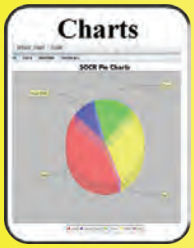
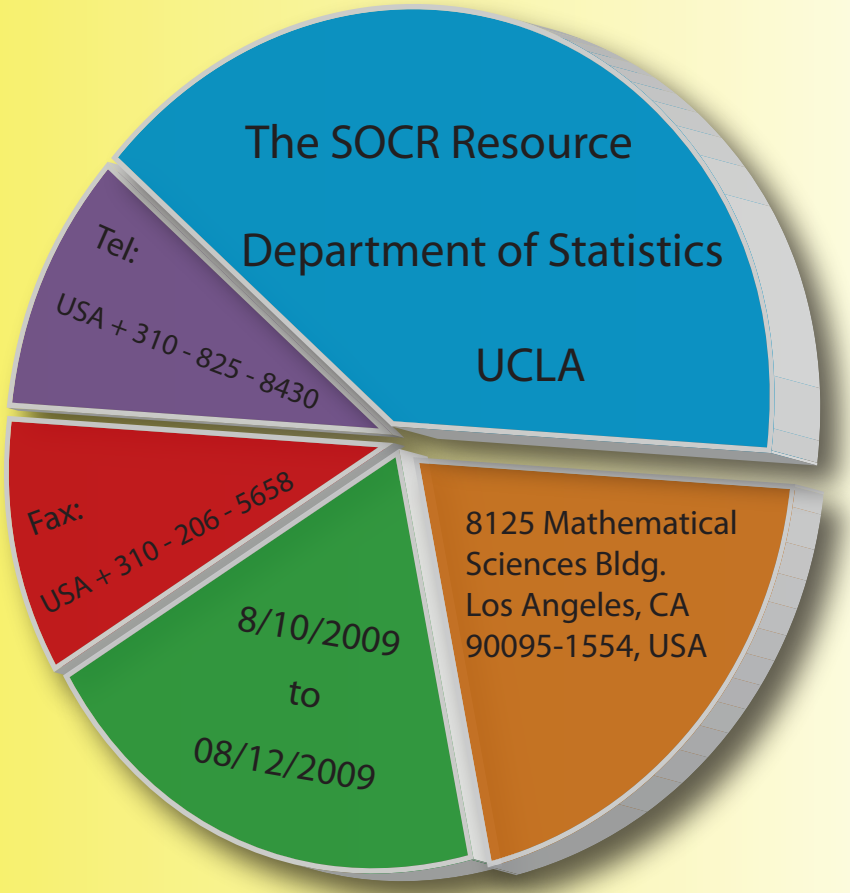
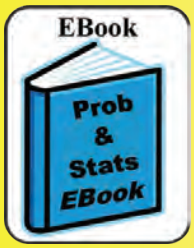
It's Online, Therefore it Exists!

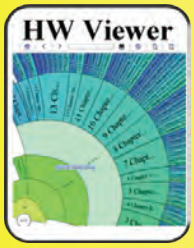
2009 SOCR Continuing Statistics Education

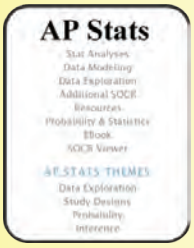
Training & Development Workshop Handbook

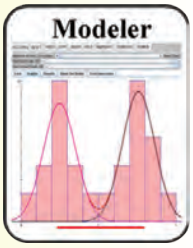












SOCR Publications Ivo D. Dinov, PhD & Nicolas Christou, PhD

Copyright Page

Cover design by: SOCR Publishing
Book design by: SOCR Publishing
Authors: Ivo D. Dinov and Nicolas Christou

SOCR Publishing, 8125 Math Sciences Bldg., Los Angeles, California 90095, USA.
www.SOCR.ucla.edu

Printed in the United States of America, 2009

Available at no cost in electronic form at different outlets including:

- http://www.SOCR.ucla.edu
- http://wiki.stat.ucla.edu/socr/index.php/SOCR_Events_Aug2009
- http://repositories.cdlib.org/socr/
- http://books.google.com/
- http://www.ebooks.com/

First Printing: August 2009

ISBN 978-0-615-30464-9

Library of Congress Control Number: 2009907564
http://lcweb.loc.gov

SOCR Continuing Statistics Education

Workshop Handbook

Table of Contents

SOCR UCLA Statistics UCLA OID NSF

www.SOCR.ucla.edu

http://wiki.stat.ucla.edu/socr/index.php/SOCR_Events_Aug2009

Preface

This book contains instructional materials, pedagogical techniques and hands-on activities for technology-enhanced science education. These educational resources were presented at the 2009 SOCR Continuing Statistics Education Workshop entitled "*It's Online, Therefore it Exists! 2009 SOCR Continuing Statistics Education Training & Development Workshop*," which took place at the University of California, Los Angeles, August 10-12, 2009.

Book purpose: This book was written to provide modern pedagogical perspective into technology-enhanced blended learning and instruction. It reflects the direction of amalgamation of knowledge from multiple disciplines with recent open networking and technological advances. This trend is anticipated to accelerate in the next decade with seamless integration of open-access data, learning materials, computational resources and collaborative environments connected via language, platform and geo-politically agnostic interfaces. There are four novel features of this book – it is community-built, completely free-open-access (in terms of use and contributions), blends concepts with technology and it is electronically available online in multilingual format. This book covers the narrow scientific area of probability and statistics education; however, the principles we promote in the book can easily be extended to all other science and technology fields.

Development process: The SOCR developments began in 2002 with the introduction of a number of Java-based web-applets for mathematics, probability and statistics education. In 2005, we began developing learning materials wrapped around the available computational libraries. This enabled us to propose a new paradigm for resource development where classroom use and student-demand drove the design, specification and implementation of new and improved web-applets. In 2007, we introduced the Probability and Statistics EBook (http://wiki.stat.ucla.edu/socr/index.php/EBook), which has had over one million users, as of June 2009. In the summer of 2007, we organized the first SOCR Continuing Statistics Education with Technology workshop (http://wiki.stat.ucla.edu/socr/index.php/SOCR_Events_Aug2007). In the next year, we received a number of constructive suggestions, critiques, evaluations and general feedback from learners and instructors that gave us specific directives on how to improve and extend these materials to improve self-learning and formal curricular training. Since 2008, we have used these materials in dozens of UCLA classes and conducted 2 IRB-approved large-scale meta-studies of the effectiveness of these resources to improve student learning. Our findings indicate that technology-enhanced blended instruction has a consistent, robust and statistically-significant effect in improving probability and statistical learning and knowledge retention in undergraduate classes (Dinov et al., 2008). In 2009, we started to work on this book which is available on the Internet at: http://wiki.stat.ucla.edu/socr/index.php/SOCR_Events_Aug2009.

Organization: This book is organized in four themes, which are presented in the text in an order according to their natural appearance in general classroom settings – data-driven *topic motivation*; interactive *tools* for data modeling, analysis and visualization; hands-on learning *activities*; and resource *navigation*. All of the materials, tools and demonstrations presented in this book may be rearranged, modified and tailored to the specific needs of the corresponding learning audiences.

Online resources: All SOCR data, tools and materials (www.SOCR.ucla.edu) are freely and openly available on the web as integrated resources (e.g., data and results may be copy-pasted from one resource to the next via simple mouse/key manipulations). Our guiding principle is that to be considered in existence, *knowledge materials need to be freely and openly accessible on the Internet via flexible and portable interfaces* (e.g., XML, Java, HTML, WSDL, etc.) SOCR develops,

validates and disseminates four types of resources: data, interactive computational Java applets and libraries, hands-on learning activities, and instructional modules and video tutorials. All of these are openly accessible via the SOCR web-page: www.SOCR.ucla.edu.

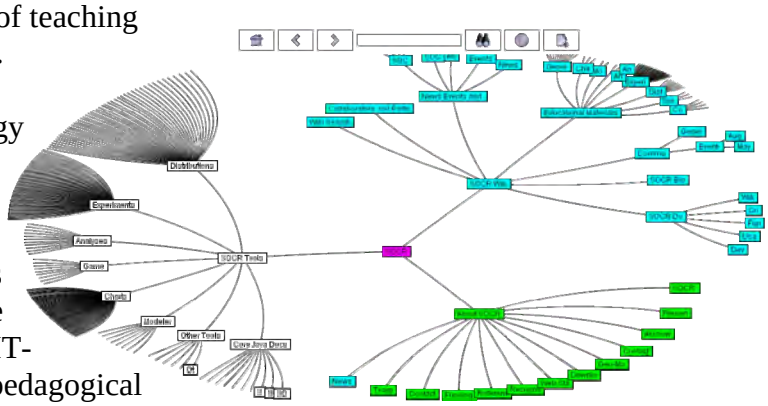
Note to learners: We recommend that self-learners and students start by going over the Probability and Statistics EBook (http://wiki.stat.ucla.edu/socr/index.php/EBook). The EBook is geared for novice audiences and provides a more systematic and consistent approach to learning the fundamental concepts in probability and statistics. This book is geared primarily for teachers and science instructors interested in technology-enhanced, blended and multidisciplinary science education.

