Effective Business Simulation: rank and salary

(Henry T.H. Tu, 14-nov-2019)

2.1. The agenda

We will generate the employee table with 10000 rows and 4 columns (code, rank, age, salary). The column code is generated directly by the given hashing formula. The column [rank] is generated by transforming [U1]. The column [age] is generated by transforming both [U3] and [U4] with the Box-Muller formula. And the column [salary] is generated by transforming [U4] and [U5] with the triangular distribution. The summary of transformations is given in table G1

code	"c" + left(regex_replace(MD5_ASCII(tostring([RowCount]) + "/code") , "\D+", ""), 7)		
U1	tonumber(left(regex_replace(MD5_ASCII(tostring([RowCount]) + "/rank") , "\D+", ""), 7))/POW(10, 7)	rank	if [U1]<0.7 then "worker" elseif [U1]<0.95 then "manager" else "director" endif
U2	tonumber(left(regex_replace(MD5_ASCII(tostring([RowCount]) + "/age") , "\D+", ""), 7))/POW(10, 7)	age	floor(if [rank] = "worker" then 25 + 2.63*sqrt(-2*log([U2])) * cos(2*pi()*[U3]) elseif [rank] = "manager" then 35 + 2.47*sqrt(-2*log([U2])) * cos(2*pi()*[U3]) elseif [rank] = "director" then 45 + 1.49*sqrt(-2*log([U2])) * cos(2*pi()*[U3]) else null() endif + 0.5)
U3	tonumber(left(regex_replace(MD5_ASCII(tostring([RowCount]) + "/age2") , "\D+", ""), 7))/POW(10, 7)		
U4	tonumber(left(regex_replace(MD5_ASCII(tostring([RowCount]) + "/salary") , "\D+", ""), 7))/POW(10, 7)	salary	floor(if [rank] = "worker" then 2500 + 800*([U4] - [U5]) elseif [rank] = "manager" then 3500 + 600*([U4] - [U5]) elseif [rank] = "director" then 5000 + 300*([U4] - [U5]) else null() endif)
U5	tonumber(left(regex_replace(MD5_ASCII(tostring([RowCount]) + "/salary2") , "\D+", ""), 7))/POW(10, 7)		

G1: the transformations to generate columns

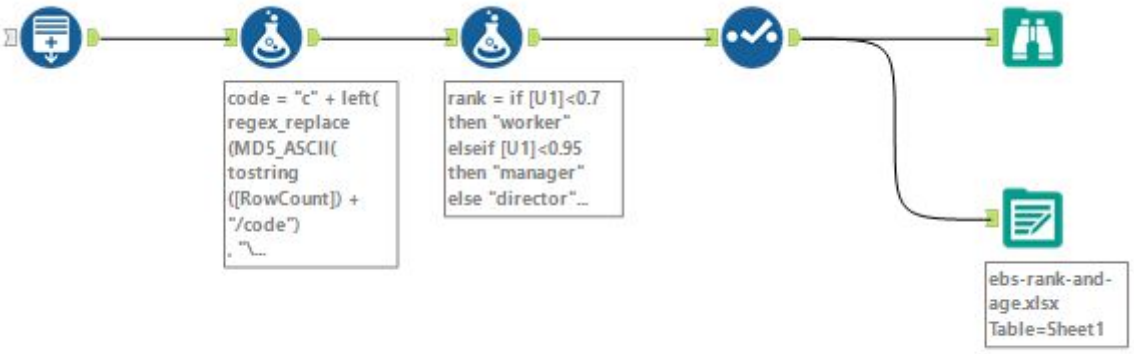
2.2. The workflow and results

Table G2 shows the generated results in which the column [RowCount] is generated first. Then the random columns [code] and [U1] to [U5] are generated by formulas. Then we generate three columns [rank] and [age] and [salary] are generated with three formulas in table G1.

