

Guidelines for Teaching and Learning Statistics using the Statistics Online Computational Resources: Taking Computer-Based Learning to the Next Step

Juana Sanchez, Nicolas Christou & Ivo D. Dinov
UCLA Department of Statistics, Los Angeles, CA 90095
www.SOCR.ucla.edu

USCOTS 2007 Breakout Session

Summary

We will introduce the Statistics Online Computational Resource (SOCR), describe the pedagogical utilization of the SOCR tools, conduct interactive activities to demonstrate their in-class use and introduce our collaborative working environment for development and exchange of tools and instructional materials. Several hands-on activities will be presented and interactively tested. The *URL page for the SOCR session* is:

http://wiki.stat.ucla.edu/socr/index.php/SOCR_Events_May2007

SOCR Activities

In this session, participants will be directly involved in some of the activities from the Statistics Online Computational Resources ([SOCR](#)), in the same way these are introduced in our Introductory Statistics and Probability classes. The three SOCR activities we plan to present are:

- **Distributions Activity:** The [Distribution Activity](#) shows how to interact with the SOCR Distributions to visualize the areas of interest and obtain probabilities and critical values for over 50+ distinct distributions.
- **Central Limit Theorem Activity:** [The Central Limit Theorem \(CLT\) Activity](#) illustrates the properties of the sampling distribution of the sample average and serves to motivate and build students' intuition.
- **Power Transform Activity:** The [Power Transform Activity](#) demonstrates the usage, effects and properties of the modified power transformation family applied to real or simulated data to reduce variation and enhance Normality. There are 4 exercises in this activity, each demonstrating the properties of the power transform in different settings for observed or simulated data: X-Y scatter plot, QQ-Normal plot, Histogram plot and Time/Index plot.

The complete activities are attached at the end and are also available online at:

http://wiki.stat.ucla.edu/socr/index.php/SOCR_Events_May2007

Participants will also learn about the available [SOCR](#) tools, how to use them and how to contribute to these activities using the [SOCR Wiki collaborative environment](#)

- **Other Interesting Activities:** If time permits, the following additional activities may be discussed:
 - [Random Number Generation \(RNG\)](#)
 - [Confidence Interval \(CI\) Activity](#)
 - [Law of Large Numbers \(LLN\)](#)
 - [Fourier Game](#)

Background

All of the [SOCR](#) resources are dynamic and are readily adaptable as the activities for the specific classrooms. The [SOCR](#) goals are to provide excellent science, technology, engineering, and mathematics (STEM) education for all undergraduate students and to involve instructors of diverse backgrounds in exchange of ideas and joint development of educational materials. We design and extend the [SOCR](#) framework so that it allows instructors to custom-design activities and tools that fit their specific course, student population and topics covered.

The efforts of the SOCR resource are to develop the foundation of tools and instances of instructional activities. These include open-source Java applets, computational probability and statistics Java library, class notes and collaborative activities. This infrastructure enables educators to go to the next level in information technology based instruction.

- **SOCR Interface:** The main objective of SOCR is to offer a homogeneous interface for online activities appropriate for the Introductory Statistics Course, Introductory Probability course and other advanced Statistics courses that rely on hands on demonstrations and simulation to illustrate Statistical concepts. A common portal for all SOCR activities is very important to minimize the amount of time that students have to spend learning the technology. SOCR materials and activities have received recognitions from several [educational technology-based initiatives and digital libraries](#).
- **SOCR Studies and Findings:** [SOCR](#) has been tested in the classroom on several occasions. Most recently an experimental study we conducted led us to conclude that using SOCR for the teaching of Introductory Statistics and Probability was effective ([Dinov, Sanchez and Christou, 2006](#)). More testing of the effectiveness of SOCR in undergraduate teaching is currently underway at UCLA during the 2006-2007 academic year. The dependence of students' performance in the SOCR-based courses will be studied as a function of their attitudes towards the subject, their learning styles and other student demographics. In this session, we propose to provide hands-on experience on how to use the SOCR resources in the best possible way to achieve the best learning outcomes for different groups of students.
- **Browsing the SOCR Resources:** [Interactive Hyperbolic Utility for Searching, Browsing and Using the SOCR Resources](#) (http://www.socr.ucla.edu/SOCR_HT_ResourceViewer.html).

APPENDIX

- [How to Contribute a New SOCR Activity?](#) (Wiki Editing Tutorial)
- [SOCR/USCOTS07 Community Portal](#)
- [SOCR Community Portal](#)
- [SOCR Location](#)
- [User Geo-Map](#)
- [SOCR references](#)
- SOCR Home page: <http://www.SOCR.ucla.edu>
- [Interactive Hyperbolic Utility for Searching, Browsing and Using the SOCR Resources](#)
- To translate this SOCR USCOTS Handbook, select a language from this list →



References (http://www.socr.ucla.edu/htmls/SOCR_References.html):

Dinov, ID, Sanchez, J. and Christou, N. *Pedagogical Utilization and Assessment of the Statistic Online Computational Resource in Introductory Probability and Statistics Courses*, in press, *Journal of Computers & Education*, 2006, <http://dx.doi.org/10.1016/j.compedu.2006.06.003>.

Dinov, ID. *SOCR: Statistics Online Computational Resource: [socr.ucla.edu](http://www.socr.ucla.edu)*, *Statistical Computing & Graphics*. Vol. 17, No. 1, 11-15, 2006.