Note to educators: Much of the materials presented in this handbook can be directly utilized in Advanced Placement (AP), undergraduate and graduate statistics curricula. However, this book is intended to demonstrate *instances* of how modern technology may be used to motivate students' learning and provide integrated resources for open and dynamic science education. There are three reason why we did not include in this book *all* learning materials and resource demonstrations that we have already developed. First, the complete book would have been over 1,000 pages long. Second, these materials are truly dynamic and constantly improved and extended, which quickly makes static versions of the materials obsolete. Third, our goal is to provide *enabling resources* that can be modified and customized by instructors to fit their course-syllabi and student-audiences. This book does *not* intend to be a complete one-size-fits-all instructional resource.

About the SOCR Resource

SOCR (www.SOCR.ucla.edu) is an NSF-funded project (DUE 0716055) that designs, implements, validates and integrates various interactive tools for statistics and probability education and computing. Many of the SOCR projects provide resources for introductory probability and statistics courses. Some SOCR resources bridge between the introductory and the more advanced computational and applied probability and statistics courses. There are four major types of SOCR users: educators, students, researchers and tool developers. The 2009 workshop is intended for educators. Course instructors and teachers will find the SOCR class notes and interactive tools useful for student motivation, concept demonstrations and for enhancing their technology-based pedagogical approaches to any study of variation and uncertainty. Students and trainees may find the SOCR class notes, analyses, computational and graphing tools extremely useful in their learning/practicing pursuits. Model developers, software programmers and other engineering, biomedical and applied researchers may find the light-weight plug-in oriented SOCR computational libraries and infrastructure useful in their algorithm designs and research efforts.

The **main objective of SOCR** is to offer a homogeneous interface for online activities appropriate for Introductory Statistics courses, Introductory Probability courses, Advanced Placement (AP Stats) courses and other 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 recognition from several international, educational and technology-based initiatives (www.socr.ucla.edu/htmls/SOCR_Recognitions.html). SOCR has been, and continues to be, tested in the classroom. Most recently, 2 large-scale experimental studies we conducted led us to conclude that using SOCR for the teaching of Introductory Statistics and Probability was effective (Dinov et al., 2008; Dinov, et al. 2009). In these studies we discovered a robust and reproducible effect of improving student performance in SOCR-based technology-enhanced probability and statistics courses. Thus, the SOCR workshop offers participants research-based knowledge on effective teaching with online learning resources, which is the best kind of teaching strategy (Richard, et al. 2002).



Until now, many technology inclined educators have adopted in their course curriculum interactive aids (e.g., applets) from diverse and heterogeneous resources. Many instructors have also created their own IT-instruments to enhance their pedagogical approaches. The *implementation of the SOCR philosophy* for developing and utilizing new IT-resources (tools) and educational materials is depicted in **Table 1**. The subdivision of all SOCR resources into tools and materials, **Table 2**, is natural as our goals are similarly dichotomous. *First*, develop libraries and foundational instruments for demonstration, motivation, visualization and statistical computing, which are platform

agnostic. And *second*, design instances of course-, topic- and student-specific educational materials (lecture notes, activities, assignments, etc.) which are agile and extensible wrappers around the available SOCR tools. Both of these categories are open-source and may be extended, revised and redistributed by others in the community. For example, technically savvy users may quickly implement a new SOCR Analysis object, add an additional SOCR Distribution, extend the functionality of a SOCR Experiment, etc., by simply implementing as a plug-in the corresponding SOCR Java object. Redistributing the new tool to the community only requires posting of the new tool on the SOCR web page. It does not require complete SOCR package rebuild or restructuring. Educators that are more interested in the application of the SOCR tools and their in-class utilization also may actively contribute to the SOCR efforts by developing new, improving existent and testing and validating the SOCR educational materials. In fact, just like Wikipedia, the entire SOCR effort is contingent upon the continued support and development efforts of the community (educators and researchers in the areas of probability, statistics, mathematics, data modeling, etc.) We simply provide the infrastructure for these developments; the user community is responsible for the rest.

SOCR PHILOSOPHY	
DEVELOPMENT OF TOOLS	**DEVELOPMENT OF EDUCATIONAL MATERIALS**
1. Tools must be freely and openly available on the Web (SOCR Motto: *It's online, therefore it exists!*) via flexible media formats.	1. Newly developed materials must increase pedagogical content knowledge, be extensible, factually correct, validated and deployed on the web.
2. Tools must be well-designed, extensible, platform-independent, open-source and connected to some activity that increases pedagogical content knowledge.	2. Materials must explicitly utilize some SOCR interactive resources (e.g., activities with applet demos) and provide the means of cross reference by various SOCR tools (e.g., applet calculations citing formulas in instructional materials).

Table 1: SOCR working philosophy in development of new tools and educational materials**.**

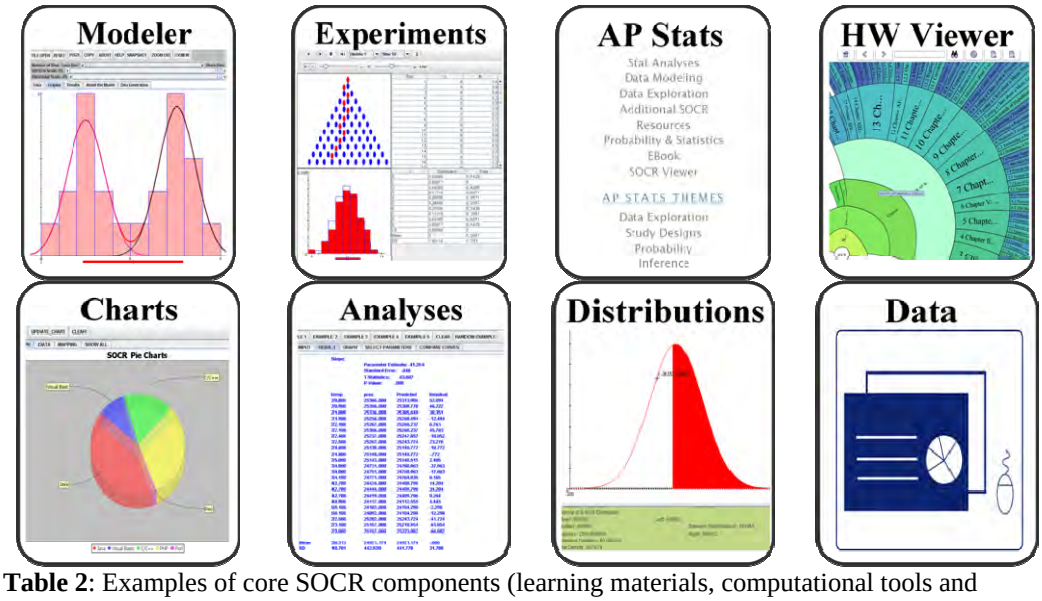


Table 2: Examples of core SOCR components (learning materials, computational tools and services).

Welcome Letter

Dear Workshop Attendees:

On behalf of the Statistics Online Computational Resource (SOCR), we would like to welcome you to the 2009 SOCR Continuing Statistics Education Workshop. The theme of this year's workshop is *It's Online, Therefore it Exists! 2009 SOCR Continuing Statistics Education Training & Development Workshop.*

We organized this workshop to achieve two major goals – provide *training for instructors* in using the latest SOCR educational resources, and at the same time, provide an open learning forum for instructors to *communicate and exchange ideas* about existent educational materials, desired resources and useful instructional tools for improving probability and statistics education at different levels.

Under different tabs in this booklet, you will find the Workshop Program; the Workshop Logistics and Goals; more about SOCR; the Complete Workshop activities; an area Map; and Workshop Evaluation Forms. Please complete, tear off and return to us the evaluation forms as instructed. We are mandated by our funding agency, the National Science Foundation, to provide quantitative and qualitative evaluation of all of our educational activities, computational resources and learning materials, including this training workshop.

Over the past several years, SOCR members have developed new, catalogued existent, and annotated a large number of instructional materials, interactive applets, internet resources and collaborative resources. We will present many of these new developments during the workshop. You may always find the complete collections of SOCR resources on the web at www.SOCR.ucla.edu.

We are looking forward to a productive and exciting session on statistics education in the next few days and hope that this exchange of ideas, instructional materials and Internet resources stimulates long-term collaborations and facilitates novel approaches to teaching with technology. Any feedback, comments and ideas from all of you are welcome throughout the workshop as well as after the conclusion of this training event.

Ivo D. Dinov — SOCR Resource PI, Professor of Statistics

Nicolas Christou — SOCR Resource Co-PI, Professor of Statistics

Workshop Logistics

There will be 35 workshop participants physically attending this 3-day event and a large number of virtual attendees viewing the live web-stream. All participants will be partially supported to attend the workshop, by the NSF-funded SOCR resource.

- **Dates**: Mon-Wed, August 10-12, 2009.
- **Times**: AM & PM Sessions (9AM - 12PM & 1PM – 4:30 PM).
- **Venue/Place**: Powell Library (CLICC Classroom B, Powell 320B, see map on the back).
- **Accommodation:** UCLA UCLA Hedrick Hall for checking in at 4 PM on Sun 08/09/09 and checking out by 11 AM on Wed 08/12/09. UCLA Catering will provide housing and accommodation to all remote participants.
- **Local Information**: Maps & local visitor information (see Handbook back cover).
- **Funding Support Details**: Participants will be staying at UCLA Hedrick Hall for 3 nights (Aug 09, 10, 11, 2009), these costs are covered by the conference organizers and will be paid directly. Only no-shows will be charged. All meals will be provided during the Workshop. There is no Workshop registration fee nor are there any charges for the Workshop materials which will be distributed. All transportation costs are the attendee's responsibilities.

