Data Analytics Unit 5 - Advanced Techniques

Vibha Masti - vibha@pesu.pes.edu

1 Latent Semantic Analysis (LSA)

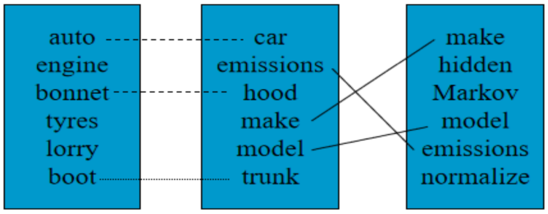
- Form of dimensionality reduction for text data (NLP)
- Topic modelling technique - extracting topics from large text documents
- Topic defined by proportions of words it contains
- Unsupervised learning technique
- Uses contextual knowledge - words are related to their surrounding words
- Automatically done with linear algebra techniques

1.1 The vector space model

- Create term (rows) by document (columns) matrix
- Each row is a vector
- Find cosine similarity between vectors

1.2 Problems with vector space model

- Polysemy (same words with different meanings) - leads to poor precision $\left(\frac{\text{TP}}{\text{TP+FP}}\right)$
  - I liked his last **novel** quite a lot
  - We would like to go for a **novel** marketing campaign
- Synonymy (different words with same meaning) - leads to poor recall $\left(\frac{\text{TP}}{\text{TP+FN}}\right)$
  - I am a **car** mechanic
  - I work in the **automobile** industry

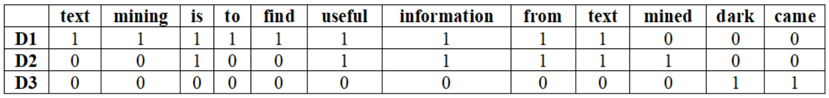


Synonymy
Will have small cosine
but are related
Not handling this leads
to poor recall

Polysemy
Will have large cosine
but not truly related
Not handling this leads
to poor precision

1.3 Steps of LSA

1. Assume we have m text documents with n unique words across them all
2. Create an $n \times m$ document-term matrix

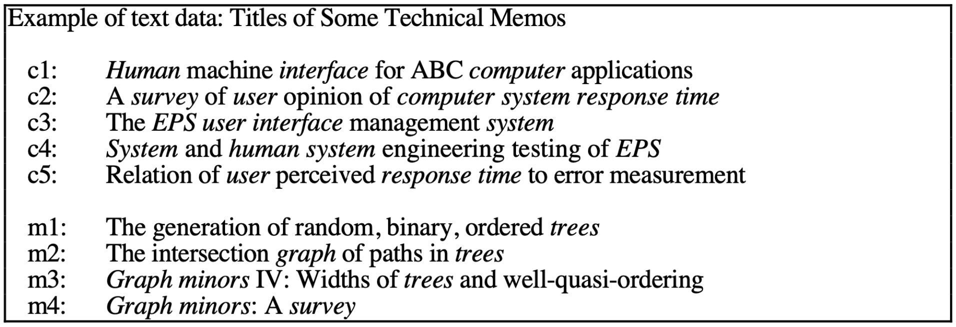


	text	mining	is	to	find	useful	information	from	text	mined	dark	came
D1	1	1	1	1	1	1	1	1	1	0	0	0
D2	0	0	1	0	0	1	1	1	1	1	0	0
D3	0	0	0	0	0	0	0	0	0	0	1	1

3. Optionally, create a TF-IDF matrix
   - TF(t, d) = Number of occurrences of term t in document d
   - DF(t) = Number of documents with the term t occurring
   - $\text{IDF}(t) = 1 + \log\left(\frac{N}{\text{DF}(t)}\right)$
     - N : number of documents in the corpus
     - Eg: suppose there are 100 documents in the corpus and 10 documents contain the term **text**
     - $\text{IDF}(text) = 1 + \log\left(\frac{100}{10}\right) = 1 + \log 10 = 1 + 1 = 2$
   - $\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t)$
4. Transpose the matrix to get a term-document matrix of dimensions $n \times m$
5. Perform SVD on the $n \times m$ matrix where the rows are the words and the columns are the documents

1.4 LSA Example

- Corpus - titles of 9 technical memoranda; 5 about human computer interaction (HCI), and 4 about mathematical graph theory, topics that are conceptually quite disjoint



Example of text data: Titles of Some Technical Memos

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

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*

- Frequency matrix - 9 columns for the 9 documents, 12 rows for the 12 words where each word appears in more than 1 document
- Find cosine similarity between *human* and *user* and *human* and *minors*

	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

  - $\text{sim}(human, user) = 0$
  - $\text{sim}(human, minor) = 0$

- Perform SVD on the term-document matrix (eg shows full SVD and not rank reduced SVD) - full code here.

```python
import numpy as np
term_doc_counts = np.array([
    [1,0,0,1,0,0,0,0,0],
    [1,0,1,0,0,0,0,0,0],
    [1,1,0,0,0,0,0,0,0],
    [0,1,1,0,1,0,0,0,0],
    [0,1,1,2,0,0,0,0,0],
    [0,1,0,0,1,0,0,0,0],
    [0,1,0,0,1,0,0,0,0],
    [0,0,1,1,0,0,0,0,0],
    [0,1,0,0,0,0,0,0,1],
    [0,0,0,0,0,1,1,1,0],
    [0,0,0,0,0,0,1,1,1],
    [0,0,0,0,0,0,0,1,1]
])
u, s, vt = np.linalg.svd(term_doc_counts, full_matrices=True)

# Taking only the k=2 most important features
up, sp, vtp = u[:, 0:2], np.diag(s[0:2]), vt[0:2, :]

# Estimate the new term-document matrix
lsa_term_doc = up @ sp @ vtp
```

