doi:10.5959/eimj.3.1.2011.er1

EDUCATIONAL RESOURCE

Creating artificial data for teaching of statistics

Wan Nor Arifin

Biostatistics and Research Methodology Unit, School of Medical Sciences, Universiti Sains Malaysia.

Abstract

For the purpose of teaching statistics, lecturers often rely on data from real studies, text book examples or painstakingly created datasets. The process of creating a dataset can be made easier with the utilization of PASW Statistics to generate random values. The objective of this article is to demonstrate the creation of data which are measured on continuous scale, using PASW Statistics menus and syntax.

Keywords: artificial data, normal distribution, random values, teaching statistics

Introduction

For the purpose of teaching statistics, datasets used are usually obtained from real studies, text book examples or created by lecturers themselves. By using data from real studies, there is an issue of violation of confidentiality of information, more so for clinical studies in which sensitive information from patients might be obtained, and their medical records are reviewed. Even though the data are made anonymous, and with most of the original data deleted and truncated, still the issue of confidentiality might be there. On the other hand, the use of textbook examples as the source of data for teaching might raise the issue of copyright of intellectual properties.

Creation of artificial datasets is a better option for the purpose of teaching. However, data creation is a painstaking process if done manually, which means the observations are keyed-in one by one based on whatever values come to the lecturer's mind, then having to adjust the values repeatedly to fulfill some distributional assumptions. Being able to utilize computer software to automate this process would be a blessing to the lecturer.

In this article, I will demonstrate the creation of datasets consisting of data measured on continuous scale using PASW Statistics version 18 [1]. I assume that readers have basic working knowledge to use this software. I will show to readers how to make the data to meet normality assumption for the purpose of parametric statistical tests, as well as to make the data to violate the assumption of normality by making the distribution skewed to the left and right.

NORMAL DISTRIBUTION

Dataset descriptions

We want to create a dataset consisting of a variable SYSTOLIC. The observations are systolic blood pressure readings of 100 subjects. The distribution of systolic blood pressure readings in this sample is normally distributed with mean of 120mmHg and standard deviation of 14.6mmHg [2]. The observation is precise up to 2mmHg.

Creating new cases

Start PASW Statistics version 18 (referred as PASW in this article) with an empty dataset. We want to use **Transform → Compute Variable** menu, but with an empty dataset we would be presented with an error pop-out box, stating that we need to have some data.

Before using the menu, the dataset must have a number of cases or observations for us to proceed. Create a new variable and rename the variable to SYSTOLIC and set the Decimals to 0. To create 100 empty cases, highlight a number of rows under SYSTOLIC variable column, then from menus choose **Edit → Insert Cases**. The rows would be filled with "." and observation numbers would appear in black instead of gray. Scroll down and repeat the same step until we obtain 100 observations.

Generating values

Once we have 100 cases, from menus choose **Transform → Compute Variable**. Set Target Variable as SYSTOLIC. For Numeric Expression, under Function group select Random Numbers, then select `Rv.Normal` under Functions and Special Variables. This function is written as `RV.NORMAL(mean,stddev)`; enter our desired mean (120) in the first option, followed by our standard deviation (14.6). The expression should look like this (**Figure 1**):

```
RV.NORMAL(120,14.6)
```

Next, we need to round our values and set the precision to 2mmHg. Highlight `RV.NORMAL(120,14.6)` whole expression. Under Function group select Arithmetic, then select `Rnd(2)` under Functions and Special Variables. This function is written as `RND(numexpr, mult)`; `numexpr` is our numeric expression

