Paper 198-31

Using the RAND Function in SAS® for Data Simulation in Clinical Trials

Wenping (Wendy) Zhang, sanofi-aventis, Malvern, PA

ABSTRACT

Often an important decision needs to be made based on anticipated data for a trial design or a determination of data handling rules. In this regard, simulation is a very useful method. For example, to prepare programs for statistical analyses and report generation before database lock, some SAS data has to be simulated. RAND, a new SAS function, is an easy-to-use general random number generator, and basically gives "standard distribution." Therefore, to obtain a random number for "non-standard distribution," some additional math work is needed to transform data from "standard" to "non-standard." This paper demonstrates a SAS macro that generates simulation SAS data for clinical trials for a variety of "standard" and "non-standard" distributions. Moreover, the data is generated with different distribution parameters and the expected sample size for each treatment group that corresponds to the trial design.

INTRODUCTION

The RAND (*'dist'*, *parm-1*, ... , *parm-k*) function was newly released in SAS® Version 9 but has been an experimental function since SAS® Version 8.2. It generates random numbers for any given distribution within a wide range of selections, such as normal, uniform, exponential, etc., but it gives only the "standard" format of some distributions. For example, the RAND function generates only random numbers with probability density function e^{-t} ($t > 0$), which is only a special case of the general exponential distribution $(1/\lambda)e^{-t/\lambda}$, hence, it is called "standard" distribution in this paper. In clinical trials, usually we assume that an endpoint follows a distribution with different parameters for different treatment groups. Therefore, in order to use the RAND function in data simulation through either SAS Version 8.2 or Version 9, transformations from "standard" data to "non-standard" data for these distributions are necessary.

Moreover, when simulating data, we do not merely simulate some isolated random numbers; instead, we simulate data which are part of a structured data set. For example, for different treatments at different visits, the random numbers – anticipated treatment effects – may be from a distribution with different parameters. The macro *rndnmb* is designed to generate a SAS data set for a given design structure and anticipated treatment effects.

The rest of this paper is organized as the description of macro *rndnmb***,** examples, limitations, and the conclusion.

DESCRIPTION MACRO *rndnmb*

The input of macro *rndnmb* consists of 17 input macro variables, 4 of which are required. The rest are optional. The output of this macro is a SAS data set corresponding to the specified visit structure and treatment effects.

The default seed of this macro is the system clock, but it can be changed by specifying the macro variable *seed*. This is implemented by using *call streaminit*. If the value of macro variable *seed* is missing, the seed is system clock; otherwise the value of macro variable *seed* is the seed.

Macro *rndnmb* can generate random numbers for all 20 distributions offered in SAS function RAND except the tabled distribution. The 20 distributions are Bernoulli, Beta, Binomial, Cauchy, Chi-square, Erlang, Exponential, F, Gamma, Geometric, Hypergeometric, Lognormal, Negative binomial, Normal, Poisson, T, Tabled, Triangular, Uniform, and Weibull. The distribution is specified by macro variable *dstb*. Some of the distributions, such as Weibull distribution, have some parameters. In this case the parameters are introduced by specifying macro variables *parma, parmb, and parmc*. The use of these macro variables will be illustrated below with some examples.

As mentioned in the introduction, the SAS RAND function gives only the "standard" normal, lognormal, and exponential distributions. To obtain random numbers for "non-standard" distributions of normal, lognormal or

exponential, we have to apply a transformation to the standard random variable from these distributions. The details are as follows:
Normal. It is well known that if $z \sim N(0, 1)$ then $x = \mu + \sigma z \sim N(\mu, \sigma^2)$. Therefore, to obtain "non-standard" normal distribution with parameters μ and σ, we may start with a standard normal variable, z, then calculate the "non-standard" normal variable, $x = \mu + \sigma z$.
Lognormal. Assume x is a general normal variable with mean μ, and standard deviation σ, then $r = e^x$ is lognormally distributed with parameters μ and σ.
Exponential. For the exponential distribution we know that if $t \sim e^{-t}$ then $x = \lambda t \sim (1/\lambda)e^{-x/\lambda}$. Hence, to obtain random numbers which follow an exponential distribution with parameter λ, we may generate a standard exponential variable t, then calculate the variable we need as $x = \lambda t$. Here $\lambda = E(x)$.

In macro *rndnmb* the transformations described above are specified by using two macro variables, *ca* and *cb*. For normal or lognormal distributions *ca* equals μ and *cb* equals σ; while for exponential distribution *ca* equals λ and *cb* should be given a missing value. Once these macro variables are specified corresponding to the intended distribution, macro *rndnmb* will automatically perform the transformation.

