ESTIMATING THE DISTRIBUTION OF DEMAND USING BOUNDED SALES DATA

Michael R. Middleton, McLaren School of Business, University of San Francisco
2130 Fulton Street, San Francisco, CA 94117-1080 415-422-6768 middleton@usfca.edu

ABSTRACT

Flight capacity can be modified by using different seating configurations or different aircraft. An airline planning department wants to estimate the distribution of demand for a specific flight as an aid for deciding on flight capacity. Historical sales data for a specific flight are available, but the number of passenger tickets sold is bounded by the number of seats available. The planners are willing to assume that underlying demand is normally distributed. This paper describes an estimation method using normal scores, a normal probability plot, and simple linear regression. Step-by-step details using Excel 5 or 7 (95) spreadsheet software are included. Finally, the results are compared with the hazard function method of censored data analysis.

BACKGROUND

The vice president for planning of a major international airline must decide on the type of aircraft and seating configuration for each scheduled flight [5]. One of the inputs to this decision is the probability distribution of demand. The airline has historical data on the number of seats sold for each flight, but no information is kept about customer demand after all seats are sold. The planners believe that demand for each flight is normally distributed, and they need a systematic method for estimating the mean and standard deviation of demand for all flights using historical sales data.

The historical data in this problem are an example of censored data where seat capacity places an upper bound on the values. Many methods for analyzing censored data have been developed for lifetime data and survival rates, including the hazard function method by Nelson [3]. Kesling has described a variety of business applications using the hazard function method for censored data analysis [2].

This paper first describes an estimation method for unbounded data using normal scores, a normal probability plot, and simple linear regression. Then the same method is modified for the bounded sales data. Last, the hazard function method is used and compared. All methods described are supported by Excel spreadsheet software.

UNBOUNDED DATA

If there were unlimited seat capacity on a flight, then the historical sales data would be the same as the demand data. The planners could check for normality using several methods: visually look for nonnormal patterns in a histogram, count the number of observations within 1, 2, and 3 standard deviations of the mean, conduct a chi-square goodness-of-fit test, or check for a straight-line pattern on a normal probability plot [1, p.264]. Since only the latter method can be extended to the situation with bounded data, that approach is described here.

Normal Scores

Assume a random sample of n = 20 values from a normal distribution. Using a bracket median approach, we can regard the lowest sample value as an estimate from the lowest 5% = 1/n range of population values and specifically as an estimate of the median of that range, which is the .025 = 1/2n fractile [6, p. 170]. (A fractile is a value below which that proportion of a distribution's values lie.) Referring to the standard normal distribution, the .025 fractile is located 1.96 standard deviations below the mean. This ideal location, –1.96, is the normal score for the lowest sample value.

Similarly, the second lowest value in a sample of 20 is a general estimate of the next 5% range of population values and a specific estimate of the .075 = 3/2n fractile, located 1.44 standard deviations below the mean, with normal score –1.44. In general, if the n sample values have ranks i = 1,...,n, with rank i =1 for the lowest value and rank i = n for the highest, the normal score using the bracket median method is associated with the (i–0.5)/n fractile.

Figure 1 shows a normal cumulative probability curve with the location of 20 equally-likely values on the horizontal axis expressed as z-scores, or standard deviation units from the mean. If a random sample of 20 values is normally distributed, we expect a dot plot of those values to have approximately the same spacing as these z-scores. A scatter plot called a normal probability plot can be used to visually compare the spacing of actual sample values with the ideal spacing of the z-scores. If the relative locations are similar, the pattern of the scatter plot will be approximately linear.

