Discrete Probability Excercises

J. Alejandro Gelves

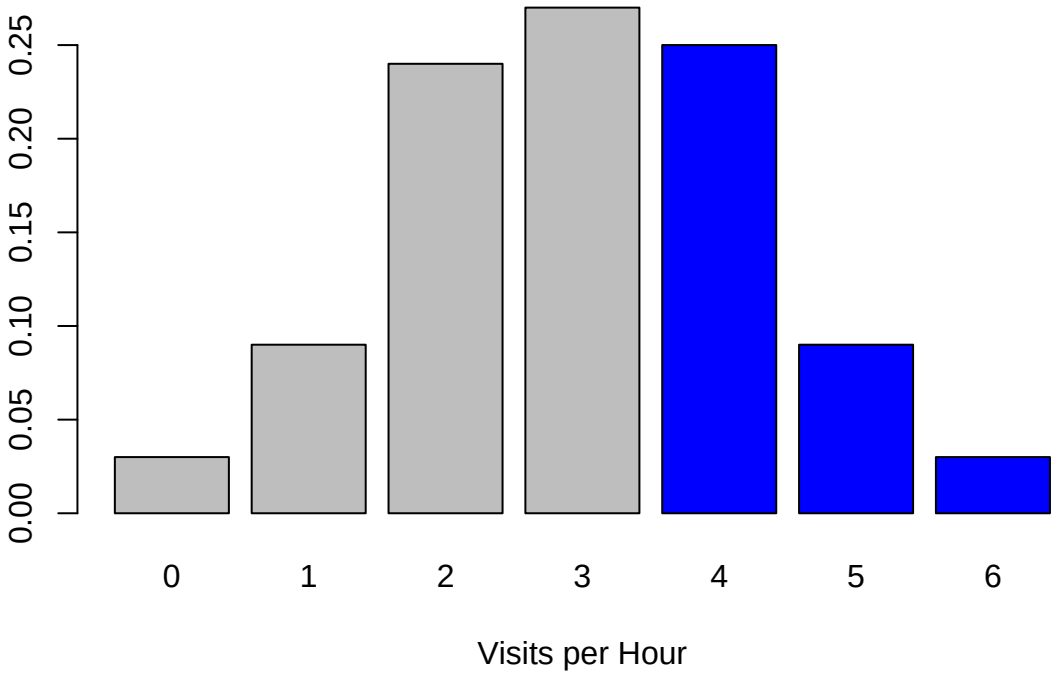
October 25, 2018

ANSWERS TO PROBLEMS IN WEB ANALYTICS.

Answer to question 1.

Without making any assumptions about the distribution of the data, the probablility that the number of visits will be between four and six (inclusive) is given by adding the probabilities associated with the 3 blue bars.

```
barplot(c(0.03, 0.09, 0.24, 0.27, 0.25, 0.09, 0.03), names.arg = c(0, 1, 2,
    3, 4, 5, 6), xlab = "Visits per Hour", col = c("grey", "grey", "grey", "grey",
    "blue", "blue", "blue"))
```



Specifically,

$$P(4 \leq NumberOfVisits \leq 6) =$$

$$\sum_{i=4}^{6} p_i = 0.25 + 0.09 + 0.03 = 0.37$$

where p_i is the probability of the i_{th} Visit per Hour bin. The probability that the number of visits will be between 4 and 6 is **0.37**.

The **mean** of the random variable is given by:

$$\bar{X} = \sum_{i=0}^{6} p_i x_i$$

$$\bar{X} = 3.01$$

This is the code:

```
prob<-c(.03,.09,.24,.27,.25,.09,.03)
visits<-c(0,1,2,3,4,5,6)
xbar<-sum(prob*visits)
```

The **variance** is given by:

$$s_x^2 = \sum_{i=0}^{6} p_i (x_i - \bar{x})^2$$

$$s_x^2 = 1.7499$$

Finding the square root yields the **standard deviation**:

$$s_x = 1.3228$$

Below the code:

```
var<- sum(prob*(visits-xbar)^2)
sd<-sqrt(var)
```

Answer to question 2.