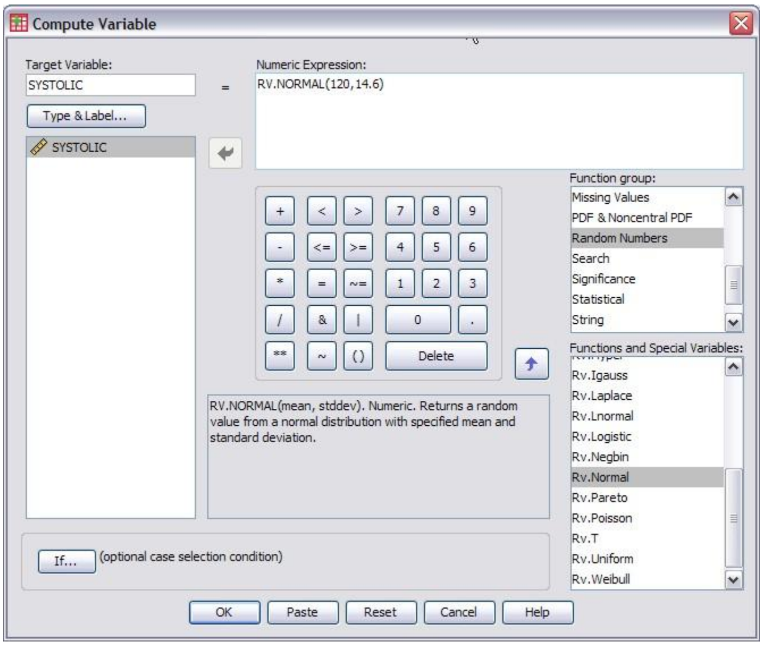

Figure 1. Compute Variable window. Write expression in Numeric expression text area.

RV.NORMAL(120,14.6), and mult is for setting rounding parameter in multiple of the number. By setting it to 1, we would have normal rounding of the values with 0 decimal place (all the values are in multiple of 1). By setting it to 2, we will round the values with precision of 2 (that is in multiple of 2). In our case we round the values to the closest 2mmHg, so we set our the Numeric Expression as (**Figure 2**):

```
RND(RV.NORMAL(120,14.6),2)
```

Click OK. With that, we already generated values systolic blood pressure for 100 subjects with mean and standard deviation of about 120mmHg and 14.6mmHg respectively. Save the dataset.

Data checking

We can check out the descriptive statistics of our data as well as checking the data for

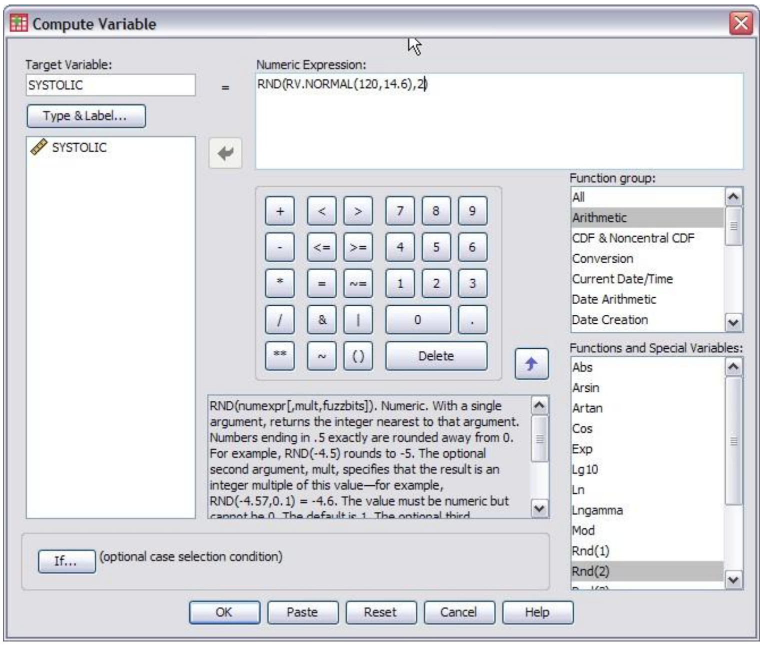

Figure 2. Rounding generated values and setting the precision.

normality. Take a look at PASW output for descriptive statistics and tests of normality (**Table 1 and Table 2**) as well as graphical checks on normality **Figure 3 and Figure 4**).

Table 1. PASW output for descriptive statistics systolic blood pressure readings.