FIGURE 1
Normal Cumulative Distribution

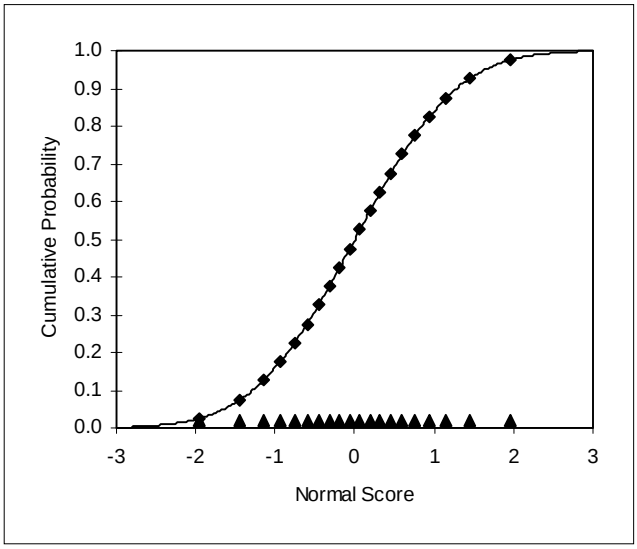


FIGURE 2
Worksheet for Unbounded Data

	A	B	C	D
1	Rank	Cumul. Prob.	Normal Score	Sorted Data
2	1	0.025	-1.960	144
3	2	0.075	-1.440	169
4	3	0.125	-1.150	174
5	4	0.175	-0.935	212
6	5	0.225	-0.755	224
7	6	0.275	-0.598	231
8	7	0.325	-0.454	235
9	8	0.375	-0.319	235
10	9	0.425	-0.189	242
11	10	0.475	-0.063	245
12	11	0.525	0.063	264
13	12	0.575	0.189	272
14	13	0.625	0.319	275
15	14	0.675	0.454	275
16	15	0.725	0.598	278
17	16	0.775	0.755	289
18	17	0.825	0.935	298
19	18	0.875	1.150	302
20	19	0.925	1.440	340
21	20	0.975	1.960	342

Normal Probability Plot

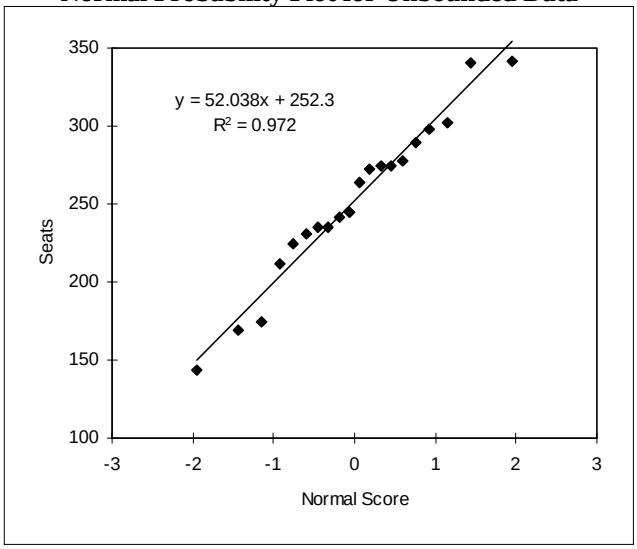
To develop a worksheet for a normal probability plot using Excel, enter the data in column D, and sort in ascending order as shown in Figure 2. (Here the data are a simulated random sample from a normal distribution with mean 250 and standard deviation 50.) Enter ranks in column A. In cell B2 enter the formula =(A2–0.5)/20, and copy to cells B3:B21. In cell C2 enter the formula =NORMSINV(B2), and copy to cells C3:C21.

To obtain the normal probability plot shown in Figure 3, select cells C2:D21 and use the Chart Wizard. (To ensure that the Chart Wizard makes appropriate assumptions, the data is arranged on the worksheet with the x-axis data in a column on the left and the y-axis data on the right.) In step 1, verify the range; in step 2, choose XY (Scatter); in step 3, choose format 1; in step 4, verify the entries; in step 5, click No for adding a legend, type "Normal Score" for the X axis title, type "Seats" for the Y axis title, and click Finish.

To format the embedded chart, double-click it to activate it for editing. Select the horizontal axis, right-click, and choose Format Axis from the shortcut menu; on the Scale tab, type –3 for Minimum, 3 for Maximum, 1 for Major Unit, and –3 for Value (Y) Axis Crosses At; on the Number tab, set Decimal Places to 0; click OK. Select the vertical axis, right-click, and choose Format Axis from the shortcut menu; on the Scale tab, type 100 for Minimum, 350 for Maximum, and 100 for Value (X) Axis Crosses At; click OK.

To insert a trendline, select the data series by clicking on one of the points. From the Insert menu, choose Trendline. Click the Type tab of the Trendline dialog box, and click the Linear icon. Click the Options tab, and click the check boxes for Display Equation on Chart and Display R-squared Value on Chart. Click OK.