Macro *rndnmb* works also for the repeated measurement situation. In the case of repeated measurements, for a given subject the measurements are in the form of $y_{visit} = \mu + \beta \times visit + \eta + \varepsilon$, where $\eta \sim N(0, \sigma_\eta^2)$ and $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ are between-subject and within-subject variabilities respectively. Users need to specify the number of visits (*nmvst*), the visit points (*visit*), and the time trend (*slope*, β); the between-subject variability is defined by the macro variables *dstb*, *ca* (μ), and *cb* (σ_η); and the within-subject standard deviation (σ_ε) is specified by the macro variable *stdwhn*. In the current version of the macro, the between-subject variability takes only normal or lognormal distributions, and the within-subject variability can only be normally distributed. These limitations may be removed in the future.

The 17 input macro variables, their usages and input formats are described in Table 1.

Table 1: Description of input macro variables

#	Macro Variable	Description	Usage	Input Parameter format
1	OTDT	Name of output data set	Optional (default: randnmb)	SAS data set name
2	OTVAR	Output variable name for created random numbers	Optional (default: t)	SAS variable name
3	DSTB	Distribution	Required	Key word of distributions (ex: NORMAL)
4	NSTD	Non-standard distribution?	Optional (default: N)	Y/N for normal, exponent
5	TTOBS	Total # of observations or subjects	Required	Numeric number
6	NTRT	Number of treatment groups	Required	Numeric number
7	PP	Proportion of the data for each of subgroups/treatments	Required	Numeric numbers ordered by treatment groups and separated by bars (sum equal to 1). Ex: 0.25| 0.25|0.50
8	CA	The first constant (A) , used for transformation from standard to general distribution in each subgroup	Optional[1] (default: Empty)	Numeric numbers ordered by treatment groups and separated by bars[1]. Ex: 0.3|4|10
9	CB	The second constant (B) , used for transformation from standard to general distribution in each subgroup	Optional[1] (default: empty)	Numeric numbers ordered by treatment groups and separated by bars[1]. Ex: 2.1|2.4|3
10	PARMA	The first parameter needed in RAND function for each of the subgroups	Optional (default: empty)	Numeric numbers ordered by treatment groups and separated by bars. Ex: Weibull Shape 0.2|3.5|11
11	PARMB	The second parameter needed in RAND function for each of subgroups	Optional (default: empty)	Numeric numbers ordered by treatment groups and separated by bars.

				Ex: Weibull Scale 4|5.6|0.9
12	PARMC	The third parameter needed in RAND function for each of subgroups	Optional[2] (default: empty)	Numeric numbers ordered by treatment groups and separated by bars. Ex: 2.1|2.4|3
13	SEED	Seed for reproducible streams of random numbers	Optional (default: empty)	Numeric number.
14	NMVIS	The total number of visits	Optional[3] (default: =1)	Numeric numbers Ex: 2
15	SLOPE	The rate of the baseline (visit=0) mean used for the calculation of the mean at each visit point in each subgroup	Optional[3]	Numeric numbers ordered by treatment groups and separated by bars. Ex: 0|-0.5|-0.6
16	VISIT	Visit points with baseline visit =0. They are meaningful numbers with the same unit corresponding to the variable SLOPE.	Optional[3]	Numeric numbers ordered by the designed multiple visits and separated by bars. Ex: 0|1, where 0 is the baseline, 1 is at the end of treatment.
17	SDTWTHN	Within-subject standard deviation in normal distribution with mean=0 for the repeated measurement variation within subject	Optional[3] (default: =1 ~ N(0, 1))	Numeric number

Note: [1]Use only when non-standard distribution is required. [2]Use only for hypergeometric distribution. [3]Use only for multiple visit design.

EXAMPLES

Example 1: In this example we want to generate a total of 200 random numbers for three treatment groups with the ratio 1:1:2. Assume that random numbers are from a Weibull distribution for each of the treatment groups, whose parameters (Shape, Scale) are (0.2, 4), (3.5, 5.6), and (11, 0.9), respectively for each of the treatment groups. The syntax of the macro call is given below, and the contents of the output data set and listing of the first 30 observations are given in Table 2 and Table 3.

```
%rndnmb(
  otdt = weibull
 ,dstb = WEIBULL
 ,nstd = n
,ttobs = 200
 ,ntrt = 3
   ,pp = 0.25|0.25|0.50
,parma =  0.2|3.5|11     /**  shape parameter (weibull);*/
,parmb =  4|5.6|0.9      /**  scale parameter (weibull);*/
,parmc =
);
```