- Find new cosine similarity between *human* and *user* and *human* and *minors*

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	0.16	0.4	0.38	0.47	0.18	-0.05	-0.12	-0.16	-0.09
interface	0.14	0.37	0.33	0.4	0.16	-0.03	-0.07	-0.1	-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	0.7	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	-0.2	-0.11
survey	0.1	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	-0.3	0.2	0.31	0.69	0.98	0.85
minors	-0.04	0.25	-0.1	-0.21	0.15	0.22	0.5	0.71	0.62

  - $\text{sim}(human, user) = 0.89$
  - $\text{sim}(human, minor) = -0.28$

1.5 Pros and Cons of LSA

Pros:

- Fast and easy
- Decent results (bettern than vector model)

Cons:

- Linear model
- LSA assumes a Gaussian distribution of the terms in the documents
- SVD is computationally intensive as more data generated

2 Concept of Hidden Variables

2.1 Bayesian network

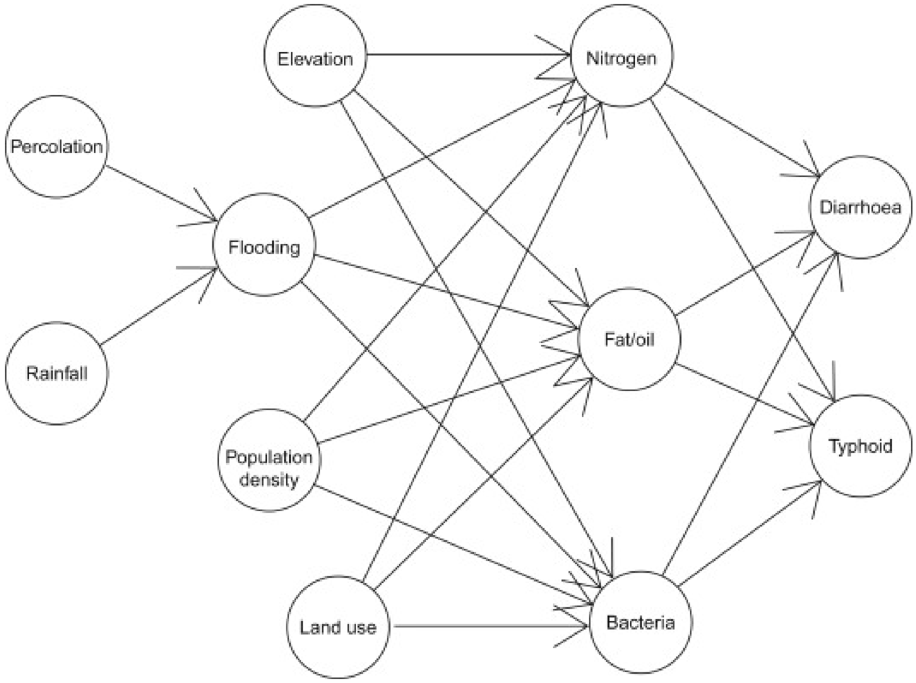
- Probabilistic graphical model for representing knowledge about an uncertain domain
- Nodes are random variables and edges are conditional probabilities
- Bayes nets/belief networks are DAGs where cycles are not permitted
- Eg: the following DAG indicates that the incidence of two waterborne diseases (diarrhoea and typhoid) depends on three indicators of water samples:
  - total nitrogen
  - fat/oil
  - bacteria count

  Each of them is influenced by another layer of nodes:

  - elevation
  - flooding
  - population density
  - land use

  Flooding may be further influenced by two factors:

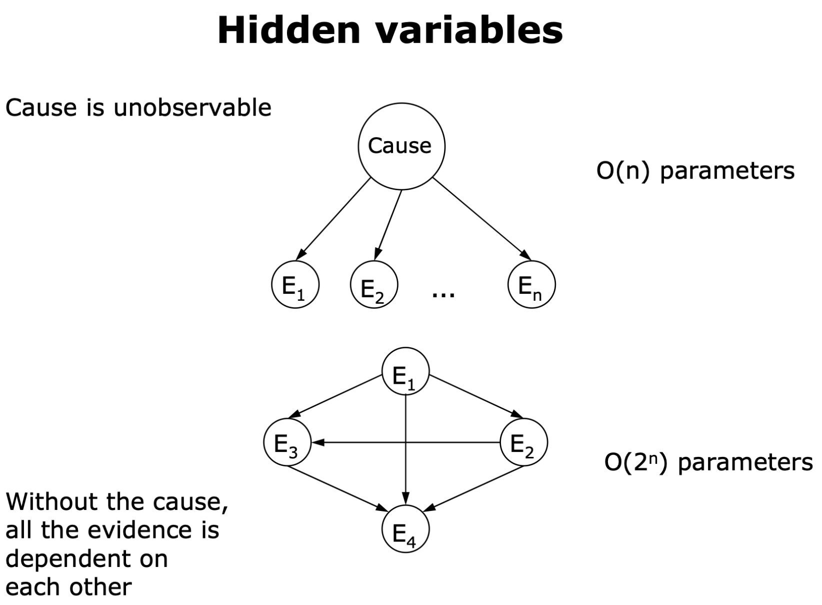
  - percolation
  - rainfall
- Overall Bayes net can be represented as



