CS544 Module2

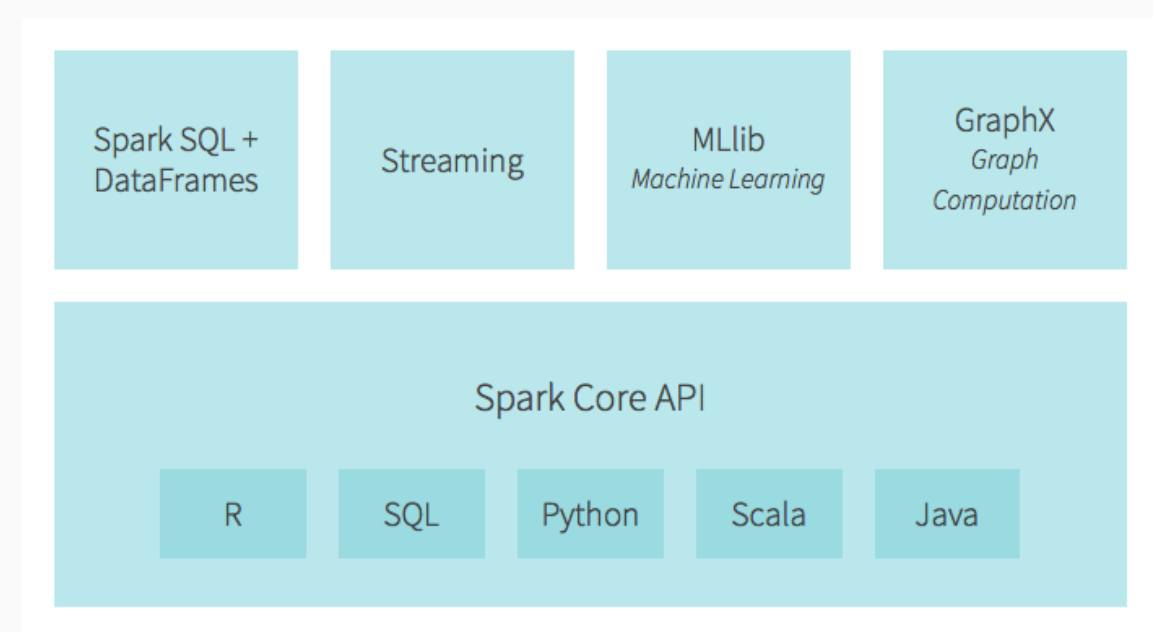
Suresh Kalathur

Module2

- Spark Overview
- Probability
- Conditional Probability
- Random Variables
- R Programming Constructs

Spark

- http://spark.apache.org/



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 using R

Package prob

```
> library('prob')
Error in library("prob") : there is no package called ‘prob’
>
```

```
> is.element("prob", installed.packages()[,"Package"])
[1] FALSE
```

```
> install.packages('prob', dependencies = TRUE)
trying URL 'https://cran.rstudio.com/bin/macosx/mavericks/contrib/3.3/prob_0.9-5.tgz'
Content type 'application/x-gzip' length 709720 bytes (693 KB)
==================================================
downloaded 693 KB


The downloaded binary packages are in
        /var/folders/s3/hy6_p79n3w1fw802t6ps40qr0000gp/T//RtmpcKGqWe/downloaded_packages
```

```
> library('prob')
Loading required package: combinat

Attaching package: ‘combinat’

The following object is masked from ‘package:utils’
```

Probability using R

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
```

…Probability using R

```
> S <- rolldie(2, makespace = TRUE)
> head(S, n = 3)
  X1 X2      probs
1  1  1 0.02777778
2  2  1 0.02777778
3  3  1 0.02777778
> tail(S, n = 3)
   X1 X2      probs
34  4  6 0.02777778
35  5  6 0.02777778
36  6  6 0.02777778
```

Prob function

- Probability of the event

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

```
> 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
```

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

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

Bayes' Rule

Rule of Total Probability

Suppose the events A_1, A_2, …, A_k are mutually exclusive and exhaustive, 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

$$P(B) = P(A_1 \text{ and } B) + P(A_2 \text{ and } B) + \ldots + P(A_k \text{ and } B)$$

Using the multiplication rule,

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

…Bayes' Rule

- Compute $P(A_1|B)$, $P(A_2|B)$, …, $P(A_k|B)$

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

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

...Probability using R

Add random variable

```
> S <- rolldie(2, makespace = TRUE)
> S <- addrv(S, U = X1 + X2)
> head(S, n = 2)
  X1 X2 U      probs
1  1  1 2 0.02777778
2  2  1 3 0.02777778
> tail(S, n = 2)
   X1 X2  U      probs
35  5  6 11 0.02777778
36  6  6 12 0.02777778
```

```
> S <- rolldie(2, makespace = TRUE)
> S <- addrv(S, FUN = sum, name = "U")
> head(S, n = 2)
  X1 X2 U      probs
1  1  1 2 0.02777778
2  2  1 3 0.02777778
> tail(S, n = 2)
   X1 X2  U      probs
35  5  6 11 0.02777778
36  6  6 12 0.02777778
```