Table 2: Table contents of simulated dataset **weibull**

#	Variable	Type	Len	Pos	Label
7	cumpp	Num	8	40	Cumulative population (%)
1	dstb	Char	20	48	Distribution
2	order	Num	8	0	Random number generation order
5	parma	Num	8	24	Assigned Dist. parameter 1
6	parmb	Num	8	32	Assigned Dist. parameter 2
4	t	Num	8	16	Rand. variable created from assigned dist.
3	trt	Num	8	8	Treatment group

Table 3: Weibull distribution simulated data **weibull** with 3 treatment groups -- the first 30 observations

Obs	Dstb	order	Trt	t	parma	parmb	Cumpp
1	WEIBULL	1	2	4.008	3.5	5.6	50
2	WEIBULL	2	1	207.235	0.2	4.0	25
3	WEIBULL	3	3	0.922	11.0	0.9	100
4	WEIBULL	4	3	0.812	11.0	0.9	100
5	WEIBULL	5	3	0.848	11.0	0.9	100
6	WEIBULL	6	2	6.304	3.5	5.6	50
7	WEIBULL	7	1	0.337	0.2	4.0	25
8	WEIBULL	8	1	1.893	0.2	4.0	25
9	WEIBULL	9	3	0.931	11.0	0.9	100
10	WEIBULL	10	3	0.913	11.0	0.9	100
11	WEIBULL	11	1	0.013	0.2	4.0	25
12	WEIBULL	12	3	0.828	11.0	0.9	100
13	WEIBULL	13	3	0.903	11.0	0.9	100
14	WEIBULL	14	1	3.060	0.2	4.0	25
15	WEIBULL	15	1	3.174	0.2	4.0	25
16	WEIBULL	16	3	0.859	11.0	0.9	100
17	WEIBULL	17	3	0.908	11.0	0.9	100
18	WEIBULL	18	3	0.799	11.0	0.9	100
19	WEIBULL	19	2	6.571	3.5	5.6	50
20	WEIBULL	20	1	32.683	0.2	4.0	25
21	WEIBULL	21	3	0.944	11.0	0.9	100
22	WEIBULL	22	2	3.333	3.5	5.6	50
23	WEIBULL	23	1	0.570	0.2	4.0	25
24	WEIBULL	24	3	0.703	11.0	0.9	100
25	WEIBULL	25	3	0.595	11.0	0.9	100
26	WEIBULL	26	2	5.274	3.5	5.6	50
27	WEIBULL	27	3	0.807	11.0	0.9	100
28	WEIBULL	28	3	0.975	11.0	0.9	100
29	WEIBULL	29	3	0.941	11.0	0.9	100
30	WEIBULL	30	3	0.871	11.0	0.9	100

Example 2: In this example we want to generate a total of 200 random numbers for two treatment groups with the ratio 1:1. Assume that random numbers are from an exponential distribution for each of the treatment groups, whose parameter λ is 0.2 and 10 respectively. The syntax of the macro call is given below, and the contents of the output data set and listing of the first 30 observations are given in Table 4 and Table 5.

```
%rndnmb(otdt = expon
 ,dstb = EXPO
 ,nstd = Y
,ttobs = 200
 ,ntrt = 2
   ,pp = 0.5|0.5
   ,ca = 0.2|10
);
```

Table 4: Table contents of simulated data set **expon**

#	Variable	Type	Len	Pos	Label
6	Ca	Num	8	32	Constant a for transformation
5	Cumpp	Num	8	24	Cumulative population (%)
1	Dstb	Char	20	40	Distribution
2	Order	Num	8	0	Random number generation order
4	T	Num	8	16	Rand. variable created from assigned dist.
3	Trt	Num	8	8	Treatment group

Table 5: EXPO distribution simulated data **expon** with 2 treatment groups -- the first 30 observations

