Multipage Testing are 4 different types of testing.

Step 5: Result Analysis and Deployment

Once the test concludes, analyze the test results by considering metrics like percentage increase, confidence level, direct and indirect impact on other metrics, etc.

If the test succeeds, deploy the winning variation.

If the test remains inconclusive, draw insights from it, and implement these in your subsequent tests.



5. What Can you A/B Test?

- Website's conversion funnel determines the fate of your business.
- Therefore every piece of content that reaches your users via your website must be optimized to its maximum potential.
- This is especially true for elements that can influence visitor behavior and conversion rate.
- When undertaking an optimization program, the following key elements should be A/B tested (the list, however, is not exhaustive):

5. What Can you A/B Test?

Copy

1. Headlines and Sub-headlines

Headline is the first thing that visitors see on any page. The headline is what defines your first impression in a visitor's eyes.

Make sure the headline catches the visitors' attention as soon as they land on the website.

Keep it short and to the point, ensuring it talks clearly about what your product or service is and its benefits.

Try A/B testing various fonts, sizes, copy, and messaging.

2. Body

The body of your website should clearly state what the visitor is getting – what's in store for them. It should also resonate with your page's headline.While writing content for your page's body, keep in mind these two parameters:

Writing style: Use the right tonality based on the target audience. Your copy should directly address the end-user and answer all their questions. It should consist of key phrases that improve usability and stylistic elements that highlight important points.

Formatting: Use relevant headlines and subheadlines, break the copy into small and easy paragraphs, and format it for skimmers using bullet points or lists.

5. What Can you A/B Test?

Design and Layout: Along with the copy, the design and layout of a page include images (product images, offer images, etc.) and videos (product videos, demo videos, advertisements, etc.).

- **Provide clear information:** . Write clear copies and provide easily noticeable size charts, color options, etc.
- **Highlight customer reviews:** Add both good and bad reviews for your products. Negative reviews add credibility to your store.
- **Write simple content:** Avoid confusing potential buyers with complicated language in the quest to decorate your content. Keep it simple and fun to read.
- **Create a sense of urgency:** Add tags like ‘only 2 left in stock’, countdowns like ‘offer ends in 2hours 15 minutes’, or highlight exclusive discounts and festive offers, etc. to nudge the prospective buyer to purchase immediately.

5. What Can you A/B Test?

Navigation

- Another element of your website that you can optimize by A/B testing is your website's navigation. It is the most crucial element when it comes to delivering excellent user experience. Each click should direct visitors to the desired page.

- For example, as an eCommerce store, you may be selling a variety of earphones and headphones. Some of them may be wired, while others may be wireless or ear-pods.

- Bucket these in such a way that when a visitor looks for earphones or headphones, they find all these varieties in one place rather than having to search for each kind separately

5. What Can you A/B Test?

Forms

- Forms are mediums through which prospective customers get in touch with you. They become even more important if they are part of your purchase funnel.
- Just as no two websites are the same, no two forms addressing the different audience is the same.
- While for some businesses, a small comprehensive form may work, for other businesses, long forms might do wonders for their lead quality.
- You can figure out which style works for your audience the best by using research tools/methods like form analysis to determine the problem area in your form and work towards optimizing it.

5. What Can you A/B Test?

CTA (Call To Action)
The CTA is where all the real action takes place – whether or not visitors finish their purchases and convert if they fill out the sign-up form or not, and more such actions that have a direct bearing on your conversion rate.

With A/B testing, you can A/B test different copies, placement, colors & sizes, etc. for your CTA till you find the winning variation – and then test the winning version further to optimize it even more.

5. What Can you A/B Test?

Social Proof

Social proof may take the form of recommendations and reviews from experts of the particular fields, from celebrities and customers themselves, or can come as testimonials, media mentions, awards and badges, certificates, and so on.

The presence of these proofs validates the claims made by your website. A/B testing can help you determine if adding social proof is a good idea, what kinds of social proof if it is a good idea and how many should be added.

6. What are the Mistakes to Avoid While A/B Testing?

Mistake #1: Not Planning your Optimization Roadmap

•**Invalid hypothesis:** In A/B testing, a hypothesis is formulated before conducting a test. All the next steps depend on it: what should be changed, why should it be changed, what the expected outcome is, and so on. If you start with the wrong hypothesis, the probability of the test succeeding decreases.

•**Taking others' word for it:** Sure, someone else changed their sign-up flow and saw a 30% uplift in conversions. But it is their test result, based on their traffic, their hypothesis, and their goals. Here's why you should not implement someone else's test results as is onto your website: no two websites are the same – what worked for them might not work for you. Their traffic will be different; their target audience might be different; their optimization method may have been different than yours, and so on.