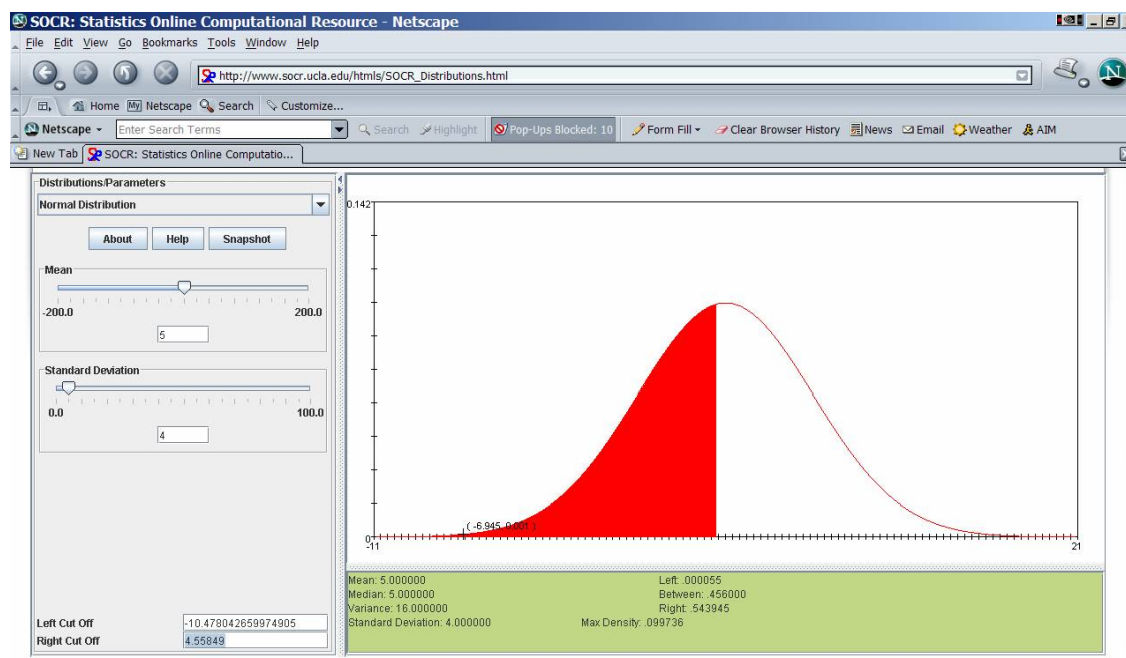
Leslie, M. *NetWatch EDUCATION: Statistics Starter Kit*, *Science Magazine*, Volume 302, Number 5651, Issue of 5 December 2003.

SOCR Distributions Activity

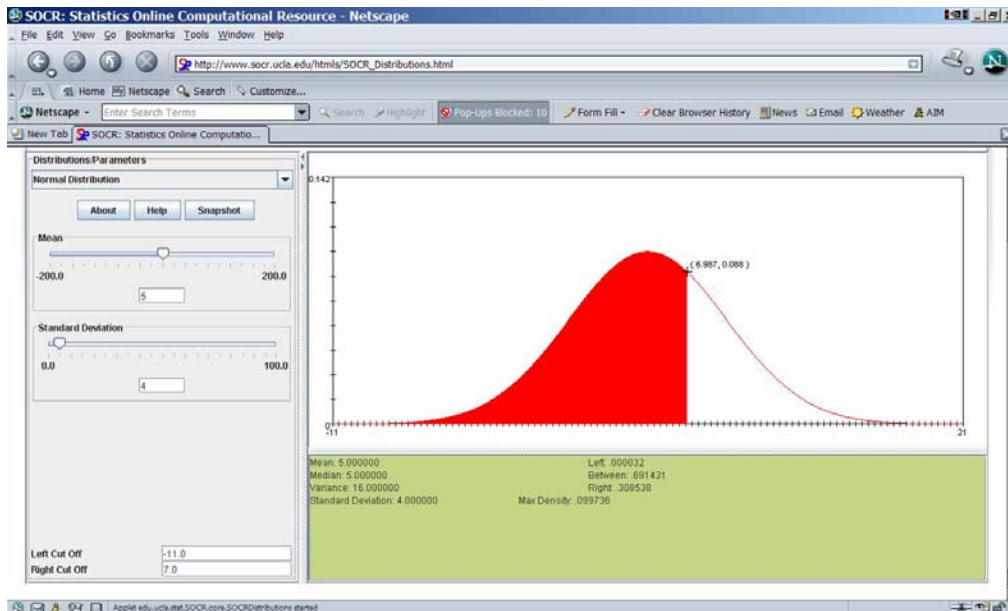
Goal: To compute probabilities or critical values for different distributions using SOCR tools and observe the shape of the distributions by varying their parameters and to explore relations among various distributions (e.g. Poisson-binomial, binomial-hypergeometric, normal-binomial, normal-Poisson).

Example: Here we use Normal distribution (mean=5, sigma=4) as an example

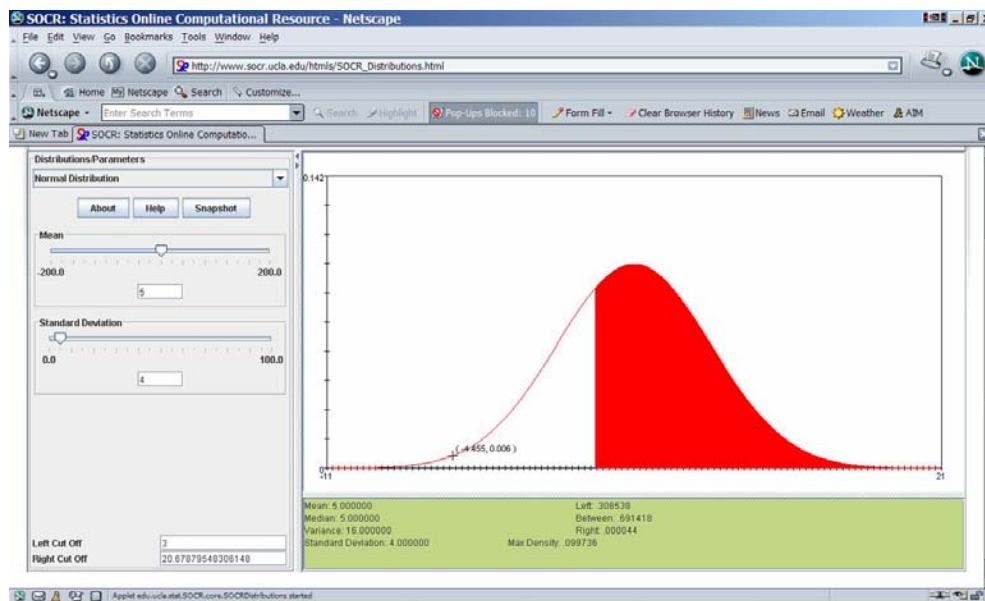
- Step 1: Set up parameters! Select the Normal Distribution from the drop-down list on the top-left. Set Mean = 5 & Standard Deviation = 4. Now the density function of $Normal(\mu=5, \sigma=4)$ will show up in the top-right with red color. See following figure.