RowCount	code	U1	U2	U3	U4	U5	rank	age	salary
1	c5001234	0.685276	0.368469	0.852406	0.58176	0.378855	worker	27	2662
2	c1063141	0.097084	0.270186	0.407089	0.873405	0.400245	worker	21	2878
3	c8805521	0.710985	0.694836	0.102336	0.521126	0.819518	manager	37	3320
4	c3287523	0.295254	0.428487	0.763373	0.823276	0.707939	worker	25	2592
5	c9383355	0.540448	0.926825	0.627806	0.329021	0.001058	worker	24	2762
6	c6058807	0.274867	0.467534	0.805387	0.587009	0.26179	worker	26	2760
7	c6848714	0.643517	0.028897	0.4298	0.539978	0.215023	worker	19	2759
8	c6335186	0.932327	0.622773	0.290514	0.394588	0.954273	manager	34	3164
9	c3293141	0.261679	0.51246	0.24886	0.265153	0.740562	worker	25	2119
10	c2165420	0.444088	0.411694	0.394305	0.860092	0.595614	worker	22	2711
11	c2071901	0.915368	0.883573	0.130254	0.667568	0.469086	manager	36	3619
12	c0365824	0.608124	0.526794	0.139025	0.559004	0.056637	worker	27	2901
13	c8807146	0.746001	0.234332	0.633402	0.140469	0.500989	manager	32	3283
14	c3822024	0.446991	0.473975	0.086493	0.845214	0.451495	worker	28	2814
15	c4941203	0.066749	0.28057	0.799512	0.809342	0.555657	worker	26	2702
16	c0788077	0.747921	0.074732	0.045415	0.797324	0.962919	manager	40	3400
17	c1380179	0.797315	0.261433	0.281225	0.969089	0.903275	manager	34	3539

G2: the generative results

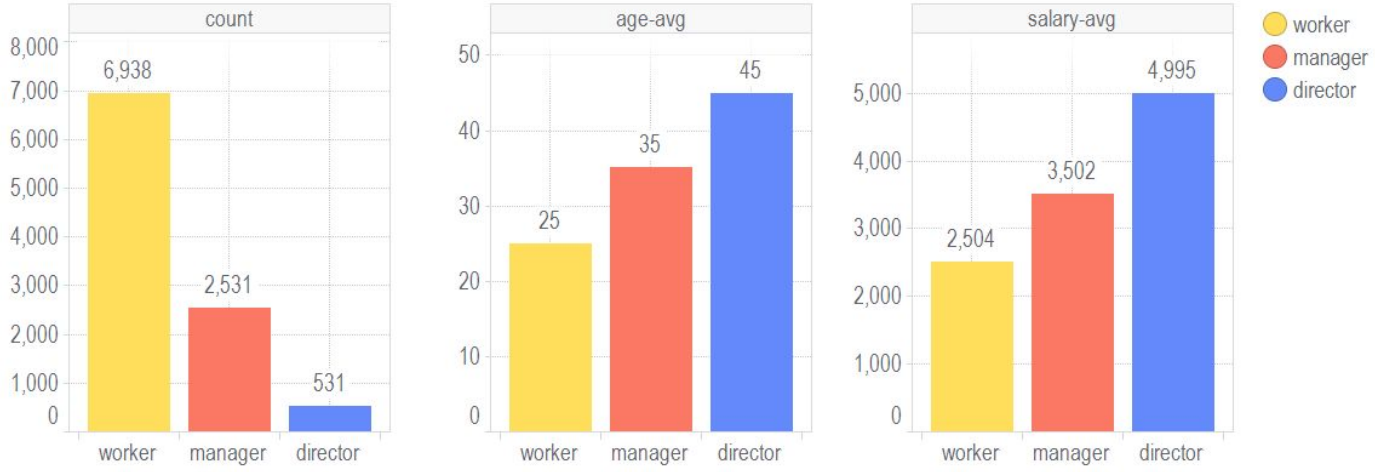
Figure G3 shows the alteryx workflow to generate the data. We start with the Generate Rows tool, then Formula tool to add random variables, then another Formula tool to add categorical and numerical variables. Finally, the select tool is to remove unwanted variables and keep only code, rank, age, and salary.



