Lab 2: Basic Probability

Quang Trinh

01.31.18

In this lab we will work through two basic probability problems, and in the process practice more with RMarkdown.

#### Packages needed:

knitr, xtable, pander

#### R Markdown: Automated intro text when creating a new .Rmd file in RStudio

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:

## Task 1: Coin flipping

This task illustrates the interpretation of a probability as the long run relative frequency of an event after a large number of trials.

Dobrow presents R code on pages 24 and 453 for simulating coin tosses. We perform the experiment of observing the number of heads after tossing a fair coin 100 times (probability of a heads on any one toss is 50%). Just like rolling a die, we can use the R function sample to flip a coin. Though recognize there are only two outcomes: heads (1) and tails (0). We will report the number of heads after 50 tosses and intermediary output. We will also graphically display the cumulative proportion of heads again the coin toss (1 to 100). The type="l" parameter in the R function plot will draw a solid line. *Always label your axes! In RMarkdown, a graph title is useful too.*

### Code set-up

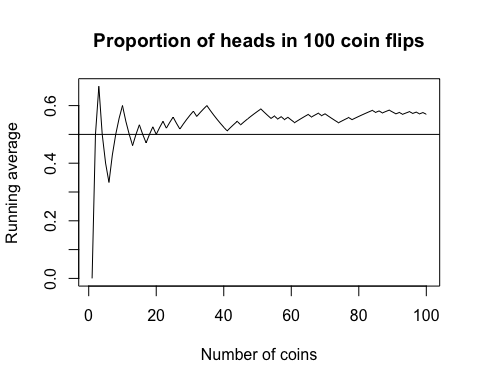
simnum = 100 # number of coin flips  
coinflips = sample(0:1, simnum, replace = TRUE) # flip the coin: heads = 1, tails = 0  
heads = cumsum(coinflips) # cumulative sum of number of heads after each coin toss  
prop = heads/(1:simnum) # running proportion of heads after each coin toss  
head(heads) # report cumulative number of heads after each of the first 6 flips

## [1] 0 1 2 2 2 2

heads[50]

## [1] 29

# running mean plot for proportion of heads.  
plot(1:simnum, prop, type="l", xlab="Number of coins", ylab="Running average", main="Proportion of heads in 100 coin flips")  
abline(h=0.5) # add a line at 50%



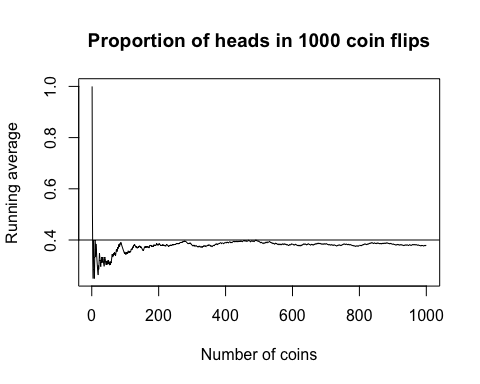
### The problem

Let us now flip a “biased” coin. Perform the experiment of observing the number of heads after tossing a coin 1000 times, with the probability of getting a heads on any one toss being 40%. To change the probability in the R function sample use the parameter prob=c(0.6,0.4); note that we need to specify the probability of a tails (0) and a heads (1) in this parameter. Note that if you do not want the code presented in your html report, use the parameter echo=FALSE in the code chunk.

#### Report the following:

* Proportion of heads after 10, 50, 100, 200, and 500 tosses (see table code chunk below under “RMarkdown presenting output”!)
* Plot of the cumulative proportion of heads vs. coin toss number (1 to 1000); label the axes and title the graphic appropriately!
* On the plot, draw a horizontal line at *y=0.40*, the probability of tossing a head for this coin

# [Place code here]  
simnum <- 1000 # number of coin flips  
coinflips <- sample(0:1, simnum, replace = TRUE, prob = c(0.6, 0.4)) # flip a coin with p(head) = 40%  
heads <- cumsum(coinflips) # number of heads after each toss  
prop <- heads/(1:simnum) # proportion of heads after each coin toss  
  
# Running mean plot for proportion of heads  
plot(1:simnum, prop, type="l", xlab="Number of coins", ylab="Running average", main="Proportion of heads in 1000 coin flips")  
abline(h=0.4) # add a line at 40%



### Questions:

* Describe the behavior of the graphic (cumulative proportion of heads) during the first 150 tosses (1-150), next 150 tosses (151-300), and then later tosses.