Descriptives

			Statistic	Std. Error
SYSTOLIC	Mean		121.34	1.442
	95% Confidence Interval for Mean	Lower Bound	118.48	
		Upper Bound	124.20	
	5% Trimmed Mean		121.31	
	Median		120.00	
	Variance		207.843	
	Std. Deviation		14.417	
	Minimum		86	
	Maximum		154	
	Range		68	
	Interquartile Range		21	
	Skewness		.106	.241
	Kurtosis		-.287	.478

Table 2. PASW output for tests of normality.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
SYSTOLIC	.057	100	.200*	.990	100	.662

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

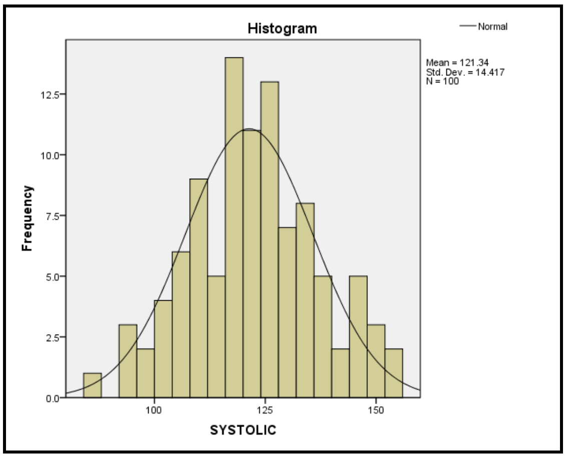

Figure 3. Histogram of systolic blood pressure readings.

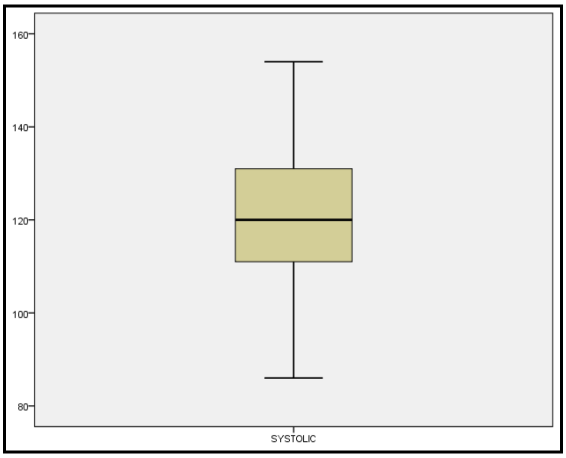

Figure 4. Box-and-Whisker plot for systolic blood pressure readings.

Generating data using syntax

A faster way to generate the data is by using syntax. In this example, we want the syntax to open a new dataset, automatically create 100 empty observations, then generate the values of systolic blood pressure just like we did previously.

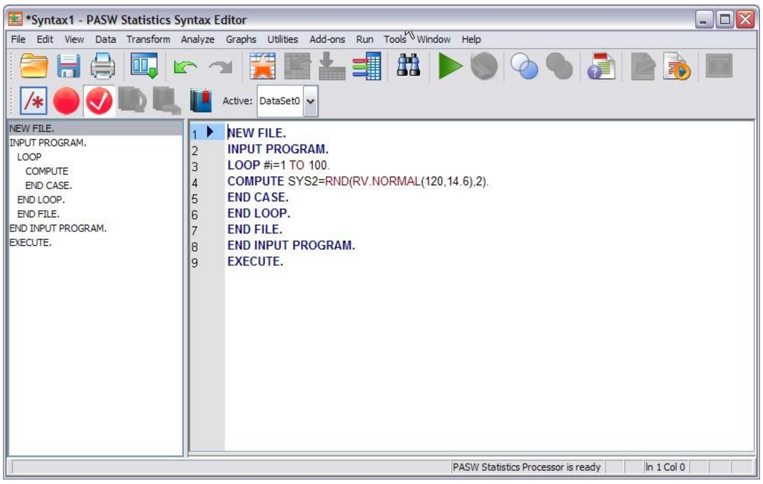

Figure 5. Syntax Editor.

From menus choose **File → New → Syntax**. This will open Syntax Editor (**Figure 5**).

Write the following syntax in editor pane of Syntax Editor window (text input area on the right):

```
NEW FILE.

INPUT PROGRAM.

LOOP #i=1 TO 100.

COMPUTE SYSTOLIC =
  RND(RV.NORMAL(120,14.6),2).

END CASE.

END LOOP.

END FILE.

END INPUT PROGRAM.

EXECUTE.
```

Please take note of the "." at the end of each line as it can be easy missed. From menus choose **Run → All.** A new dataset will be open, together with 100 observations in SYSTOLIC variable.

