**Probability: Computing by Hand and**

**Experimenting with R**

In this laboratory you will work on some probability exercises that do not require R, along with some experiments in R that allow you to verify your computations. The Assignment asks you to record key answers to the four parts of this laboratory.

**Part 1: Favorite Music**

Students enrolled in STAT 200 at Pennsylvania State University in the Spring 2006 term were surveyed about a number of topics, including their favorite music. The results are recorded in the following table.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sex** | **Country** | **Other** | **Pop** | **Rap/Hip Hop** | **Rock** |
| **Female** | 75 | 107 | 107 | 97 | 133 |
| **Male** | 32 | 72 | 14 | 102 | 166 |

As you did in Laboratory #4, you can create this table of counts in R using the command

>FavMusic=matrix(c(75, 107, 107, 97, 133, 32, 72, 14, 102, 166), nrow=2,byrow=T)

and you can ask R to compute the column proportions

>prop.table(FavMusic,2)

and the row proportions

>prop.table(FavMusic,1)

**a)** If you were to randomly sample one individual from this dataset, determine the probability of each of the following events.

Probability that the person likes Rock the best:

Probability the person likes Rock best if they are female:

Are “likes Rock best” and “female” independent?

Probability the person is female:

Probability the person is female if they like Country best:

Are “female” and “like Country best” independent?

**b)** Give an example of two events that are disjoint.

**c)** Are the events you picked in part (b) independent?

**d)** Just to keep in practice, run the hypothesis test

>chisq.test(FavMusic)

Does the resulting p-value give good evidence that musical taste is influenced by the sex of the individual?

You are now ready to work on the first question on the assignment sheet.

**Part 2: Rating Top 40**

In the same study the students were asked to rate how much they like various kinds of music on a scale of 1 (don’t like it at all) to 6 (like it very much). Below are probability distributions for ratings of Top 40 music, with distributions given separately for males and females in the class. The probability for a rating of 6 by males is intentionally blank.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Females** |  |  |  |  |  |  |
| Rating | 1 | 2 | 3 | 4 | 5 | 6 |
| Probability | .04 | .05 | .09 | .24 | .32 | .26 |
| **Males** |  |  |  |  |  |  |
| Rating | 1 | 2 | 3 | 4 | 5 | 6 |
| Probability | .09 | .13 | .17 | .28 | .23 |  |

**a)** What is the value of the missing probability?

**b)** Suppose that we randomly sample one person in the class. Let event A = the rating given by this person is a 6. What event is the **complement** of event A? This is often denoted A**C**.

**c)** Refer to part (b), where event A = rating is a 6. Let event F be the individual is female.

What is P(A)?

What is the value of P(A|F)?

What is P(A**C**)?

What is P(A**C**|F)?

Is “female” independent of “rating is a 6”?

**d)** Consider randomly selecting one female from the class. Are the two events A = rating is a 6 and B = rating is a 5 mutually exclusive? Why?

**e)** For each sex separately, determine the probability that the rating given by a randomly selected person is either a 5 (event A) or a 6 (event B).

P(A or B|F) =

P(A or B|M) =

**f)** Using the values found in part f, calculate the probability that the rating is 4 or less for each sex separately.

For females, probability that rating is 4 or less =

For males, probability that rating is 4 or less =

**g)** Using the values found in parts (e) and (f), explain why we can say that the rating of Top 40 music and sex of the student are dependent (related) characteristics.

You are now ready to work on the second question on the assignment sheet.

**Part 3: Random Digit Dialing**

As we have seen in a number of real life examples, polling organizations use random digit dialing to conduct their surveys. The telephone polling firm Zogby International reports that the probability that a call reaches a live person is 0.2. (*New York Times*, November 9, 1999)

**a)** If Zogby’s makes four calls, what is the probability that none of them reaches a person?

**b)** If Zogby’s makes four calls, what is the probability that at least one of them reaches a person?

**c)** You can confirm your computations by asking R to do a small experiment for you. The following code asks R to choose four random numbers from the set {1, 2, 3, 4, 5}, where numbers can be repeated.

> sample(5,4,replace=TRUE)

Since there is a 20% chance that R will pick the number 1, you can think of a successful call as corresponding to R picking the number 1. Your sample may contain a successful call, multiple successful calls, or no successful calls.

To have R repeat this process ten times, add in the **replicate** function:

> replicate(10,sample(5,4,replace=TRUE))

You should get a table whose columns correspond to a sample of six calls.

How many columns contain the number 1?

