Exploring Power Transformations

1. What is a power transformation?

Statisticians using linear models for analysis of experimental data have long been transforming response variables for such varied purposes as stabilizing variance, restoring normality, and removing nonadditivity. Empirical studies may indicate that a transformation might improve the analysis of a given set of data. Basic texts on statistical methods provide many such examples.

A more formal method, which allows using statistical procedures for determining an appropriate transformation was proposed by Box and Cox. They proposed a *parametric family of power transformations*:

$$x^{(\lambda_1,\lambda_2)} = \begin{cases} \frac{(x+\lambda_2)^{\lambda_1}-1}{\lambda_1} & \lambda_1 \neq 0 \\ \log\,(x+\lambda_2) & \lambda_1 = 0, \quad x+\lambda_2 > 0 \end{cases}$$

where λ_1 is the *power* parameter, and λ_2 is a shift parameter that defines a transformation of the variable x. The purpose of the transformation is to determine values of λ_1 and λ_2 such that $E[X^{(\lambda_1,\lambda_2)}]$ will fit the model better and that $X^{(\lambda_1,\lambda_2)}$ will satisfy the assumptions of constancy of variance, independence and additivity.

2. Objectives

- The power transformation is effective in making skewed distributions more symmetric, and hopefully more normal.
- Graphical methods can be used to roughly gauge the appropriate transformations to normality (values of the shift and power parameters) for samples from particular distributions, and that samples from some distributions, e.g. Cauchy, cannot be transformed to normality.
- Graphical methods can also be used to examine the effect of power transformations on actual data sets.

3. Startup Instructions

On a Vincent workstation

```
% add lisp
% add stat
% tr_module
```

On a PC

- Click on the Lisp-Stat icon in the program manager window
- Click on the `tr.lsp` icon in the Lisp-Stat window

On a Macintosh

- Start up xlispstat, by clicking on the XLispStat icon
- Pull down the `File` menu and select `Load`
- Select the folder `Teach`
- Select `tr.lsp`

4. The module interface

The TR module has four windows (transformation controls, probability plot, box plot, and histogram) displayed on start up. The plot windows are empty until the student selects a parent distribution (`Distributions`) or a data set (`Select Data`). In Figure 1 the parent distribution is a Chi-square with 1 degree of freedom and the sample size is 20. The plot in the control window shows the density function of the parent distribution or a description of the data (if a data set is selected. Clicking on the `power` slide-bar changes the power parameter (λ_1), and likewise the `shift` slide-bar controls the shift parameter (λ_2) in the transformation

$$x^{(\lambda_1,\lambda_2)} = \begin{cases} \frac{(x+\lambda_2)^{\lambda_1}-1}{\lambda_1} & \lambda_1 \neq 0 \\ \log\,(x+\lambda_2) & \lambda_1 = 0, \quad x+\lambda_2 > 0 \end{cases}$$

For the sample shown in Figure 1 a power of 0.3 and shift close to 0 appear reasonable: the points in the probability plot lie approximately on a straight line, the boxplot is close to symmetric and the histogram is reasonably bell-shaped.

5. Warm-ups

To gain some familiarity with the `tr_module`, try the following:

- Use the `distribution` button to explore the shapes of the different `parent distributions`.
- For each parent distribution, notice the effect that changing the parameter values [done by moving the appropriate slide bar(s)] has on distribution shape.
- Choose a distribution from the list (e.g. a chi-square with 1 degree of freedom) from which to sample. Choose a sample size (say, 15) using the `Sample Size:` slider. Then peruse the three plot (probability plot, box plot, and histogram) windows. Does this sample appear to you to be one from the parent distribution you selected ?
- Now click on the `power` slide-bar to change the power parameter (λ_1). Keep doing this until the resulting plots appear to indicate a more symmetric distribution (for e.g., the normal probability plot appears to be linear).
- In the histogram window experiment with the number of histogram cells and with the `GaussKerDen` smoothing constant value in the histogram window to see if you can find a most visually appealing smoothed density estimate.
- Click the `Show Statistics` button on in the box plot window. The values shown will disappear if you click elsewhere on the window.

