CS544 Module2

Suresh Kalathur

Module2

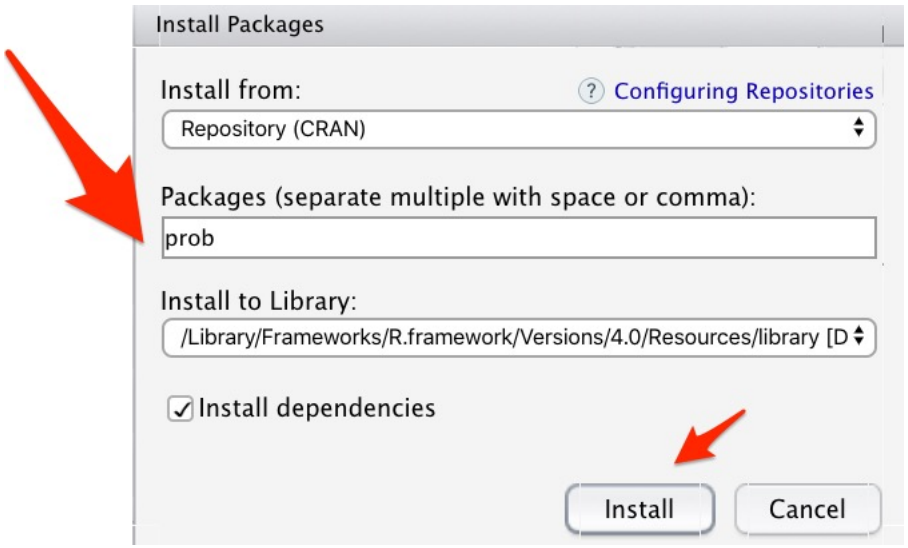
- Probability
- Conditional Probability
- Bayes Theorem
- R Programming Constructs
- Reading and Writing Data

Probability

- Random Experiment
- Sample Space
  - Set of all possible outcomes
- “prob” package of R
  - Common sample spaces
    - Tossing coins, rolling dice, cards, etc.
- Sampling from an Urn
- Event
  - Subset of sample space
  - Probability of events

Probability using R

- Install R package (prob)
  - RStudio
    - Tools -> Install Packages…



    - If prompted for compile from sources, type **no**

Probability using R

Use Package *prob*

```
> library(prob)
```

```
> S <- tosscoin(3, makespace = TRUE)
> S
  toss1 toss2 toss3 probs
1     H     H     H 0.125
2     T     H     H 0.125
3     H     T     H 0.125
4     T     T     H 0.125
5     H     H     T 0.125
6     T     H     T 0.125
7     H     T     T 0.125
8     T     T     T 0.125
```

```
> subset(S, toss1 == 'H' & toss3 == 'H')
  toss1 toss2 toss3 probs
1     H     H     H 0.125
3     H     T     H 0.125
> Prob(S, toss1 == 'H' & toss3 == 'H')
[1] 0.25
```

```
> subset(S, toss1 == 'H' | toss3 == 'H')
  toss1 toss2 toss3 probs
1     H     H     H 0.125
2     T     H     H 0.125
3     H     T     H 0.125
4     T     T     H 0.125
5     H     H     T 0.125
7     H     T     T 0.125
> Prob(S, toss1 == 'H' | toss3 == 'H')
[1] 0.75
```

```
> subset(S, isin(S, c('H', 'T')))
  toss1 toss2 toss3 probs
2     T     H     H 0.125
3     H     T     H 0.125
4     T     T     H 0.125
5     H     H     T 0.125
6     T     H     T 0.125
7     H     T     T 0.125
```

```
> subset(S, isin(S, c('H', 'T'), ordered = TRUE))
  toss1 toss2 toss3 probs
3     H     T     H 0.125
5     H     H     T 0.125
6     T     H     T 0.125
7     H     T     T 0.125
```