Workshop Goals

The overarching goals of this workshop are to provide continuing education and training for instructors using the latest SOCR educational resources and at the same time, provide an open learning environment for attendees to communicate and exchange ideas about existent validated educational materials, desired new resources and useful pedagogical techniques and instruments.

In particular, we will discuss the diverse SOCR Internet resources, their design, usage, evaluation, extensibility and classroom utilization. Among these are the SOCR Java applets for distributions, experiments, analysis, modeling and data exploration, various activities for hands-on demonstrations, as well as, students' and instructors' Internet resources. The SOCR philosophy is that *one-size-does-not-fit-all!* This means that we provide tools, data, materials and infrastructure for technology enhanced science education. However, it's ultimately the instructor's responsibility to wrap these resources into a coherent set of materials appropriate for their concrete classes, students' maturity and course syllabi.

Workshop Attendees

http://wiki.stat.ucla.edu/socr/index.php/SOCR_Events_Aug2009_Attendees

Workshop Attendees	
Attendee	**Affiliation**
Alisher Abdullayev	American River College, CA
Amit Agarwal	LAUSD, CA
Robert Baker	University Senior High School, CA
Leyla Batakci	Elizabethtown College, PA
Pete Bouzar	Golden West College, CA
Rebecca Cajucom	Pierce College, CA
Nasser Dastrange	Buena Vista University, IA
Dan Debevec	Mira Costa High School, CA
John Egenolf	Oregon State University, OR
Peter Esperanza,	Barstow Unified School District, CA
Todd Frost	Flintridge Preparatory School, CA
Ming-Lun Ho	Chabot College, CA
Wei Huang	University of Dundee, United Kingdom
Joseph Kazimir	East Los Angeles College, CA
Leah Klugman	Westridge School, CA
Kari Kooker	Sunny Hills High School, CA
Lee Kucera	Capistrano Valley High School, CA
Julie Margala	Chino High School, CA
Mary Martin	Cosumnes River College, CA
Rahila Munshi	West Adams Prep, CA
Said Ngobi	Victor Valley College, CA
Radoslav Nickolov	Fayetteville State University, NC
Tedja Oepomo	West Los Angeles College, CA
Patrick O'Sullivan	Mary Immaculate College Limerick, Ireland
Colleen Ryan	California Lutheran University, CA
John Stedl	Chicago State University, IL
Mike Wade	The Community School, ID
Anita Wah	Chabot College, CA
Jerimi Walker	Moraine Valley Community College, IL
Vonda Walsh	Virginia Military Institute, VA
Angela Wang	California State University Northridge, CA
Lam Wong	Cal Poly Pomona, CA
Lawrence Yee	George Washington High School, CA
Bee Yew	Fayetteville State University, NC

The list above only includes the limited invited physical attendees of the workshop. The entire workshop was also streamed live and will be archived online at the California Digital Library as a permanent web-cast for others to see.

Workshop Program At-A-Glance

DAY 1 (MON 08/10/09): MORNING SESSION – 9AM - 12PM		
OPEN, DIVERSE, MOTIVATIONAL, INTERACTIVE AND WEB-BASED SOCR DATASETS		
TIME	PRESENTER	TOPIC
7:00-8:00 AM	*Breakfast, UCLA Cafeteria*	
8:00-9:00 AM	*Registration and Coffee*	
9:00-9:10 AM	Ivo Dinov	Welcome
9:10-9:20 AM	Everyone	Participant Introductions
9:20-9:30 AM	Ivo Dinov	Guest Accounts
9:30-9:40 AM	Ivo Dinov	The State of the SOCR Resource
9:40-10:40AM	Nicolas Christou	SOCR Open Motivational Datasets • Research-derived data • Multi-disciplinary data understanding
10:40-10:50AM	*Morning Break*	
10:50-11:30AM	Ivo Dinov	SOCR Open Motivational Datasets (cont.) • Simulated Data (RNG)
11:30-12:00PM	Everyone	Interactive Discussion on generating data and curricular integration of datasets
12:00-1:00 PM	*Lunch Break, UCLA Cafeteria*	
AFTERNOON SESSION – 1PM – 4:30PM		
SOCR TOOLS		
1:00-1:30 PM	Ivo Dinov	SOCR Distributions
1:30-2:00 PM	Ivo Dinov	SOCR Experiments
2:00-2:30 PM	Ivo Dinov	SOCR Games
2:30-2:45 PM	*Afternoon Break*	
2:45-3:15PM	Annie Chu	SOCR Analyses
3:15-3:45PM	Ivo Dinov	SOCR Modeler
3:45-4:15PM	Ivo Dinov	SOCR Charts
4:15-4:30PM	Everyone	Interactive Group Discussion on Tools for Probability & Stats Education - What works, what doesn't, how to extend the collection and enhance the experiences of others?
6:30-8:00PM	*Dinner, UCLA Cafeteria*	

DAY 2 (TUE 08/11/09): MORNING SESSION – 9AM - 12PM

SOCR ACTIVITIES		
TIME	**PRESENTER**	**TOPIC**
7:00-8:30 AM	*Breakfast, UCLA Cafeteria*	
9:00-10:00 AM	Annie Chu Ivo Dinov	Analysis Activities • ANOVA • Simple Linear Regression
10:00-10:30 AM	Ivo Dinov	Modeler Activities • SOCR Normal & Beta Distribution Model Fitting
10:30-10:40AM	*Morning Break*	
10:40-11:40AM	Nicolas Christou	Distribution Activities • Normal Distribution Activity • Relations between distributions
11:40-12:00PM	Everyone	Group Interactive Discussion on Hands-on Activities - What works, what doesn't, how to extend the collection and how to improve teaching of statistical analysis methodologies?
12:00-1:00 PM	*Lunch Break, UCLA Cafeteria*	

AFTERNOON SESSION – 1PM – 4:30PM

SOCR ACTIVITIES (CONT.)		
1:00-2:15 PM	Ivo Dinov	Central Limit Theorem Activity
2:15-2:45 PM	Ivo Dinov	SOCR Resource Navigation • Hyperbolic/Carousel Viewers, Data Import
2:45-3:00 PM	*Afternoon Break*	
3:00-3:45 PM	Nicolas Christou	Confidence Intervals Activity
3:45-4:30 PM	Nicolas Christou	SOCR Application Activities • Portfolio Risk Management
4:30-4:45 PM	Everyone	Interactive Group Discussion on Hands-on Activities - What works, what doesn't, how to extend the collection and enhance the experiences of others?
6:30-8:00PM	*Dinner, UCLA Cafeteria*	

DAY 3 (WED 08/12/09): MORNING SESSION – 9AM - 12PM		
SOCR ACTIVITIES		
TIME	PRESENTER	TOPIC
7:00-8:30 AM	*Breakfast, UCLA Cafeteria*	
9:00-10:15 AM	Ivo Dinov	EBook and Exploratory Data Analyses (EDA)
		SOCR Charts and Activities
		SOCR MotionCharts
		SOCR AP Statistics Materials
10:15-10:30AM	*Morning Break*	
10:30-11:30AM	Ivo Dinov	Law of Large Numbers (LLN) Activity
11:30-11:45AM	Everyone	Group Interactive Discussion on Hands-on Activities – What works, what doesn't, how to extend the collection and how to improve teaching of statistical analysis methodologies?
11:45-12:00 PM	*Workshop Evaluation by Participants*	
12:00-1:00 PM	*Lunch Break, UCLA Cafeteria*	
AFTERNOON SESSION – 1PM – 4:30PM		
VISIT TO THE J. PAUL GETTY CENTER		

Workshop Activities and Materials

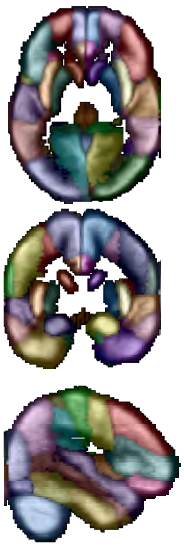
Day 1: Mon 08/10/09

Morning Session: Open, Diverse, Motivational, Interactive and Web-Based SOCR Datasets

SOCR Open Motivational Datasets (http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data)
The SOCR resource provides a number of mechanisms to simulate data using computer random-number generation. The most commonly used SOCR generators of simulated data are: *SOCR Experiments* - each experiment reports random outcomes, sample and population distributions and summary statistics*; SOCR random-number generator* - enables sampling of any size from any of the SOCR Distributions; and, *SOCR Analyses* - all of the SOCR analyses allow random sampling from various populations appropriate for the user-specified analysis.

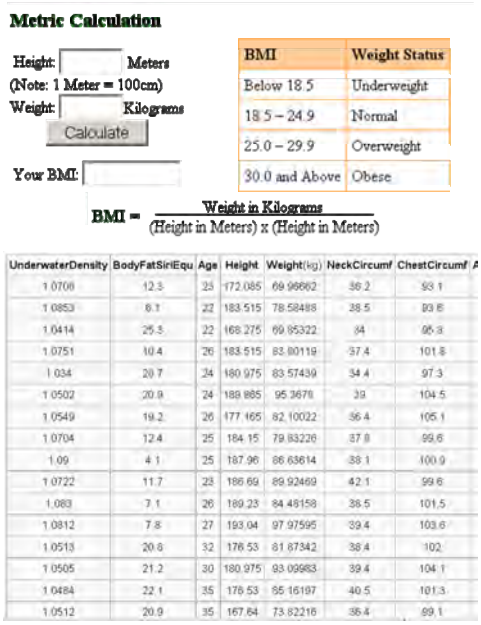
- *Classroom use guidelines*:
  - Show an appropriate dataset (spreadsheet, summary statistics, and simple graphs) before introducing a new probability or statistics concept.
  - Use the data (its properties, characteristics and appearance) to motivate the need for another concept or technique.
  - If possible, provide hands-on calculations on parts of the data to justify the methodology.

