CREATING REALISTIC DATA SETS WITH SPECIFIED PROPERTIES VIA SIMULATION

Robert Goldman and John D. McKenzie, Jr.
Simmons College, Boston, MA 02115 and Babson College, Babson Park, MA 02457
robert.goldman@simmons.edu and mckenzie@babson.edu

Abstract

There are many situations in which an instructor is confronted with a summarized data set. For example, he or she may discover an interesting data set in which only its mean, standard deviation, and sample size are given. This summary may be found in a newspaper or journal article or in a textbook example or exercise. But, without the actual or raw data, the instructor may decide not to use the data set because it is impossible to present a visual display such as a graph or a table, something that the statistical community considers to be essential for a complete data analysis. Nor can he or she illustrate the far more common situation of data analysis with raw data if only summarized data are given.

This paper will describe how to generate a raw data set with specified characteristics by using simulation. Some introductory applied statistics courses include an introduction to simulation by showing how to generate a set of random data from a normal distribution. However, the creation of realistic data sets is almost never present in such courses because most instructors are unaware of the ease in which one may generate such data with the use of statistical software, such as Minitab.

1. Introduction

1.1 Types of Situations

In our presentation at the ICTCM, we explained how to create data sets to illustrate the following three topics present in most introductory applied statistics courses:

1. One variable with specified $\overline{X}$, S, n for a one-sample t procedure
2. Two variables with specified cell frequencies for contingency table chi-square test
3. Two variables with specified correlation coefficient, means, and standard deviations.

The three examples we used to illustrate these situations were taken from De Veaux, Velleman, and Bock (DVB) (2006)

We have found Minitab to be an excellent tool for simulating data in the fashion described in this report. It is the statistical software most commonly used in the first statistics course. A convenient guide to the Minitab software is McKenzie and Goldman (2005).

The GAISE College Report [Garfield (2005)] urges instructors to take advantage of technology in teaching statistics. These methodologies are consistent with that recommendation.

1.2 How Can I Use these Results?

The results of these simulations can be used in a number of ways. First, one can create raw data when you have only summarized results for your favorite problems. Or for textbook exercises in which only summarized data are present. Second, such simulations may be used to generate raw data for an instructor to obtain graphs and to test assumptions for such problems and exercises. Third, one may create their own problems with specified characteristics, including ones with specified outcomes. For instance, a set of raw data in which a null hypothesis is rejected when α is .05 but not rejected when α is .01. Finally, these techniques allow an instructor to produce examinations and quizzes with different data for different students to either illustrate a point (or prevent unauthorized collaboration).

1.3 Previous Work on These Situations

Surprisingly, there appears to be almost nothing in the literature about such simulations. Of the various situations we reviewed for this paper, the only situation that appears in the literature is where the aim is to generate bivariate quantitative data with specified characteristics. There is a substantial literature on generating bivariate quantitative data from bivariate distributions (usually the bivariate normal) with specified parameters. A recent, elementary treatment can be found in Hunt (2001), for example. The only previous technique for generating bivariate quantitative data with specified characteristics is iterative, complex (based upon number theory), but not random. See Searle and Firey (1980).

Because of space limitations, in this paper we cover only the first of these three situations. Information about the other two methodologies may be obtained from the authors.

2. One Variable with Specified $\overline{X}$, S, and n for a One-Sample t Procedure.

In this situation our example is taken from DVB, p. 541 Q. 12.

Portable phones *A manufacturer claims that a new design for a portable phone has increased the range to 150 feet, allowing many customers to use the phone throughout their homes and yards. An independent testing laboratory found that a random sample of 44 of these phones worked over an average of 142 feet, with a standard deviation of 12 feet. Is there evidence that the manufacturer's claim is false?*

Here we create a raw data set with a specified mean, standard deviation, and sample size. First, we name the four variables we will need for this example. The "Base" command ensures that the same result occurs each time the sequence of commands is issued. We next randomly generated 44 observations from a normal distribution with a mean of 126 and a standard deviation of 15.

```
MTB > name c1 "1stDis" c2 "1stDis_Stan" c3 "2ndDis" c4 "Dis"
MTB > Base 2312
MTB > Random 44 c1;
SUBC>   Normal 126 15.
```

But, as the following descriptive statistics indicate, the mean and standard deviation are close to, but not equal to, our desired values. Hence, we standardized our observations so that they had a mean of 0 and a standard deviation of 1, multiplied the resulting column by 15, and added 126 to it. If we had issued a "Describe" command at this point we would have had a column with the desired characteristics. But, we decided to round the values to two decimal places, to make these values be more realistic. (Note that to obtain the exact mean and standard deviation values, this procedure may have to repeated more than once due to round-off difficulties.)

```
MTB > Describe  '1stDis';
SUBC>   Mean;
SUBC>   StDeviation;
SUBC>   N.
```

Descriptive Statistics: 1stDis

```
Variable   N    Mean  StDev
1stDis    44  127.40  13.54
```

```
MTB > Center '1stDis'  '1stDis_Stan'.
MTB > Let '2ndDis' = '1stDis_Stan'*15 + 126
MTB > Let 'Dis' = ROUND('2ndDis',2)
MTB > Describe  'Dis';
SUBC>   Mean;
SUBC>   StDeviation;
SUBC>   N.
```