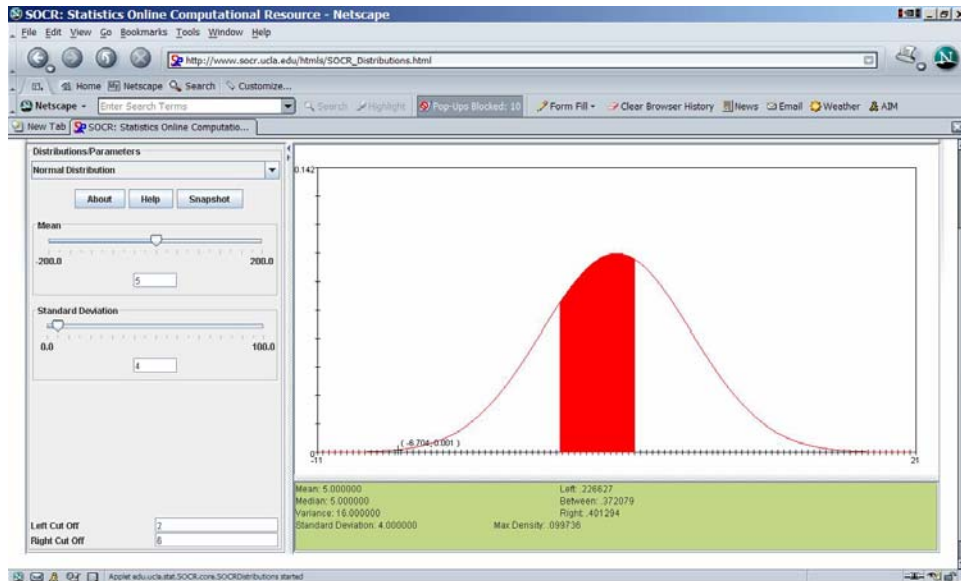
- Step 2: Compute $P(X < 7)$! Move cursor from 21, the most right side, to the right until 7. Now the red area means $X < 7$. And the probability of red area is always represented by the value of "Between". Hence $P(X < 7) = 0.691431$. See figure below. The same result can be obtained by using the *Left Cut-Off* and *Right Cut-Off* boxes on the bottom-left corner of the applet.



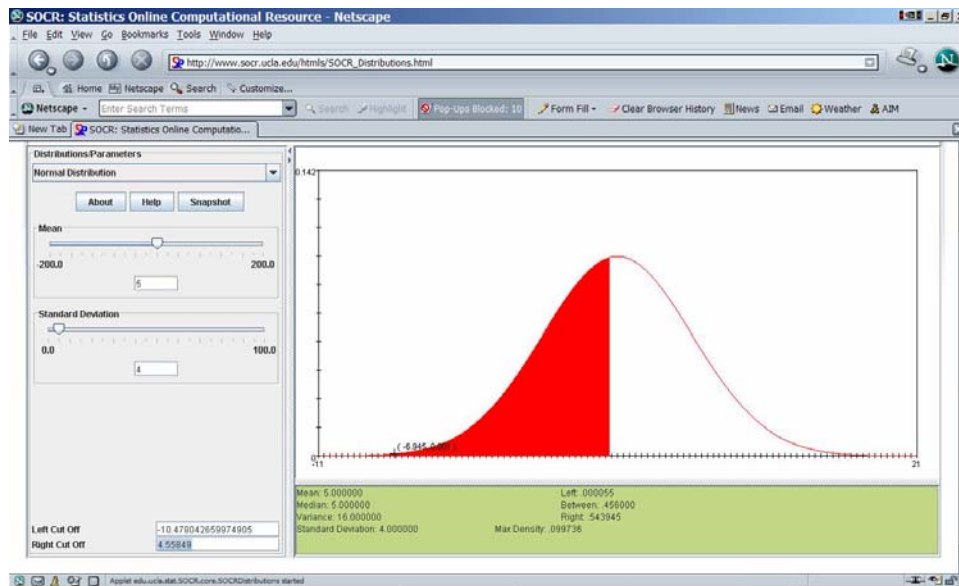
- To get a snapshot of this result, click on Snapshot. Then save the file with a filename and extension .jpg , for example, HW1_Image2.jpg. When done, you can open that new .jpeg file and go to edit→copy. Open separately a Word file and paste into that file. You can reduce the image, but make sure that it is big enough for us to see the numbers.
- Step 3: Compute $P(X>3)$! Move cursor from -11, the most right side, to the left until 3. Red area means $X>3$. And the probability of red area is always recorded in the value of “Between”. Therefore, $P(X>3)=0.691431$. See figure below. Is the result the same as in Step 2, by chance? Explain!



- Step 4: Compute $P(2<X<6)$! First, move cursor from -11, the most left side, to the right until 2. Then move cursor from 21, the most right side, to the left until 6. Find “Between” in the bottom-right widow, which is just the probability of red area. Then $P(2<X<6)=0.372079$. See attached figure.



- Similarly, we can use SOCR to compute the probability in other known distribution, such as uniform distribution, exponential distribution, beta distribution, etc.
- Step 5: Now find the critical value \mathbf{d} , so that $P(X < \mathbf{d}) = 0.456$; \mathbf{d} also represents the 45.6th percentile of the $N(\mu=5, \sigma=4)$ distribution! Move the left vertical limit on the graph to the left (close to -11). Start the right limit from 11 and move it down until you reach 4.55849, monitor the “between” red area and its probability value (0.456). You may also experiment with setting the limits numerically using the Left and Right Cut-Off text-fields on the bottom-left of the applet. These allow more accurate vertical limit setting.

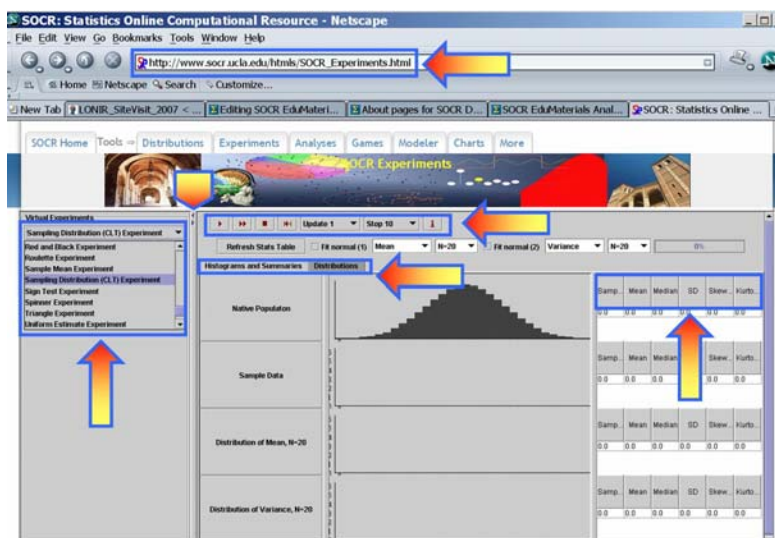


- **Questions:**
 - What is the area bound below the Normal density curve on the interval $[\mu - 2\sigma, \mu + 2\sigma]$? Does it depend on the values of μ & σ ?
 - What is the critical value t_o , that gives $P(t > t_o) = 0.025$, when $t \sim T_{df=15}$?
 - Which one is larger t_o or t_1 , where $P(T_o > t_o) = P(T_1 > t_1) = 0.025$, and $T_o \sim T_{df=15}$, $T_1 \sim T_{df=5}$? What is the limit of t_o as $df \rightarrow \infty$?