*From (1-150), the proportions of heads strongly fluctuates around the 40% line. From (151-300), the proportions of heads fall close to the 40% line. And from (300-1000), the proportions of heads are very much close to the 40% line, indicate that the probability to get a head is very much 40%.*

* What do you notice about the limiting value of the curve in your plot?

*The limiting value of the curve fall between 0 and 1.*

* Why would you expect the behavior you discuss in the previous two bullets?

*Because the coin is a biased coin with the probability of getting heads is 40%. Thus, we expect to see the overal proportion of heads is close to 40% (n=1000 is large enough). Also, since we are talking about probability, then 0<= p(heads) <=1.*

#### RMarkdown presenting output:

Below is code to present a table for the proportion of heads after 10, 50, and 100 tosses.

Reminders:

The echo=FALSE parameter prevents printing of code from a code chunk. The include=FALSE parameter prevents printing of output from a code chunk. The results=asis allows the LaTeX code produced by xtable to be compiled and output.

# we will create a table using xtable and pander  
library(knitr)  
library(xtable)  
library(pander)  
# output desired summary statistics  
# formatC used so integer coin tosses do not have a decimal place in the figure!  
numtoss = formatC(c(10, 50, 100, 200, 500), digits=0, format="d", flag="#")  
num.heads = c(heads[10], heads[50], heads[100], heads[200], heads[500])  
num.heads = formatC(num.heads, digits=0, format="d", flag="#")  
# formatC used here so proportions have exactly two decimal places (including zeros at the end)!  
prop.heads = c(prop[10], prop[50], prop[100], prop[200], prop[500])  
prop.heads = formatC(signif(prop.heads,digits=6), digits=2, format="f", flag="#")  
table.elts = rbind(numtoss, num.heads, prop.heads)  
row.names(table.elts) = cbind("# coin tosses", "Number of heads", "Proportion of heads")  
lab1.table = xtable(table.elts, caption = "Proportion of heads for a given number of tosses of a biased coin.", label="cointoss", align = "|l|rrrrr|")  
pander(lab1.table)

Proportion of heads for a given number of tosses of a biased coin.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 |
| **# coin tosses** | 10 | 50 | 100 | 200 | 500 |
| **Number of heads** | 4 | 16 | 35 | 77 | 197 |
| **Proportion of heads** | 0.40 | 0.32 | 0.35 | 0.39 | 0.39 |

The table shows the proportion of heads after a certain number of tosses in a single simulation experiment. These proportions provide empirical estimates of the probability of a head for specified simulation sample sizes.

### The problem

Directly in the code chunk above, add columns for 200 tosses and 500 tosses. Hint: you will need to append these two elements to numtoss, num.heads, and prop.heads. Also, note that in the xtable function, the align is only for 3 right-justified columns; need to augment that to 5 right-justified columns.

## Task 2: Divisibility probability

This task provides a probability problem to explore if-then statements and functions in R. Consider an interger drawn uniformly at random from the numbers {1, 2, …, 1000} such that each number is equally likely. We wish to simulate the probability that the number drawn is divisible by 3, 5, or 6.

### Code set-up

Dobrow presents R code on page 25 for simulating this experiment, and provides the exact probability calculation in Example 1.20. The code presents a slick application of the replicate R function, one we used in the R Introduction lab. In particular, a function is written which draws the number at random from the integers 1 to 1000 and then checks if it is divisible by 3, 5, or 6. It uses modular arithmetic, x%%n being *x mod n* in R. For example, if the remainder of the number divided by 3 (modulus) is 0, then the number is divisible by 3! We will then repeat the function (experiment) 1000 times to get the empirical probability.