This line of syntax:

```
LOOP #i=1 TO 100.
```

can be changed accordingly to adjust for the number of observations required. For example we want to have 500 observations in SYSTOLIC variable, then just change "100" to "500":

```
LOOP #i=1 TO 500.
```

The name of resulting variable can also be changed. For example, instead of naming the new variable as SYSTOLIC, you can change it to another name for example SYS_BP:

```
COMPUTE SYS_BP =
  RND(RV.NORMAL(120,14.6),2).
```

The expression on the right hand side of the equal sign ("=") can also be changed accordingly. Just make sure that you tested the syntax beforehand to make sure that the syntax would run smoothly as well as the resulting observations are as required.

Deciding on standard deviation value

After deciding on the value of the mean, the standard deviation can be decided based on literature or from experience. However, a logical standard deviation can also be estimated based on minimum and maximum value that we decide on and number of subjects.

For example, we want to add a new variable for diastolic blood pressure. We decided that the smallest diastolic blood pressure reading is about 50mmHg, the largest is about 110mmHg, with a mean of 80mmHg. We want to generate values for 250 subjects. The range of the values is 110 minus 50, which is 60. We want to have 50 as the smallest value, while 110 as the largest value at opposite tails of normal distribution.

Using z distribution as the base for calculation, for two-tailed, 99.9% of the area lies between -3.29 to 3.29 standard deviation of z distribution. So, approximately 1 standard deviation of diastolic blood pressure reading is:

60 / (3.29 x 2) = 9.12

On the other hand, to be more precise since we generate values for a sample not a population, we can base our calculation on t distribution instead. For two-tailed, 99.9% of the area of the distribution lie between -3.33 to 3.33 standard deviation of t distribution. So, approximately 1 standard deviation of diastolic blood pressure reading is:

60 / (3.33 x 2) = 9.00

Please refer **Appendix 1** for calculation of standard deviation of distributions.

The following is the descriptive results for dataset generated for diastolic blood pressure for 250 observations:

Table 3. PASW output for descriptive statistics for diastolic blood pressure readings.

Descriptives

			Statistic	Std. Error
DIASTOLIC	Mean		80.3520	.59956
	95% Confidence Interval for Mean	Lower Bound	79.1711	
		Upper Bound	81.5329	
	5% Trimmed Mean		80.3822	
	Median		80.0000	
	Variance		89.868	
	Std. Deviation		9.47985	
	Minimum		54.00	
	Maximum		108.00	
	Range		54.00	
	Interquartile Range		12.00	
	Skewness		-.024	.154
	Kurtosis		.001	.307

Note that with random number generator in PASW, we would not get the exact mean, standard deviation, maximum and minimum values, and range that we specified. We can repeatedly run **Calculate Variable** repeatedly until we have an acceptable dataset.

SKEWED DISTRIBUTION

Dataset descriptions

We want to create a dataset consisting of a variable SYS_RIGHT. The variable observations are systolic blood pressure readings of 150 subjects. The distribution of systolic blood pressure readings in this sample is skewed to the right with smallest value of about 80mmHg, with standard deviation of about 8mmHg. The observation is precise up to 2mmHg.

Skewed to the right

By following the steps explained previously, create 150 cases in an empty dataset. You can also use syntax for that purpose. Choose **Transform → Compute Variable**, and then select Random Numbers under Function group, followed by `Rv.Halfnrm` under Functions and Special Variables. This function is written as `RV.HALFNRM(mean, stddev)`; enter our smallest value for `mean` (80), and about one half to two times the value of our standard deviation for `stddev` (we use 15, guess work). Our expression should look like this:

```
RV.HALFNRM(80,15)
```

Then round the value with RND function:

```
RND(RV.HALFNRM(80,15), 2)
```

Take a look at descriptive statistics (**Table 4**) and histogram (**Figure 6**) for the generated values that I obtained.

Table 4. PASW output for descriptive statistics for skewed distribution of blood pressure readings.