…Probability using R

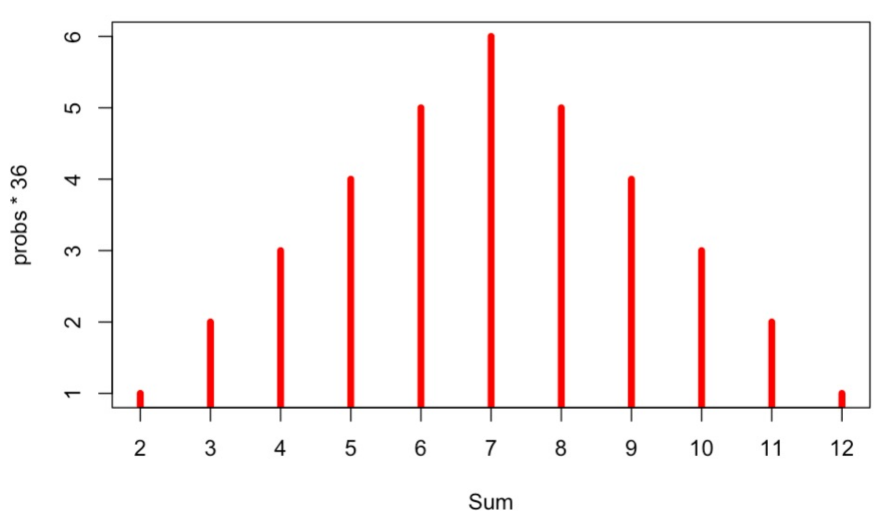
```
> S <- rolldie(2, makespace = TRUE); S
   X1 X2      probs
1   1  1 0.02777778
2   2  1 0.02777778
3   3  1 0.02777778
4   4  1 0.02777778
5   5  1 0.02777778
6   6  1 0.02777778
7   1  2 0.02777778
8   2  2 0.02777778
9   3  2 0.02777778
10  4  2 0.02777778
11  5  2 0.02777778
12  6  2 0.02777778
13  1  3 0.02777778
14  2  3 0.02777778
15  3  3 0.02777778
16  4  3 0.02777778
17  5  3 0.02777778
18  6  3 0.02777778
19  1  4 0.02777778
20  2  4 0.02777778
21  3  4 0.02777778
22  4  4 0.02777778
23  5  4 0.02777778
24  6  4 0.02777778
25  1  5 0.02777778
26  2  5 0.02777778
27  3  5 0.02777778
28  4  5 0.02777778
29  5  5 0.02777778
30  6  5 0.02777778
31  1  6 0.02777778
32  2  6 0.02777778
33  3  6 0.02777778
34  4  6 0.02777778
35  5  6 0.02777778
36  6  6 0.02777778
```

```
> subset(S, X1 == X2)
   X1 X2      probs
1   1  1 0.02777778
8   2  2 0.02777778
15  3  3 0.02777778
22  4  4 0.02777778
29  5  5 0.02777778
36  6  6 0.02777778
> Prob(S, X1 == X2)
[1] 0.1666667
```

```
> subset(S, X1 + X2 >= 10)
   X1 X2      probs
24  6  4 0.02777778
29  5  5 0.02777778
30  6  5 0.02777778
34  4  6 0.02777778
35  5  6 0.02777778
36  6  6 0.02777778
> Prob(S, X1 + X2 >= 10)
[1] 0.1666667
```

```
> subset(S, X1 + X2 == 10)
   X1 X2      probs
24  6  4 0.02777778
29  5  5 0.02777778
34  4  6 0.02777778
> Prob(S, X1 + X2 == 10)
[1] 0.08333333
```

```
> subset(S, X1 + X2 == 7)
   X1 X2      probs
6   6  1 0.02777778
11  5  2 0.02777778
16  4  3 0.02777778
21  3  4 0.02777778
26  2  5 0.02777778
31  1  6 0.02777778
> Prob(S, X1 + X2 == 7)
[1] 0.1666667
```



```
> S <- cards(makespace = TRUE)
```

...Prob function

```
> nrow(S)
[1] 52
> head(S, n = 2)
  rank suit      probs
1    2 Club 0.01923077
2    3 Club 0.01923077
> tail(S, n = 2)
   rank  suit      probs
51    K Spade 0.01923077
52    A Spade 0.01923077
```

```
> subset(S, rank %in% 2:4)
   rank    suit      probs
1     2    Club 0.01923077
2     3    Club 0.01923077
3     4    Club 0.01923077
14    2 Diamond 0.01923077
15    3 Diamond 0.01923077
16    4 Diamond 0.01923077
27    2   Heart 0.01923077
28    3   Heart 0.01923077
29    4   Heart 0.01923077
40    2   Spade 0.01923077
41    3   Spade 0.01923077
42    4   Spade 0.01923077
> Prob(S, rank %in% 2:4)
[1] 0.2307692
```

```
> A <- subset(S, rank == "Q")
> A
   rank    suit      probs
11    Q    Club 0.01923077
24    Q Diamond 0.01923077
37    Q   Heart 0.01923077
50    Q   Spade 0.01923077
```

```
> Prob(A)
[1] 0.07692308
>
> Prob(S, rank == "Q")
[1] 0.07692308
```

```
> subset(S, rank %in% c(10, "Q") & suit %in% c('Diamond', 'Spade'))
   rank    suit      probs
22   10 Diamond 0.01923077
24    Q Diamond 0.01923077
48   10   Spade 0.01923077
50    Q   Spade 0.01923077
>
> Prob(S, rank %in% c(10, "Q") & suit %in% c('Diamond', 'Spade'))
[1] 0.07692308
```

