Problem Set1

Mashhood Syed

October 1, 2014

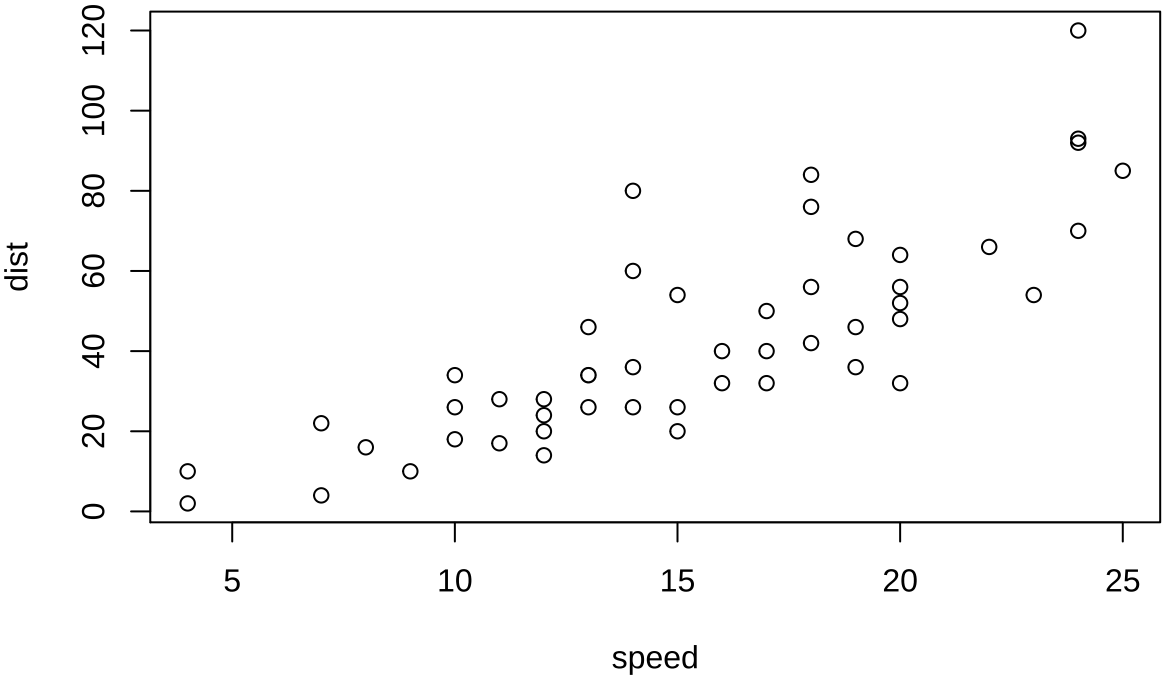
This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2  
##  1st Qu.:12.0   1st Qu.: 26  
##  Median :15.0   Median : 36  
##  Mean   :15.4   Mean   : 43  
##  3rd Qu.:19.0   3rd Qu.: 56  
##  Max.   :25.0   Max.   :120  
```

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

Q1.

Data Set1 Title: “Random Acts of Pizza”.
Source: http://cs.stanford.edu/~althoff/raop-dataset Summary: In this data set, users of the popular Reddit website made a request to the Reddit

community for a free pizza. The researchers used a total of 5671 observations to output the likelihood that the requestor would receive a free pizza. If we view the data set as a state, say Sn, then we can assume that there were previous states Sn-1, Sn-2, . . . , Sn-n. The stochastic process that generated this data can be described as the bodys response to satisfying a basic human need. In a previous state, a user of a website community was looking for a way to satisfy his/her need. Going back further in the state machine, they could have been financially solvent and were able to satiate their hunger. Prior to this state, they may have not have had the internet as an outlet and therefore asked friends and family for help. Because the states could have unfolded in infinitely different ways, we can stop the description here.

Based on the research paper success depends upon: -who is asking -how they are asking -when they are asking -what is being requested

Variables: There were a total of 33 variables per observation. The variables ranged from: Binary (True or False, ie Did the requestor get the pizza or not) Counts (The number of "upvotes" and "downvotes" that the posting received) Ordinal (The number of days that had passed from date of account creation to posting) Categorical (A list of the different subreddits/groups that the user had membership in) Datetime (The timestamp of the request)

Data Set2 Title: "Adult Data Set".
Source: http://archive.ics.uci.edu/ml/datasets/Adult Summary: The data was used in a machine learning experiment to predict whether income exceeds $50k/yr based on US census data. If we view the data set as a state, say Sn, then we can assume that there were
previous states Sn-1, Sn-2, . . . , Sn-n. The process that generated the data set can be viewed as human evolution. Between Sn-n and Sn there was a change from one state to the next. For example, one of the observations got married, another observation got a college degree, and so on. All of these events caused the state to change for that observation which ultimately changes the outcome or probability of the original question. Going even further back in state, we know that all of these observations had to be born in order to get to the current state. Prior to this state, the concept of payment for labor needed to be accepted by society. Going back even further, there must have been a demand for a particular type of work such that someone was willing to pay for that work to be done. Because the previous states could have transformed in infinitely different ways, we can stop the description at this point. The description of the process that led up to the data is just one instance of a set of state transitions.

Variables: There were a total of 14 variables per observation. The variables ranged from: Binary (sex: M or F) Ordinal (age, education number) Categorical (Workclass, Education, marital status, occupation, relationship, race, native country) money (capital-gain, capital-loss) time (hours per week worked)

Q2.