6.What are the Mistakes to Avoid While A/B Testing?

Mistake #2: Testing too Many Elements Together

Industry experts caution against running too many tests at the same time. Testing too many elements of a website together makes it difficult to pinpoint which element influenced the success or failure of the test most. Apart from this, more the elements tested, more needs to be the traffic on that page to justify statistically significant testing. Thus, prioritization of tests is indispensable for successful A/B testing.

6.What are the Mistakes to Avoid While A/B Testing?

Mistake #3: Ignoring Statistical Significance

If gut feelings or personal opinions find a way into hypothesis formulation or while you are setting the A/B test goals, it is most likely to fail. Irrespective of everything, whether the test succeeds or fails, you must let it run through its entire course so that it reaches its statistical significance. For a reason, that test results, no matter good or bad, will give you valuable insights and help you plan your upcoming test in a better manner.

6. What are the Mistakes to Avoid While A/B Testing?

Mistake #4: Using Unbalanced Traffic

Businesses often end up testing unbalanced traffic. A/B testing should be done with the appropriate traffic to get significant results. Using lower or higher traffic than required for testing increases the chances of your campaign failing or generating inconclusive results.

Mistake #5: Testing for Incorrect Duration

- Based on your traffic and goals, run A/B tests for a certain length of time for it to achieve statistical significance.
- Running a test for too long or too short a period can result in the test failing or producing insignificant results.
- Because one version of your website appears to be winning within the first few days of starting the test does not mean that you should call it off before time and declare a winner.
- The duration for which you need to run your test depends on various factors like existing traffic, existing conversion rate, expected improvement, and so on.

What are the Mistakes to Avoid While A/B Testing?

Mistake #6: Failing to Follow an Iterative Process

- A/B testing is an iterative process, with each test building upon the results of the previous tests. Businesses give up on A/B testing after their first test fails. But to improve the chances of your next test succeeding, you should draw insights from your last tests while planning and deploying your next test.

- This improves the probability of your test, succeeding with statistically significant results.

- Additionally, do not stop testing after a successful one. Test each element repetitively to produce the most optimized version of it even if they are a product of a successful campaign.

6. What are the Mistakes to Avoid While A/B Testing?

Mistake #7: Failing to consider external factors

Tests should be run in comparable periods to produce meaningful results. It is wrong to compare website traffic on the days when it gets the highest traffic to the days when it witnesses the lowest traffic because of external factors such as sale, holidays, and so on.

Mistake #8: Using the Wrong Tools

With A/B testing gaining popularity, multiple low-cost tools have also come up. Not all of these tools are equally good. Some tools drastically slow down your site, while others are not closely integrated with necessary qualitative tools (heatmaps, session recordings, and so on), leading to data deterioration.

Mistake #9: Sticking to Plain Vanilla A/B Testing Method

In the long run, sticking to plain vanilla A/B testing method will not work wonders. For instance, if you are planning to revamp one of your website's pages entirely, you ought to make use of split testing. Meanwhile, if you wish to test a series of permutations of CTA buttons, their color, the text and image of your page's banner, you must use multivariate testing.

7. What are the Challenges of A/B Testing?

The ROI from A/B testing can be huge and positive. It helps you direct your marketing efforts to the most valuable elements by pinpointing exact problem areas. The 6 primary challenges are as follows:

Challenge #1: Deciding What to Test

You can't just wake up one day and decide to test certain elements of your choice. A bitter reality that marketers are now coming to realize is that not all small changes that are easy to implement are always the best when you consider your business goals and often fail to prove significant. The same goes for complex tests. This is where website data and visitor analysis data come into play. These data points help you overcome the challenge of 'not knowing what to test' out of your unending backlog by generally pointing to the elements which may have the most impact on your conversion rates or by directing you to pages with the highest traffic.

7. What are the Challenges of A/B Testing?

Challenge #2: Formulating Hypotheses

In great resonance with the first challenge is the second challenge: formulating a hypothesis. This is where the importance of having scientific data at your disposal comes in handy. If you are testing without proper data, you might as well be gambling away your business. With the help of data gathered in the first step (i.e., research) of A/B testing, you need to discover where the problems lie with your site and come up with a hypothesis. This will not be possible unless you follow a well structured and planned A/B testing program.

Challenge #3: Locking in on Sample Size

Not many marketers are statisticians. We often make the mistake of calling conclusive results too quickly because we are more often than not after quick results. As marketers, we need to learn about sample sizes, in particular, how large should our testing sample size be based on our web page's traffic.