- *Research-derived data*:
  There are a number of research acquired datasets available on the SOCR Data web-page. Some of these examples include:
  - *Neuroimaging study of prefrontal cortex volume across species & tissue types* (http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_April2009_ID_NI)
    The prefrontal cortex is the anterior part of the frontal lobes of the brain in front of the premotor areas. Prefrontal cortex includes cytoarchitectonic layer IV and includes three regions: orbitofrontal (OFC), dorsolateral prefrontal cortex (PFC), anterior and ventral cingulate cortex. Human brains are much distinct from the brains of other primates and apes specifically in the prefrontal cortex. These structural differences induce significant functional abilities which may account for the significant associating, planning and strategic thinking in humans, compared to other primates. The study below investigated the quantitative differences between the PFC volumes across species and tissue types.



	Cerebrum ICV (mm3)			PFC (mm3)			% PFC/ICV
	Total	GM	WM	Total	GM	WM	
Human	1,206,129	535,956	670,174	156,443	82,346	74,097	12.97067
Bonobo	272,837	125,469	147,368	25,374	15,382	9,993	9.300058
Chimpanzee	277,843	123,858	153,985	30,324	17,234	13,091	10.91408
Gorilla	389,681	170,505	219,176	41,655	23,419	18,235	10.68951
Orangutan	333,075	154,293	178,781	33,763	21,193	12,570	10.13676
Gibbon	65,938	32,418	33,520	6,302	3,994	2,307	9.557463
Mangabey	82,495	37,985	44,510	6,599	3,804	2,794	7.999273
Baboon	130,600	58,615	71,985	11,660	6,530	5,130	8.928025
Macaca	69,875	31,841	38,034	5,579	2,813	2,766	7.984258
Capuchin	58,581	29,358	29,223	5,835	3,293	2,542	9.960567
Squirrel(monkey)	21,542	9,785	11,756	1,820	1,094	727	8.448612

- *SOCR Body Density Data* (http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_BMI_Regression)
  This is a comprehensive dataset that lists estimates of the percentage of body fat determined by underwater weighing and various body circumference measurements for 252 men. This data set can be used to illustrate multiple regression techniques. Accurate measurement of body fat is inconvenient/costly and it is desirable to have easy methods of estimating body fat that are cost-effective and convenient.



- Body Density
- Percent body fat
- Age (years)
- Weight (kg)
- Height (cm)
- Neck circumference (cm)
- Chest circumference (cm)
- Abdomen 2 circumference
- Hip circumference (cm)
- Thigh circumference (cm)
- Knee circumference (cm)
- Ankle circumference (cm)
- Biceps (cm)
- Forearm circumference (cm)
- Wrist circumference (cm)

- *Multi-disciplinary data understanding - Mercury Contamination in Fish*
  Visualization, understanding and interpreting real data may be challenging because of noise in the data, data complexity, multiple variables, hidden relations between variables and large variation. This NISER activity demonstrates how to use free Internet-based IT-tools and resources to solve problems that arise in the

areas of biological, chemical, medical and social research. These data may be used to:

- demonstrate the typical research investigation pipeline - from problem formulation, to data collection, visualization, analysis and interpretation;
- illustrate the variety of portable freely available Internet-based Java tools, computational resources and learning materials for solving practical problems;
- provide a hands-on example of interdisciplinary training, cross-over of research techniques, data, models and expertise to enhance contemporary science education;
- promote interactions between different science education areas and stimulate the development of new and synergistic learning materials and course curricula across disciplines.

	A	B	C	D	E	F	G
1	Lake ID	Lake	Alkalinity	pH	Calcium	Chlorophyll	Avg_Mercury
2	2	Annie	3.5	5.1	1.9	3.2	1.33
3	8	Apopka	116	9.1	44.1	128.3	0.04
4	4	Blue Cypr	39.4	6.9	16.4	3.5	0.44
5	15	Farm-13	120	7.6	38.5	71.1	0.05
6	25	Jackson	12.8	6.1	3.7	21	0.41
7	29	Miccasuke	8.5	4.9	7.7	14.9	0.6
8	33	Ocean Pon	5.9	3.9	1.6	3.2	0.87
9	34	Ocheese P	4.5	4.4	1.1	3.2	0.56
10	35	Parker	53	8.4	45.0	152.4	0.04
11	47	Trafford	81.5	8.0	20.5	9.6	0.27
12	1	Alligator	5.9	6.1	3	0.7	1.33
13	5	Brick	2.5	4.6	2.9	1.8	0.2
14	6	Bryant	19.6	7.3	4.8	44.1	0.27
15	7	Cherry	5.2	6.4	2.6	3.4	0.48
16	9	Crescent	71.4	8.1	35.2	33.7	0.19
17	9	Deer Point	26.4	5.8	9.5	1.6	0.83
18	10	Dias	4.8	6.4	4.6	22.5	0.81
19	11	Dorr	6.6	5.4	2.7	14.9	0.71

Largemouth bass were studied in 53 different Florida lakes to examine the factors that influence the level of mercury contamination. Water samples were collected from the surface of the middle of each lake in August 1990 and then again in March 1991. The pH level, the amount of chlorophyll, calcium, and alkalinity were measured in each sample. The average of the August and March values were used in the analysis. Next, a sample of fish was taken from each lake with sample sizes ranging from 4 to 44 fish. The age of each fish and mercury concentration in the muscle tissue was measured.

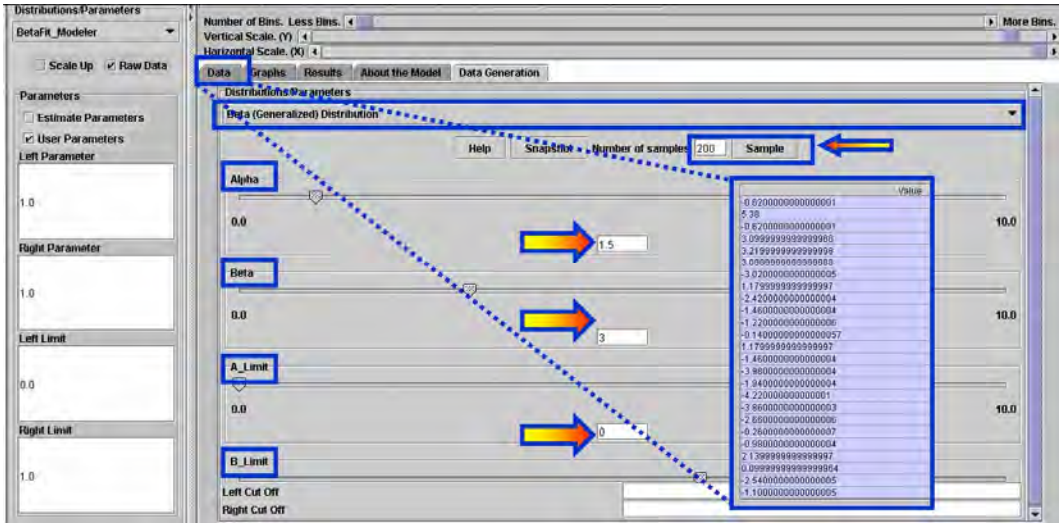
- *Simulated Data – Random Number Generation (RNG)*
  http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_Activities_RNG

  - *Are real-life natural processes deterministic and do they have exact mathematical closed-form descriptions?*
    - Arrival times to school each day?
    - Motion of the Moon around the Earth?
    - The computer CPU?
    - The atomic clock?

    It is an unsettling paradox that all natural phenomena we observe are stochastic in nature. Yet, we do not know how to replicate any of them exactly. There are good computational strategies to approximate natural processes using analytical mathematical models; however, upon careful review one always finds out a deterministic pattern in all purely computationally generated processes.

- *Two strategies to generate random numbers*.
  - One approach relies on observing a physical process which is expected to be random.
  - Another approach is to use computational algorithms that produce long sequences of apparently random results, which are in fact determined by a shorter initial seed. Random number generators based on physical processes may be based on random particles' momentum or position or any of the three fundamental physical forces. Examples of such processes are the Atari gaming console (noise from analog circuits to generate true random numbers), radioactive decay, thermal noise, shot noise and clock drift. A random number generator (RNG) based solely on deterministic computation is referred to pseudo-random number generator. There are various techniques for obtaining computational (pseudo)random numbers. Virtually all RNG's used in practice are pseudo-RNGs. To distinguish real random numbers from the pseudo-random numbers is a very difficult problem.

- *Why do we need random samples?* Random number generators have several important applications in statistical modeling, computer simulation, cryptography, etc. For example, data collection is often very expensive. Hence, to do appropriate inference on datasets of smaller sizes, we may consider simulating repeatedly from appropriate distributions instead of using real observations. Another example of why random number generators are so important comes from cryptography. It is a commonly held misconception that every encryption method can be broken. Claude Shannon, Bell Labs, 1948, proved that the one-time pad cipher is unbreakable, provided the secret key is truly random and of length equal or greater than the length of the encoded message. Monte Carlo simulations are also based on RNGs and are used for finding numerical solutions to (multi-dimensional) mathematical problems that cannot easily be solved exactly. For example, integration, differentiation, root-finding, etc.