- To assume a binomial experiment, college student visits and non-college student visits should be mutually exclusive events. The number of visitors should be fixed at 11. The probability that a visitor is a college student should be constant. Visits must be independent. That is, the probability that a college student visits the site is unaffected by the event of a non-college student visiting (or not visiting) the site.

- p, the probability of "success", is **68%** and n, the number of trials, is **11**.

- The probability of 7 successes is given by:

$$P(X = 7) =$$

$$\frac{11!}{7!4!} 0.68^7 0.32^4 = 0.2326$$

0.2326

Here is the code:

```
p7formula<-factorial(11)/(factorial(7)*factorial(4))*0.68^7*0.32^4
p7function<-dbinom(7,11,0.68)
```

- The **mean** of the binomial distribution μ_x is given by np.

$$\mu_x = 0.68 * 11 = 7.48$$

7.48

- The variance of the binomial distribution σ_x^2 is given by $np(1-p)$.

$$\sigma_x^2 = 0.68 * 11 * 0.32 = 2.3936$$

Yielding a **standard deviation** of **1.5471**.

Answer to question 3.

Let "mobile" consumers be successes (X) and "desktop" consumers be failures (Y). So the probability that 8 of the 20 randomly selected users are "mobile" is given by:

$$P(X = 8) =$$

$$\frac{20!}{8!12!}0.38^8 0.62^{12} = 0.1767$$

The code is given by:

```
p8formula<-factorial(20)/(factorial(8)*factorial(12))*0.38^8*0.62^(12)
p8function<-dbinom(8,20,0.38)
```

The probability that between 12-14 (inclusive) of the 20 randomly selected are mobile consumers is given by:

$$P(12 \leq Y \leq 14) =$$

$$P(Y \leq 14) - P(Y < 12) = 0.4988$$

Below, the code that calculates the answer.

```
p12_14<-sum(dbinom(0:14,20,0.62))-sum(dbinom(0:11,20,0.62))
p12_14cu<-pbinom(14,20,0.62)-pbinom(11,20,0.62)
```

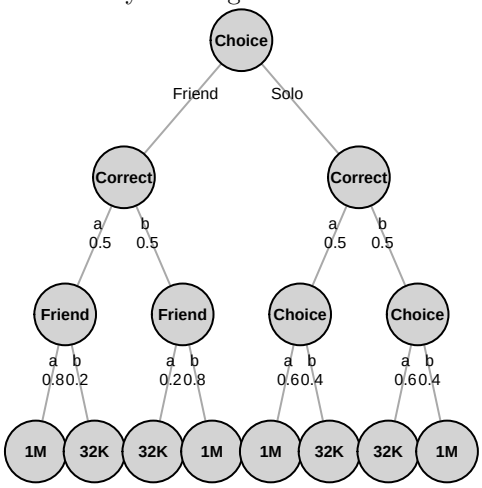
Finally, the probability that 3 or fewer randomly selected consumers are desktop user is:

$$P(Y \leq 3) = 0.00002164$$

```
p3cu<-pbinom(3,20,0.62)
```

ANSWERS TO PROBLEMS IN A GAME OF CHANCE.

I started by drawing the decision tree.



The decision tree models the outocomes of calling a friend or continuing alone. The outcomes will be compared to a third option of quiting with 500K. For both "call a friend" or "solo" there is complete uncertainty on the correct answer. So the assigned probabilities are 50/50 for a and b. In both situations. if the suggested answer aligns with the correct answer 1M is won, else we leave with 32K. The three decisions yield the following expected values:

Quit: $500,000 Solo: $516,000 Friend: $806,400

* The probability that my friend says a is:

$$P(a) = (0.5)0.8 + (0.5)0.2 = 0.5$$

* The probability that my friend says b is:

$$P(b) = (0.5)0.2 + (0.5)0.8 = 0.5$$

The best strategy would be to call a friend. It makes intuitive sense as my friend has a high probability of getting the answer right. This strategy yields the highest expected value at **$806,400**.

I discussed with my group and they suggested another solution. The actual probabilities that a and b are right are 60% and 40% respectively (my friend knows this as well). Ultimately, this would not affect my recommendation or the procedure to find the answers. But the probabilities of my friend saying a would be 0.56 and b 0.44. These are found by just substituting 0.6 and 0.4 in place of the 0.5 in the calculations above.