Counting Methods

- Sampling from an Urn (pick **k** objects)
  - Distinguishable objects (out of **n** objects)
- Four options
  - Ordered sampling with replacement n^k
  - Ordered sampling without replacement $\frac{n!}{(n-k)!}$
  - Unordered sampling without replacement

$$\frac{n!}{k!(n-k)!} = \binom{n}{k} = \binom{n}{n-k}$$

  - Unordered sampling with replacement

$$\binom{n+k-1}{k} = \binom{n+k-1}{n-1} = \frac{(n+k-1)!}{k!(n-1)!}$$

Unordered Sampling Without Replacement

Combinations

```
> urnsamples(1:5, size = 3)
   X1 X2 X3
1   1  2  3
2   1  2  4
3   1  2  5
4   1  3  4
5   1  3  5
6   1  4  5
7   2  3  4
8   2  3  5
9   2  4  5
10  3  4  5
```

```
> urnsamples(1:5, size = 2)
   X1 X2
1   1  2
2   1  3
3   1  4
4   1  5
5   2  3
6   2  4
7   2  5
8   3  4
9   3  5
10  4  5
```

Ordered Sampling Without Replacement

Permutations

```
> urnsamples(1:5, size = 2,
+            replace = FALSE, ordered = TRUE)
   X1 X2
1   1  2
2   2  1
3   1  3
4   3  1
5   1  4
6   4  1
7   1  5
8   5  1
9   2  3
10  3  2
11  2  4
12  4  2
13  2  5
14  5  2
15  3  4
16  4  3
17  3  5
18  5  3
19  4  5
20  5  4
```

```
> urnsamples(1:5, size = 3,
+            replace = FALSE, ordered = TRUE)
   X1 X2 X3
1   1  2  3
2   1  3  2
3   3  1  2
4   3  2  1
5   2  3  1
6   2  1  3
7   1  2  4
8   1  4  2
9   4  1  2
10  4  2  1
11  2  4  1
12  2  1  4
13  1  2  5
14  1  5  2
15  5  1  2
16  5  2  1
17  2  5  1
18  2  1  5
19  1  3  4
20  1  4  3
21  4  1  3
22  4  3  1
23  3  4  1
24  3  1  4
25  1  3  5
26  1  5  3
27  5  1  3
28  5  3  1
29  3  5  1
30  3  1  5
31  1  4  5
32  1  5  4
33  5  1  4
34  5  4  1
35  4  5  1
36  4  1  5
37  2  3  4
38  2  4  3
39  4  2  3
40  4  3  2
41  3  4  2
42  3  2  4
43  2  3  5
44  2  5  3
45  5  2  3
46  5  3  2
47  3  5  2
48  3  2  5
49  2  4  5
50  2  5  4
51  5  2  4
52  5  4  2
53  4  5  2
54  4  2  5
55  3  4  5
56  3  5  4
57  5  3  4
58  5  4  3
59  4  5  3
60  4  3  5
```

Picking 3 out of 3

`

```
> urnsamples(1:3, size = 3,
+              replace = FALSE, ordered = FALSE)
  X1 X2 X3
1  1  2  3
>
> urnsamples(1:3, size = 3,
+              replace = FALSE, ordered = TRUE)
  X1 X2 X3
1  1  2  3
2  1  3  2
3  3  1  2
4  3  2  1
5  2  3  1
6  2  1  3
```

```
> urnsamples(1:3, size = 3,
+              replace = TRUE, ordered = FALSE)
   X1 X2 X3
1   1  1  1
2   1  1  2
3   1  1  3
4   1  2  2
5   1  2  3
6   1  3  3
7   2  2  2
8   2  2  3
9   2  3  3
10  3  3  3
```

```
> urnsamples(1:3, size = 3,
+              replace = TRUE, ordered = TRUE)
   X1 X2 X3
1   1  1  1
2   2  1  1
3   3  1  1
4   1  2  1
5   2  2  1
6   3  2  1
7   1  3  1
8   2  3  1
9   3  3  1
10  1  1  2
11  2  1  2
12  3  1  2
13  1  2  2
14  2  2  2
15  3  2  2
16  1  3  2
17  2  3  2
18  3  3  2
19  1  1  3
20  2  1  3
21  3  1  3
22  1  2  3
23  2  2  3
24  3  2  3
25  1  3  3
26  2  3  3
27  3  3  3
```