FIGURE 3
Normal Probability Plot for Unbounded Data

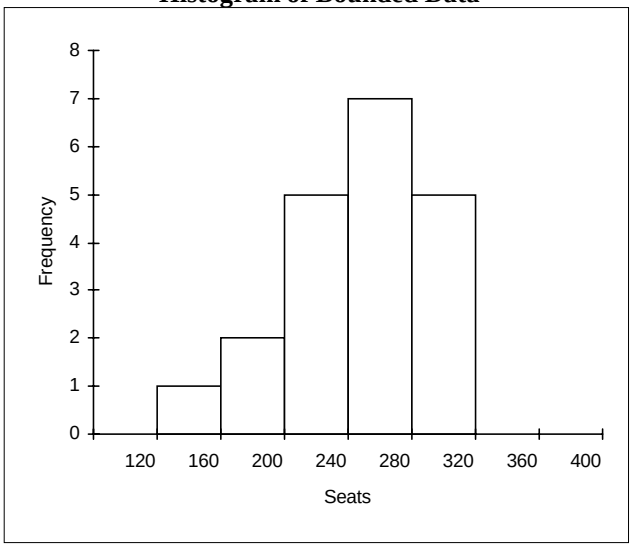


The results are shown in Figure 3. The normal probability plot allows visual verification of normality, and the linear trendline produces estimates of the mean 252.3 (the intercept, when the normal score is zero) and standard deviation 52.0 (the slope, or the change in seats for a unit change in the normal score). For unbounded data, these estimates could be obtained directly from the sample data; the advantage of the method shown in Figure 3 is that it can also be applied to the case of bounded data.

BOUNDED DATA

Figure 4 shows a histogram of simulated data for 20 trips of the same flight. In actual practice the observations should be restricted to indistinguishable, stationary situations, e.g., a Boeing 747 with 285 coach seats on non-holiday Thursday afternoon flights from Hong Kong to Tokyo [6, p. 262]. The seating capacity obscures demand exceeding 285. The sorted data are shown in column D of Figure 5.

FIGURE 4
Histogram of Bounded Data



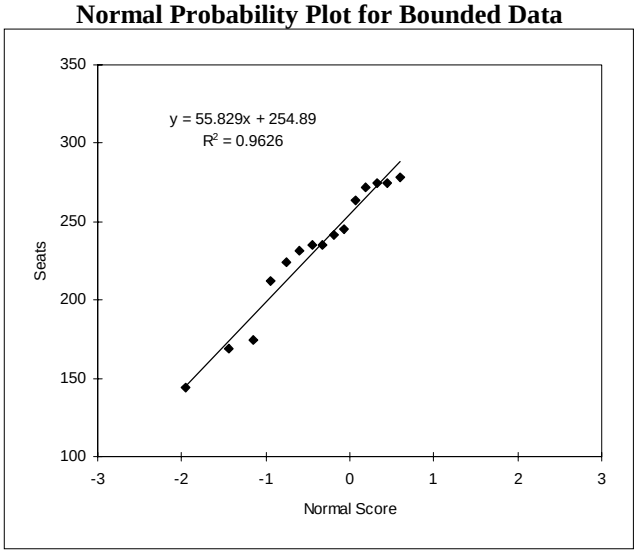
To analyze the bounded data using Excel, prepare the worksheet as previously described. The cumulative probabilities and normal scores are not needed for the censored data. The results are shown in rows 1 through 21 of Figure 5.

To obtain the normal probability plot shown in Figure 6, select cells C2:D16, use the Chart Wizard, format the chart, and insert a trendline as previously described. For the unbounded data, the trendline used all observations; for these bounded data, the trendline is based on only the unbounded observations.

FIGURE 5
Worksheet for Bounded Data

	A	B	C	D
1	Rank	Cumul. Prob.	Normal Score	Sorted Data
2	1	0.025	-1.960	144
3	2	0.075	-1.440	169
4	3	0.125	-1.150	174
5	4	0.175	-0.935	212
6	5	0.225	-0.755	224
7	6	0.275	-0.598	231
8	7	0.325	-0.454	235
9	8	0.375	-0.319	235
10	9	0.425	-0.189	242
11	10	0.475	-0.063	245
12	11	0.525	0.063	264
13	12	0.575	0.189	272
14	13	0.625	0.319	275
15	14	0.675	0.454	275
16	15	0.725	0.598	278
17	16			285
18	17			285
19	18			285
20	19			285
21	20			285
22				
23		Mean	Intercept	254.89
24		StDev	Slope	55.83

FIGURE 6
Normal Probability Plot for Bounded Data



To obtain the regression results on the worksheet, enter the labels in cells B23:C24 as shown in Figure 5. In cell D23, enter the formula =INTERCEPT(D2:D16,C2:C16); in cell D24, enter the formula =SLOPE(D2:D16,C2:C16).

The results may be interpreted like those for unbounded data. The estimate of normally distributed seat demand is Seats = 254.89 + 55.83*Z, where Z is the normal score. The mean demand, 254.89, corresponds to Z = 0; the standard deviation of demand, 55.83, corresponds to the change in demand for a unit change in Z.