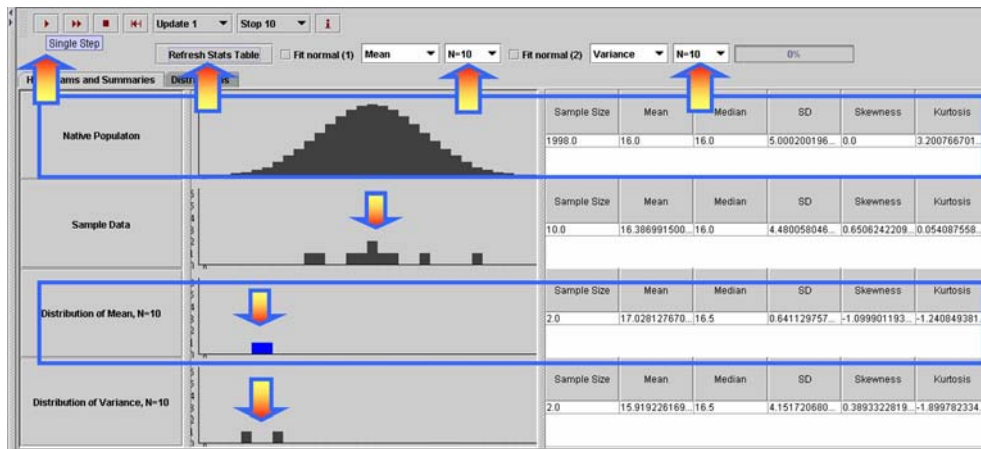
SOCR General Central Limit Theorem (CLT) Activity

This activity represents a very general demonstration of the effects of the [Central Limit Theorem \(CLT\)](#). The activity is based on the [SOCR Sampling Distribution CLT Experiment](#). This experiment builds upon a [RVLS CLT applet](#) by extending the applet functionality and providing the capability of sampling from any [SOCR Distribution](#).

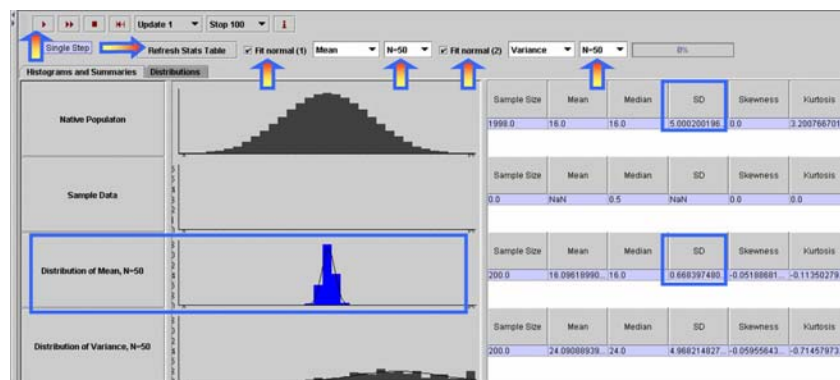
- **Goals:** The aims of this activity are to
 - provide intuitive notion of sampling from any process with a well-defined distribution
 - motivate and facilitate learning of the [central limit theorem](#)
 - empirically validate that sample-averages of random observations (most processes) follow approximately [normal distribution](#)
 - empirically demonstrate that the *sample-average* is special and other [sample statistics](#) (e.g., median, variance, range, etc.) generally do not have distributions that are normal
 - illustrate that the expectation of the sample-average equals the population mean (and the sample-average is typically a good measure of centrality for a population/process)
 - show that the variation of the sampling distribution of the mean rapidly decreases as the sample size increases ($\frac{1}{\sqrt{n}}$).
 - reinforce the concepts of a native distribution, sample, sample distribution, sampling distribution, parameter estimator and data-driven numerical parameter estimate.
- The **SOCR CLT Experiment:** To start this Experiment, go to [SOCR Experiments](#) and select the SOCR Sampling Distribution CLT Experiment from the drop-down list of experiments in the left panel. The image below shows the interface to this experiment. Notice the main control widgets on this image (boxed in blue and pointed to by arrows). The generic control buttons on the top allow you to do one or multiple steps/runs, stop and reset this experiment. The two tabs in the main frame provide graphical access to the results of the experiment (Histograms and Summaries) or the Distribution selection panel (Distributions). Remember that choosing sample-sizes ≤ 16 will animate the samples (second graphing row), whereas larger sample-sizes ($N > 20$) will only show the updates of the sampling distributions (bottom two graphing rows).



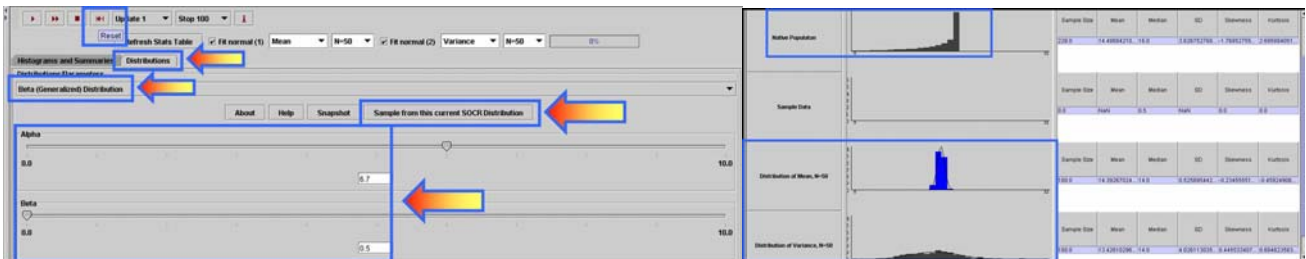
- Experiment 1:** Expand your Experiment panel (right panel) by clicking/dragging the vertical split-pane bar. Choose the two sample sizes for the two statistics to be 10. Press the **step**-button a few of times (2-5) to see the experiment run several times. Notice how data is being sampled from the native population (the distribution of the process on the top). For each step, the process of sampling 2 samples of 10 observations will generate 2 sample statistics of the 2 population parameters of interest (these are defaulted to *mean* and *variance*). At each step, you can see the plots of all sample values, as well as the computed sample statistics for each parameter. The sample values are shown on the second row graph, below the distribution of the process, and the two sample statistics are plotted on the bottom two rows. If we run this experiment many times, the bottom two graphs become the histograms of the corresponding sample statistics. If we did this infinitely many times these two graphs become the sampling distributions of the chosen sample statistics (as the observations/measurements are independent within each sample and between samples). Finally, press the **Refresh Stats Table** button on the top to see the sample summary statistics for the native population distribution (row 1), last sample (row 2) and the two sampling distributions, in this case mean and variance (rows 3 and 4).



