Mat 354

Homework 5

Kenny Roffo

Due October 2, 2015

The distribution function is defined $F(y) = P(Y \leq y)$

1. For a random variable Y that is integer valued, $F(12) = 0.2035$ and $F(11) = 0.0658$. Determine $p(y) = P(Y = 12)$.

$$\begin{aligned} P(Y = y) &= F(y) - F(y-1) \\ P(Y = 12) &= F(12) - F(11) \\ &= 0.2035 - 0.0658 \\ &= 0.1377 \end{aligned}$$

2. For the case of a simple random sample of 4 integers from $1, 2, ..., 10$ (one of each), and taking Y to be the maximum:

   a) Tabulate the distribution function $F(y)$ for values $1, 2, ..., 10$. (If the maximum is no greater than y, then all four values must be no greater than y. How many ways can you choose four values no greater than y? In how many ways can 4 values be chosen? Divide.)

   For a given value n, the probability no integer will be greater than n is given by $\frac{n^4}{10^4}$. This $F(Y)$ is given by:

Y	$F(Y)$
1	0.0001
2	0.0016
3	0.0081
4	0.0256
5	0.0625
6	0.1296
7	0.2401
8	0.4096
9	0.6561
10	1.00000

b) Give values for $F(-3), F(0)F(7.5)$ and $F(11.7)$.

$$\begin{aligned} F(-3) &= 0 \\ F(0) &= 0 \\ F(7.5) &= 0.2401 \\ F(11.7) &= 1 \end{aligned}$$

c) Tabulate the probability function $p(y) = P(Y = y)$ for values $1, 2, ..., 10$. (Use F to help with this. This is the third assignment drilling on this point.)

We do this by using $P(Y = y) = F(y) - F(y - 1)$:

y	$0(y)$
1	0.0001
2	0.0015
3	0.0075
4	0.0175
5	0.0369
6	0.0671
7	0.1105
8	0.1695
9	0.2465
10	0.3439

d) Determine the expected value, variance and standard deviation of Y:

Expected Value: For the expected value we use $E(Y) = \sum_{ally} yp(y)$

$$\begin{aligned} E(Y) &= 1(0.0001) + 2(0.0015) + 3(0.0075) + ... + 10(0.3439) \\ &= 8.4667 \end{aligned}$$

Variance: The Variance is calculated using $E(Y^2) - E(Y)^2$

$$\begin{aligned} \sigma^2 &= 1^2(0) + 2^2(0) + 3^2(0) + 4^2(0.00476) + ... + 9^2(0.26667) + 10^2(0.40000) - (8.4667)^2 \\ &= 2.6167 \end{aligned}$$

Standard Deviation: The standard deviation is simply the square root of the variance:

$$\begin{aligned} \sigma &= \sqrt{2.6167} \\ &= 1.6176 \end{aligned}$$

3. For the case of a random sample (i.e. with replacement) of 4 integers from $1, 2, ..., 10$ (one of each), and taking Y to be the maximum, we found, in class, that $E[Y] = 8.4667$. Suppose we instead sample from $1, 2, ..., k$.:

a) Try $k = 5$ and obtain $E[Y]$ numerically. Is $E[Y] = 8.4667/2 = 4.23335$?

Using the same methods as for the previous problem, we obtain a value for $E(Y) = 4.4336$ (I am doing these calculations in R in very few lines, so I am not actually seeing the intermediate values, so I don't have them to put in here.). Thus, we cannot scale E(Y) as the amount of numbers change.

b) (You can quickly glean that a simple general expression will not apply by considering $k = 1$.) Challenge? Give a general expression for $E[Y]$ in terms of k:

$$\sum_{i=1}^{k} \left[i \left(\frac{i^4}{k^4} - \frac{(i-1)^4}{k^4} \right) \right]$$