7. What are the Challenges of A/B Testing?

Challenge #4: Analyzing Test Results

With A/B testing, you will witness success and failure at each step. This challenge, however, is pertinent to both successful and failed tests:

•**Successful campaigns:** Interpreting test results after they conclude is extremely important to understand why the test succeeded. A fundamental question to be asked is – why? Why did customers behave the way they did? Why did they react a certain way with one version and not with the other versions? What visitor insights did you gather, and how can you use them? Many marketers often struggle or fail to answer these questions, which not only help you make sense of the current test but also provide inputs for future tests.

•**Failed campaigns:** Sometimes, marketers don't even look back at failed tests. They either have a hard time dealing with them, for example, while telling the team about the failed tests or have no clue what to do with them. No failed test is unsuccessful unless you fail to draw learnings from them. Failed campaigns should be treated like pillars that would ultimately lead you to success.

7. What are the Challenges of A/B Testing?

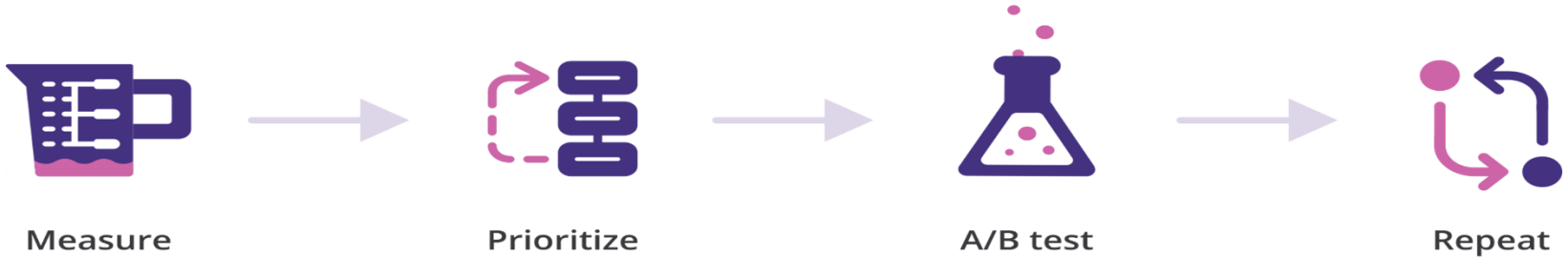
Challenge #5: Maintaining a Testing Culture

One of the most crucial characteristics of optimization programs like CRO and A/B testing is that it is an iterative process. This is also one of the major obstacles that businesses and marketers face. For your optimization efforts to be fruitful in the long run, they should form a cycle that roughly starts with research and ends in research.

Challenge #6: Changing Experiment Settings in the Middle of an A/B Test

When you launch an experiment, you must commit to it completely. Try and not change your experiment settings, edit or omit your test goals, or play with the design of the control or the variation while the test is running. Moreso, do not try and change the traffic allocations to variations as well because doing so will not only alter the sampling size of your returning visitors but massively skew your test results as well.

8. How To Make an A/B Testing Calendar – Plan & Prioritize



Stage 1: Measure	Stage 2: Prioritize	Stage 3: A/B Test	Stage 4: Repeat

8. How To Make an A/B Testing Calendar – Plan & Prioritize

Stage 1: Measure

This stage is the planning stage of your A/B testing program.

Everything that goes on in your website should correspond to your business goals. So before everything else, you need to be sure what your business goal/s is (are).

Tools like Google Analytics can help you measure your goals. Once you have clearly defined goals, set up GA for your website and define your key performance indicators.

1. Define your business objectives.
2. Define your website goals.
3. Define your Key Performance Indicators.
4. Define your target metrics.

8. How To Make an A/B Testing Calendar – Plan & Prioritize

Stage 2: Prioritize

- The next stage involves prioritizing your test opportunities.
- Prioritizing helps you scientifically sort the multiple hypotheses.
- By now, you should be fully equipped with website data, visitor data, and be clear on your goals.
- With the backlog you prepared in the first stage along with the hypothesis ready for each candidate, you are halfway there on your optimization roadmap.
- Now comes the main task of this stage: prioritizing.

8. How To Make an A/B Testing Calendar – Plan & Prioritize

Stage 3: A/B Test

- The third and most crucial stage is the testing stage.
- After the prioritization stage, you will have all the required data and a prioritized backlog.
- Once you have formulated hypotheses that align to your goal and prioritized them, create variations, and flag off the test.
- While your test is running, make sure it meets every requirement to produce statistically significant results before closure, like testing on accurate traffic, not testing too many elements together, testing for the correct amount of duration, and so on.
-

8. How To Make an A/B Testing Calendar – Plan & Prioritize

Stage 4: Repeat

This stage is all about learning from your past and current test and applying them in future tests.