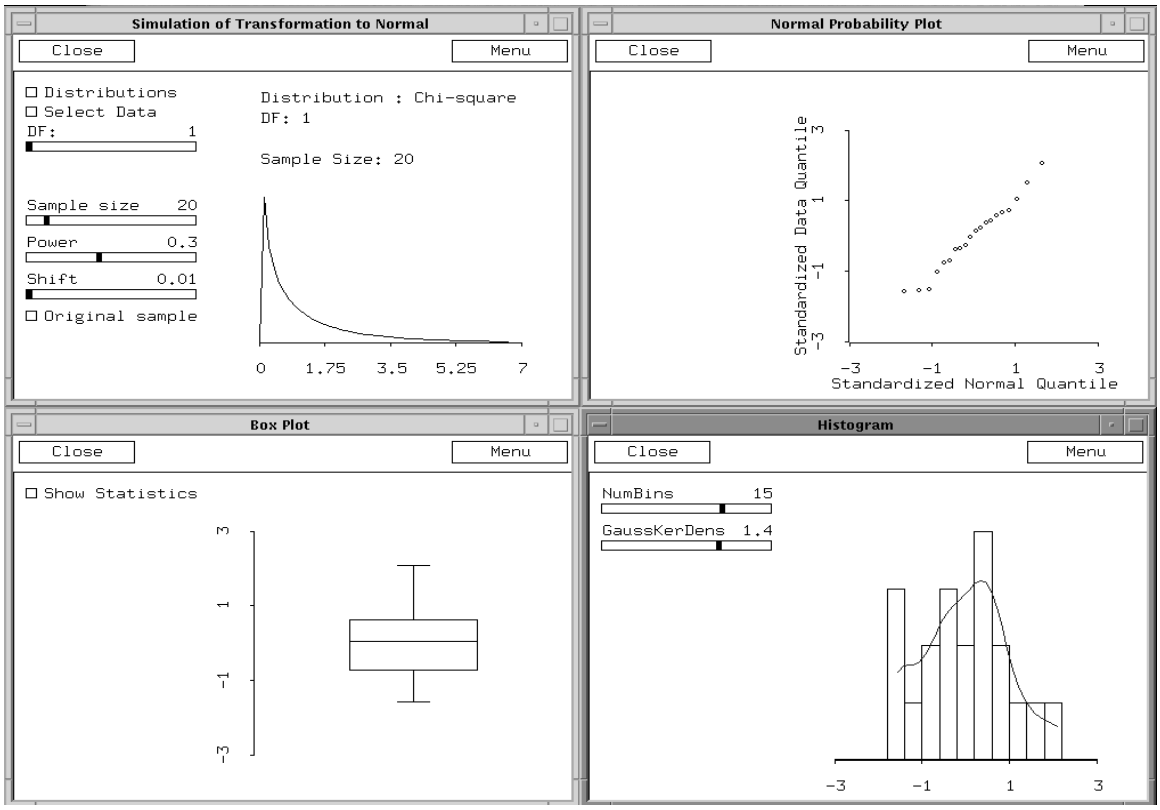

Figure 1: Frames of the Transformations Module Windows

- Go back to the first window and click on the `Original sample` button. This would restore all the plot windows to their original status (i.e.,prior to the transformation).
- Click on the `Select data` button to select a real data set and peruse the three plot (probability plot, box plot, and histogram) windows.

6. Exercises

In the following exercises, choose the indicated distribution or dataset as your first step. Then, for a selected sample size, use the `power` slide-bar to change the power parameter λ_1, and the `shift` slide-bar to change the shift parameter λ_2, if necessary. Study the 3 plot windows thus obtained to answer the questions. You may want to repeat each experiment more than once to check the consistency of your conclusions.

1. Select a normal distribution with a mean $\mu = 5$ and $\sigma = 4$. Use the `Sample size` slide-bar to change sample sizes from small to large. What shape does the Normal probability plot take? Does the shape seem to depend on the sample size? Use the `power` slide-bar to change the power parameter λ_1.

2. Select an exponential distribution with a mean $\theta = 5$. Use the `Sample size` slide-bar to change sample size to 20. What shape does the Normal probability plot take? Use

the `power` slide-bar to change the power parameter λ_1, until the shape of the normal probability plot appears reasonably close to a straight line. What value of λ_1 seems satisfactory ?

3. Click on the `Select Data` button and select the CO data set. Use the `power` slide-bar to change power parameter so that the Normal probability plot appears to be closest to a straight line.

4. Click on the `Select Data` button and select the HC data set. Use the `power` slide-bar to change power parameter so that the Normal probability plot appears to be closest to a straight line.

5. Select a normal distribution with a mean $\mu = 0$ and $\sigma = 1$. Use the `Sample size` slide-bar to select sample size 30. Use the `power` slide-bar to change power to 2. What is the shape of the distribution of the transformed data ? Does your observation agree with statistical theory? What is the shape of the Normal probability plot ?

7. Solutions to Exercises

1. The normal probability plot of the pseud-random variables generated from normal distributions appear to be approximately linear even for small sample sizes. However, the frequency of deviation from a straight line seems smaller for larger sample sizes.

2. The normal probability plot is clearly not linear for the sample generated from the exponential distribution. The shape appears closest to being approximately linear for values of the power parameter near .3.

3. The normal probability plot is clearly not linear for the original sample. It appears to straighten out around a value of .5 for λ_1, so that a square root transformation is suggested in this case.

4. Again the normal probability plot shows a distinct curvature. The shape appears closest to being approximately linear for values of the power parameter near .3 and therfore a cuberoot transformation may be appropriate.

5. Theoretically, the data from a standard normal distribution will be transformed to chi-squared variables with 1 d.f. when they are squared. Thus the normal probability plot will be curved inwards at the lower end showing skewness in the resulting data.