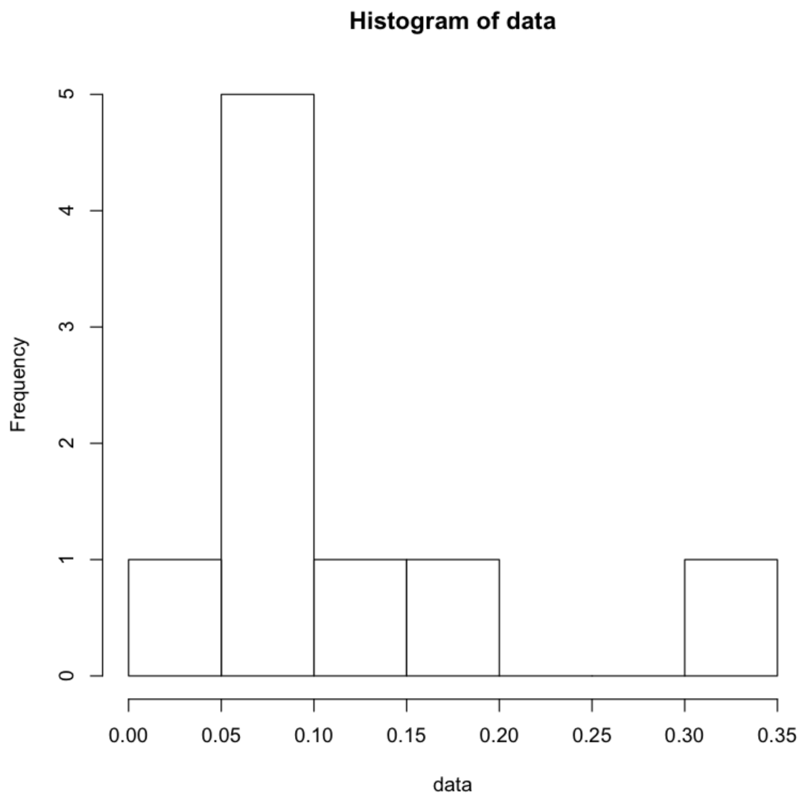
This was actually pretty easy to come up with considering I had been using it in about three separate steps in R.

4. Suppose $p(y)$ is defined for $y = 1, 2, ..., 9$ (the non-0 digits) as follows:

$$p(y) = \log_{10}(y + 1) - \log_{10}(y)$$

a) Construct a table of values y and probabilities $p(y)$. Draw a probability histogram.

y	$p(y)$
1	0.30103
2	0.1761
3	0.1249
4	0.0969
5	0.0792
6	0.0669
7	0.0580
8	0.0512
9	0.0458



b) For a non-zero digit y, give the (simplest) formula for $F(y) = P(Y \leq y)$.

$$P(Y \leq y) = \sum_{i=1}^{y} p(i)$$

c) Demonstrate that probabilities given by $p(y)$ sum to exactly 1. That is, show $\sum_{y=1}^{9} p(y) = 1$

Notice that just using the formula we see that many of the logs cancel out, and we are left with $\log(10) - \log(1)$, where $\log(1) = 0$ and $\log(10) = 1$:

$$\begin{aligned}\sum &= \big(\log(2) - \log(1)\big) + \big(\log(3) - \log(2)\big) + ... + \big(\log(10) - \log(9)\big) \\ &= -\log(1) + \big(\log(2) - \log(2)\big) + ... + \big(\log(9) - \log(9)\big) + \log(10) \\ &= 0 + 0 + ... + 0 + 1 \\ &= 1\end{aligned}$$

d) Determine $E(Y)$, the expected (or mean) value of Y. First obtain a decimal approximation. Then see if you can obtain an exact value (a decimal approximation no matter how precise is suboptimal).

Applying $E(Y) = yp(y)$ using the values from part a, we get $E(Y) = 3.440237$.

Using algebra, we can attempt to find an exact value:

$$\begin{aligned}E(Y) = &\big(\log(2) - \log(1)\big) + 2\big(\log(3) - \log(2)\big) + 3\big(\log(4) - \log(3)\big) + 4\big(\log(5) - \log(4)\big) \\ &+ 5\big(\log(6) - \log(5)\big) + 6\big(\log(7) - \log(6)\big) + 7\big(\log(8) - \log(7)\big) + 8\big(\log(9) - \log(8)\big) \\ &+ 9\big(\log(10) - \log(9)\big) \\ &= -\log(2) - \log(3) - \log(4) - \log(5) - \log(6) - \log(7) - \log(8) - \log(9) + 9\log(10)\end{aligned}$$

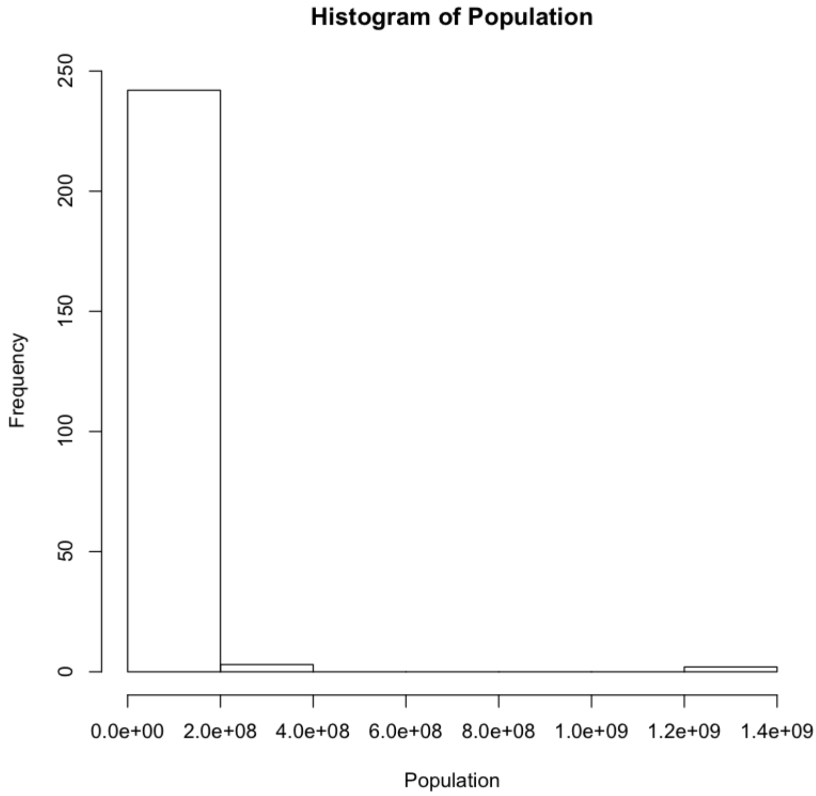
5. Wikipedia lists populations of all the countries in the world.