- The number of parameters to specify for the joint probability distribution of a network of n binary random variables is in the order of $O(2^n)$

2.2 Hidden variable

- Can simplify the complex model by introducing a common *cause* node



Hidden variables

Cause is unobservable

O(n) parameters

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

O(2^n) parameters

3 Simpson's Paradox

- A trend that is observed in data reverses when the data are separated into different groups
- Resolved when confounding variables and causal relations are appropriately addressed in the statistical modelling
- Most famous example: gender bias in UC Berkeley graduate admissions of 1973
- Arises due to presence of lurking variables that split the data up into groups

3.1 UC Berkeley Graduate Admissions

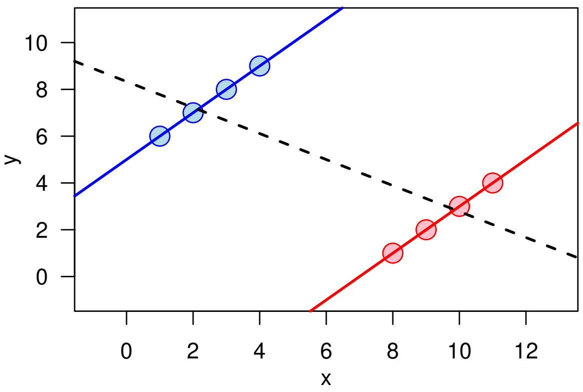
- Data shown: men far more likely to get accepted into programs than women

	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
Total	12,763	41%	8,442	**44%**	4,321	35%

- Data of top 6 departments, small but statistically significant bias in favour of women

Department	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
A	933	64%	*825*	62%	108	**82%**
B	585	63%	*560*	63%	25	**68%**
C	918	35%	325	**37%**	*593*	34%
D	792	34%	417	33%	375	**35%**
E	584	25%	191	**28%**	*393*	24%
F	714	6%	373	6%	341	**7%**
Total	4526	39%	2691	45%	1835	30%

- Graphical representation of Simpson's Paradox



3.2 Soft Drinks Industry - Simple Example

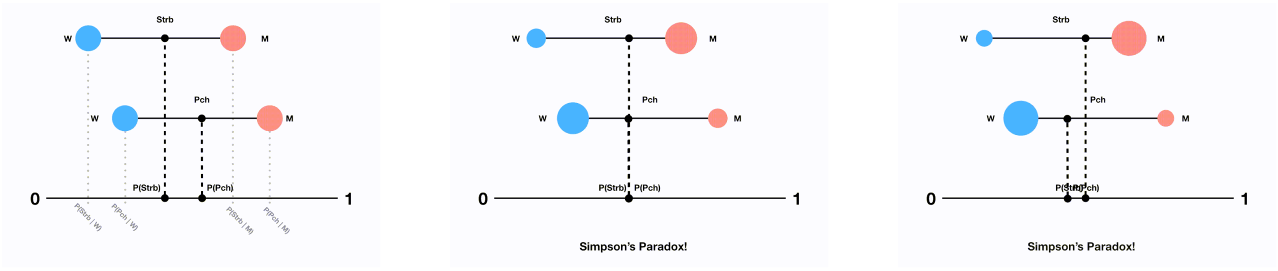
- Soft drinks industry trying to decide between two new flavours in a public survey

Flavour	Sample Size	No. who liked flavour
Sinful Strawberry	1000	800
Passionate Peach	1000	750

- 80% of the people preferred *Sinful Strawberry*
- 75% of the people preferred *Passionate Peach*
- Splitting up by sex

Flavour	Men	No. who liked flavour	Women	No. who liked flavour
Sinful Strawberry	900	760	100	40
Passionate Peach	700	600	300	150

- 84.4% of men and 40% of women preferred *Sinful Strawberry*
- 85.7% of men and 50% of women preferred *Passionate Peach*
- Both men and women separately preferred *Passionate Peach* even though overall *Sinful Strawberry* was preferred
- Position of the centre of each circle corresponds to that group's probability of liking the flavour
- Size of each circle: proportion of population as men or women
- Both groups lie further to the right - have higher probability for liking Peach
- The marginal distributions shift and switch as samples become weighted with respect to the lurking variable (sex)



(a)

(b)

(c)

4 Confounding Variables

- Unknown, unaccounted for variable that can suggest correlation between variables when it does not exist
- Example:
  - Independent variable - lack of exercise

  - Dependent variable - weight gain
  - Confounding variable - age/sex/food intake

- Confounding variables
  - Increase variance
  - Introduce bias

5 Hidden variables vs Confounding variables

- **Hidden variables:** connects two variables that are spuriously correlated
  - Ice cream sales and bank robberies
- **Confounding variables:** related to the two variables that are not spuriously correlated
  - Lack of exercise and increase in weight

5.1 Confounding bias

- Result of having confounding variables in a model
- Usually due to errors in data collection or measurement methods
- Direction - depending on under or over estimation of the correlation
  - **Positive confounding:** observed association biased away from null (overestimates the effect)
  - **Negative confounding:** observed association biased towards null (underestimates the effect)

Techniques to reduce effect:

- Random sampling to reduce bias
- Control variables (constants for the data collection) like age
- Counterbalancing for paired designs

Causes of omitted variable bias:

- Omitted variable must correlate with dependent variable
- Omitted variable must correlate with at least one independent variable
- OLS assumptions violated (correlation of residuals with independent variable)

Example

- Define 3 variables: included (independent and included) - I, dependent (dependent and included) - D, and omitted (not included) - O
- Correlations summarised in table

	I and O: -ve correlation	I and O: +ve correlation
I and D: -ve correlation	Positive bias	Negative bias
I and D: +ve correlation	Negative bias	Positive bias

6 Stochastic Models

- Problems that are dynamic in nature
- Defined as a collection of random variables $\{X_n | n \geq 0\}$ indexed by n (typically time)
- Value of X_n is called the state of the stochastic process at index n
- State space: set of all possible values of X
- Examples of stochastic processes:
  1. Poisson process
  2. Compound Poisson process
  3. Markov process and Markov chain
  4. Markov decision process
  5. Partially observable MDP
  6. Semi-Markov process

7. Random walk
8. Brownian motion process
9. Autoregressive and moving average processes

6.1 Poisson Process

- Really good videos by khan academy: part 1 and part 2
- Count the number of events that occur over a period of time
- λ is the number of events that occur in a defined period of time (eg: 1 hour)

Poisson distribution (probability of k events occuring over time period t) is given by $\frac{(\lambda t)^k}{k!} e^{-\lambda t}$

- Homogeneous Poisson Process (HPP) is a stochastic counting process $N(t)$
- $N(0) = 0$ (number of event that have occured at time 0)
- $N(t)$ has independent increments ($N(t_1) - N(t_0)$ is independent from $N(t_2) - N(t_2)$)
- Number of events by time t is the cumulative distribution of a Poisson distribution

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

- Mean and variance of Poisson process $N(t)$ are λt
- MLE of λ is defined as $\hat{\lambda} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} X_i}$ where X_i are the mean times between failure/demand

Time between events

- Time between events follows exponential distribution
- Density function of time between events: $f(t) = \lambda e^{-\lambda t}$
- CDF of time between events: $F(t) = 1 - e^{-\lambda t}$

6.1.1 Example - textbook

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 (mea- sured in number of days) for hydraulic pumps are shown in Table 16.1.

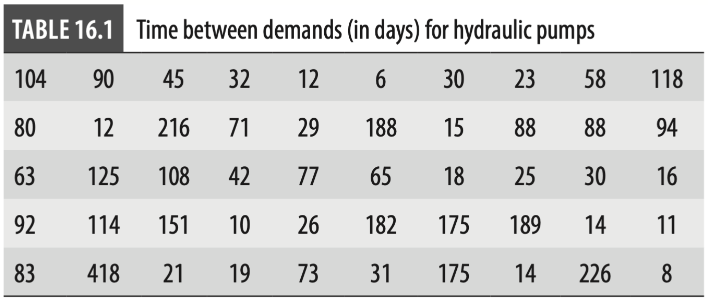


TABLE 16.1 Time between demands (in days) for hydraulic pumps

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

1. Calculate the expected number of demand for hydraulic pump spares for next two years.
2. 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.

6.1.2 Answer

1. Expected number of hydraulic pump spares for 2 years
   - Estimate λ with MLE
   - $\hat{\lambda} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} X_i} = 0.0125$

- $E[N(t)]$ - expected number of demand for spares for 2 years ($2 \times 365 = 730$ days) is $E[N(730)] = \hat{\lambda}t = 0.0125 \times 730 = 9.125$

2. Demand is met 90% of the time - probability that k parts are needed in time 730 days is 90% where k is inventory size
   - Calculate smallest k such that $\sum_{k=0}^{n} \frac{(\lambda t)^k}{k!} e^{-\lambda t} \geq 0.90$
   - Values of CDF and PDF for different values of k

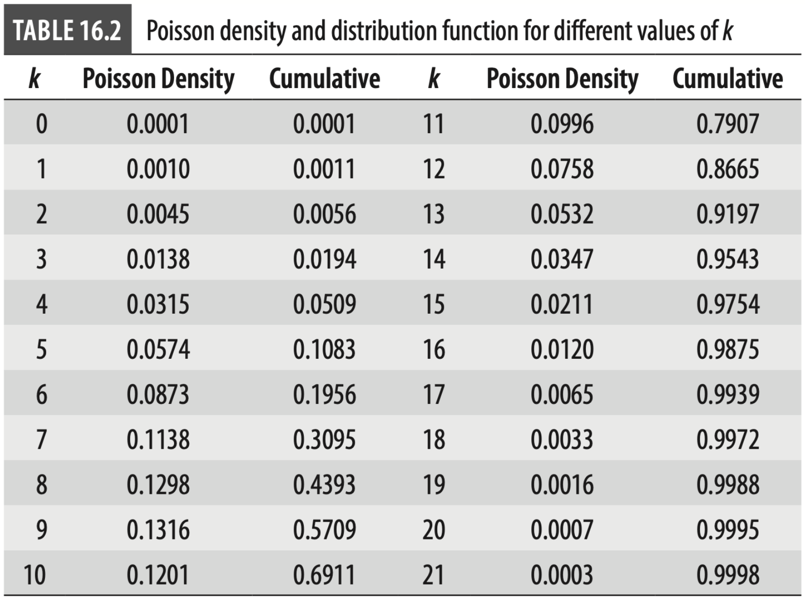
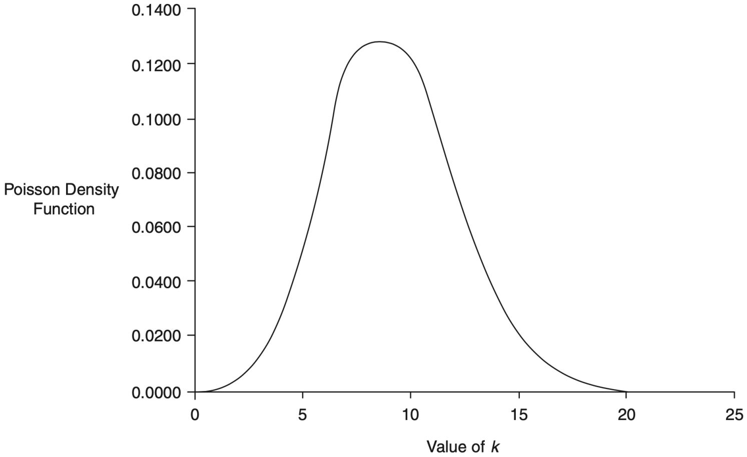