Conditional Probability

- P(B|A)

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

- Multiplication Rule

$$P(A \cap B) = P(A \text{ and } B) = P(A) \cdot P(B|A)$$

- Independent Events

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

Conditional Probability

```
S <- rolldie(2, makespace = TRUE)
> nrow(S)
[1] 36
```

Rolling a pair of dice

Event **A** – the two rolls are same

```
> A <- subset(S, X1 == X2)
> A
   X1 X2      probs
1   1  1 0.02777778
8   2  2 0.02777778
15  3  3 0.02777778
22  4  4 0.02777778
29  5  5 0.02777778
36  6  6 0.02777778
> Prob(A)
[1] 0.1666667
```

```
> Prob(S, X1 == X2)
[1] 0.1666667
```

```
> B <- subset(S, X1 + X2 >= 9)
> B
   X1 X2      probs
18  6  3 0.02777778
23  5  4 0.02777778
24  6  4 0.02777778
28  4  5 0.02777778
29  5  5 0.02777778
30  6  5 0.02777778
33  3  6 0.02777778
34  4  6 0.02777778
35  5  6 0.02777778
36  6  6 0.02777778
> Prob(B)
[1] 0.2777778
```

```
> Prob(S, X1 + X2 >= 9)
[1] 0.2777778
```

Event **B** – the sum is at least 9

```
> Prob(B, given = A)
[1] 0.3333333
```

```
> Prob(A, given = B)
[1] 0.2
```

```
> subset(A, X1 + X2 >= 9)
   X1 X2      probs
29  5  5 0.02777778
36  6  6 0.02777778
```

```
> subset(B, X1 == X2)
   X1 X2      probs
29  5  5 0.02777778
36  6  6 0.02777778
```

Same as

```
> subset(S, (X1 == X2) & (X1 + X2 >= 9))
   X1 X2      probs
29  5  5 0.02777778
36  6  6 0.02777778
```

```
> probspace(A)
   X1 X2     probs
1   1  1 0.1666667
8   2  2 0.1666667
15  3  3 0.1666667
22  4  4 0.1666667
29  5  5 0.1666667
36  6  6 0.1666667
>
> Prob(probspace(A), X1 + X2 >= 9)
[1] 0.3333333
```

```
> probspace(B)
   X1 X2 probs
18  6  3   0.1
23  5  4   0.1
24  6  4   0.1
28  4  5   0.1
29  5  5   0.1
30  6  5   0.1
33  3  6   0.1
34  4  6   0.1
35  5  6   0.1
36  6  6   0.1
>
> Prob(probspace(B), X1 == X2)
[1] 0.2
```

Bayes Theorem

- Developed by Reverend Bayes
  - To infer the existence of God
- Historical
  - Cracking the infamous Nazi Enigma code in WWII (Alan Turing)
- Finance & Business
  - Evaluating interest rates
  - Managing net income streams
  - Lending Credit
- Insurance Companies
  - Risk of flooding in coastal areas
- Health
  - Probability of having disease X given that test Y is positive
- AI - Driverless vehicles
  - Improving decision making using probabilities on road conditions
- AI – Robots
  - Robot’s next step given the steps it already has executed
- Others
  - Sort spam from e-mail