   You will, for each country, determine the lead digit of the population. For China that digit is 1; for the U.S. that digit is 3. etc.

   Download the data set Pops.csv. Read the file into a data frame in R. (Here Ive used the identifier data for that data frame.) Attach the data frame. You can determine lead digits using the code thats supplied. (Its easier in Excel, where you can use the =LEFT function.)

```
data <- read.csv([path\_to\_file], header=TRUE)
attach(data)
s <- Population/10^floor(log10(Population))
ld <- floor(s)
```

a) Obtain a histogram of the Populations. What does this tell you about the populations of countries of the world?



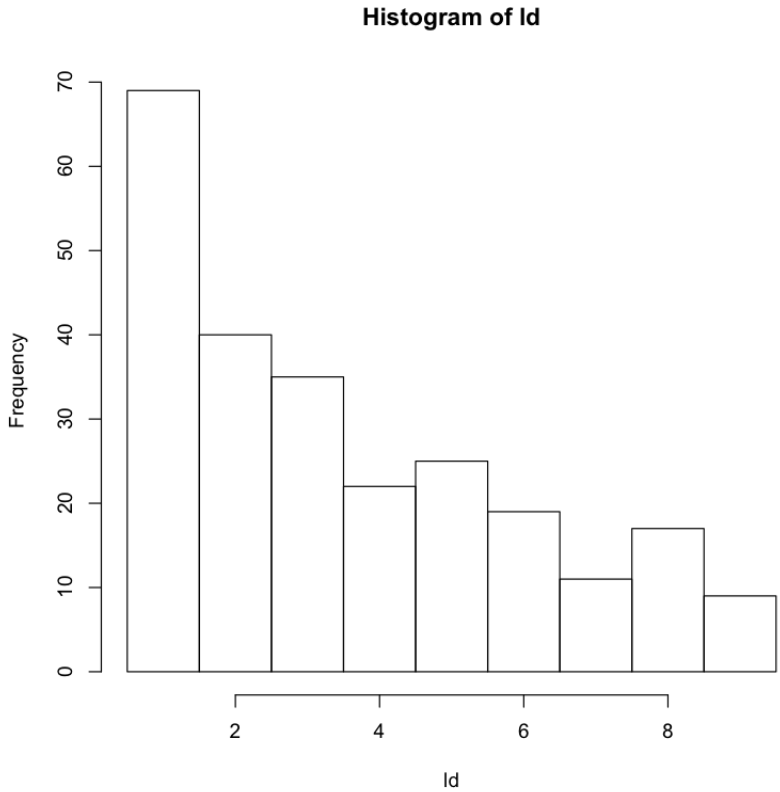
b) Whats the population of the world? (You may not be able to use sum directly on the Population variable. Ill explain later in class.) To find out, first determine the mean population of all countries (use mean); then how many (you can use length(Population) to help).

7145097761 People

c) Obtain a histogram of the lead digits.
Left out due to the requirements of part d.

d) Notice that left to its own devices, R pools all the 1s and 2s. You dont want to do this. So improve your histogram as follows. Include this histogram only with your assignment.

```
hist(ld, breaks=seq(.5,9.5,1))
```



e) Determine the mean of the lead digits.
   3.54251

6. - How does this work?

```
s <- Population/10^floor(log10(Population))
ld <- floor(s)
```

  The right side of the first statement first finds the order (power of 10) each population is, then gets rid of any decimal parts. Then, 10 is raised to this number to give nicely rounded off populations, and the population for the given country is divided by the result to get a number that only has one digit before the decimal. Basically, it just puts a decimal after the first digit of each population. The second statment then floors the results to give only the first digits of all the populations.

- What do questions 5 and 6 have to do with each other?
  5 is to show how to do something, and 6 is to make sure I understand what's going on.