```
> S <- rolldie(2, makespace = TRUE)
> mySum <- function(data) { data[1] + data[2] }
> S <- addrv(S, FUN = mySum, name = "U")
> head(S, n = 2)
  X1 X2 U      probs
1  1  1 2 0.02777778
2  2  1 3 0.02777778
> tail(S, n = 2)
   X1 X2  U      probs
35  5  6 11 0.02777778
36  6  6 12 0.02777778
```

...Probability using R

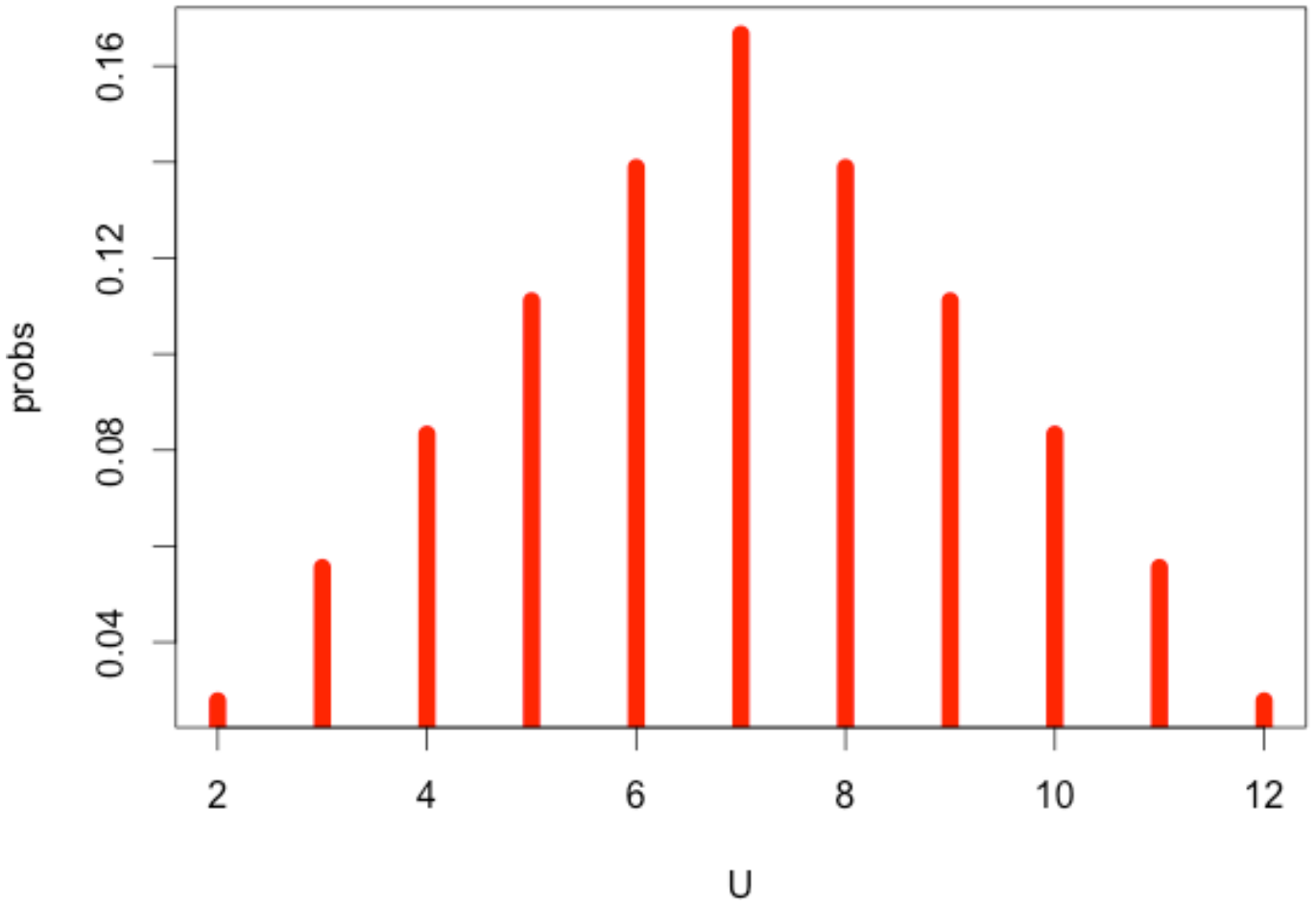
Marginal distribution

```
> S <- marginal(S, vars = "U")
> S
    U      probs
1   2 0.02777778
2   3 0.05555556
3   4 0.08333333
4   5 0.11111111
5   6 0.13888889
6   7 0.16666667
7   8 0.13888889
8   9 0.11111111
9  10 0.08333333
10 11 0.05555556
11 12 0.02777778
```

...R

Plot

```
> plot(probs ~ U, S, type='h', col = "red", lwd=10)
```



R Programming Constructs

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