Статистические испытания в R Studio

Задача

Для того чтобы продемонстрировать работу данного метода смоделируем задачу.

Вам необходимо успеть на самолёт. Для этого нужно ехать сначала на автобусе, затем на метро и далее на экспрессе. Дабы уменьшить число переменных, не потеряв при этом наглядность примера, не станем учитывать время затраченное на то, чтобы сменить вид транспорта.

Предположим, что интервал следования автобуса равен 10 минутам, а время в пути составляет в среднем 20 минут. Среднее время, проведённое в метро – 10 минут. Экспресс отправляется каждые 30 минут, время в пусти постоянно и составляет 45 минут. Положим, что есть основания полагать (например, в результате анализа проведённых ранее экспериментов), что время потраченное на перемещение автобусом и на метро распределено по нормальному закону со стандартным отклонением 5 и 2 минуты соответственно.

Стоит задача описать распределение времени затраченного на дорогу и вычислить вероятность, с которой Вы успеваете на самолёт, если заложили полтора часа на дорогу.

Решение

Схема передвижения и ожидания представляется следующей:

ожидание автобуса → время в автобусе → время в метро → ожидание экспресса → время в поезде → аэропорт.

Исследуемая величина - время в пути T будет функцией пяти переменных, четыре из которых — случайные величины:

T_a = U(0,10)[1] — время ожидания автобуса;

T_{bus} = N(20,5)[2] — время в автобусе;

T_{sub} = N(10,2) — время в метро;

T_{a2} = U(0,30) — ожидание экспресса;

T_{ex} = 45 мин. — время в экспрессе.

Общее время в пути

$$T = T_a + T_{bus} + T_{sub} + T_{a2} + T_{ex} \quad .$$

Далее проведём статистические испытания, сгенерировав (разыграв) по N = 10000 значений случайных величин T_a, T_{bus} , T_{sub} , T_a в соответствии с их законами распределения, чтобы в конце концов собрать статистику общего времени T.

```
N = 100000
Ta = runif(N, 0, 10)
Tbus = rnorm(N, 20, 5)
Tsub = rnorm(N, 10, 2)
Ta2 = runif(N, 0, 30)
```

Теперь значения случайных величин можно представить в виде таблицы из N строк.

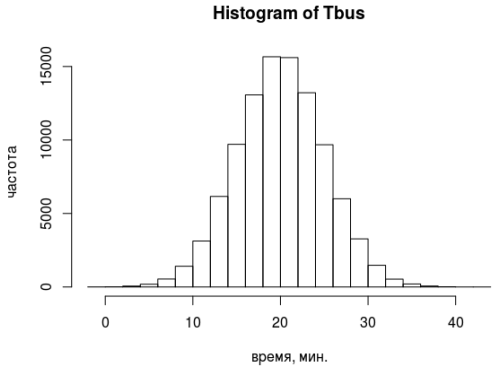
```
      Ta      Tbus     Tsub       Ta2
1 4.396670 27.621645 9.360712  6.554547
2 2.053775 19.681378 7.528912 13.929733
3 8.338687 18.256175 9.705486  5.103797
4 5.670176 16.176994 9.844666 19.214915
5 ...
```

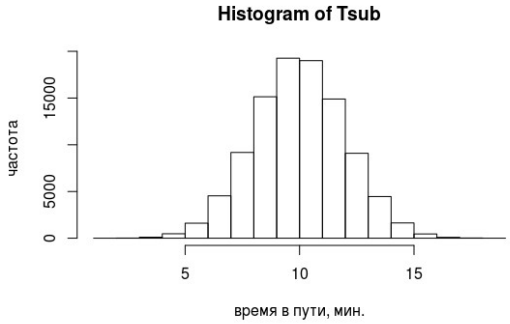
1 U(0,10) означает случайную величину распределённую равномерно на интервале от 0 до 10.