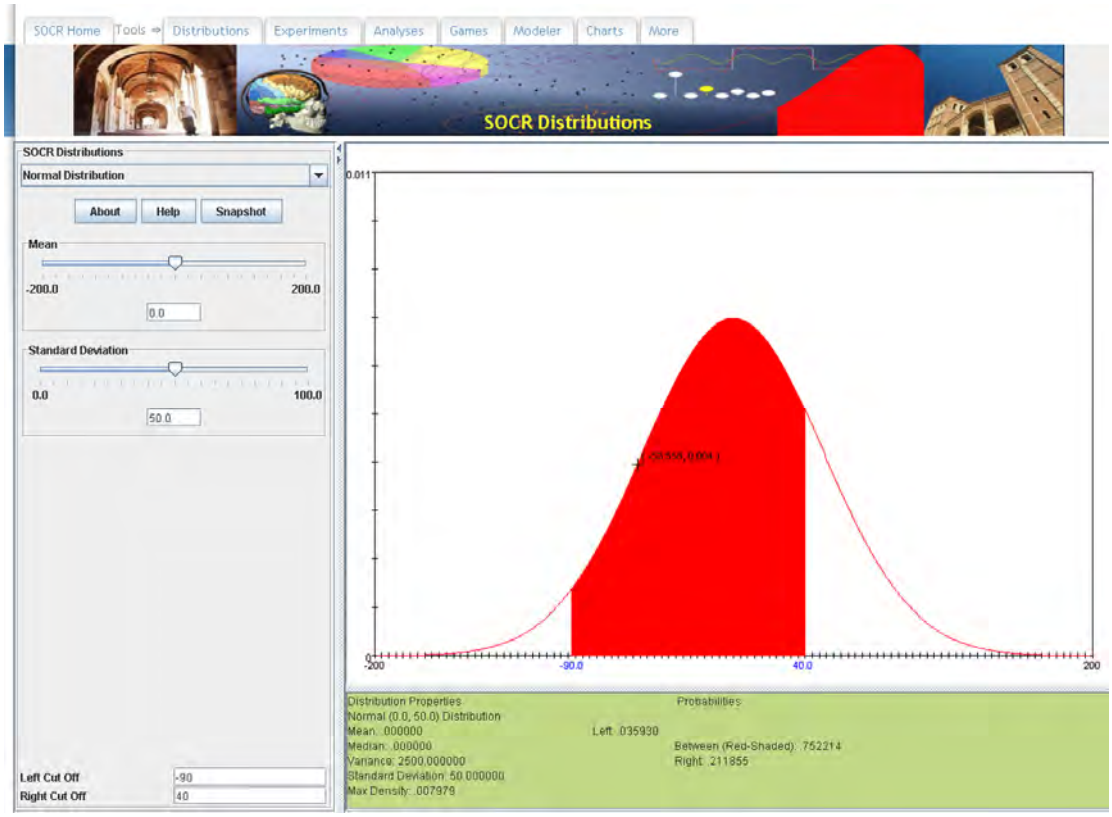


Afternoon Session: SOCR Tools

SOCR Distributions (www.socr.ucla.edu/htmls/SOCR_Distributions.html)

The core of many SOCR resources is the ability to compute, sample and model using various probability distributions. SOCR Distributions provides one of the largest open and graphically accessible collections of over 65 different distributions. Each of these distributions includes:

- applets: www.socr.ucla.edu/htmls/dist
- activities: http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_DistributionsActivities
- computational libraries: www.socr.ucla.edu/htmls/SOCR_Download.html
- usage documentation: www.socr.ucla.edu/docs/



Basic Operations:

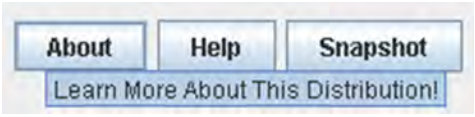
The basic operations allowed via the SOCR Distribution applets are:

- Selection of a distribution of interest, and its corresponding parameters.
- Calculation of probability values for a (graphically or numerically) specified interval.
- Calculation of critical values for specified probability values.

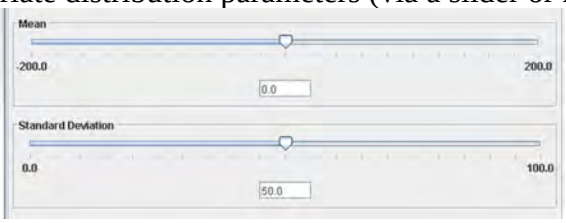
Controls:

SOCR distributions have the following controls:

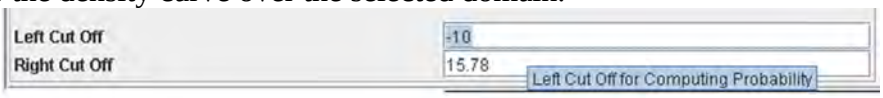
- General controls, which provide more information about each distribution and enables taking a snapshot of the state of the applet.



- Distribution parameter settings facilitate the specification of user-defined values for appropriate distribution parameters (via a slider or numerically).



- Limit settings enable interval specification for computing the probability under the density curve over the selected domain.

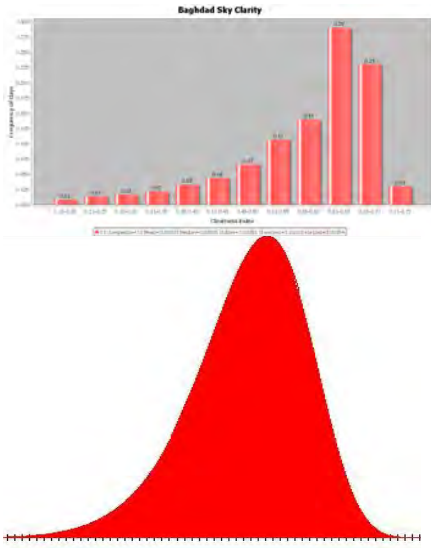


Classroom use examples:

- Visually choose a model distribution for a given dataset using SOCR Distributions. Compute and compare the corresponding data-driven and model probabilities for several ranges in the data domain.
- Compute the critical value for a given distribution and a specified probability value.
- Compute the probability value for a given critical value (or statistics).
- Illustrate the parallels between discrete distribution tables and the corresponding SOCR distribution calculations.

Clearness Index	Number of Days	Relative Freq	Cumulative Rel Freq	Weibull Model Prob's
0.16-0.20	3	0.008219	0.0082191	0
0.21-0.25	5	0.013698	0.0219178	0
0.26-0.30	6	0.016438	0.0383561	.0005
0.31-0.35	8	0.021917	0.0602739	.0022
0.36-0.40	12	0.032876	0.0931506	.0076
0.41-0.45	16	0.043835	0.1369863	.0233
0.46-0.50	24	0.065753	0.2027397	.0605
0.51-0.55	39	0.106849	0.3095890	.1332
0.56-0.60	51	0.139726	0.4493150	.2367
0.61-0.65	106	0.290410	0.7397260	.2811
0.66-0.70	84	0.230136	0.9698630	.2000
0.71-0.75	11	0.030136	1	.0540

Data relative frequency vs. model, Weibull(Shape=9, Scale =0.63), cumulative distribution values.

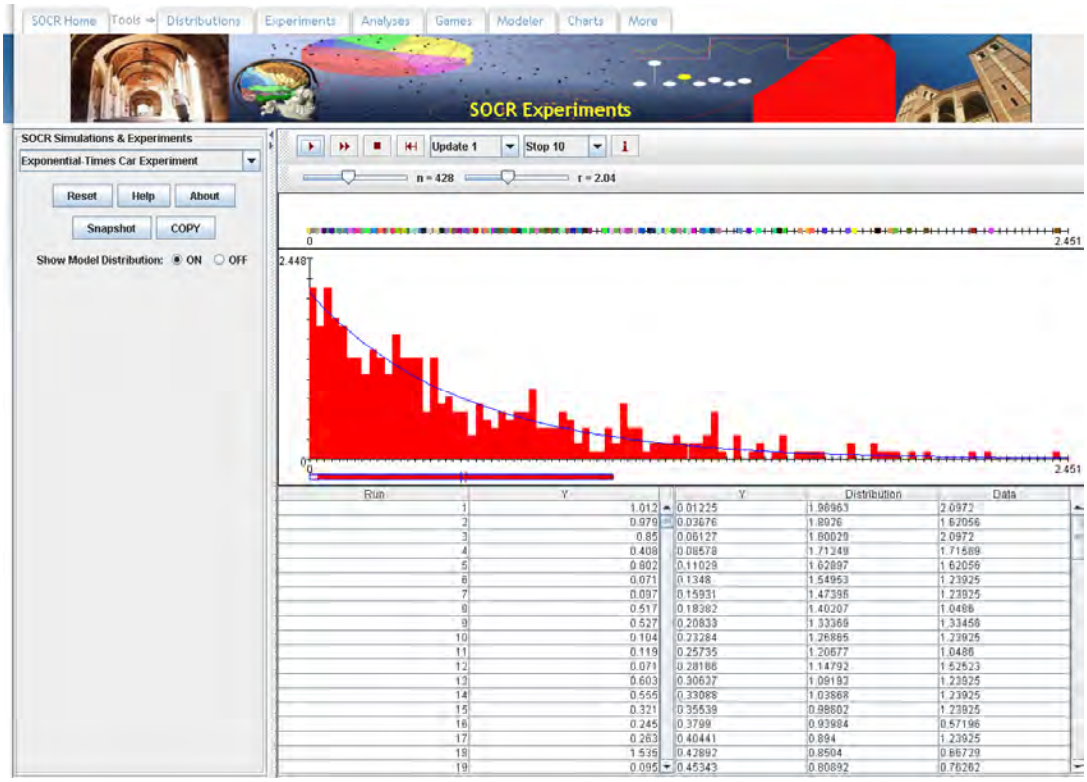


Comparison between the data histogram and the graph of the model density function (Weibull).