Obs	Dstb	order	trt	t	cumpp	Ca
1	EXPO	1	1	0.0344	50	0.2
2	EXPO	2	1	0.3325	50	0.2
3	EXPO	3	1	0.0049	50	0.2
4	EXPO	4	1	0.0813	50	0.2
5	EXPO	5	2	17.8980	100	10.0
6	EXPO	6	2	4.1650	100	10.0
7	EXPO	7	1	0.0722	50	0.2
8	EXPO	8	1	0.2665	50	0.2
9	EXPO	9	1	0.1205	50	0.2
10	EXPO	10	1	0.1185	50	0.2
11	EXPO	11	1	0.0986	50	0.2
12	EXPO	12	2	0.5446	100	10.0
13	EXPO	13	2	12.1806	100	10.0
14	EXPO	14	1	0.8407	50	0.2
15	EXPO	15	1	0.0339	50	0.2
16	EXPO	16	1	0.0911	50	0.2
17	EXPO	17	1	0.0599	50	0.2
18	EXPO	18	1	0.2487	50	0.2
19	EXPO	19	2	13.3287	100	10.0
20	EXPO	20	1	0.0948	50	0.2
21	EXPO	21	2	5.4754	100	10.0
22	EXPO	22	1	0.3799	50	0.2
23	EXPO	23	2	5.9215	100	10.0
24	EXPO	24	2	2.8745	100	10.0
25	EXPO	25	2	21.9198	100	10.0
26	EXPO	26	1	0.0529	50	0.2
27	EXPO	27	1	0.1156	50	0.2
28	EXPO	28	1	0.3120	50	0.2
29	EXPO	29	2	12.4514	100	10.0
30	EXPO	30	1	0.1186	50	0.2

Example 3: In this example, we consider the generation of random numbers from NORMAL distribution for 64 subjects who are randomized into 4 treatment groups with the ratio 1:1:1:1. Each of the 64 subjects will have 2 visits, the baseline visit and the end of treatment visit. So a total of 128 observations are expected. The time trends of the four treatment groups, specified by macro variable *slope*, are 0, -0.5, -0.6, and -0.65 respectively. The between-subject variabilities for the four treatment groups are defined as *dstb* = normal, *ca* = 41|40|43|42, and *cb* = 12|13.4|15|16. The within-subject standard deviation is given as *stdwhn* = 2.

The syntax of the macro call is given below, and the contents of the output data set and listing of the first 50 observations are given in Table 6 and Table 7.

```
%rndnmb(
  otdt  = Pltaggsm
 ,otvar = pltagg
  ,dstb = NORMAL
  ,nstd = Y
 ,ttobs = 64
  ,ntrt = 4
    ,pp = 0.25|0.25|0.25|0.25
    ,ca = 41|40|43|42
    ,cb = 12|13.4|15|16
 ,parma =
 ,parmb =
/**** repeat measurements ***/
  ,nmvis = 2
```

```
  ,slope = 0|-0.5|-0.6|-0.65
  ,visit = 0|1
,stdwthn = 2
);
```

Table 6: Table contents of simulated dataset **Pltaggsm** (total # obs = 128)

#	Variable	Type	Len	Pos	Label
6	ca	Num	8	32	First constant (A)
7	cb	Num	8	40	Second constant (B)
4	cumpp	Num	8	16	Cumulative population (%)
1	dstb	Char	15	72	Distribution
5	pltagg	Num	8	24	Rand. variable created from assigned dist.
8	sdtwthn	Num	8	48	Within subject std
9	slope	Num	8	56	Slope
2	subject	Num	8	0	Random number generation order (subject)
3	trt	Num	8	8	Treatment group
10	visit	Num	8	64	Visit point

Table 7: NORMAL distribution simulated data **Pltaggsm** with 4 treatment groups and 2 visits -- the first 50 observations