##### The true probability that a randomly drawn integer between 1 and 1000 is divisible by 3, 5, or 6 is 0.467.

# simdivis() simulates one trial  
simdivis = function(){  
 num = sample(1:1000, 1) # draw a number at random from the integers 1 to 1000  
 # determine if the number is divisible by 3, 5 or 6 by checking if the remainder is 0  
 if (num%%3==0 || num%%5==0 || num%%6==0) 1 else 0  
}  
simlist = replicate(1000, simdivis()) # replicate the experiment 1000 times  
mean(simlist) # compute the estimated probability as the proportion of times the number is divisible by 3, 5, or 6

## [1] 0.463

### The problem

Simulate the probablity that a random integer between 1 and 5000 is divisible by 4, 7, or 10.

##### The true probability that a randomly drawn integer between 1 and 1000 is divisible by 3, 5, or 6 is 0.40.

# [Place code here]  
# simdivis() simulates one trial  
simdivis = function(){  
 num = sample(1:5000, 1) # draw a number at radom from the integers 1 to 5000  
 # determine if the number is divisible by 4, 7, or 10 by checking if the remainder is 0  
 if (num%%4==0 || num%%7==0 || num%%10==0) 1 else 0  
}  
simlist1 = replicate(100, simdivis()) # replicate the experiment 100 times  
simlist2 = replicate(1000, simdivis()) # replicate the experiment 1000 times  
simlist3 = replicate(10000, simdivis()) # replicate the experiment 10000 times  
simlist4 = replicate(100000, simdivis()) # replicate the experiment 100000 times  
mean(simlist1)

## [1] 0.42

mean(simlist2)

## [1] 0.393

mean(simlist3)

## [1] 0.399

mean(simlist4)

## [1] 0.3979

# we will create a table using xtable and pander  
library(knitr)  
library(xtable)  
library(pander)  
numExperiments = formatC(c(100,1000,10000,100000), digits=0, format="d", flag="#")  
prob = c(mean(simlist1), mean(simlist2), mean(simlist3), mean(simlist4))  
prob = formatC(signif(prob,digits=6), digits=2, format="f", flag="#")  
table.elts = rbind(numExperiments, prob)  
row.names(table.elts) = cbind("Number of reps", "Probability")  
lab1.table = xtable(table.elts, caption = "Probability of number which is divisible by 4,7 or 10.", label=" ", align = "|l|rrrr|")  
pander(lab1.table)

Probability of number which is divisible by 4,7 or 10.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 |
| **Number of reps** | 100 | 1000 | 10000 | 100000 |
| **Probability** | 0.42 | 0.39 | 0.40 | 0.40 |

### Questions:

* Present the empirical probability based on repeating the experiment 100, 1000, 10000, and 100000 times. Consider using the xtable code chunk from the first task to build a table for these values.

*The more experiment we perform, the closer that the probability to 40%.*

* How do these values compare to the truth?

*These values are very much close to the truth, which is 40%.* *The values converge to the truth of 40% as the number of experiments grows. The empirical probability is accurate to one decimal place even by 100*.

* Extra credit: show that the true probability that a random integer between 1 and 5000 is divisible by 4, 7, or 10 is 40%?

*Let D4, D7, D10 denote the events that the number drawn is divisible by 4,7, and 10 respectively. By inclusion-exclusion, P(D4 U D7 U D10) = P(D4) + P(D7) + P(D10) - P(D4D7) - P(D4D10) - P(D7D10) + P(D4D7D10).*

*P(D4) = floor(5000/4)/5000 = 0.25, P(D7) = floor(5000/7)/5000 = 0.14, P(D10) = floor(5000/10)/5000 = 0.1.*

*P(D4D7) = floor(5000/28)/5000 = 0.04.*

*P(D4D10) = floor(5000/40)/5000 = 0.025.*

*P(D7D10) = floor(5000/70)/5000 = 0.01.*

*P(D4D7D10) = floor(5000/280)/5000 = 0.0034.*

*Thus, P(D4 U D7 U D10) = 0.25 + 0.14 + 0.1 - 0.04 - 0.025 - 0.01 + 0.0034 = 0.4184.*