TABLE 16.2 Poisson density and distribution function for different values of *k*

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

- JS should stock 13 spares to meet the demand 90% of the time
- PDF for $\lambda = 9.125$ for different values of k



6.2 Compound Poisson Process

- Regular Poisson process: we are interested in the number of occurrences of events, not the values of the events themselves
- Example: the number of spare parts needed, number of cars that cross an intersection, etc
- Compound Poisson process: we are interested in the values of each event
- Example: the amount of money drawn from an ATM each time a customer visits it
- The distribution of customers visiting the ATM is a Poisson process $X(t)$

- The distribution of the amount of money withdrawn is another independent and identically distributed (IID) random variable (could be a normal distribution) Y_i
- $N(t)$ is the number of events that have occurred until time t in a Poisson process with mean λ
- Compound Poisson process $X(t) = \sum_{k=1}^{N(t)} Y_k$

Expected values - mean and variance

- Mean: $E[X(t)] = \mu_{X(t)} = \lambda t \times E(Y_i)$
- Variance: $Var[X(t)] = \sigma^2_{X(t)} = \lambda t \times E(Y_i^2) = \lambda t \times (VAR(Y_i) + (E[Y_i])^2)$

6.2.1 Example - textbook

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?

6.2.2 Answer

- $\lambda = 12$ per hour
- $E[Y_i] = 4200$, $Var[Y_i] = 250,000$
- Mean of compound process is $\mu_{X(t)} = \lambda t \times E[Y_i] = 12 \times 10 \times 4200 = 504,000$
- Variance of compound process is $\sigma^2_{X(t)} = \lambda t \times (VAR(Y_i) + (E[Y_i])^2) = 12 \times 10 \times (250,000 + 4200^2) = 21468 \times 10^5$
- Standard deviation of compound process is $\sigma = \sqrt{21468 \times 10^5} = 46333.57$
- Probability that cash withdrawn is greater than 600,000 is obtained from normal cumulative distribution
- Z-score $= \dfrac{600,000 - 504,000}{46333.57} = 2.0719$
- $P(Z \geq 2.0719) = 0.01914 = 1.91\% \simeq 2\%$ chance

6.3 Markov Chains

- Let a sequence of RVs be $X_n, n = 1, 2, ..., n$
- The conditional probability that $X_{n+1} = j$ given a state sequence $X_0 = i_0, X_1 = i_1, ..., X_n = i_n$ for a first order Markov process only depends on the value of X_n
- $P(X_{n+1} = j | X_0 = i_0, X_1 = i_1, ..., X_n = i) = P(X_{n+1} = j | X_n = i) = P_{ij}$
- One-step transition probabilities P_{ij} represented in a matrix

$$\begin{bmatrix} P_{11} & P_{12} & ... & P_{1n} \\ P_{21} & P_{22} & ... & P_{2n} \\ ... & ... & ... & ... \\ P_{n1} & P_{n2} & ... & P_{nn} \end{bmatrix}$$

- s-Step transition probability (probability of reaching state j from state i in s steps) is given by

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

- Estimation of 1-step probabilities (from MLE) is given by

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

where m is the total number of states, N_{ij} are the number of transitions from i to j, N_{ik} are the number of transitions from i to k.

6.3.1 Anderson-Goodman Test

- Hypothesis test
  - H_0 : The sequence of RVs $(X_1, X_2, ..., X_n)$ are independent (zero-order Markov chain)

  - H_1 : The sequence of RVs $(X_1, X_2, ..., X_n)$ are dependent (first-order Markov chain)

- Test statistic (χ^2 test)

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

  where O_{ij} is the observed number of transitions, E_{ij} is the expected number of transitions assuming independence

6.3.2 Likelihood Ratio Test

- To check whether the transition probability matrices are time homogeneous
- In other words, are the transition probabilities constant
  - $H_0 : P_{ij}(t) = P_{ij}, t = 1, 2, ..., T$ (not a function of time; constant)
  - $H_1 : P_{ij}(t) \neq P_{ij}, t = 1, 2, ..., T$ (function of time)

  where $P_{ij}(t)$ is the estimated value of transition probability between state i and j from time t to $t + 1$
- Test statistic: likelihood ratio test statistic