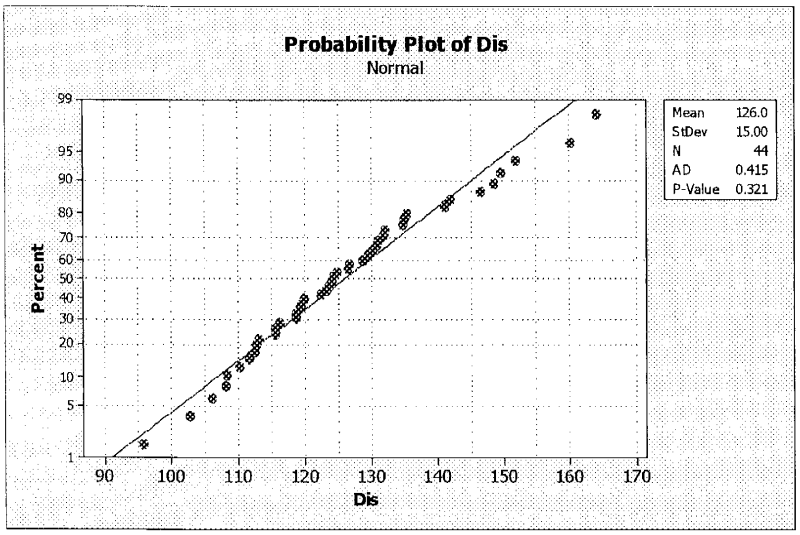
Descriptive Statistics: Dis

```
Variable   N    Mean  StDev
Dis       44  126.00  15.00
```

The next step is to check whether the generated data set satisfies the assumption of normality required to use the appropriate one-sample t test for a single mean. Based upon the Anderson-Darling test for normality associated with the normal probability plot ($p = 0.321$), there does not appear to be any evidence to negate the validity of this assumption.

```
MTB > NormTest 'Dis'.
```

Probability Plot of Dis



One variable with specified $\bar{X}$, S, and n for a one-sample t procedure—with skewed (Exponential) data

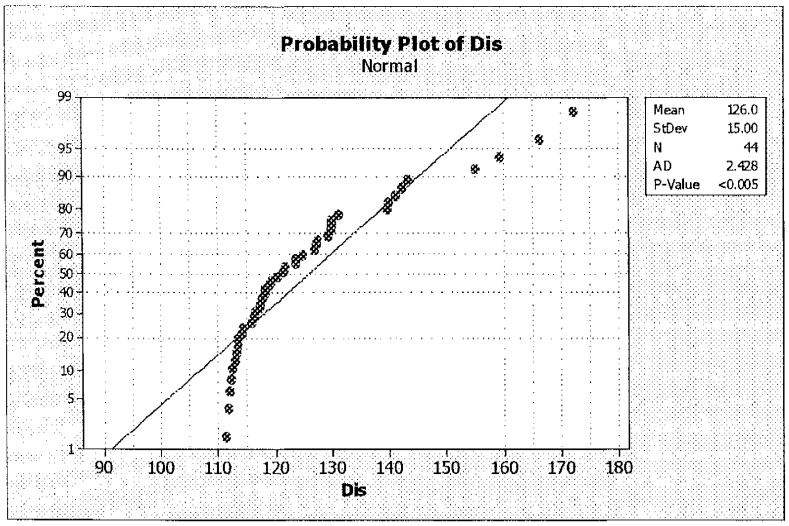
The only difference between this revised example and the one above is that we generated the 44 observations from an exponential distribution with a mean of 1. It results in a data set that does not meet the normality condition required for this t-test.

```
MTB > name c1 "1stDis" c2 "1stDis_Stan" c3 "2ndDis" c4 "Dis"
MTB > Base 2312
MTB > Random 44 c1;
SUBC>   Exponential 1.0.
 :    :     :     :     :     :     :

 :    :     :     :     :     :     :
MTB > Describe  'Dis';
SUBC>   Mean;
SUBC>   StDeviation;
SUBC>   N.
```

Descriptive Statistics: Dis

```
Variable   N    Mean  StDev
Dis       44  126.00  15.00

MTB > NormTest 'Dis'.
```

Probability Plot of Dis



3. Conclusion

The examples of the three situations presented at the ICTCM illustrate a range of simulation techniques. They show how an instructor can use statistical software to create a raw data set from summarized data. The raw data can then be used to obtain a visual display for either a presentation or to check a condition necessary for inference.

It is straightforward to adapt these techniques to simulating data for a paired t-test, for a one-way analysis of variance, and for a chi-square test of fit. The methods we used in the first and third situations for quantitative data are flexible enough to produce a variety of data shapes, as well as outlying values.

References

De Veaux, R.D., Velleman, P. F., and Bock, D. E., (2006). *Intro Stats* (Second Edition). Boston: Addison-Wesley.

Garfield, J. (Chair), et al (2005). *The Guidelines for Assessment and Instruction in Statistical Education (GAISE) College Report.* Alexandria, Virginia, USA: American Statistical Association. www.amstat.org/education/gaise.

Hunt, N. (2001).Generating Multivariate Normal Data in Excel. *Teaching Statistics* (pp 58-59) Vol. 23, No. 2.

McKenzie, J. D., Jr. and Goldman, R. (2005). *The Student Guide to Minitab, Release 14*, Boston: Addison-Wesley.

Searle, S. R. and Firey, P.A., . (1980). Computer Generation of Data Sets for Homework Exercises in Simple Regression. *The American Statistician* pp. 51-54, Vol. 34 No. 1. Alexandria, Virginia, USA: American Statistical Association.