SOCR Experiments

SOCR Experiments include a large number of virtual trials, computer games and simulated studies that demonstrate specific real-world processes, and can be used for data simulation, model fitting and assessment. SOCR Experiments extend the Virtual Laboratory in Probability and Statistics classes (www.math.uah.edu/STAT). As with SOCR Distributions, Experiments have 4 types of associated resources:

- applets: www.socr.ucla.edu/htmls/SOCR_Experiments.html
- activities: http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_ExperimentsActivities
- computational libraries: www.socr.ucla.edu/htmls/SOCR_Download.html
- usage documentation: www.socr.ucla.edu/docs/



Basic Operations:

The basic operations allowed via the SOCR Experiments applet are:

- Selection of an experiment of interest, and the corresponding parameters.
- Conducting a single (step) or a series (run) of experiments.
- Studying the relation between the model and data distributions.
- Graphical observation of the theoretical and sample distributions.

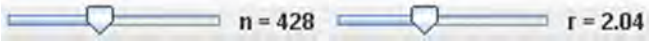
Controls:

All SOCR experiments have the following controls:

- General controls, which allow the user to reset the experiment, obtain more information and help about the experiment, take a snapshot of an experiment applet, copy data and results from the experiment to the mouse buffer (clipboard) and enable the display of the model distribution.

- Action controls allow performing the experiment once or many times and controlling the frequency of the reported outcomes of the experiment.

In addition, each individual experiment contains specific parameter settings and controls that effect only the concrete process modeled by the experiment. These controls vary widely between different experiments.



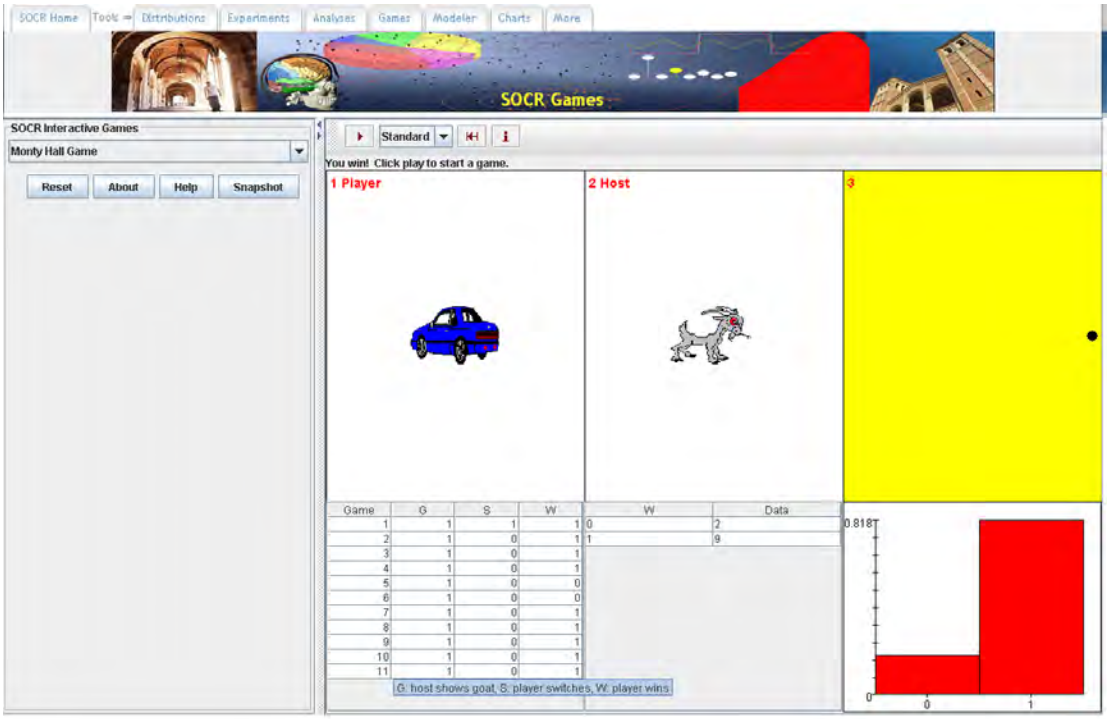
Classroom use examples:

- Run a SOCR Experiment 100 times (e.g., virtual Roulette experiment). Formulate an event of interest (e.g., a red outcome). Compare the number of observed and expected number of outcomes in this experiment. Do these numbers become more or less similar as the number of experiments increases?
- Run a SOCR Experiment 1,000 times and record the results (simulated data). Study the results using graphical exploratory data analysis (EDA) techniques.
- Chose an appropriate experiment, propose reasonable research hypotheses about the outcome of this experiment and try to empirically validate or disprove these hypotheses by running the virtual experiment a large number of times.

SOCR Games

SOCR Games consist of a set of applets that illustrate interactively some real-world games, e.g., Monty Hall (three-door) game. As with SOCR Distributions, the Games applet also has 4 types of associated resources:

- applets: www.socr.ucla.edu/htmls/SOCR_Games.html
- activities: http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_GamesActivities
- computational libraries: www.socr.ucla.edu/htmls/SOCR_Download.html
- usage documentation: www.socr.ucla.edu/docs/



Basic Operations:

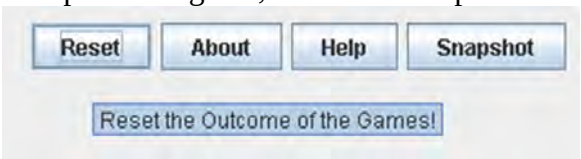
The basic operations allowed in the SOCR Games applet are:

- Selection of a game.
- Playing a game.
- Graphical observation of the game outcomes.

Controls:

All SOCR games have the following controls:

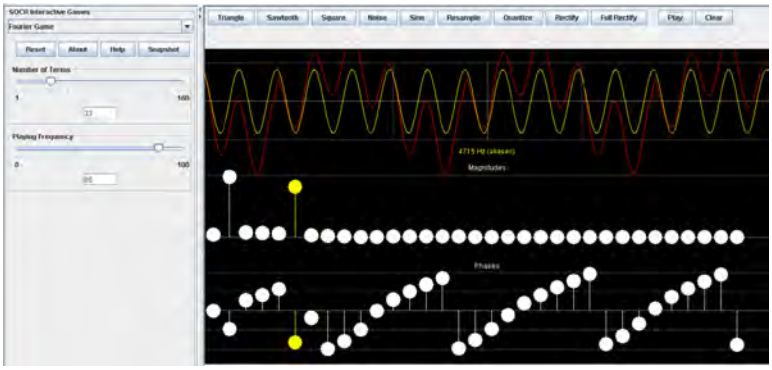
- General controls, which allow the user to reset a game, obtain more information and help about a game, and take a snapshot of the state of a game.



- Action controls vary among the different types of games, based on their functionality.

Classroom use examples:

- Interactively run a SOCR game a few times to understand the process of interest. Then extend this understanding by automatically running its analogous SOCR experiment (e.g., Monty Hall Game/Experiment).
- Understanding the interactive effects of Fourier and spectral methods for data representation.



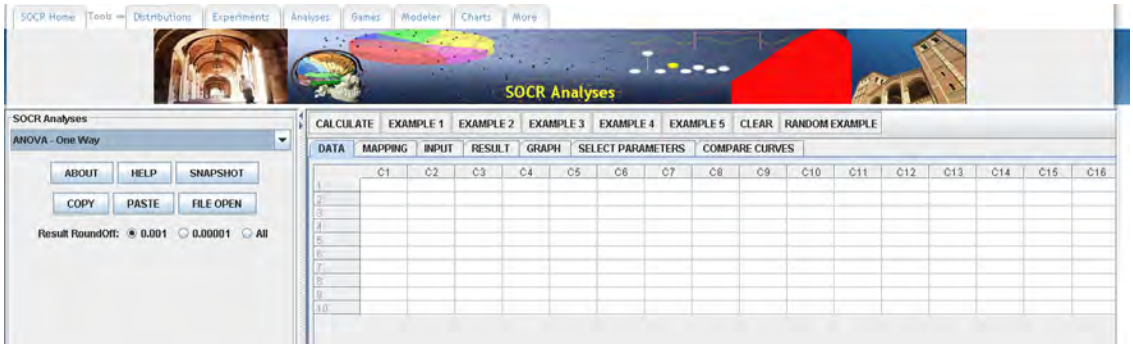
SOCR Analyses

SOCR Analyses (Che et al., 2009) applets have four core components:

- Linear models: simple linear regression, multiple linear regression, one-way and two-way ANOVA.
- Tests for sample comparisons: parametric t-test, and the non-parametric Wilcoxon rank sum test, Kruskal-Wallis test, Friedman's test, Kolmogorov-Smirnoff test and Fligner-Killeen test.
- Hypothesis testing models: contingency tables, Friedman's test and Fisher's exact test.
- Power Analysis: utility for computing sample sizes for the Normal distribution.