G3: the generative workflow

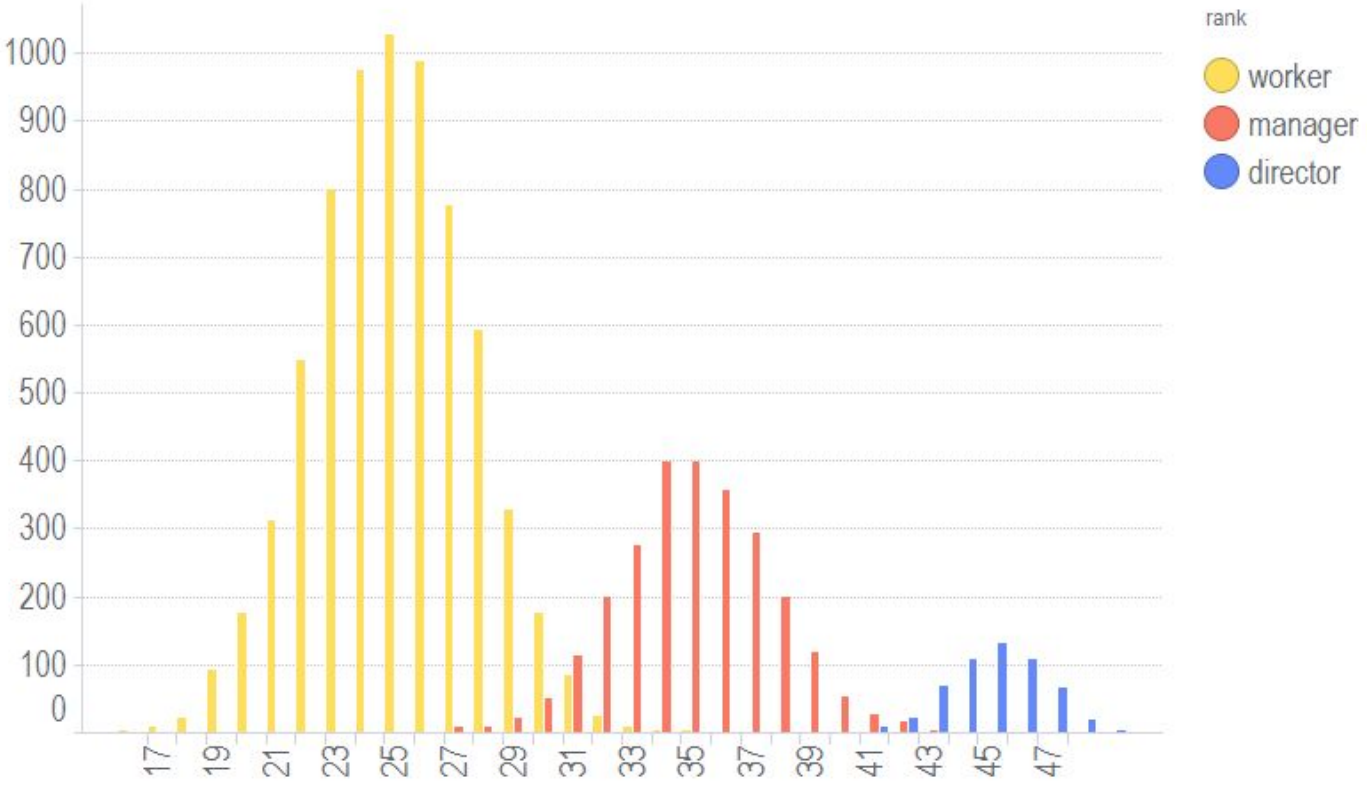
2.3. The statistics

Figure G4 summarizes the data table. We have 6938 workers (70% of 10000) and 2531 managers (25% of 10000) and 531 directors (5% of 10000). The average age of workers is 25 and that of manager is 35 and that of director is 45. The average salary of workers is 2504 usd, that of managers is 3502 usd, and that of directors is 4995 usd