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

  where $n_{ij}(t)$ is frequency of transition between states i and j at time t

6.3.3 Markov Chains in Predictive Analytics

- Predict X_n given X_1 and transition matrix P
- Let $P_I = (450, 225, 175, 150)$
- Let one step transition matrix P be

$$P = \begin{bmatrix} 0.8189 & 0.0882 & 0.0472 & 0.0457 \\ 0.1128 & 0.7180 & 0.0902 & 0.0789 \\ 0.2077 & 0.0849 & 0.6011 & 0.0929 \\ 0.0663 & 0.0964 & 0.0964 & 0.7410 \end{bmatrix}$$

- Distribution of X after n time periods is given by $P_I \times P$

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

6.3.4 Stationary Distribution in a Markov Chain

- For a large n, $P_I \times P^n$ converges to a value known as the steady state or stationary distribution of the Markov chain and is denoted as $\pi = (\pi_1, \pi_2, ..., \pi_n)$
- Consider brand switching between two brands (B_1, B_2) and let the initial market share be $P_I = (0.2, 0.8)$

	Brand 1	Brand 2
Brand 1	0.8	0.2
Brand 2	0.25	0.75

- Can be represented by the state transition diagram

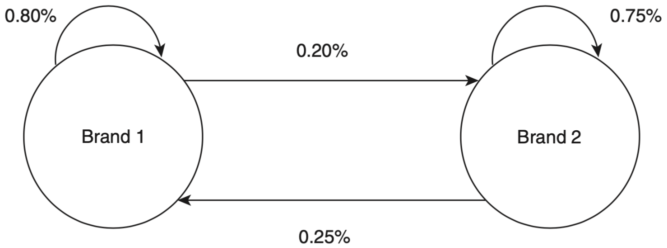
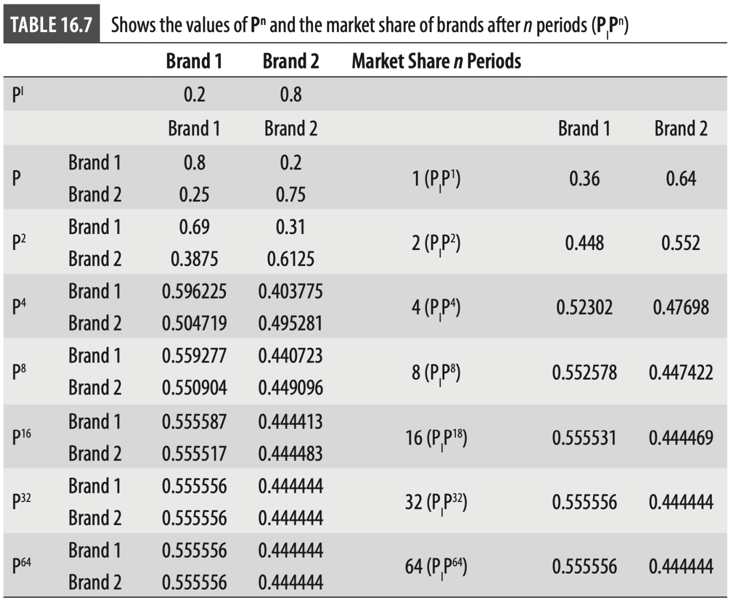


FIGURE 16.3 State transition diagram between brands.

- Convergence of P_n

TABLE 16.7 Shows the values of $\mathbf{P}^n$ and the market share of brands after *n* periods ($\mathbf{P}_I\mathbf{P}^n$)



		Brand 1	Brand 2	Market Share *n* Periods		
P^I		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 probability of Markov chain are $(0.555556, 0.444444)$

Let $\pi = (\pi_1, \pi_2, ..., \pi_n)$ be a stationary distribution. It satisfies the following equations

1. $\pi_j = \sum\limits_{k=1}^{m} \pi_k \times P_{kj}$ which can be rewritten as $\pi = \pi P$
2. $\sum\limits_{k=1}^{m} \pi_k = 1$

6.3.5 Regular Matrix

If a matrix has all nonzero entries, it is regular. A regular matrix will have stationary distribution and satisfy the stationary system of equations.

6.3.6 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.

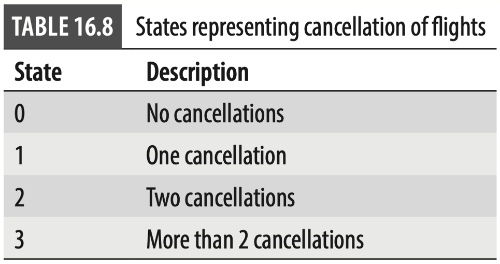


TABLE 16.8 States representing cancellation of flights

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

The revenue loss (in millions of rupees) due to cancellation of flights in various states is given in Table 16.9.

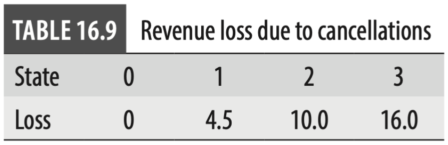


TABLE 16.9 Revenue loss due to cancellations

State	0	1	2	3
Loss	0	4.5	10.0	16.0

The transition probability matrix between states is shown in Table 16.10.

