STU22004 Applied Probability I

Lab 1

Instructions

- Download the template file *STU22004_ lab1.xlsx* from Blackboard.
- Address the points below by adapting what was done in the template (where the same points are addressed in the simpler "2 from 4 passwords" case).
- Work individually or in small groups.
- If you do not complete the instructions today, you can work on it during the week (if any doubt arises, ask next Monday).
- The last page of this file contains a short summary of the functions used in the template.

Lab 1: generating random passwords

1. **Generate 10 random passwords, each one composed by six letters from the English alphabet. Repetitions are allowed, e.g. QWERTQ is an admissible password.**

   a. Store $A, \ldots, Z$ in a set of cells, naming the set *Symbols* (Use *Define Name* under *Formulas Tab*)

   b. Use RANDBETWEEN to create six index values in $1, 2, \ldots, 26$

   c. Use INDEX(Symbols,$*$) to choose the member of the cells holding $A, \ldots, Z$ in position $*$, where each $*$ is a cell holding an index value generated at step b.

   d. Use the function CONCATENATE to concatenate these 6 symbols into one word

   e. Make 10 copies (recall you can type a formula into one or more cells, and then copy formula to the other cells, but note Excel's conventions on relative, e.g. A5, and absolute, e.g. A5, addressing)

   Could you equivalently use the following commands instead of RANDBETWEEN used in step b?

   b′. 1+INT(26*RAND())

   b″. ROUNDUP(26*RAND(),0)

2. **Generate 10 random passwords, each one composed by six letters from the English alphabet. Repetitions are not allowed, e.g. QWERTQ is not an admissible password, QWERTY is an admissible password. Try with both methods.**

   First method: rejection. Generate passwords allowing for repetitions and then reject passwords that show repetitions.

   a. Generate ten passwords as in point 1

b. For each letter, count the number the letter occurs. You can use the function COUNTIF.

c. Reject the password if any letter is repeated. You can use the function IF to return the accepted password if no letter is repeated and to return nothing if there are repetitions.

Second method: sampling without replacement. Develop a sampling strategy that does not allow for repetitions.

a. Use RAND to generate 26 random numbers in (0,1)

b. Use the function RANK to find the rank of the first 6 generated numbers

c. Use INDEX(Symbols,$*$) to choose the member of the cells holding $A, \ldots, Z$ in position $*$, where each $*$ is a cell holding an index value generated at step b.

3. After generating passwords as indicated in points 1 and 2, compute the following quantities.

a. The proportion of passwords containing the symbol "A" (do not include the rejected one in your analysis)

b. The proportion of passwords with repetitions (for passwords generated as in point 1)

c. The proportion of accepted passwords (for passwords generated as in point 2, first method)

4. Which of the two methods considered at point 2 is more efficient in the long run, in terms of the average number of calls to =RAND() to generate a password?

We can study questions like this by replicating the password generation n times and computing the average number of calls to =RAND(). In the template $n = 100$.

See the work sheet "100 Reps" for syntax. These steps refer to the simple "2 from 4 passwords" that you can find in the template. Adapt the same steps for the "6 from 26 passwords".

a. Extend the calculations made at point 2 to 100 rows.

b. Use an =IF() statement to return 1 if there is a duplicate, and 0 otherwise.

c. Compute the running sum and the running average.

d. Make a line plot of the running sum and the running average. Do this by:

   i. selecting the column of cells in, e.g., the sum column.

   ii. clicking on Excel's (top level) INSERT Tab.

   iii. selecting Line from the plotting options.

   iv. Repeat with the running average.

e. compute the average number of calls to =RAND() per one valid generated password for the two methods considered at point 2.

The running average can be thought of as a running proportion and it suggests convergence to some value. In the "2 from 4 passwords" case, the long run average is in fact $1/4$. Why?

List of used functions:

- RAND() returns an evenly distributed random real number in $(0, 1)$
- RANDBETWEEN(a,b) returns a random integer number between a and b
- CONCATENATE("A","B") joins the text strings A and B into one string AB
- INT rounds a number down to the nearest integer, e.g. $\text{INT}(8.9) = 8$
- ROUND rounds a number to a specified number of digits, e.g. $\text{ROUND}(8.793, 1) = 8.7$
- ROUNDUP rounds a number up, away from 0, e.g. $\text{ROUND}(8.793, 1) = 8.8$
- COUNTIF counts the number of cells that meet a criterion
- IF makes logical comparison between two values. E.g. $\text{IF}(\text{C2} = 1, \text{"Yes"}, \text{"No"})$ returns "yes" if the value at cell C2 is equal to 1, "no" otherwise
- MAX returns the largest value in a set of values
- RANK returns the rank of a number in a list of numbers
- AVERAGE returns the arithmetic mean of the arguments
- INDEX returns a value from within a table or range