2 N(20,5) означает случайную величину распределенную по нормальному закону с м.о. 20 и стандартным отклонением 5.

Построим гистограммы для всех случайных величин, чтобы наглядно представить распределения

```
hist(Ta, xlab = 'время ожидания, мин.', ylab='частота')
hist(Tbus, xlab = 'время, мин.', ylab='частота')
hist(Tsub, xlab = 'время в пути, мин.', ylab='частота')
 hist(Ta2, xlab = 'время ожидания, мин.', ylab='частота')
```





Вычислим время затраченное на весь путь. В результату в переменной T будет содержатся N значений времени.

```
T = Ta + Tbus + Tsub + Ta2 + 45
```

Отделим те их них, которые удовлетворяют нашему условию T < 90 мин

```
Tm = T[T < 90]
```

и сосчитаем их

```
m = length(Tm)
```

Вычислим и напечатаем вероятность потратить на дорогу в аэропорт менее полутора часов

```
P = m/N
print(P)
```

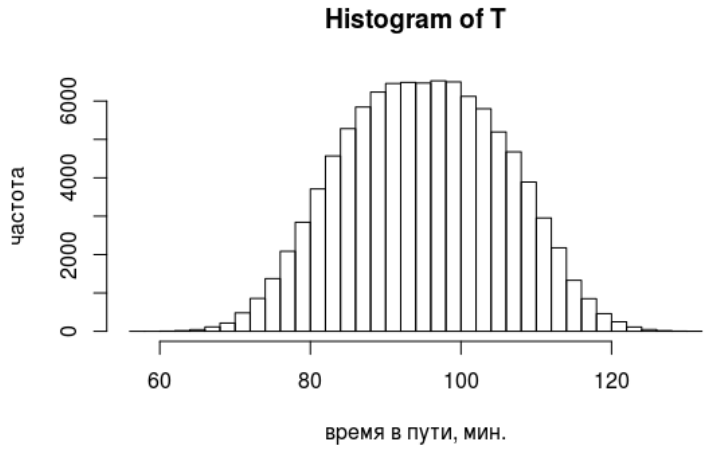
В итоге имеем ответ:

```
[1] 0.33677
```

Необходимо убедится в устойчивости результата. Для этого проведём эксперимент несколько раз. В результате получим следующие вероятности: `0.33522, 0.339, .33507, 0.33774`. Как видно первые два знака после запятой не изменяются, сочтём этот результат удовлетворительным.

Наконец построим гистограмму распределения времени T

```
hist(T, xlab = 'время в пути, мин.', ylab='частота', breaks = 40)
```



В итоге вероятность добраться до аэропорта за полтора часа примерно равна 0.34.

Однако, не станем останавливаться на достигнутом ответе, а поставим своей целью использовать полученный данные для предсказания вероятности уложится в заранее заданное время пользуясь исключительно собранными данными. Получим эмпирическую функцию распределения используя полученную статистику

```
F <- ecdf(T)
```

Теперь можно делать предсказания для разных значений времени, например

```
> F(120)
[1] 0.996
```

Выходит, что отправившись в аэропорт за 120 минут до предполагаемого времени прибытия, с вероятностью 0.996 можно добраться без опозданий, т.е. потратить 120 минут или меньше. Остальные (>120 минут) временные затраты менее вероятны, на них приходится вероятность равная 0.004.

Решим обратную задачу: определить время которое необходимо потратить на дорогу задав для себя допустимую вероятность прибыть вовремя

```
quantile(T, 0.99)
```

В результате выполнения функции будет напечатано следующее:

```
     99%
117.7139
```

Таким образом, если вероятность опоздать в 1 случае из ста считается удовлетворительной, то следует закладывать на дорогу около 118 минут.

В заключении следует отметить, что метод статистического моделирования даёт только приближенный результат.

Ссылки

1. R Studio, сайт разработчика https://www.rstudio.com/products/rstudio/download/
2. R для статистической обработки данных (Учебно-методическое пособие)

http://kpfu.ru/docs/F407025247/metodichka_R_2.pdf

3. Мастицкий С.Э., Шитиков В.К. (2014) Статистический анализ и визуализация данных с помощью R. – Электронная книга http://www.ievbras.ru/ecostat/Kiril/R/Mastitsky%20and%20Shitikov%202014.pdf

4. Блог посвящённый языку программирования R http://r-analytics.blogspot.ru/p/rstudio.html