Descriptives

			Statistic	Std. Error
SYS_RIGHT	Mean		91.0533	.71612
	95% Confidence Interval for Mean	Lower Bound	89.6383	
		Upper Bound	92.4684	
	5% Trimmed Mean		90.3926	
	Median		90.0000	
	Variance		76.923	
	Std. Deviation		8.77059	
	Minimum		80.00	
	Maximum		122.00	
	Range		42.00	
	Interquartile Range		12.00	
	Skewness		1.117	.198
	Kurtosis		1.033	.394

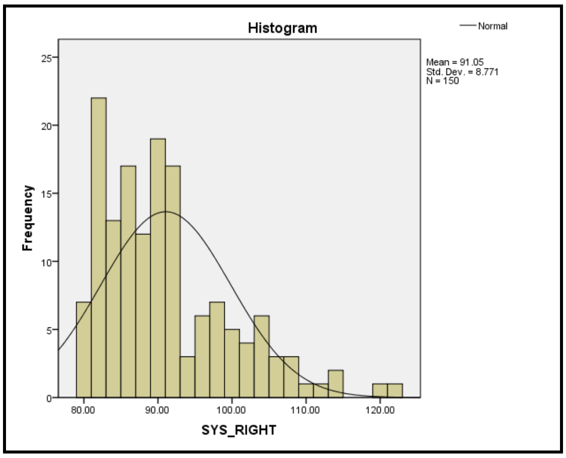

Figure 6. Histogram of systolic blood pressure readings. It is skewed to the right.

Skewed to the left

To create observations with distribution that is skewed to the left, repeat the steps similar to creating a skewed to the right dataset outlined previously. Then, mirror image the observations by turning all the values to negative values by deducting the values from 0 by choosing **Transform → Compute Variable**, then name a new variable as SYS_LEFT in Target Variable box, and write:

```
0 - SYS_RIGHT
```

in Numeric Expression.

Check the minimum value in descriptive statistics, then add up all the values to the positive of the minimum value plus the desired lower limit (80). For example, in my generated dataset, the minimum is -122, so compute:

```
SYS_LEFT + 122 + 80
```

in Numeric Expression.

Conclusion

Creating an artificial dataset with observations measured on continuous scale is easy and frill free with the utilization of PASW functions, as opposed to manually keying in values and making adjustments to have the desired distributions with the observations. Apart from using data from real studies or from text book examples, creating artificial datasets for teaching purposes is a viable option for statistics lecturers as it allows creation of data specific to the objectives of a lecture. It is hoped that this article would give a clear idea as to the creation of artificial datasets.

References

1. SPSS Inc. PASW Statistics 18. Chicago IL: SPSS Inc.; 2009.

2. Lim T, Morad Z. Prevalence, awareness, treatment and control of hypertension in the Malaysian adult population: Results from the national health and morbidity survey 1996. Singapore medical journal. 2004;45(1):20-7.

3. StataCorp LP. Stata/SE 11.2 for Windows. Texas: StataCorp LP; 2009.

Further readings

1. EFOmarko. MathKB Online forum: Synthetic data set. Advenet LLC; 2005 [June 17, 2011]; Available from: http://www.mathkb.com/Uwe/Forum.aspx/spss/1306/Synthetic-data-set.

2. Pickering A. Computer Use Classes Accompanying Statistics Lectures 2008-2009. London2008 [cited 2011 June 17, 2011]; Available from: http://homepages.gold.ac.uk/aphome/cc5work.doc.

3. SPSS Inc. PASW Statistics 18 online help. Chicago IL: SPSS Inc.; 2009.

Appendix 1

The calculation for the standard deviations of distribution at given probability or confidence interval were calculated with STATA 11 [3]. The commands used are:

For z distribution:

```
. di invnorm(1-.0005)
3.2905267
```

At 99.95% confidence interval, two tailed.

For z distribution:

```
. di invttail(249, 0.0005)
3.3300269
```

At 99.95% confidence interval, two tailed. Degree of freedom for t distribution is sample size, 250 minus 1, which is equal to 249.

Corresponding Author: Dr. Wan Nor Arifin, Biostatistics and Research Methodology Unit, School of Medical Sciences, Universiti Sains Malaysia, 16150 Kubang Kerian, Kelantan, Malaysia. Email: wnarifin@yahoo.com

Accepted: June 2011 Published: June 2011