There can be 3 outcomes of your test:

1. Your variation or one of your variations will have won with statistical significance.
2. Your control was the better version and won over the variation/s.
3. Your test failed and produced insignificant results. Determine the significance of your test results with the help of tools like the A/B test significance calculator.

Summary

1. **What is A/B testing definition?** A/B testing is the process of comparing two variations of a page element, usually by testing users' response to variant A vs variant B, and concluding which of the two variants is more effective.
2. **What is A/B testing in digital marketing?** In digital marketing, A/B testing is the process of showing two versions of the same web page to different segments of website visitors at the same time, and then comparing which version improves website conversions.
3. **Why do we do A/B testing?** There are various reasons why we do A/B testing. A few of them include solving visitor pain points, increasing website conversions or leads, and decreasing the bounce rate. Read our guide to know the rest of the reasons.
4. **What is A/B testing and multivariate testing?** In A/B testing, traffic is split amongst two or more completely different versions of a webpage. In multivariate testing, multiple combinations of a few key elements of a page are tested against each other to figure out which combination works best for the goal of the test.

DATA ANALYTICS

Additional References

1. https://vwo.com/ab-testing/
2. http://www.datascienceassn.org/sites/default/files/AB%20Testing%20Guide.pdf
3. https://www.tutorialspoint.com/ab_testing/ab_testing_tutorial.pdf
4. https://en.wikipedia.org/wiki/A/B_testing

THANK YOU

Prof Bharathi R

Associate Professor, Department of Computer Science

rbharathi@pes.edu

DATA ANALYTICS

Unit 5:Interpreting business value

Bharathi R
Department of Computer Science and Engineering

DATA ANALYTICS

Unit 5:Interpreting business value

Bharathi R

Department of Computer Science and Engineering

Data Analytics to Drive Business Value

- The metamorphosis taking place in the analytics world has been rapid and transformational.
- The analytics space has moved from business intelligence to the age of big data analytics.
- These transformations have been affected in the recent years as more and more companies are opting for digitization.
- The technology, infrastructure and the philosophy of organizations have changed as well.
- It is now evident, that organizations will have to keep pace with the changes or perish.
- It has all the more become important to embrace data analytics to develop an agile IT framework and build a strong base for data science.

ROI of Analytics: Explaining the Business Value of Measurement

- Analytics is a critical part of building and growing a successful business.
- With all of the intense discussion around big data, artificial intelligence and the advance of analytics tools, it seems obvious that having an annual analytics budget just *makes sense*.
- And yet, when it comes time to secure funding for next year's analytics budget, often get unexpected push back and resistance. That's because we struggle to explain and demonstrate the business value and the return on investment (ROI) of analytics.
- And when senior executives don't have a clear understanding of the business value of analytics, the budget goes to higher priority areas where the business value is clear and demonstrable.

Questioning the Business Value of Data Analytics

According to the Digital Analytics Association, "44% of analytics teams spend more than half their time accessing and preparing data rather than doing actual analysis."

So why should a business invest more budget in data analytics if less than 50% of an analyst's time is going to doing the actual business value of analyzing the data and improving the value of the business?

The following are five of the most important business drivers where you can easily demonstrate the ROI of your analytics efforts.

1. Analytics Business Value #1: Acquire More of the RIGHT Customers
2. Analytics Business Value #2: Marketing Attribution and Media Mix Modeling
3. Analytics Business Value #3: More Revenue from Current Customers
4. Analytics Business Value #4: Growth From A/B Testing
5. Analytics Business Value #5: Moving at the Speed of Business

Analytics Business Value #1: Acquire More of the RIGHT Customers

Without the support of rich data and robust analytics tools, businesses are left to measure the most cursory top line and bottom line impact of any marketing effort. These are questions such as:

1. How many customers have we acquired?
2. How much do these customers purchase?
3. What is our overall profitability?

The ROI of acquiring more of the right customers means that even with a higher cost per acquisition:

1. The repeat purchase rates are higher.
2. The customer churn rates are lower.
3. The overall lifetime value of the right customers will generally outperform customers acquired at the lowest possible price

DATA ANALYTICS

Analytics Business Value #2: Marketing Attribution and Media Mix Modeling

- When you combine the right analysts with powerful data analytics tools, you solve the John Wanamaker problem.

- Wanamaker was the department-store magnate, who once said, “Half the money I spend on advertising is wasted; the trouble is, I don't know which half.”

- At least in the area of digital marketing combined with powerful analytics tools and professionals, the data will illuminate the ideal marketing mix ratio by allowing you to test a combination of marketing options that produce the maximum results for the minimum investment.