Obs	dstb	subject	trt	cumpp	pltagg	ca	cb	sdtwthn	slope	Visit
1	NORMAL	1	3	75	46.3562	43	15.0	2	-0.60	0
2	NORMAL	1	3	75	17.7852	43	15.0	2	-0.60	1
3	NORMAL	2	1	25	50.3898	41	12.0	2	0.00	0
4	NORMAL	2	1	25	48.7384	41	12.0	2	0.00	1
5	NORMAL	3	2	50	43.2321	40	13.4	2	-0.50	0
6	NORMAL	3	2	50	23.8520	40	13.4	2	-0.50	1
7	NORMAL	4	1	25	25.5237	41	12.0	2	0.00	0
8	NORMAL	4	1	25	31.2820	41	12.0	2	0.00	1
9	NORMAL	5	3	75	26.3575	43	15.0	2	-0.60	0
10	NORMAL	5	3	75	-2.4979	43	15.0	2	-0.60	1
11	NORMAL	6	3	75	39.7846	43	15.0	2	-0.60	0
12	NORMAL	6	3	75	15.7053	43	15.0	2	-0.60	1
13	NORMAL	7	3	75	57.0582	43	15.0	2	-0.60	0
14	NORMAL	7	3	75	27.6449	43	15.0	2	-0.60	1
15	NORMAL	8	2	50	54.9894	40	13.4	2	-0.50	0
16	NORMAL	8	2	50	38.2921	40	13.4	2	-0.50	1
17	NORMAL	9	2	50	21.9375	40	13.4	2	-0.50	0
18	NORMAL	9	2	50	1.8877	40	13.4	2	-0.50	1
19	NORMAL	10	3	75	33.8610	43	15.0	2	-0.60	0
20	NORMAL	10	3	75	12.4173	43	15.0	2	-0.60	1
21	NORMAL	11	2	50	60.0145	40	13.4	2	-0.50	0
22	NORMAL	11	2	50	43.6534	40	13.4	2	-0.50	1
23	NORMAL	12	2	50	67.3616	40	13.4	2	-0.50	0
24	NORMAL	12	2	50	47.4374	40	13.4	2	-0.50	1
25	NORMAL	13	2	50	32.0181	40	13.4	2	-0.50	0
26	NORMAL	13	2	50	8.9966	40	13.4	2	-0.50	1
27	NORMAL	14	3	75	59.0277	43	15.0	2	-0.60	0
28	NORMAL	14	3	75	34.0410	43	15.0	2	-0.60	1
29	NORMAL	15	1	25	31.8500	41	12.0	2	0.00	0
30	NORMAL	15	1	25	32.1791	41	12.0	2	0.00	1
31	NORMAL	16	3	75	65.8011	43	15.0	2	-0.60	0
32	NORMAL	16	3	75	40.0990	43	15.0	2	-0.60	1
33	NORMAL	17	4	100	42.4319	42	16.0	2	-0.65	0
34	NORMAL	17	4	100	17.8474	42	16.0	2	-0.65	1
35	NORMAL	18	4	100	61.8720	42	16.0	2	-0.65	0
36	NORMAL	18	4	100	37.0609	42	16.0	2	-0.65	1
37	NORMAL	19	4	100	42.7876	42	16.0	2	-0.65	0

38	NORMAL	19	4	100	15.0941	42	16.0	2	-0.65	1
39	NORMAL	20	2	50	17.9762	40	13.4	2	-0.50	0
40	NORMAL	20	2	50	-2.5215	40	13.4	2	-0.50	1
41	NORMAL	21	1	25	43.1537	41	12.0	2	0.00	0
42	NORMAL	21	1	25	44.5833	41	12.0	2	0.00	1
43	NORMAL	22	3	75	34.5868	43	15.0	2	-0.60	0
44	NORMAL	22	3	75	6.8111	43	15.0	2	-0.60	1
45	NORMAL	23	3	75	43.4553	43	15.0	2	-0.60	0
46	NORMAL	23	3	75	15.5165	43	15.0	2	-0.60	1
47	NORMAL	24	4	100	38.4346	42	16.0	2	-0.65	0
48	NORMAL	24	4	100	15.0635	42	16.0	2	-0.65	1
49	NORMAL	25	4	100	55.3236	42	16.0	2	-0.65	0
50	NORMAL	25	4	100	30.2963	42	16.0	2	-0.65	1

The distribution of the simulated dataset **Pltaggsm** can be verified through SAS procedure RELIABILITY. The sample SAS code is the following:

```
proc reliability ;
  distribution normal;
  analyze pltagg;
  by trt visit;
run;
```

LIMITATION

All twenty valid distributions in the SAS RAND function can be specified in this macro except the Tabled distribution. The macro call in creating multiple visits for repeated measurements is applicable only for the normal distribution or lognormal distributions, and the within-subject variability is limited to the normal distribution only. Data simulations for crossover designs are difficult to be covered in such a general SAS macro.

CONCLUSION

The macro demonstrated in this paper can easily be used to create simulation data to meet various needs in scientific research and development. It is an especially useful tool in preparing programming work before the actual data becomes available.

REFERENCES

SAS Institute Inc. *SAS® 9.1 Language Reference: Dictionary, Volumes 1 and 2.* Cary, NC: SAS Institute Inc.

Walck, C. (2001). *Hand-book on Statistical Distributions for Experimentalists.* Internal Report SUF-PFY96-01, University of Stockholms.

CONTACT INFORMATION

Comments and questions are encouraged. Please contact the author at:

Wenping (Wendy) Zhang
Biostatistics and Programming, Sanofi-aventis
9 Great Valley pkwy
Malvern, PA 19355
Phone: 610-889-6655
Fax: 610-889-6932
e-mail: wendy.zhang@sanofi-aventis.com