A Simple Dataset for Demonstrating Common Distributions

Peter K. Dunn
University of Southern Queensland

Journal of Statistics Education v.7, n.3 (1999)

Key Words: Binomial; Births; Classroom data; Exponential; Geometric; Normal; Poisson.

Abstract

The baby boom dataset contains the time of birth, sex, and birth weight for 44 babies born in one 24-hour period at a hospital in Brisbane, Australia. The data can be used to demonstrate that some common distributions -- the normal, binomial, geometric, Poisson, and exponential -- can be used to model real situations. Because the dataset is small and easily understood, it provides a useful classroom example for discussing these distributions.

1. Introduction

1 Useful datasets may come from surprising places. A dataset that appeared in a local tabloid newspaper proved to be of great value in teaching some common distributions. The dataset is easy to understand and small enough that students can manipulate the data in class, yet it can be used for demonstrating the normal, binomial, geometric, Poisson, and exponential distributions. We do not claim that the distributions fitted to particular variables are necessarily the 'best' fitting distributions in any sense. Rather, we show that the data follow distributions that they could reasonably be expected to follow. In a course introducing common distributions, the dataset demonstrates that these distributions can have practical applications. The dataset can also be used to illustrate hypothesis tests about proportions, comparisons of birth weights by gender, the runs test of randomness of gender, and skewed data.

2. The Baby Boom Data

2 These data appeared in an article entitled "Babies by the Dozen for Christmas: 24-Hour Baby Boom" in the newspaper *The Sunday Mail* on December 21, 1997 (Steele 1997). According to the article, a record 44 babies were born in one 24-hour period at the Mater Mothers' Hospital, Brisbane, Australia, on December 18, 1997. The article listed the time of birth, the sex, and the weight in grams for each of the 44 babies.

3 We use these data in a unit in which we discuss, among other things, various discrete and continuous distributions, along with the theory of distributions. Where possible, we try to use real datasets. The present dataset provides a common source of examples for use in a number of situations.

4 The dataset contains four variables: TIME, the time each birth occurred, given on the 24-hour clock; SEX, the sex of the baby, coded using 1 for a girl and 2 for a boy; WEIGHT,

the birth weight in grams; and MINSMID, the number of minutes after midnight that each birth occurred.

3. The Baby Boom in the Classroom

5 These data have been used in various situations, which are considered separately below. The fits are generally quite good for such a small dataset. In the classroom, a copy of the original article (which includes the data) is given to each student.

6 In this article, *p*-values are given for the goodness of fit of certain distributions. These *p*-values have been calculated with the Kolmogorov-Smirnov test in S-Plus (MathSoft 1997) using `ks.gof`. In the case of the normal and exponential distributions, the null hypotheses are composite; that is, the parameters of the distribution have been estimated from the sample. For the other distributions, this facility is unavailable, and so the *p*-values are computed under the assumptions that the parameters of the distribution are known. This test also assumes continuous data. Together with the small sample sizes, the given *p*-values are only of limited use, and are given only as a general indication of the suitability of the distributions. (The χ^2 goodness-of-fit test is difficult to apply in many cases because of small counts.) For further information, see, for example, Daniel (1990, chap. 8).

3.1 Binomial Distribution

7 The 44 babies comprised 18 girls and 26 boys. A simple application of the binomial distribution is to determine the probability that such an imbalance would occur if indeed the probability of giving birth to a boy equaled the probability of giving birth to a girl. (This ignores the well-known fact that the probability of a boy's being born is slightly higher than the probability of a girl's being born. Such information could, of course, be easily included.) The probability of observing at least 26 babies of the same sex in 44 births is 0.2912 under this assumption. It is therefore not that unusual.

3.2 Geometric Distribution

8 The geometric distribution can be used to model the number of births up to (and including) a boy's birth. With the small dataset, this exercise takes less than ten minutes. We need to ignore the last three births (which were all girls) to keep matters simple, and use the estimate $\hat{p}$ = 26/41 ≈0.634 as the probability that a boy was born during this time. The number of births until a boy was born can then be tallied, restarting the count after each boy's birth. The results are given in Table 1. The fit appears to be reasonable. The *p*-value of the Kolmogorov-Smirnov test is approximately 0.27, indicating that there is insufficient evidence to reject the hypothesis of the geometric distribution as a possible model.

Table 1. Fitting the Geometric Distribution

Births Until Boy Born	Tally	Empirical Probability	Theoretical Probability
1	18	0.692	0.634

2	3	0.115	0.232
3	4	0.154	0.085
4	0	0.000	0.031
5+	1	0.040	0.018
Total	26	1.001	1.000

3.3 Poisson Distribution

9 The Poisson distribution can be used to model the number of births each hour over the 24-hour period. Again, the theoretical probabilities (using a mean of 44/24 ≈1.83 births per hour) and experimental probabilities are calculated and compared. The task again takes about ten minutes. The results are given in Table 2. The approximate *p*-value of the Kolmogorov-Smirnov test is found to be 0.64, indicating that there is insufficient evidence to reject the Poisson distribution as a possible model.