Bayes Theorem

$$P(A|B) = \frac{P(A)\,P(B|A)}{P(B)}$$

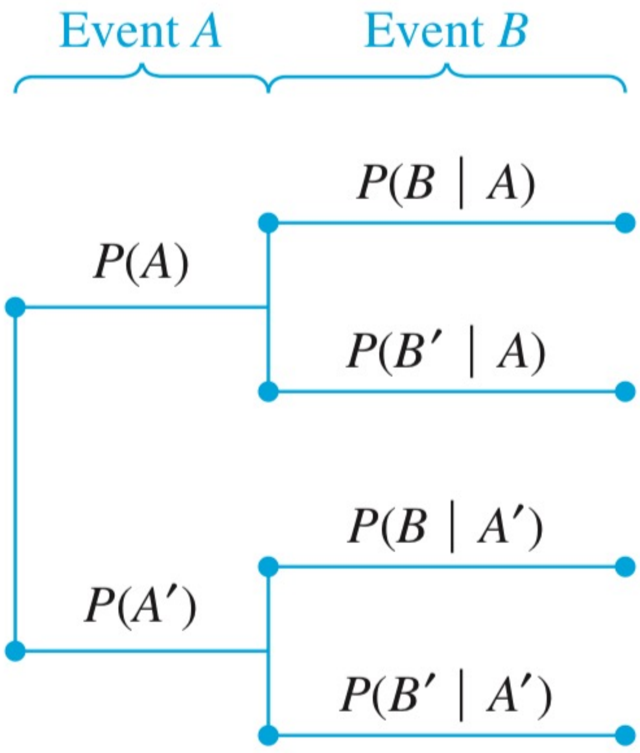
- Do a search for
  - Automatic shoe laces movie
- Result
  - Back to the future

- What we know
  - P(A) – how likely A is on its own
  - P(B) – how likely B is on its own
  - P(B|A) – how often B happens given that A happens
- What the theorem tells us?
  - How often A happens given that B happens , **P(A|B)**

Example (Fire and Smoke)

- P(Fire) – how often there is a fire
- P(Smoke) – how often we see smoke
- P(Smoke|Fire) – how often we can see smoke given there is fire
- P(Fire|Smoke) – how often there is fire given we can see smoke
- Given that dangerous fires are rare (1%), smoke is fairly common (10%), and that 90% of dangerous fires make smoke
  - P(Fire) = 0.01, P(Smoke) = 0.10, P(Smoke|Fire) = 0.90
- What is the probability of a dangerous fire given that we see a smoke?
  - P(Fire|Smoke) = $\frac{P(Fire) * P(Smoke|Fire)}{P(Smoke)} = \frac{0.01 * 0.90}{0.10}$ = 0.09
- Answer: 9% probability of a dangerous fire given we sighted smoke

More Examples: https://www.mathsisfun.com/data/bayes-theorem.html



Bayes Theorem…

- Forward looking probability
  - Probability that event B will occur given event A occurred
  - Given for us
- Backward looking probability
  - Probability that event A has occurred given event B has occurred

Rule of Total Probability

Rule of Total Probability

Suppose the events $A_1, A_2, \ldots, A_k$ are **mutually exclusive** and **exhaus**tive, i.e., exactly one of these events will occur and they cover the entire sample space.

For any event B, the events (A_1 and B), (A_2 and B), …, (A_k and B) are mutually exclusive, and hence P(B) =

P(A_1 and B) + P(A_2 and B) + … + P(A_k and B)

Using the multiplication rule,

$$P(B) = P(B|A_1)*P(A_1) + P(B|A_2)*P(A_2) + \ldots + P(B|A_k)*P(A_k)$$

$$P(B) = \sum_{j=1}^{k} P(B|A_j) * P(A_j)$$

Bayes' Theorem

Bayes' Theorem:

Suppose the events $A_1, A_2, \ldots, A_n$ are mutually exclusive and exhaustive. Let B be any event.