- Effective marketing attribution and media mix modeling reduces significant advertising spend and that business value can be measured and attributed to both data analytics and the analysts doing the work.

DATA ANALYTICS

Analytics Business Value #3: More Revenue from Current Customers

- Acquiring a new customer can cost 5 times more than retaining an existing customer.

- In addition to validating that industry accepted stat with your own data, your analytics will support you to more effectively upsell, cross-sell, and increase the frequency of purchases.

- The data, when properly analyzed, will support a better understanding of what your current customers need as well as understanding what other customers purchased.

DATA ANALYTICS

Analytics Business Value #4: Growth From A/B Testing

Analytics, when properly deployed, deliver incredible value in their ability to isolate variables and test assumptions from the top of the funnel activities all the way to the point of conversion.

”Investing in analytics is the foundation for both running statistically significant A/B tests to increase website conversion, and for running top of funnel A/B tests on advertising vehicles. Without proper analytics in place, you are essentially flying blind and will undoubtedly get left behind by those who use the data to continually improve and better connect with their customers” by Sean Lee, Vice President of Digital Marketing and eCommerce at Pure Romance

Analytics Business Value #5: Moving at the Speed of Business

- Analytics also leverage data to do things faster.

- By using machine learning and pattern recognition, we can begin to drive repeat purchases faster without the manual analysis.

- Today, speed is often the difference between closing the sale and losing the sale. There is both demonstrable opportunity cost for business lag (without analytics) and increased purchase frequency by removing any barriers to completing the current transaction.

DATA ANALYTICS

Case study : Analytics: A blueprint for value

Converting big data and analytics insights into results

- **In today's competitive marketplace,** executive leaders are racing to convert data-driven insights into meaningful results.

- Successful leaders are infusing analytics throughout their organizations to drive smarter decisions, enable faster actions and optimize outcomes.

- These are among the key findings from the 2013 IBM Institute for Business Value research study on how organizations around the globe are leveraging key capabilities to amplify their ability to create value from big data and analytics.

DATA ANALYTICS

Case study : Analytics: A blueprint for value

Converting big data and analytics insights into results

- To discover how to achieve this alignment of strategy, technology and structure, IBM surveyed 900 business and IT executives from 70 countries.

- They asked more than 50 questions to an analytics-savvy group of executives, senior managers and managers, along with analytics experts, business and data analysts, and others within organizations large and small.

- The questions were designed to reveal how to translate high-level concepts associated with delivering exceptional business value through analytics into actions that can truly deliver value.

- Through the research, IBM identified nine levers that enable organizations to create value from an ever-growing volume of data from a variety of sources – value that results from insights derived and actions taken at every level of the organization.

DATA ANALYTICS

Case study : Analytics: A blueprint for value

Converting big data and analytics insights into results

Nine levers of differentiation

These nine levers represent the sets of capabilities that most differentiated Leaders from other respondents:

1. ***Culture:*** Availability and use of data and analytics within an organization
2. ***Data:*** Structure and formality of the organization's data governance process and the security of its data
3. ***Expertise:*** Development of and access to data management and analytic skills and capabilities
4. ***Funding:*** Financial rigor in the analytics funding process
5. ***Measurement:*** Evaluating the impact on business outcomes
6. ***Platform:*** Integrated capabilities delivered by hardware and software
7. ***Source of value:*** Actions and decisions that generate results
8. ***Sponsorship:*** Executive support and involvement
9. ***Trust:*** Organizational confidence

DATA ANALYTICS

Case study : Analytics: A blueprint for value

Converting big data and analytics insights into results

Organizations that invest in these nine levers – with particular attention to the symbiotic relationships that exist – can accelerate value creation, simplify analytics implementation and realize value from analytic investments.

They identified three levels of value impact among the nine levers:

1. *Enable levers form the basis for big data and analytics;*
2. *Drive levers are needed to realize value; and*
3. *Amplify levers boost value creation.*

Reference

Interpreting business value
https://www.ibm.com/downloads/cas/4WBWGBJL

https://infotrust.com/articles/roi-of-analytics-explaining-the-business-value-of-measurement/

Additional references:
https://www.slideshare.net/jamet123/delivering-the-business-value-of-analytics,

https://medium.com/analytics-for-humans/how-to-use-analytics-to-identify-the-business-value-of-your-website-c13f9c9675c7,

https://www.searchenginejournal.com/content-marketing-kpis/business-conversion/#close

https://zivanta-analytics.com/data-analytics-to-drive-business-value/

THANK YOU

Prof Bharathi R
Department of Computer Science
rbharathi@pes.edu