HAZARD FUNCTION

The hazard function method uses a different technique for determining the cumulative probabilities and a different orientation for the normal probability plot. Figure 7 arranges the calculations described by Kesling [2].

FIGURE 7
Worksheet for Hazard Function

	A	B	C	D	E	F
1	Sorted Data	Rank	Hazard	Cumul. Hazard	Cumul. Prob.	Normal Score
2	144	1	0.0500	0.0500	0.0488	-1.657
3	169	2	0.0526	0.1026	0.0975	-1.296
4	174	3	0.0556	0.1582	0.1463	-1.052
5	212	4	0.0588	0.2170	0.1951	-0.859
6	224	5	0.0625	0.2795	0.2438	-0.694
7	231	6	0.0667	0.3462	0.2926	-0.546
8	235	7	0.0714	0.4176	0.3414	-0.409
9	235	8	0.0769	0.4945	0.3901	-0.279
10	242	9	0.0833	0.5779	0.4389	-0.154
11	245	10	0.0909	0.6688	0.4877	-0.031
12	264	11	0.1000	0.7688	0.5364	0.091
13	272	12	0.1111	0.8799	0.5852	0.215
14	275	13	0.1250	1.0049	0.6339	0.342
15	275	14	0.1429	1.1477	0.6826	0.475
16	278	15	0.1667	1.3144	0.7314	0.617
17	285	16	0			
18	285	17	0			
19	285	18	0			
20	285	19	0			
21	285	20	0			
22						
23	Intercept	-3.9872		Mean	253.89	
24	Slope	0.0157		StDev	63.68	

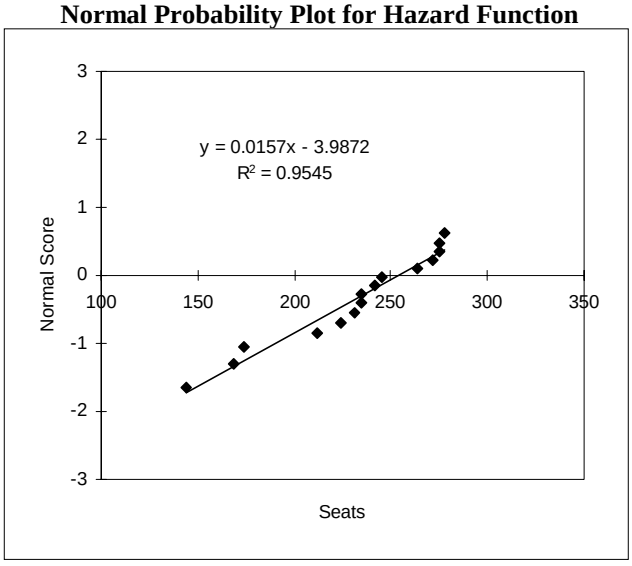
To develop the worksheet for the hazard function method, enter the sorted data and ranks in columns A and B. For each observed value, the hazard function in column C equals the reciprocal of the number of data values greater than or equal to the observed value; for suspended values, the hazard function is zero. In cell C2, enter the formula =1/(21–B2), and copy to cells C3:C16. To obtain cumulative hazard, in cell D2 enter the formula =C2, in cell D3 enter the formula =D2+C3, and copy the formula from D3 to cells D4:D16. To obtain cumulative probability, in cell E2 enter the formula =1–EXP(–D2), and copy to cells E3:E16. To obtain normal scores, in cell F2 enter the formula =NORMSINV(E2), and copy to cells F3:F16.

To obtain the regression results, enter the labels in rows 23 and 24 as shown in Figure 7. In cell B23, enter the formula =INTERCEPT(F2:F16,A2:A16); in cell B24, enter the formula =SLOPE(F2:F16,A2:A16). These are coefficients for linear regression of fitted normal score depending on seats sold.

The estimated equation is Z = –3.9872 + 0.157*Seats. The mean of normally distributed demand corresponds to a normal score of zero, so the mean of Seats equals the negative of the intercept divided by the slope. The slope in the equation is change in Z per seat, so the standard deviation of demand equals the reciprocal of the slope. To obtain these values, in cell E23 enter the formula =–B23/B24, and in cell E24 enter the formula =1/B24. The hazard function method produces a mean of 254 and a standard deviation of 64 for normally distributed demand.

To construct a normal probability plot, first select A2:A16, hold down the Control key while selecting F2:F16, and use the Chart Wizard. After inserting a trendline and formatting, the results are shown in Figure 8.

FIGURE 8
Normal Probability Plot for Hazard Function



DISCUSSION

The two methods produce nearly identical estimates of the mean: 254 from the hazard function method and 255 from the direct method proposed here. The estimates of the standard deviation show a larger difference: 64 vs. 56. Even though the hazard function method uses less extreme normal scores, the estimate of standard deviation is larger.