Given

Prior probabilities: $P(A_1), P(A_2), \ldots, P(A_n)$, and

Conditional probabilities: $P(B|A_1), P(B|A_2), \ldots, P(B|A_n)$

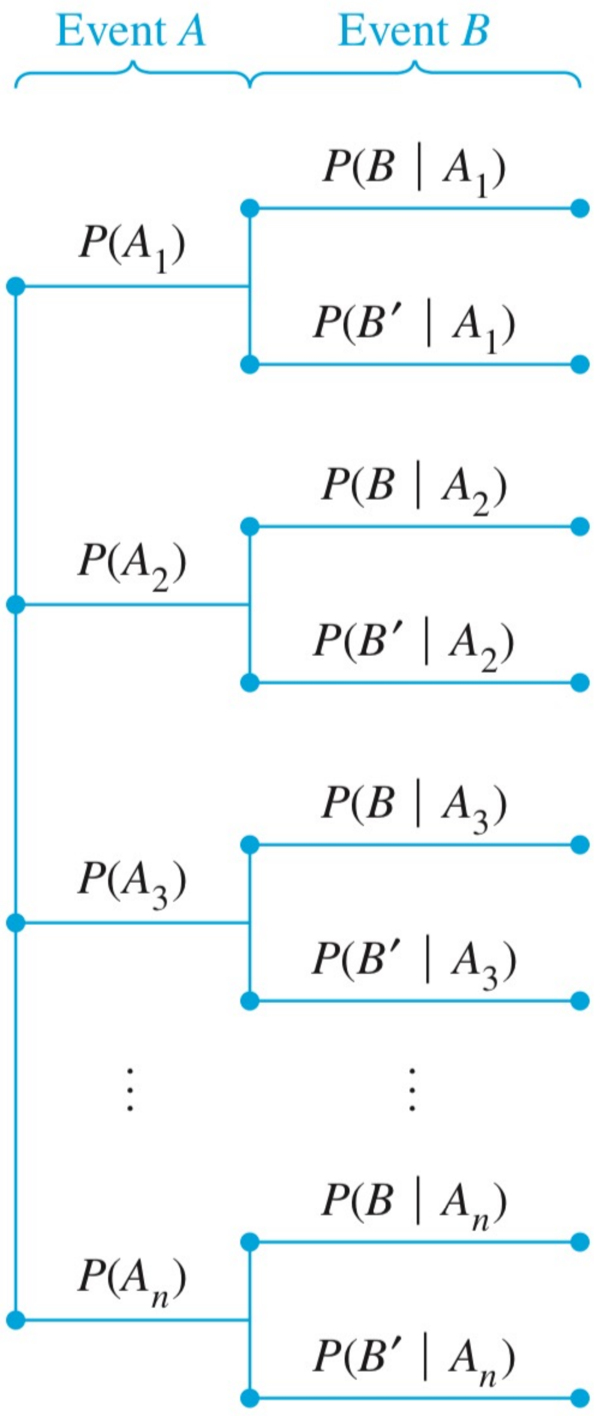
Determine

Posterior probabilities: $P(A_1|B), P(A_2|B), \ldots, P(A_n|B)$

$$P(A_i|B) = \frac{P(A_i \text{ and } B)}{P(B)} = \frac{P(B|A_i) * P(A_i)}{P(B)}$$

$$P(A_i|B) = \frac{P(B|A_i) * P(A_i)}{\sum_{j=1}^{n} P(B|A_j) * P(A_j)}$$

Bayes Theorem…



$P(B) = P(A_1)*P(B|A_1) + P(A_2)*P(B|A_2) + \ldots + P(A_n)*P(B|A_n)$

$$P(A_1|B) = \frac{P(A_1)*P(B|A_1)}{P(B)}$$

…

$$P(A_n|B) = \frac{P(A_n)*P(B|An)}{P(B)}$$

Example1 – Rule of Total Probability

Example: In an university, 60% are undergraduate students, 35% are graduate students, and 5% are postdocs. 55% of undergraduates are female, 15% of graduate students are female, and 10% of postdocs are female.
What is the probability that a randomly selected student is female?

Type	Percentage of college students	Percentage females
Undergraduate	60	55
Graduate	35	15
Postdoc	5	10
	100%	

Event B = Selected student is a female

Event A1 = Selected student is an undergraduate
Event A2 = Selected student is a graduate
Event A3 = Selected student is a postdoc

A1, A2, and A3 are mutually exclusive and exhaustive

P(B) = P(A1 and B) + P(A2 and B) + P(A3 and B)

P(A1) = 0.60	P(B|A1) = 0.55
P(A2) = 0.35	P(B|A2) = 0.15
P(A3) = 0.05	P(B|A3) = 0.10

P(B) = P(B|A1)*P(A1) + P(B|A2)*P(A2) + P(B|A3)*P(A3)
= 0.55*0.60 + 0.15*0.35 + 0.10*0.05
= 0.3875

With a probability of 0.3875, a randomly selected student is a female

Example1 - Bayes' Theorem

Example: In an university, 60% are undergraduate students, 35% are graduate students, and 5% are postdocs. 55% of undergraduates are female, 15% of graduate students are female, and 10% of postdocs are female. What is the probability that a randomly selected female student is:

an undergraduate? a graduate? A postdoc?