ANSWERS TO PROBLEMS IN HEALTHCARE ANALYTICS.

- Let D be the event that a person has the disease and ND be the event that the person does not have the disease. Let N be the event that the classifier is negative and O the event that the classifier is positive. D and ND are mutually exlusive as well as N and O. If someone has the disease, the probability that the classifier would identify the person as having the disease is given by P(O/D).
  Since
  $$P(N/D) = 0.06$$
  the probability
  $$P(O/D) = 1 - P(N/D) = 0.94$$
- It is given that the three individuals do not have the disease. The probability that the classifier will classify any single one of them as not having the disease is:

$$P(N/ND) = 0.91$$

which means that it will wrongly classify them as having the disease with a probability of:

$$P(O/ND) = 0.09$$

We can now use the binomial distribution to answer the question. That is, in three trails, we want to find the probability that at least one will be identified as having the disease, given that we know that they do not have the disease. This is given by, $P(x \geq 1) = 0.2464$ where X is the number of negative tests.

```
p3<-1-dbinom(0,3,0.09)
```

- A joint probability table is a good way to think about this problem. This table has joint probabilities in the middle and marginal probabilities at the edges (margins).

```
library(knitr)
tablep <- matrix(c("P(O and D)", "P(N and D)", "P(D)","P(O and ND)","P(N and ND)","P(ND)","P(O)","P(N)"
rownames(tablep) <- c( "O", "N","Marginal")
colnames(tablep) <- c( "D", "ND", "Marginal")
kable(tablep)
```

	D	ND	Marginal
O	P(O and D)	P(O and ND)	P(O)
N	P(N and D)	P(N and ND)	P(N)
Marginal	P(D)	P(ND)	1

The problem provides us with $P(D) = 0.17$, $P(N/D) = 0.06$ and $P(O/ND) = 0.91$. With this information the rest of the table can be filled as seen below.

```
table <- matrix(c(0.1598,0.0102,0.17,0.7553,0.0747,0.83,0.9151,0.0849,1), nrow = 3, ncol = 3)
rownames(table) <- c( "O", "N","Marginal")
colnames(table) <- c( "D", "ND", "Marginal")
kable(table)
```

	D	ND	Marginal
O	0.1598	0.7553	0.9151
N	0.0102	0.0747	0.0849
Marginal	0.1700	0.8300	1.0000

The problem asks: If a random person from this area is tested and the results indicate that the person has the disease, what is the chance the the person actually has the disease? That is, what is $P(D/O)$? We know that:

$$P(D/O) = \frac{P(O \cap D)}{P(O)}$$

Using the information in the table:

$$P(D/O) = \frac{0.1598}{0.9151}$$
$$P(D/O) = 0.1746$$

ANSWERS TO PROBLEMS IN AIRLINE ANALYTICS.

- I used the binomial distribution. Let "showing up"" be success. The probability of showing up is 80% and the number of trials is 11. Let X be the number of people that show up (successes). The probability that at most 5 people show up is given by:

$$P(0 \leq X \leq 5) = 0.0117$$

```
p5<-pbinom(5,11,0.8)
```

- The probability that exactly 10 of the persons who purchased first class tickets show up for the flight is:

$$P(X = 10) = 0.2362$$

```
p10<-dbinom(10,11,0.8)
```

- The expected profit can be calculated as:

$$1200 \sum_{i=0}^{10} X_i P(X_i) + 9000 P(11) = 11333.09$$

$11,333.09

```
e10<-seq(0,10,1)
for (i in 0:length(e10)){
e10[i]=1200*dbinom(i,11,0.8)*i
}
EV<-sum(e10)+dbinom(11,11,0.8)*9000
```

- If we assume instead that only 10 tickets are sold, the expected profit is given by:

$$1200 \sum_{i=0}^{10} X_i P(X_i) = 9600$$

$9,600
Since the cost of overbooking is only $3,000 it might be wise for the company to overbook. This might change if the cost went up to $5,000.

```
e102<-seq(0,10,1)
for (i in 0:length(e102)){
e102[i]=1200*dbinom(i,10,0.8)*i
}
EV1<-sum(e102)
```

- It does. Specifically, if a passenger does not show up it significantly affects the probability that their partner will not show up as well.