Table 2. Fitting the Poisson Distribution

Births per Hour	Tally	Empirical Probability	Theoretical Probability
0	3	0.125	0.160
1	8	0.333	0.293
2	6	0.250	0.269
3	4	0.167	0.164
4	3	0.125	0.075
5+	0	0.000	0.039
Total	24	1.000	1.000

3.4 Exponential Distribution

10 After fitting the Poisson distribution to the arrival times, it is natural to try to model the times between births using an exponential distribution. This is a little more tedious, and although it has been done in class, it may be more suited to a tutorial exercise. The first birth occurred at 0005, and the last birth in the 24-hour period at 2355. Thus the 43 inter-birth times happened over a 1430-minute period, giving a theoretical mean of 1430/43 = 33.26 minutes between births. The theoretical probabilities (calculated before the class in this situation) are found by integrating between the class boundaries. That is, the theoretical

probability for the first class is found by integrating between 0.0 and 19.5, for the second class by integrating between 19.5 and 39.5, and so on. The results are given in Table 3. The fit in this case is again quite good; the *p*-value using the Kolmogorov-Smirnov test for composite exponentiality is 0.24.

Table 3. Fitting the Exponential Distribution

Time Between Births (minutes)	Tally	Empirical Probability	Theoretical Probability
00-19	18	0.419	0.444
20-39	12	0.279	0.251
40-59	6	0.140	0.138
60-79	5	0.116	0.076
80+	2	0.047	0.092
Total	43	1.001	1.001

3.5 The Normal Distribution

11 Another modeling exercise is to model the birth weights using a normal distribution. Not surprisingly, the birth weights are not modeled very well by a normal distribution. Babies born prematurely or with illnesses, and many other factors contribute to the non-normality. One would generally expect, given these factors, that the distribution would tend to be skewed to the left.

12 Before displaying the data, these issues have been discussed with the students. Thus, before seeing the data, they expect to see a negatively skewed distribution. The Q-Q plot for the data is shown in Figure 1. The *p*-value for a Kolmogorov-Smirnov test of composite normality is 0.0007. The students are not surprised by such a plot, and they suggest that grouping the babies by gender would be sensible. The corresponding Q-Q plots are shown in Figure 2, and the *p*-values for Kolmogorov-Smirnov tests of composite normality are 0.0283 (girls) and 0.106 (boys). We can conclude, then, that the distributions by gender remain negatively skewed; separating the dataset by gender does not greatly improve the fit of the normal distribution.

Figure 1 (7.7K gif)

Figure 1. Q-Q Plot of the Birth Weights (in Grams).

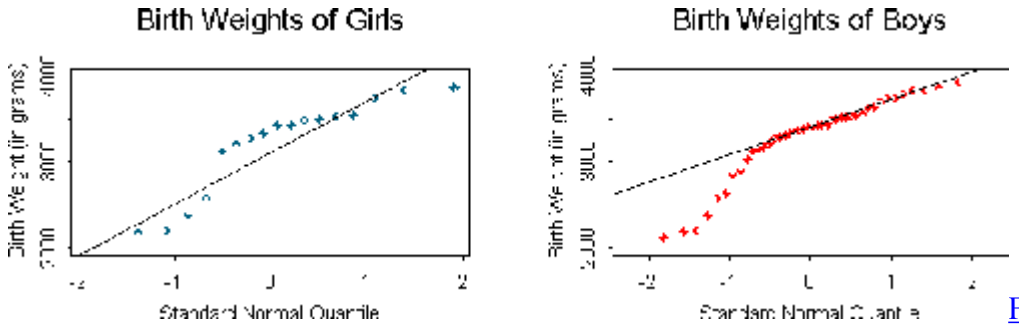

Figure 2 (4.9K gif)

Figure 2. Q-Q Plot of the Birth Weights (in Grams) by Gender.

---

4. Conclusion

13 The simple dataset given in Steele (1997) is useful and constructive in the classroom. Because the dataset is small and easily understood, distributions can be fitted to the data in the class by the students. Despite its small size, the dataset demonstrates useful applications of many common distributions, even if the fitted distributions are not necessarily optimal.

5. Getting The Data

14 The file babyboom.dat.txt contains the raw data. The file babyboom.txt is a documentation file containing a brief description of the dataset.

Acknowledgments

The author wishes to thank the editor and the reviewers for their constructive comments which led to many improvements, and for pointing out an error in calculation.

Appendix - Key To Variables in babyboom.dat.txt

```
Columns
 1 -  8  Time of birth recorded on the 24-hour clock
```

```
 9 - 16  Sex of the child (1 = girl, 2 = boy)
17 - 24  Birth weight in grams
25 - 32  Number of minutes after midnight of each birth
```

Values are aligned and delimited by blanks. There are no missing values.

References

Daniel, W. W. (1990), *Applied Nonparametric Statistics* (2nd ed.), Boston: PWS-KENT Publishing Company.

MathSoft (1997), *S-PLUS 4 Guide to Statistics*, Data Analysis Products Division, Seattle: Author.

Steele, S. (December 21, 1997), "Babies by the Dozen for Christmas: 24-Hour Baby Boom," *The Sunday Mail* (Brisbane), p. 7.

Peter K. Dunn
Department of Mathematics and Computing
University of Southern Queensland
Toowoomba, Queensland, Australia 4350

dunn@usq.edu.au