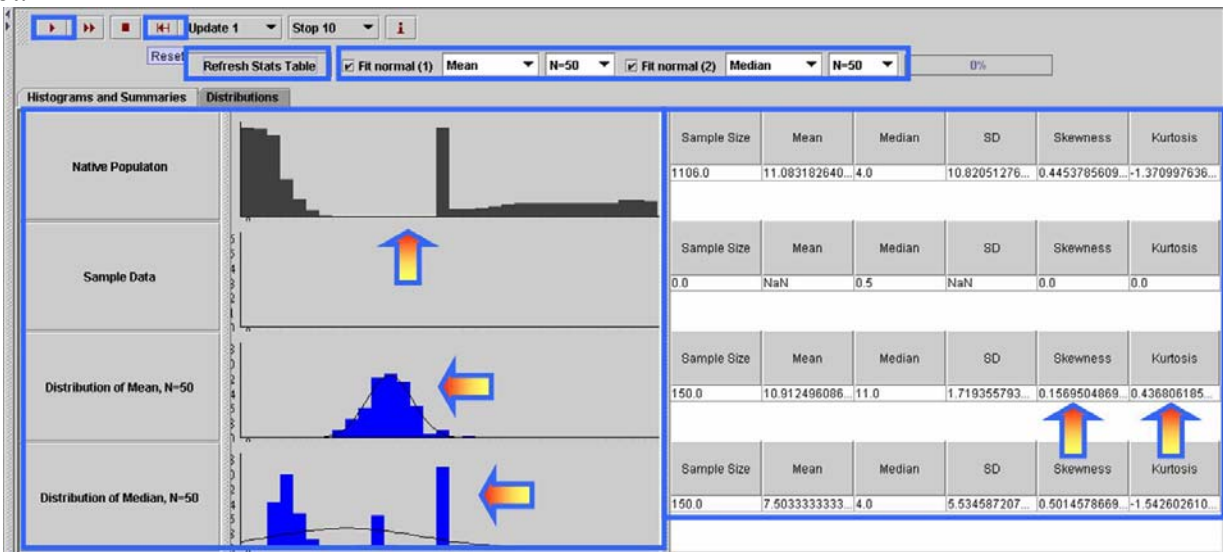
- Experiment 2:** For this experiment we'll look at the mean, standard deviation, skewness and kurtosis of the sample-average and the sample-variance (these two sample data-driven statistical estimates). Choose sample-sizes of 50, for both estimates (mean and variance). Select the **Fit Normal Curve** check-boxes for both sample distributions. **Step** through the experiment a few times (by clicking the Run button) and then click **Refresh Stats Table** button on the top to see the sample summary statistics. Try to understand and relate these sample-distribution statistics to their analogues from the native population (on the top row). For example, the mean of the multiple sample-averages is about the same as the mean of the native population, but the standard deviation of the sampling distribution of the average is about $\frac{\sigma}{\sqrt{n}}$, where σ is the standard deviation of the original native process/distribution.



- Experiment 3:** Now let's select any of the [SOCR Distributions](#), sample from it repeatedly and see if the central limit theorem is valid for the process we have selected. Try Normal, Poisson, Beta, Gamma, Cauchy and other continuous or discrete distributions. Are our empirical results in agreement with the CLT? Go to the **Distributions** tab on the top of the graphing panel. Reset the experiments panel (button on the top). Select a distribution from the drop-down list of distributions in this list. Choose appropriate parameters for your distribution, if any, and click the **Sample from this Current Distribution** button to send this distribution to the graphing panel in the **Histograms and Summaries** tab. Go to this panel and again run the experiment several times. Notice how we now sample from a Non-Normal Distribution for the first time. In this case we had chosen the Beta distribution ($\alpha = 6.7, \beta = 0.5$).



- Experiment 4:** Suppose the distribution we want to sample from is not included in the list of [SOCR Distributions](#), under the **Distributions** tab. We can then draw a shape for a hypothetical distribution by clicking and dragging the mouse in the top graphing canvas (Histograms and Summaries tab panel). This way you can construct contiguous and discontinuous, symmetric and asymmetric, unimodal and multi-modal, leptokurtic and mesokurtic and other [types of distributions](#). In the figure below, we had demonstrated this functionality to study differences between two data-driven estimates for the population center - sample [mean](#) and sample [median](#). Look how the sampling distribution of the sample-average is very close to Normal, whereas the sampling distribution of the sample median is not.



- Questions:**
 - What are the effects of asymmetry, gaps and continuity of the native distribution on the CLT?
 - When can we reasonably expect statistics, other than the sample mean, to have CLT properties?
 - If a native process has $\sigma_X=10$ and we take a sample of size 10, what will be $\sigma_{\bar{X}}$? Does it depend on the shape of the original process? How large should the sample-size be so that

$$\sigma_{\bar{X}} = \frac{2}{3} \sigma_X ?$$

SOCR Power Transformation Family Graphing Activity

Summary

This activity demonstrates the usage, effects and properties of the modified power transformation family applied to real or simulated data to reduce variation and enhance Normality. There are 4 exercises each demonstrating the properties of the power transform in different settings for observed or simulated data: X-Y scatter plot, QQ-Normal plot, Histogram plot and Time/Index plot.