The Analyses applets also have 4 types of associated resources:

- applets: www.socr.ucla.edu/htmls/SOCR_Analyses.html
- activities: http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_AnalysesActivities
- computational libraries: www.socr.ucla.edu/htmls/SOCR_Download.html
- usage documentation: www.socr.ucla.edu/docs/



Basic Operations:

The basic operations allowed in the SOCR Analyses applet are:

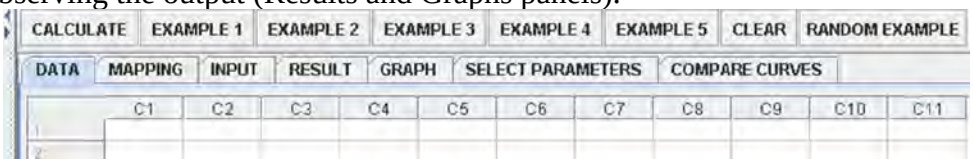
- Selection of an analysis.
- Specification of the input data spreadsheet.
- Mapping the input data (e.g., clearly delineating the dependent and independent variables as columns).
- Calculating the results of the analysis.
- Observing tabular and graphical output, as appropriate.

Controls:

All SOCR analyses have the following controls:

- General controls, which allow the user to get information about the analysis, take a snapshot of the applet, import/export data and specify the result precision.

- Action controls are similar for most analyses, including the functionality to specify the appropriate data for the analysis, computing the analysis and observing the output (Results and Graphs panels).



Some Analyses have a very different schema for data entry and result interpretation (e.g., Normal Power Analysis).

Data Input and Result Output:

SOCR Analyses allow 2 types of data input – spreadsheet copy-paste data from external sources in to the SOCR Analysis data table (using the Copy/Paste buttons), and importing data from a local file. The latter requires an ASCII input file that has a format similar to any of the example datasets provided for this specific type of analysis. Column heading row should start with the # symbol.

# Days	Eth	Sex	Age	Lrn
2.0	A	M	F0	SL
11.0	A	M	F0	SL
14.0	A	M	F0	SL
5.0	A	M	F0	AL
5.0	A	M	F0	AL
13.0	A	M	F0	AL
20.0	A	M	F0	AL
22.0	A	M	F0	AL
6.0	A	M	F1	SL
6.0	A	M	F1	SL
15.0	A	M	F1	SL
7.0	A	M	F1	AL
14.0	A	M	F1	AL
6.0	A	M	F2	SL
32.0	A	M	F2	SL
53.0	A	M	F2	SL
57.0	A	M	F2	SL
14.0	A	M	F2	AL
16.0	A	M	F2	AL
16.0	A	M	F2	AL
17.0	A	M	F2	AL
40.0	A	M	F2	AL
43.0	A	M	F2	AL
46.0	A	M	F2	AL
8.0	A	M	F3	AL

Output results can be obtained from the Results and Graph tab, by copy-and-paste, drag-and-drop or right-click-and-save functionality. There are some platform dependencies on these behaviors (especially for Apple Macintosh computers).

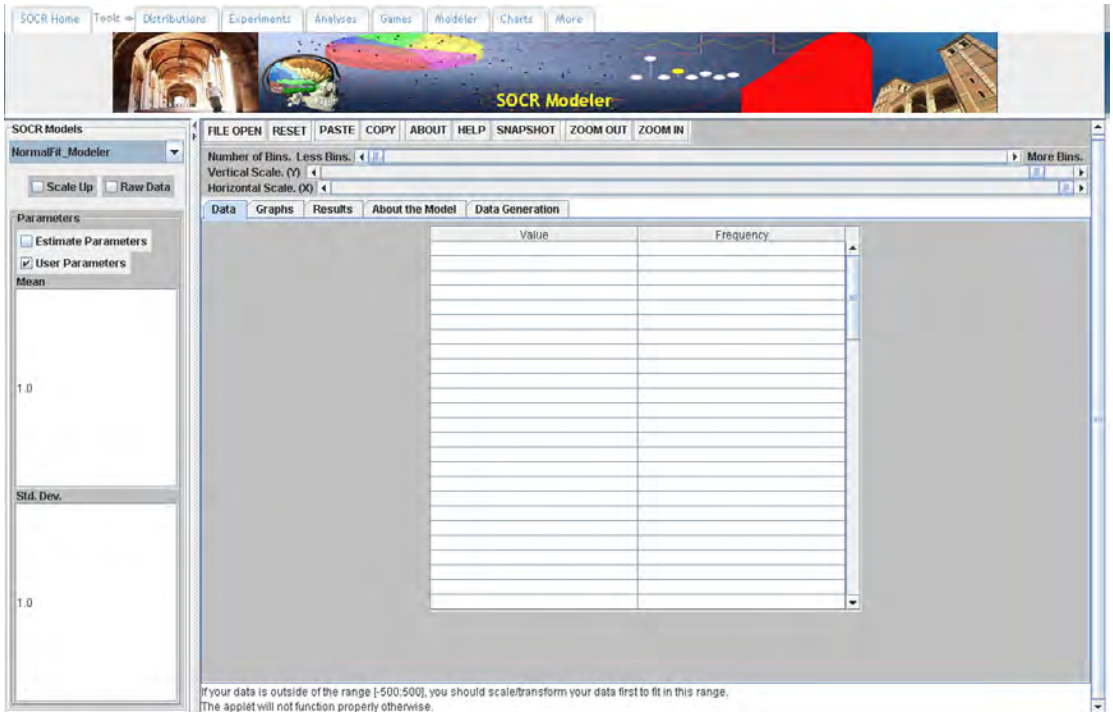
Classroom use examples:

- Select a SOCR dataset (http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data) appropriate for the specific type of analysis. Import the data into SOCR Analysis spreadsheet. Map the corresponding columns, as appropriate, and click CALCULATE to compute the analysis and obtain the results. The data choice should be made based on the course goals, student maturity level, technical expertise and interests.
- Generate random data, as discussed above, and plug it into an analysis – show that there should not be any significant data effects, unless the random sampling was chosen so as to demonstrate a specific effect.
- Complete the results of parametric and non-parametric test analogues on the same datasets. Discuss the need and validation of parametric assumptions in terms of the power of the tests.
- Demonstrate the agreement in the results between manually- and SOCR-computed analyses (e.g., regression or t-test).

SOCR Modeler

SOCR Modeler provides two core functions – the ability to sample from any SOCR Distribution, and the utility to fit polynomial, distribution of spectral models to user provided data. The Modeler applet also has 4 types of associated resources:

- applets: www.socr.ucla.edu/htmls/SOCR_Modeler.html
- activities: http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_ModelerActivities
- computational libraries: www.socr.ucla.edu/htmls/SOCR_Download.html
- usage documentation: www.socr.ucla.edu/docs/



Basic Operations:

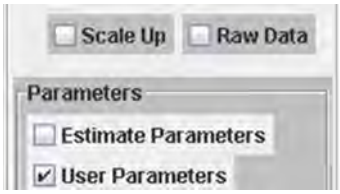
The basic operations allowed via the SOCR Modeler applet are:

- Data simulation, sampling, from any of the SOCR Distributions.
- Fitting models to data specified numerically, graphically or by random sampling.
- Exploring graphically and quantitatively the quality of model fitting.

Controls:

The SOCR Modeler applet has the following controls:

- General controls, which allow the user to scale up the model distribution (for a better graphical fit with the data histogram), specify if the data represents raw measurements of frequencies, and select automated (estimated) or manual (user-specified) model parameters.



- Action controls allow entering data (via copy-paste, by graphical mouse clicks in the Graph canvas, or from a local file), taking a snapshot of the modeler applet state, and sliders for controlling the graphical appearance of the data and model fit.

Some modeler applets have a very different appearance, because of the nature of their data manipulations, e.g., Fourier and Wavelet modelers.

Data import and result export:

The Modeler allows 4 modes of data import:

- Random number generation (Data Generation tab), where the use specifies a desired distribution model and the sample-size.
- Manual mouse clicks in the Graphing canvas will generate frequency distributions according to the user's input.
- Copy-and-paste data into the Data tab from other SOCR applets, web-pages or spreadsheets.
- File input using the FILE OPEN button.

The results of the model fitting (e.g., maximum likelihood estimates for the distribution parameters) are reported in the Results tab and can be copy-pasted in external documents. The Graph panel contains a visual representation of the quality of the model fit. The SNAPSHOT button can be used to save the results as a static image to a local file.