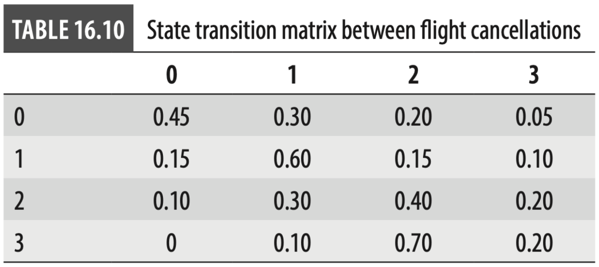


TABLE 16.10 State transition matrix between flight cancellations

	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

1. If there are no cancellations initially, what is the probability that there will be at least one cancellation after 2 days?
2. Calculate the steady-state expected loss due to cancellation of flights.

6.3.7 Answer

1. Probability of at least one cancellation
   - Initial state: $(1, 0, 0, 0)$
   - State after 2 days:

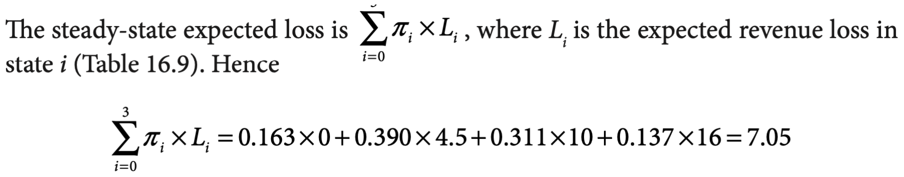
$$(1,0,0,0) \times \begin{bmatrix} 0.45 & 0.30 & 0.20 & 0.05 \\ 0.15 & 0.60 & 0.15 & 0.10 \\ 0.10 & 0.30 & 0.40 & 0.20 \\ 0 & 0.10 & 0.70 & 0.20 \end{bmatrix}^2 = (0.2675, 0.38, 0.25, 0.1025)$$

   - Probability of at least one cancellation $= 0.38 + 0.25 + 0.1025 = 0.7325$
2. Steady-state expected loss
   - System of equations

$$\begin{aligned} \pi_0 &= 0.45\pi_0 + 0.15\pi_1 + 0.10\pi_2 \\ \pi_1 &= 0.30\pi_0 + 0.60\pi_1 + 0.30\pi_2 + 0.10\pi_3 \\ \pi_2 &= 0.20\pi_0 + 0.15\pi_1 + 0.40\pi_2 + 0.70\pi_3 \\ \pi_3 &= 0.05\pi_0 + 0.10\pi_1 + 0.20\pi_2 + 0.20\pi_3 \\ \pi_0 &+ \pi_1 + \pi_2 + \pi_3 = 1 \end{aligned}$$

   - Solution: $\pi = (0.163, 0.390, 0.311, 0.137)$

- Expected loss:



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

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

6.3.8 Classification of States in Markov Chain

1. **Accessible state**: State j is accessible from state i if there exists an n such that $P^n_{ij} > 0$ (there exists a path from i to j)
2. **Communicating states**: State i and j are communicating if there exists n and m such that $P^n_{ij} > 0$ and $P^m_{ji} > 0$
3. **Irreducible Markov chain**: all states are communicating
4. **Recurrent state**: if $\sum\limits_{n=1}^{\infty} P^n_{ii} = \infty$
   - 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
5. **Transient state**: if $\sum\limits_{n=1}^{\infty} P^n_{ii} < \infty$
   - Markov chain will visit state i finite number of times in the long run
6. **First passage time**: $f^n_{ii} = P[X_n = 1 = i, X_k \neq i, k = 1, 2, ..., n-1, |X_0 = i]$
   - Probability that Markov chain will enter state i exactly n steps after leaving state i without entering it before that
   - Let $F_{ii} = \sum\limits_{n=1}^{\infty} f^n_{ii}$
   - For recurrent state, $F_{ii} = 1$
   - For transient state, $F_{ii} < 1$
7. **Mean recurrence time**: $\mu_{ii} = \sum\limits_{n=1}^{\infty} n \times f^n_{ii}$
   - If mean recurrence time is finite, recurrent state is called a **positive recurrent state**
   - If infinite, it is called a **null-recurrent state**
8. **Periodic state**: $d(i)$ is the GCD of n such that $P^n_{ii} > 0$
   - Special case of recurrent state
   - $d(i) \geq 2$
9. **Aperiodic state**:$d(i) = 1$
10. **Ergodic Markov chain**:
    - State i is ergodic if it is positive recurrent and aperiodic
    - Chain is ergodic if all states are ergodic
    - Satisfies system of equations

    $$\pi = \pi \times P \sum_{k=1}^{m} \pi_k = 1$$
11. **Limiting probability**:
    - $\lim\limits_{n\to\infty} p^n_{ij}$
    - Unlike stationary distribution, might not be unique and independent of inital state

6.3.9 Markov Chains with Absorbing States

- State i is absorbing if $P_{ii} = 1$
- If system enters this state, it will remain in this state
- Absorbing state Markov chain: Markov chain with at least one absorbing state
- Non-absorbing states in an absorbing state Markov chain are transient states

- Transition matrix not regular - no stationary distribution
- Questions to answer
    1. The probability of eventual absorption to a specific absorbing state from various transient states
    2. The expected time to absorption from a transient state to absorbing states

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

- Group absorbing and non-absorbing (transient) states and arrange rows in the matrix
- Top rows: absorbing, bottom rows: transient
- Matrix P divided into 4 matrices: I, 0, R, Q

$$P = \begin{bmatrix} & A & T \\ A & I & 0 \\ T & R & Q \end{bmatrix}$$

1. I : identity matrix
2. 0 : zero matrix
3. R : probability of absorption from transient state to absorbing state
4. Q : transition between transient states