Background

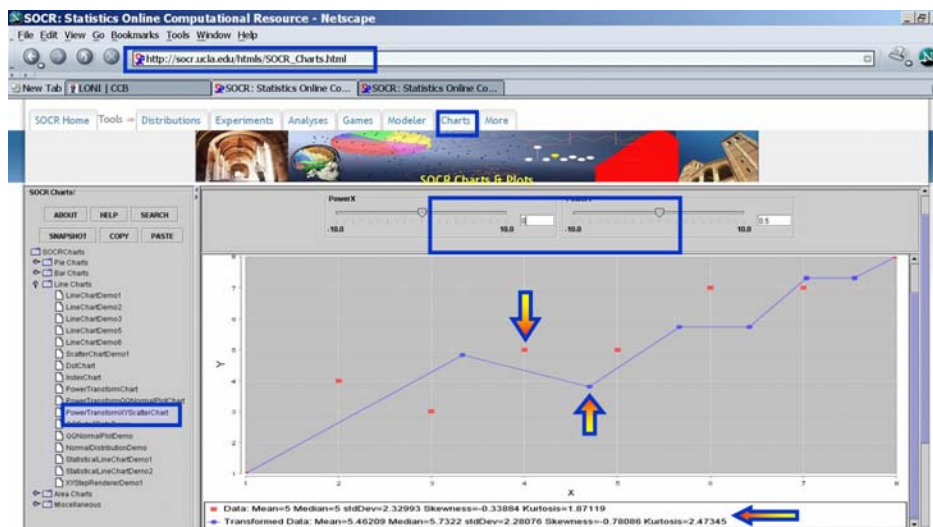
The **power transformation family** is often used for transforming data for the purpose of making it more Normal-like. The power transformation is continuously varying with respect to the power parameter λ and defined, as continuous piece-wise function, for all $y > 0$ by

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log(y), & \lambda = 0 \end{cases}$$

Exercises

Exercise 1: Power Transformation Family in a X-Y scatter Plot Setting

- This exercise demonstrates the characteristics of the power-transform when applied independently to the two processes in an X-Y scatter plot setting. In this situation, one observed paired (X,Y) observations which are typically plotted X vs. Y in the 2D plane. We are interested in studying the effects of independently applying the power transforms to the X and Y processes. How and why would the corresponding scatter plot change as we vary the power parameters for X and Y?



- First, point your browser to [SOCR Charts](#) and select the **PowerTransformXYStatterChart** (Line-Charts -> PowerTransformXYStatterChart). Then either use the default data provided for this chart, enter your own data (remember to **MAP** the data before your **UPDATE** the chart) or obtain SOCR simulated data from the **Data-Generation** tab of the [SOCR Modeler](#) (an example is shown later in Exercise 4). As shown on the image below, try changing the power parameters for the X and Y power-transforms and observe the graphical behavior of the transformed scatter-plot (blue points connected by a thin line) versus the native (original) data (red color points). We have applied a linear rescaling to the power-transform data to map it in the same space as the original data. This is done purely for

visualization purposes, as without this rescaling it will be difficult to see the correspondence of the transformed and original data. Also note the changes of the numerical summaries for the transformed data (bottom text area) as you update the power parameters. What power parameters would you suggest that make the X-Y relation most linear?

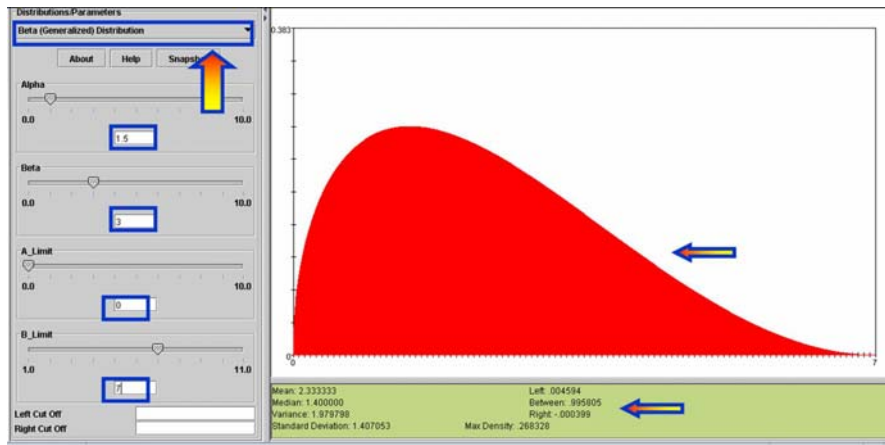
Exercise 2: Power Transformation Family in a QQ-Normal Plot Setting

- The second exercise demonstrates the effects of the power-transform applied to data in a QQ-Normal plot setting. We are interested in studying the effects of power transforming the native (original) data on the quantiles, relative the Normal quantiles (i.e., QQ-Normal plot effects). How and why do you expect the QQ-Normal plot to change as we vary the power parameter?
- Again go to [SOCR Charts](#) and select the **PowerTransformQQNormalPlotChart** (Line-Charts -> PowerTransformQQNormalPlotChart). You can use different data for this experiment - either use the default data provided with the QQ-Normal chart, enter your own data (remember to **MAP** the data before your **UPDATE** the chart) or obtain SOCR simulated data from the **Data-Generation** tab of the [SOCR Modeler](#) (an example is shown later in Exercise 4). Change the power-transform parameter (using the slider or the by typing into the text area) and observe the graphical behavior of the transformed data in the QQ-Normal plot (green points connected by a thin line) versus the plot of the native data (red color points). What power parameter would you suggest that make the (transformed) data quantiles similar to those of the Normal distribution? Why?



Exercise 3: Power Transformation Family in a Histogram Plot Setting

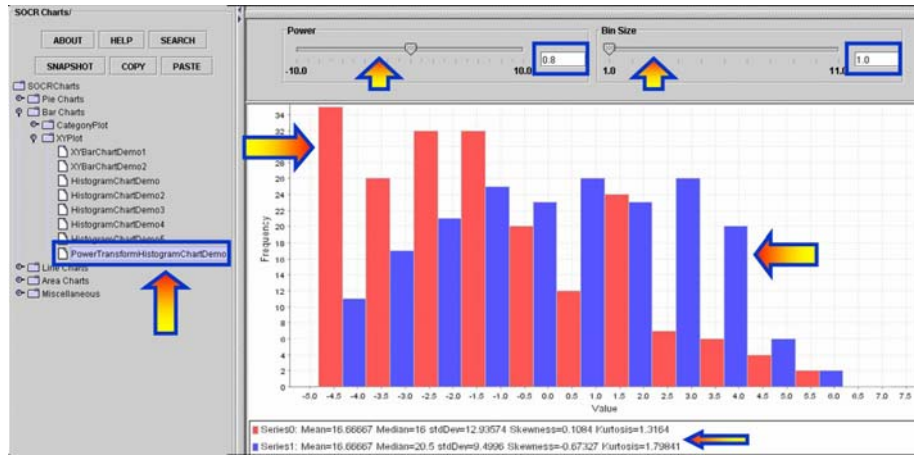
- This exercise demonstrates the effects on the histogram distribution after applying the power-transform to the (observed or simulated) data. In this experiment, we want to see whether we can reduce the variance of a dataset and make its histogram more symmetric, unimodal and bell-shaped.
- Again go to [SOCR Charts](#) and select the **PowerTransformHistogramChart** (Bar-Charts -> XYPlot -> PowerTransformHistogramChart). We will use SOCR simulated data from the **Data-Generation** tab of the [SOCR Modeler](#), however you may choose to use the default data for this chart or enter your own data. The image below shows you the [Generalized Beta Distribution](#) using [SOCR Distributions](#).