```
Conditional probability for 2 events A and B = P(A | B) or P(B | A)
In the case of: P(A | B) = [P(A) * P(B)] / P(B)
In the case of: P(B | A) = [P(B) * P(A)] / P(A)

Conditional probability is the probability of a event (A) given that another event (B) has already
occurred.  Event B could be related or unrelated to Event A.

Marginal probability is the likelihood of a single event A occurring. It can be stated as P(A).

Prior probability is the original probability of an outcome (already known based on past knowledge)
which will be updated to create posterior probabilities (based on Bayes Rule).

Ex 1: Given 5 marbles in a bag, 2 Blue and 3 Red (w/o replacement)

      Marginal probability says whats the P(1 Blue)?

      Conditional probabiility says whats the P(1 Red | 1 Blue)?
      Prior probability says P(1 Red) for the past 1000 trials = .62

Ex 2: Let P(A) = A random Hard Drive crashing = .05
      Let P(B) = The computer is manufactured by HP = .28

      Marginal probability says what is P(A)?

      Conditional probability says what is P(A | B)?

      Prior probability says P(A) happening in 2013 = .06

Ex 3: Let P(A) = drawing a King on 1st draw from a standard 52 card deck = 1/13
      Let P(B) = drawing a King on the 2nd draw from a standard 52 card deck = 3/51

      Marginal probability says what is P(A)?

      Conditional probability says what is P(A | B)?

      Prior probability says P(A) after 1000 trials = .072
```

Q3.

a. `A      B      C    Totals`

   1 X 4.56% 24.94% 11.90% 41.40% 2 Y 7.94% 17.30% 15.51% 40.76% 3 Z 1.79% 11.86% 4.19% 17.84% 4 Totals 14.30% 54.11% 31.60% 100.00%

b. P(V1 = x) = 41.40% P(V1 = y) = 40.76% P(V1 = z) = 17.84%

c. P(V2 = a) = 14.30% P(V2 = b) = 54.11% P(V2 = c) = 31.60%

d. P(V1 | V2 = b) = [P(V1) * P(V2 = b)] / P(V2 = b) P(V1 = x | V2 = b) = .2494/.5411 = .4609 ~ 46.09% P(V1 = y | V2 = b) = .1730/.5411 = .3197 ~ 31.97% P(V1 = Z | V2 = b) = .1186/.5411 = .2192 ~ 21.92%

e. P(V2 | V1 = z) = [P(V2) * P(V1 = z) / P(V1 = z) P(V2 = a | V1 = z) = .0179/.1784 = .1003 ~ 10.03% P(V2 = b | V1 = z) = .1186/.1784 = .6648 ~ 66.48% P(V2 = c | V1 = z) = .0419/.1784 = .2349 ~ 23.49%

Q4.

Legend: GC = Genetic Condition T = Test Pos. = Positive

a. P(GC = Yes) = 2.3% P(GC = No) = 97.7% P(T = Pos.) = P(GC = Yes)P(T = Pos. | GC = Yes) + P(GC = No)P(T = Pos. | GC = No) P(T = Pos. | GC = Yes) = 72.1% P(T = Pos. | GC = No) = 20.3%

```
P(GC = Yes | T = Pos.) = P(T = Pos. | GC = Yes)P(GC = Yes) / P(T = Pos.)
                       = (.721)(.023) / (.023)(.721) + (.977)(.203)
                       = .0165 / .2149
                       = .0767 ~ 7.67%
```

b. P(GC = Yes) = 2.3% P(GC = No) = 97.7% P(T = Pos.) = P(GC = Yes)P(T = Pos. | GC = Yes) + P(GC = No)P(T = Pos. | GC = No) P(T = Pos. | GC = Yes) = 97.3% P(T = Pos. | GC = No) = 37.2%

```
P(GC = Yes | T = Pos.) = P(T = Pos. | GC = Yes)P(GC = Yes) / P(T = Pos.)
                       = (.973)(.023) / (.023)(.973) + (.977)(.372)
                       = .0223 / .3858
                       = .0578 ~ 5.78%
```