Future research could investigate the reasons for these different results. One factor is the different orientation of the normal probability plots. The direct method plots seats vs. normal score, so the regression line minimizes sum of squared deviations in terms of seats. Conversely, the hazard function method plots normal score vs. seats, so the deviations that are minimized are fitted normal score minus actual normal score. The rationale for choosing between these two orientations should be explored.

Another factor is the different methods for determining cumulative probabilities, which affect the normal scores used in the regression. The bracket median method is relatively easy to explain; the rationale for the hazard function method is more obscure. These methods and two others are shown in Figure 9. The method labeled Cryer uses i/(n+1) to determine the cumulative probability for rank i [1, p. 266]. The method labeled Neter uses (i–3/8)/(n+1/4), which is described as yielding a good approximation based on statistical theory [4, p. 107]. The Neter method produces normal scores almost as extreme as the bracket median method. The hazard method with 5 censored values in a sample of 20 generates the least extreme normal scores of the four methods.

FIGURE 9
Cumulative Probabilities and Normal Scores

	Bracket Median		Hazard (15 of 20)		Cryer		Neter	
Rank	Prob.	Score	Prob.	Score	Prob.	Score	Prob.	Score
1	0.025	-1.960	0.049	-1.657	0.048	-1.668	0.031	-1.868
2	0.075	-1.440	0.098	-1.296	0.095	-1.309	0.080	-1.403
3	0.125	-1.150	0.146	-1.052	0.143	-1.068	0.130	-1.128
4	0.175	-0.935	0.195	-0.859	0.190	-0.876	0.179	-0.919
5	0.225	-0.755	0.244	-0.694	0.238	-0.712	0.228	-0.744
6	0.275	-0.598	0.293	-0.546	0.286	-0.566	0.278	-0.589
7	0.325	-0.454	0.341	-0.409	0.333	-0.431	0.327	-0.448
8	0.375	-0.319	0.390	-0.279	0.381	-0.303	0.377	-0.315
9	0.425	-0.189	0.439	-0.154	0.429	-0.180	0.426	-0.187
10	0.475	-0.063	0.488	-0.031	0.476	-0.060	0.475	-0.062
11	0.525	0.063	0.536	0.091	0.524	0.060	0.525	0.062
12	0.575	0.189	0.585	0.215	0.571	0.180	0.574	0.187
13	0.625	0.319	0.634	0.342	0.619	0.303	0.623	0.315
14	0.675	0.454	0.683	0.475	0.667	0.431	0.673	0.448
15	0.725	0.598	0.731	0.617	0.714	0.566	0.722	0.589
16	0.775	0.755			0.762	0.712	0.772	0.744
17	0.825	0.935			0.810	0.876	0.821	0.919
18	0.875	1.150			0.857	1.068	0.870	1.128
19	0.925	1.440			0.905	1.309	0.920	1.403
20	0.975	1.960			0.952	1.668	0.969	1.868

The direct method (using bracket median, Cryer, or Neter cumulative probabilities) and the hazard function method can be applied to data with missing values at the low end, interspersed throughout, or at the high end. These methods can also be applied to nonnormal distributions for which the inverse cumulative function is available. In addition to the standard normal inverse and normal inverse, Excel has inverse functions for the beta, chi-square, exponential, F, gamma, lognormal, and t distributions. Since the computations are easy to perform with spreadsheet software, it is likely that these methods will become more widely used in business decision analysis.

REFERENCES

[1] Cryer, J.D., and Miller, R.B. Statistics for business: Data analysis and modeling, 2nd ed. Belmont, CA: Duxbury, 1994.

[2] Kesling, G.D. Censored data analysis in business using the hazard function method. Proceedings of Western Decision Sciences Institute, 1994, 685.

[3] Nelson, W. Theory and applications of hazard plotting for censored failure data. Technometrics, 1972, 14(4), 945-966.

[4] Neter, J., Kutner, M.H., Nachtsheim, C.J., and Wasserman, W. Applied linear statistical models, 4th ed. Chicago, IL: Irwin, 1996.

[5] Slosar, J. Personal communication, 1993.

[6] Vatter, P.A., Bradley, S.P., Frey, S.C., and Jackson, B.B. Quantitative methods in management: Text and cases. Homewood, IL: Irwin, 1978.

This paper was presented at the annual meeting of the Western Decision Sciences Institute, March 25-29, 1997, and published in the conference *Proceedings*, pp. 658-662.