- Go to the [SOCR Modeler](#) and select 200 observations from the [Generalized Beta Distribution](#) ($\alpha = 1.5; \beta = 3; A = 0; B = 7$), as shown on the image below. Copy these 200 values in your mouse buffer (CNT-C) and paste them in the **Data** tab of the **PowerTransformHistogramChart**. Then *map* this column to *XYValue* (under the **MAP** tab) and click **Update_Chart**. This will generate the histogram of the 200 observations. Indeed, this graph should look like a discrete analog of the [Generalized Beta](#) density curve above.

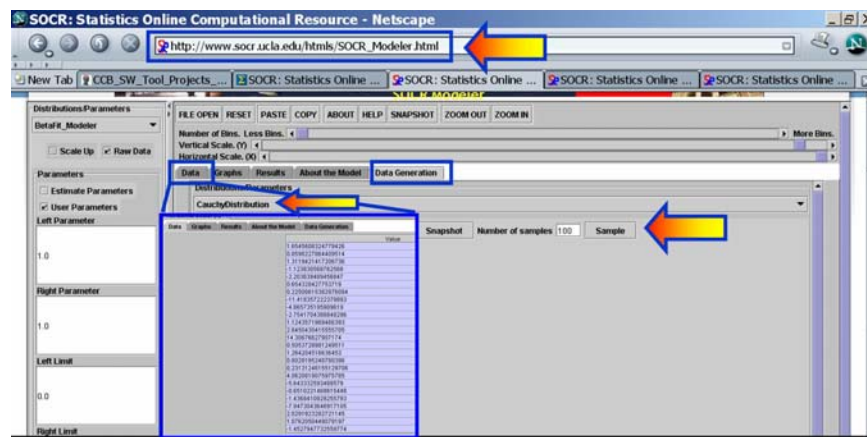


- In the **Graph** tab of the **PowerTransformHistogramChart**, change the power-transform parameter (using the slider on the top). All SOCR Histogram charts allow you to choose the width of the histogram bins, using the second slider on the top. Observe the graphical behavior of the **histogram** of the transformed data (blue bins) and compare it to the **histogram** of the native data (red bins). What power parameter would you suggest that make the **histogram** of the power-transformed data better? Why?

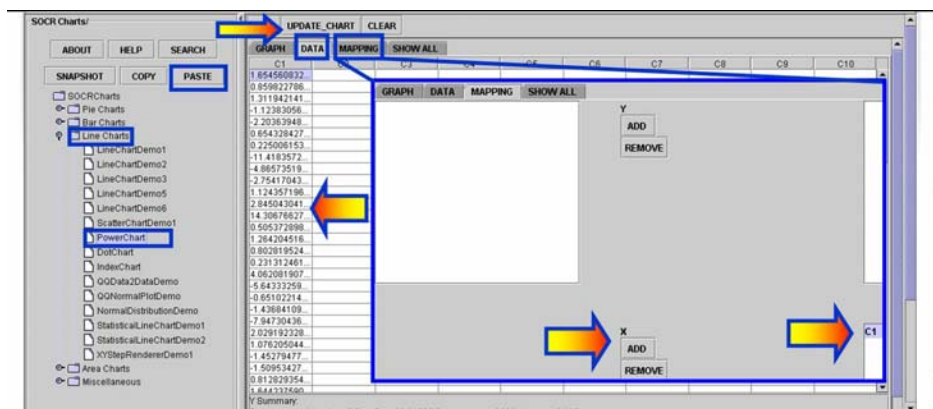


Exercise 4: Power Transformation Family in a Time/Index Plot Setting

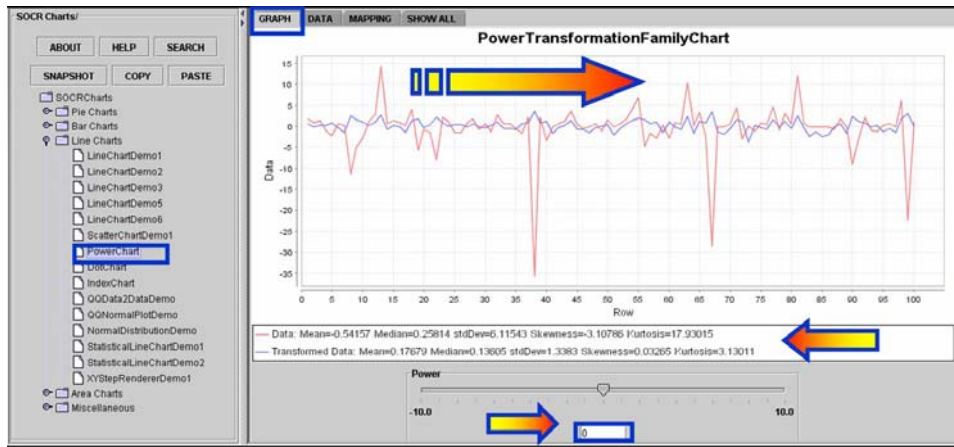
- Let's first get some data: Go to [SOCR Modeler](http://www.socr.ucla.edu/htmls/SOCR_Modeler.html) and generate 100 Cauchy Distributed variables. Copy these data in your mouse buffer (CNT-C). Of course, you may use your own data throughout. We choose Cauchy data to demonstrate how the Power Transform Family allows us to normalize data that is far from being Normal-like.