Type	Percentage of college students	Percentage females
Undergraduate	60	55
Graduate	35	15
Postdoc	5	10
	100%	

Event B = Selected student is a female

Event A1 = Selected student is an undergraduate

Event A2 = Selected student is a graduate

Event A3 = Selected student is a postdoc

P(A1) = 0.60	P(B|A1) = 0.55
P(A2) = 0.35	P(B|A2) = 0.15
P(A3) = 0.05	P(B|A3) = 0.10

$P(B) = P(B|A1)*P(A1) + P(B|A2)*P(A2) + P(B|A3)*P(A3)$
$= 0.55*0.60 + 0.15*0.35 + 0.10*0.05$
$= 0.3875$

$P(A1|B) = P(B|A1)*P(A1)/P(B) = 0.55*0.60/0.3875 = 0.85$
$P(A2|B) = P(B|A2)*P(A2)/P(B) = 0.15*0.35/0.3875 = 0.14$
$P(A3|B) = P(B|A3)*P(A3)/P(B) = 0.10*0.05/0.3875 = 0.01$

With a probability of 0.85, a randomly selected female student is an Undergraduate.

Example2 – Rule of Total Probability

Example: A company orders parts from three different suppliers, *Supplier1*, *Supplier2*, and *Supplier3*. From historical records, 3% of parts provided by *Supplier1* are defective, 5% of parts provided by *Supplier2* are defective, and 4% of parts provided by *Supplier3* are defective. The current inventory consists of 5000 units from *Suppler1*, 3500 units from *Supplier2*, and 2000 units from *Supplier3*.

What is the probability that a randomly selected part is defective?

Type	Inventory	Percentage Defective
Supplier1	5000	3
Supplier2	3500	5
Supplier3	2000	4
	10500	

Event D = Selected part is a defective one

Event S1 = Selected part is from Supplier1

Event S2 = Selected part is from Supplier2

Event S3 = Selected part is from Supplier3

S1, S2, and S3 are mutually exclusive and exhaustive

$P(S1) = \frac{50}{105} = 0.48$	$P(D\|S1) = 0.03$
$P(S2) = \frac{35}{105} = 0.33$	$P(D\|S2) = 0.05$
$P(S3) = \frac{20}{105} = 0.19$	$P(D\|S3) = 0.04$

P(D) = P(S1 and D) + P(S2 and D) + P(S3 and D)

P(D) = P(D|S1)*P(S1) + P(D|S2)*P(S2) + P(D|S3)*P(S3)
= 0.03*0.48 + 0.05*0.33 + 0.04*0.19
= 0.039

So, there is a 4% chance that a randomly selected part is a defective

Example2 – Bayes Theorem

Example: A company orders parts from three different suppliers, *Supplier1*, *Supplier2*, and *Supplier3*. From historical records, 3% of parts provided by *Supplier1* are defective, 5% of parts provided by *Supplier2* are defective, and 4% of parts provided by *Supplier3* are defective. The current inventory consists of 5000 units from *Suppler1*, 3500 units from *Supplier2*, and 2000 units from *Supplier3*.

What is the probability that a randomly selected defective part: came from *Supplier1*? Came from *Supplier2*? Came from *Supplier3*?

Type	Inventory	Percentage Defective
Supplier1	5000	3
Supplier2	3500	5
Supplier3	2000	4
	10500	

Event D = Selected part is a defective one

Event S1 = Selected part is from *Supplier1*

Event S2 = Selected part is from *Supplier2*

Event S3 = Selected part is from *Supplier3*

$P(S1) = \frac{50}{105} = 0.48$	$P(D\|S1) = 0.03$
$P(S2) = \frac{35}{105} = 0.33$	$P(D\|S2) = 0.05$
$P(S3) = \frac{20}{105} = 0.19$	$P(D\|S3) = 0.04$

$P(D) = P(D|S1)*P(S1) + P(D|S2)*P(S2) + P(D|S3)*P(S3)$
$= 0.03*0.48 + 0.05*0.33 + 0.04*0.19 = 0.039$

$P(S1|D) = P(D|S1)*P(S1)/P(D) = 0.03*0.48/0.039 = 0.37$
$P(S2|D) = P(D|S2)*P(S2)/P(D) = 0.05*0.33/0.039 = 0.43$
$P(S3|D) = P(D|S3)*P(S3)/P(D) = 0.04*0.19/0.039 = 0.20$

So, there is a 37% chance that a randomly selected defective part came from *Supplier1*.

R Programming Constructs

- Functions
- Scope of variables
- Control structures
    - if-else, for, while, repeat
- Reading and Writing Data