Does this proportion match your computation from part (b)?

You are now ready to work on the third question on the assignment sheet.

**Part 4: The Birthday Problem**

A classic probability problem asks: What is the least number of people in a room for there to be a greater than 50% chance that two people share the same birthday? Clearly having 366 people in the room will work, but in fact it doesn’t take anywhere near that many. It only takes 23! If you want we can work through the exact calculation in class, but for the laboratory, we’ll make use of our ability to run simulations in R.

The following code will create a sample of 23 birthdays (numbers 1 to 365) where birthdays are allowed to repeat.

> b=sample(365,23,replace = TRUE)

When I ran this command, and then asked R to display b, I got the following list of birthdays:

[1] 16 48 42 86 343 100 291 310 81 319 134 334 317 170 245 289 351 279 365 164 187 158 19

It doesn’t look like there are any repeats, but just to be sure, I then used

> length(unique(b))

and discovered that indeed there were 23 distinct birthdays. So I got another sample from R:

[1] 317 69 309 19 84 90 8 345 92 59 326 119 350 306 249 350 277 272 107 49 250 145 204

This one has a repeated birthday, which I discovered using the length function as above.

To see just how often a group of 23 individuals will have a shared birthday, we will ask R to run 1000 simulations just like the one above. To do this, we are going to use the **for** statement. This repeats given code for the number of times specified in the “for” statement. The general structure is:

for(*statement that indicates how many times the code should be executed* ) {

*code that needs to be repeatedly executed*

}

We first need a place to store the results of our 1000 experiments, so we will set up “experiments”

>experiments<-numeric(1000)

And the for statement that will give us a 1000 simulations is

>for (i in 1:1000) {

b <- sample(365, 23, replace = TRUE)

experiments[i] <- 23 – length(unique(b))

}

where I got the line feeds using shift+enter (holding down the shift key while pressing enter).

You can see the results of your experiment by simply typing

>experiments

What you will see is 1000 integers. A zero indicates that no two people shared a birthday, as this means “23 – length(unique(b))” is zero, so there were 23 unique birthdays. However a 1 means that there were only 22 unique birthdays, so one birthday was shared.

To count the number of times you see a particular entry among the 1000 experiments, use the table command.

>table(experiments)

It should be the case that you had just a bit fewer than 500 entries that were 0, meaning that the majority of the time there were at least two people with the same birthday.

You are now ready to work on the fourth question on the assignment sheet.

**Assignment! Practicing Probability** Name:

Question #1 Let A be the event that the individual prefers Country, Other, or Rap/Hip Hop. Let F be the event that the individual is female. What is the probability P(A)? What is the conditional probability P(A|F)? Are events A and F independent?

Question #2 If X and Y are two events contained in a sample space, we know

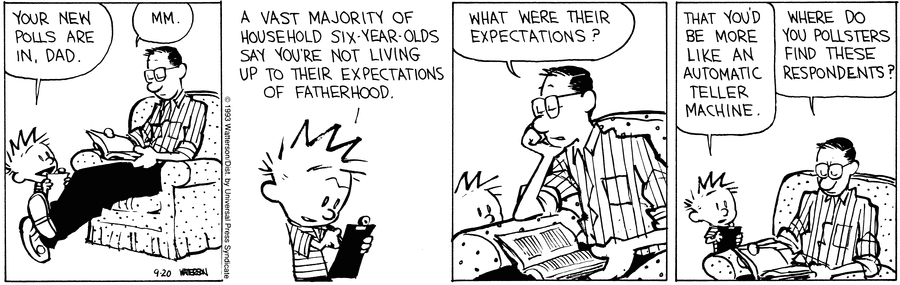
P(X and Y) = P(X)P(Y|X).

Illustrate this formula using the data from Part 2 of this laboratory. Clearly state what your events X and Y are.

Question #3 You know how to compute the probability that none of the four calls is successful and the probability that at least one call is successful. It is a bit more complicated to compute the probability of exactly one successful call, or exactly two successful calls, etc. Do an R experiment, with at least 50 trials, to fill in the following table. You are encouraged to work on this with others at your pod!

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **# Successful Calls:** | 0 | 1 | 2 | 3 | 4 |
| **Probability:** |  |  |  |  |  |

Question #4 Our class has 34 students in it. What is the probability that in a class with 34 students there will be at least two who share a birthday? Be ambitious, and try this with 10,000 experiments, so that your answer will likely be correct to three decimal places.



© Bill Watterson