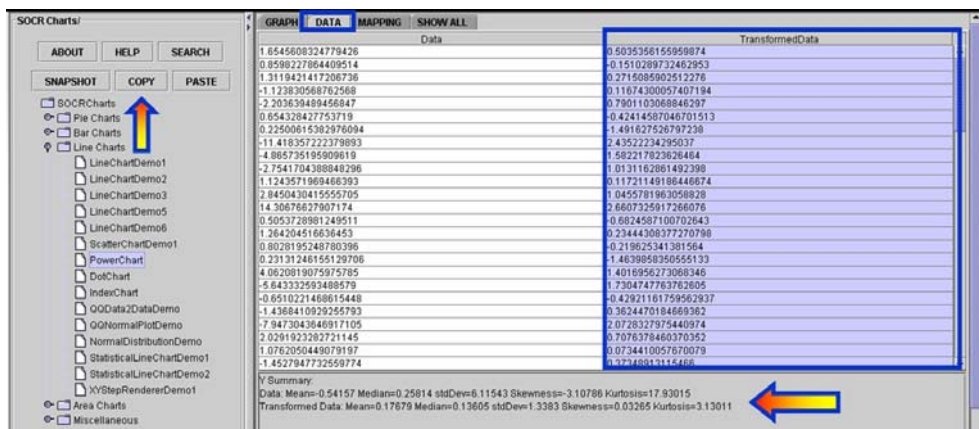
- Next, paste (CNT-V) these 100 observations in [SOCR Charts](#) (Line-Charts -> Power Transform Chart). Click **Update Chart** to see the index plot of this data in RED!



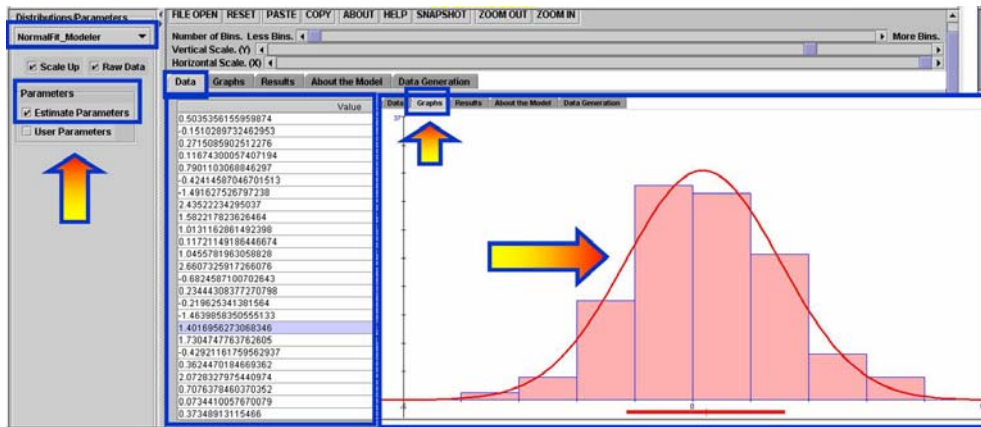
- Now go to the **Graph Tab-Pane** and choose $\lambda = 0$ (the power parameter). Why is $\lambda = 0$ the best choice for this data? Try experimenting with different values of λ . Observe the variability in the Graph of the transformed data in Blue (relative to the variability of the native data in Red).



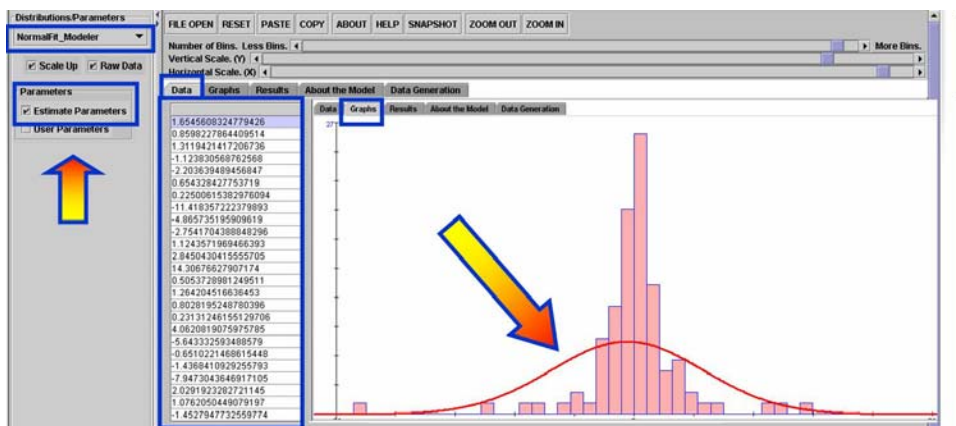
- Then go back to the **Data Tab-Pane** and copy in your mouse buffer the transformed data. We will compare how well does [Normal distribution](#) fit the histograms of the raw data ([Cauchy distribution](#)) and the transformed data. One can experiment with other powers of λ , as well! In the case of $\lambda = 0$, the power transform reduces to a **log transform**, which is generally a good way to make the histogram of a data set well approximated by a Normal Distribution. In our case, the histogram of the original data is close to Cauchy distribution, which is heavy tailed and far from Normal (Recall that the $T(df)$ distribution provides a 1-parameter homotopy between Cauchy and Normal).



- Now copy in your mouse buffer the transformed data and paste it in the [SOCR Modeler](#). Check the **Estimate Parameters** check-box on the top-left. This will allow you to fit a Normal curve to the histogram of the (log) Power Family Transformed Data. You see that Normal Distribution is a great fit to the histogram of the transformed Data. Be sure to check the parameters of the Normal Distribution (these are estimated using least squares and reported in the **Results Tab-Pane**). In this case, these parameters are: $Mean = 0.177$, $Variance = 1.77$, however, these will vary, in general.



- Let's try to fit a Normal model to the histogram of the native data (recall that this histogram should be shaped as Cauchy, as we sampled from Cauchy distribution – therefore, we would not expect a Normal Distribution to be a good fit for these data. This fact, by itself, demonstrates the importance of the Power Transformation Family. Basically we were able to *Normalize* a significantly Non-Normal data set. Go back to the original [SOCR Modeler](#), where you sampled the 100 Cauchy observations. Select **NormalFit_Modeler** from the drop-down list of models in the top-left and click on the **Graphs** and **Results** Tab-Panes to see the graphical results of the histogram of the native (heavy-tailed) data and the parameters of its best Normal Fit. Clearly, as expected, we do not have a got match.



- Questions:**
 - Try experimenting with other (real or simulated) data sets and different Power parameters (λ). What are the general effects of increasing/decreasing λ in any of these domains $[-10;0]$, $[0;1]$ and $[1;10]$?
 - For each of the exercises (X-Y scatter-plot, QQ-Normal plot, Histogram plot and Time/Index plot) empirically study the effects of the power transform as a tool for *normalizing* the data. You can take samples of size 100 from Student's T-distribution (low df) and determine appropriate levels of λ for which the transformed data is (visually) well approximated by a Normal Distribution.