- Eventual absorption probability: long-run (limiting probability) value of R

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

- For large n, $\left(\sum_{k=0}^{n-1} Q^k\right) R$ gives the probability of eventual absorption to an absorbing state
- As $n \to \infty$, $\sum_{k=0}^{n-1} Q^k = F = (I - Q)^{-1}$
- F is called fundamental matrix
- Expected time to absorption is Fc where c is a unit vector

6.3.11 Example

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:

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

Transition probabilities

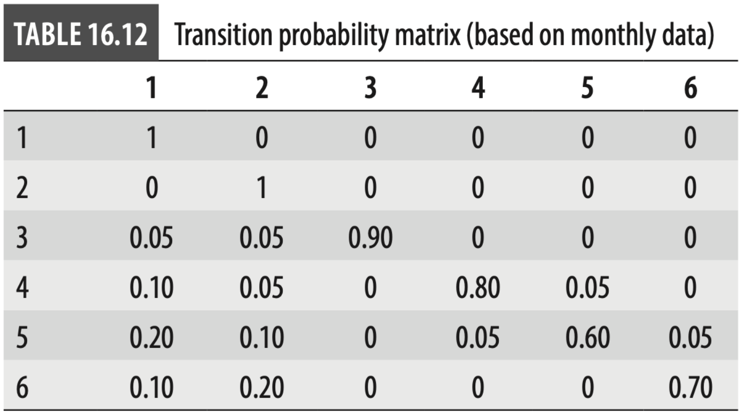


TABLE 16.12 Transition probability matrix (based on monthly data)

	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

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

6.3.12 Answer

1. Probability of eventual absorption in state 2
   - Start state $P_I = (0, 0, 0, 0, 0, 1)$
   - $Q = P[3:6][3:6]$

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

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

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

   - Probability of absorption FR

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

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

2. Expected value of time taken to absorption if the current state is 4

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

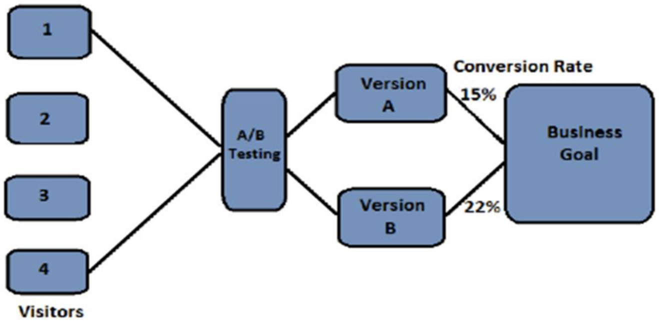
Expected value of time to absorption when the current state is 4 is 5.91 months.

7 A/B Testing

- Bucket testing or split run testing
- User experience research methodology
- Randomised experiment with 2 variants A and B
- Two-sample hypothesis testing
- 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

Conversion Rate

- The instance when any visitor on your website performs a desired action



A/B testing in popular industries

- Media & publishing industry - Netflix
- eCommerce industry - Amazon
- Travel industry - Booking.com
- B2B/SaaS Industry - POSist

7.1 How to perform A/B test

1. Research
   - How the website is performing now
   - Collect data (number of users, pages with most traffic, conversion goal)
   - Tools: Google Analytics, Omniture, Mixpanel etc
2. Observe and Formulate Hypothesis
   - Data backed hypotheses
   - Ease of setup, impact on macro goals
3. Create Variations
   - Control and variations
   - Test multiple variations
   - Eg: try shorted form with less fields
4. Run test (testing methods)
   - A/B testing
   - Multivariate testing
   - Split URL testing
   - Multipage testing
5. Result analysis and deployment
   - Metrics
   - Percentage increase, confidence level
   - Deploy if successful

7.2 What to A/B test

1. Headlines and sub-headlines
   - Short, to-the-point
   - Fonts, sizes, copy, messaging
2. Body
   - Writing style
   - Formatting
3. Design and layout
4. Navigation

5. Forms
6. CTA (Call to action)
7. Social proof

7.3 Mistakes to avoid

1. Not planning your optimisation roadmap
2. Testing too many elements together
3. Ignoring statistical significance
4. Using unbalanced traffic
5. Testing for incorrect duration
6. Failing to follow an iterative process
7. Failing to consider external factors
8. Using the wrong tools
9. Sticking to plain vanilla A/B testing method

7.4 Challenges of A/B testing

1. Deciding what to test
2. Formulating hypotheses
3. Locking in on sample size
4. Analysing test results
5. Maintaining a testing culture
6. Changing Experiment Settings in the Middle of an A/B Test (avoid)

7.5 How to make a testing calendar

1. Measure
   - Plan according to business goals
   - Define business objectives, website goals, performance indicators, target metrics
2. Prioritise
   - Scientifically sort hypotheses
3. A/B test
4. Repeat
   - Three outcomes
     (a) One of your variations would have won with statistical significance
     (b) Control was the better version
     (c) Test failed and produces insignificant results
5. Maintaining a testing culture
6. Changing Experiment Settings in the Middle of an A/B Test (avoid)

8 Business Value of Data Analytics

Five important drivers to demonstrate ROI of analytics efforts

1. Acquire more of the right customers
2. Marketing Attribution and Media Mix Modeling
   - Wanamaker problem: half the money spent on advertising is wasted
   - Do not know which half
3. More Revenue from Current Customers
4. Growth From A/B Testing
5. Moving at the speed of business