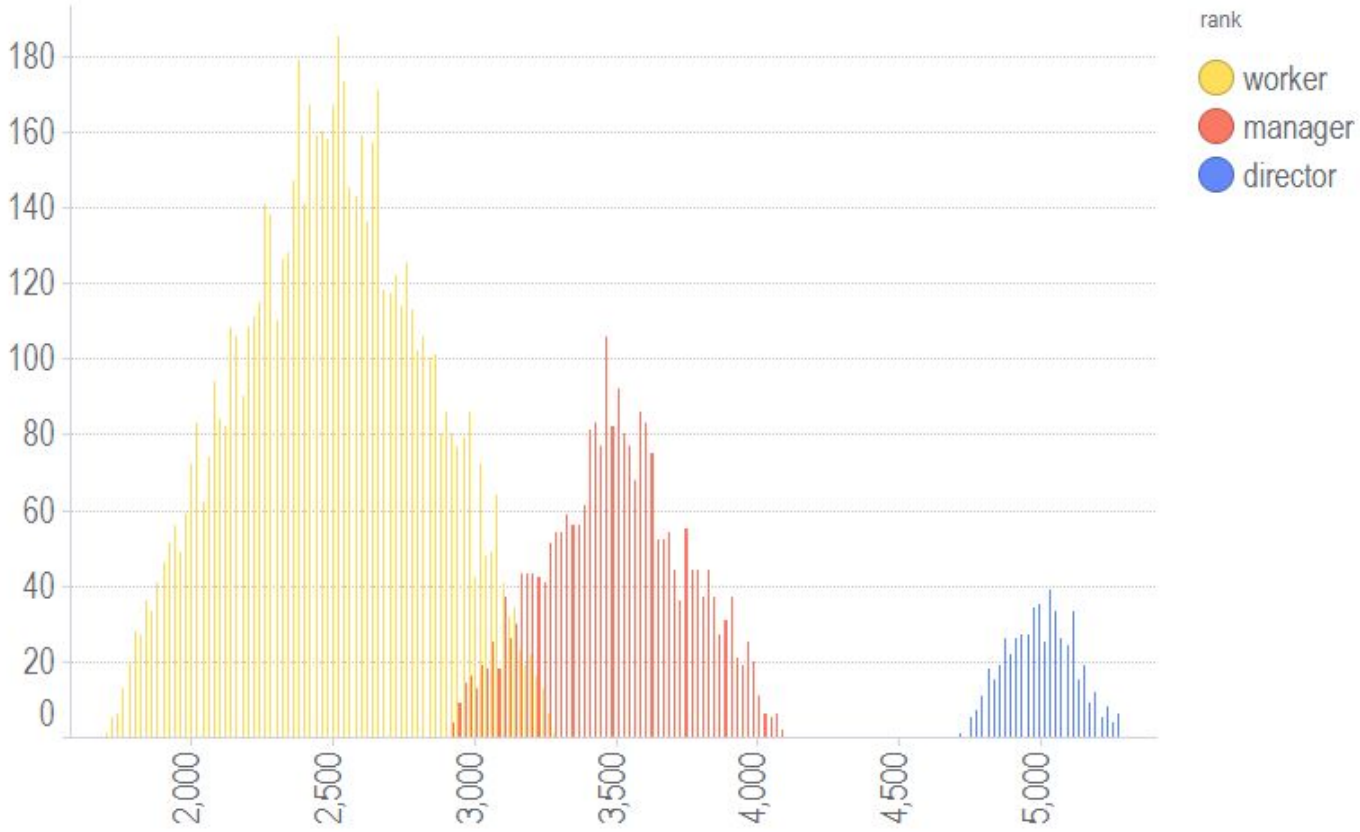
G4: the data summary

Figure G5 shows the age distribution of the employees. As we use different Box-Muller transformations for different rank, we can see the ages for works are around 25 but they spread from 17 to 35. And the ages for manages are centered around 35 but they range from 27 to 43



G5: age distribution

Figure G6 shows the salary distribution. The triangular formulas were used instead of Box-Muller transformation. Therefore, we can see the triangular shapes for each rank (each component). The work salary is ranging from below 1700 to 3400 with very high number of employees. The salary values for manage range from 3000 to 4000.



G6: salary distribution

2.4. Summary

In this simulation, we show how to capture dependencies between columns (age depends on rank, salary depends on rank) and mixture distributions (mixture of gaussian, mixture of triangulars) and some typical transformation method (box-muller transformation, triangular transformation).