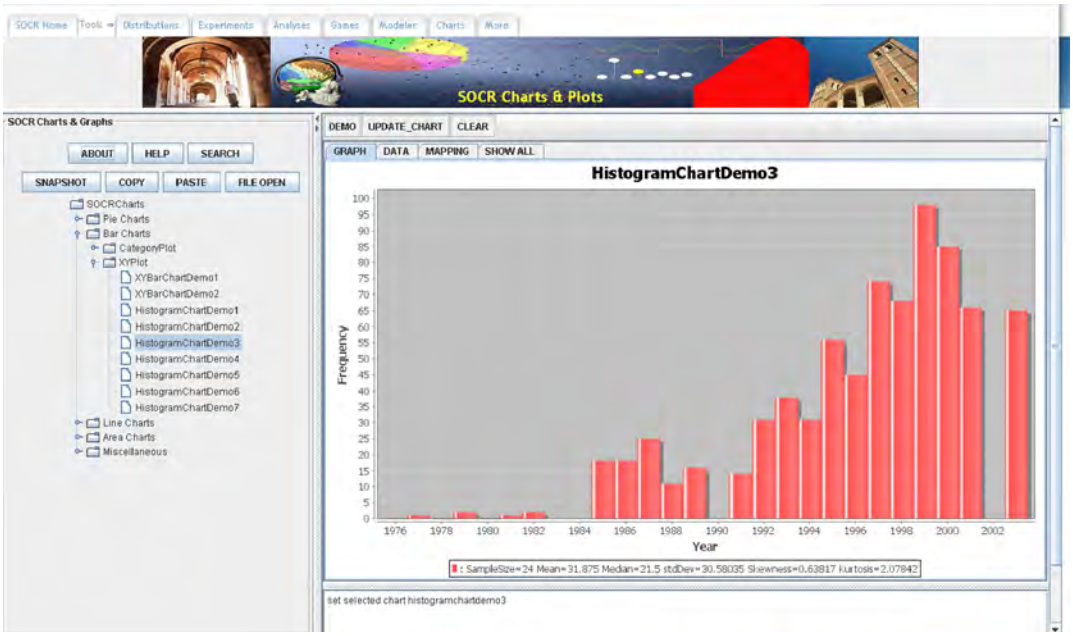
Classroom use examples:

- Generate a random sample of size 1,000 from Normal (μ=2.3, σ^2=9) distribution and fit a Normal or a Beta model to these data.
- Generate 1,000 random Cauchy observations and fit a Normal model to this heavy-tail sample. Discuss the problems with the model fit.
- Run a SOCR Experiment 1,000 times and record the results (simulated data). Fit an appropriate SOCR model to these data and provide analytical (model-based) estimations for various events of interest. Contrast these model estimates with the corresponding empirical estimates.
- Manually click on the Graph canvas to generate data and test various model fits.
- Try the SOCR Polynomial model fit (http://www.socr.ucla.edu/Applets.dir/SOCRCurveFitter.html).
- Illustrate the SOCR Mixture model (www.socr.ucla.edu/htmls/mod/MixFit_Modeler.html) on various SOCR dataset (http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data).

SOCR Charts

SOCR Charts include over 70 different types of graphs, charts and plots which are useful in exploratory data analysis (EDA), model validation and data understanding. SOCR Charts also have 4 types of associated resources:

- applets: www.socr.ucla.edu/htmls/SOCR_Charts.html
- activities: http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_ChartsActivities
- computational libraries: www.socr.ucla.edu/htmls/SOCR_Download.html
- usage documentation: www.socr.ucla.edu/docs/



Basic Operations:

The basic SOCR Charts operations include:

- Plotting, graphing and charting univariate and multivariate data for one or more series.
- Computing the summary statistics for multivariate data.
- Graphical exploration of the relations within one variable or between several variables.

Controls:

All SOCR Charts have the following controls:

- General controls, enable users to find out more about a specific type of chart, navigate and discover appropriate charts, take a snapshot of the state of a chart and import or export data (from a file or via the mouse buffer).

- Action controls include the selection of a default demonstration dataset, appropriate for the specific chart, refreshing/redrawing/updating the chart and resetting/clearing the chart and data.

Some Charts have additional action controls like sliders and buttons that allow the user to connect to external activities, set additional parameters (e.g., histogram bin-size) or manipulate the data (e.g., power-transformation). In addition, each chart is interactive and includes pop-up controls for setting the chart appearance and controls. SOCR Charts are based on JFreeCharts (www.jfree.org/jfreechart/).

Data import and result export:

SOCR Charts provide 2 modes of data import:

- Copy-and-paste data into the Charts Data tab from other SOCR applets, web-pages or external spreadsheets.
- File input using the FILE OPEN button.

The SOCR Charts results include mostly graphs and plots, but occasionally also include processed data (e.g., power transformation). As with the other SOCR applets, graphs can be saved using either the SNAPSHOT button, or via right-click-and-save in the Graphing canvas.

Classroom use examples:

The widely diverse types of SOCR Charts make them useful for multiple purposes, which oftentimes arise in probability and statistics classes. Some examples of these include:

- Validation of parametric assumptions on the data – use the QQ plot.
- Demonstrate dot-plots, box-and-whisker plots, line and index charts, histograms, bubble plots, etc.
- The SOCR Expectation maximization chart provides a nice 2D example of model fitting and data classification of special data.
- General exploratory data analysis.
- Plotting of residuals to assess quality of linear models.
- Compute summary statistics for multivariate data sets.

Day 2: Tue 08/11/09

Morning Session: SOCR Activities

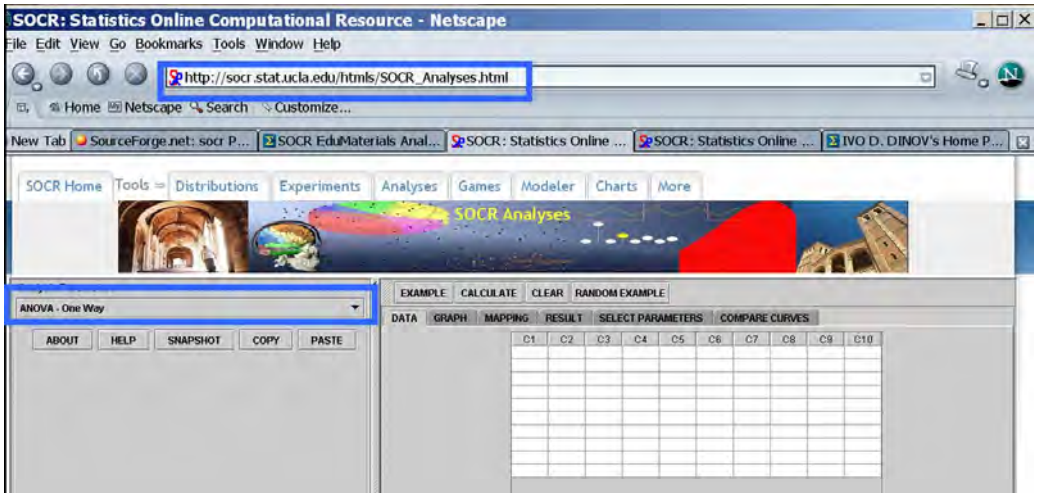
Analysis Activities

- Analysis of Variance (ANOVA)

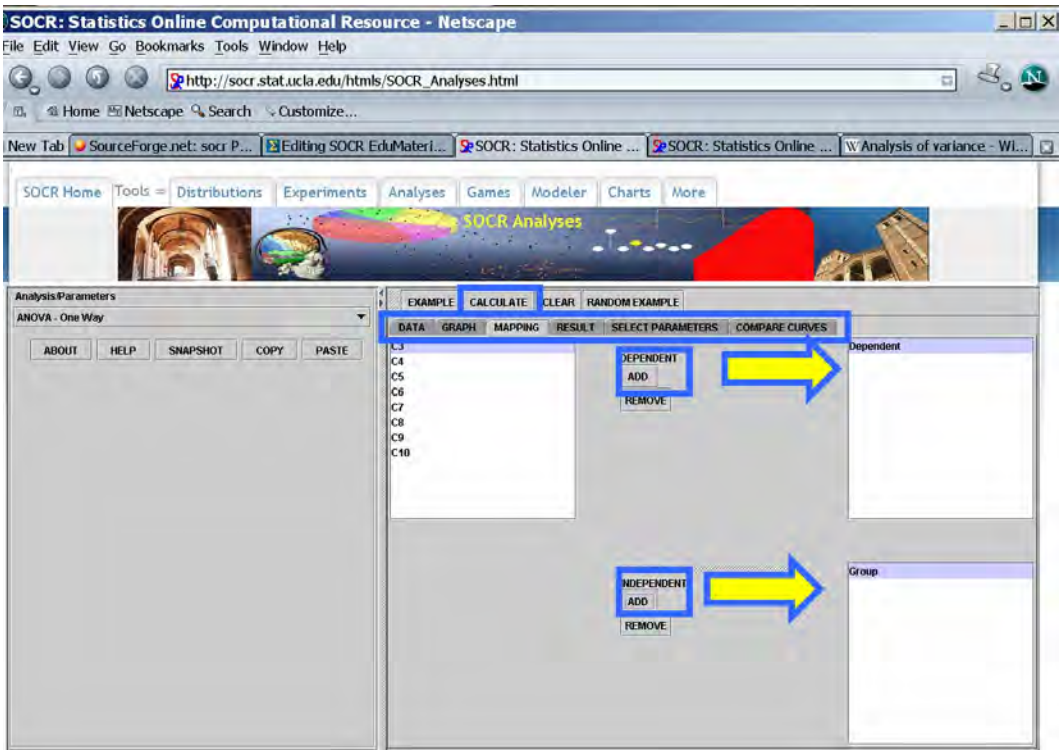
  This SOCR Activity demonstrates the utilization of the SOCR Analyses package for statistical Computing. In particular, it shows how to use Analysis of Variance (ANOVA) and how to interpret the results.

  - **ANOVA Background**: Analysis of variance (ANOVA) is a class of statistical analysis models and procedures, which compare means by splitting the overall observed variance into different parts. The initial techniques of the analysis of variance were pioneered by the statistician and geneticist R. A. Fisher in the 1920s and 1930s, and are sometimes known as Fisher's ANOVA or Fisher's analysis of variance, due to the use of Fisher's F-distribution as part of the test of statistical significance. Read more about ANOVA here (http://en.wikipedia.org/wiki/ANOVA).

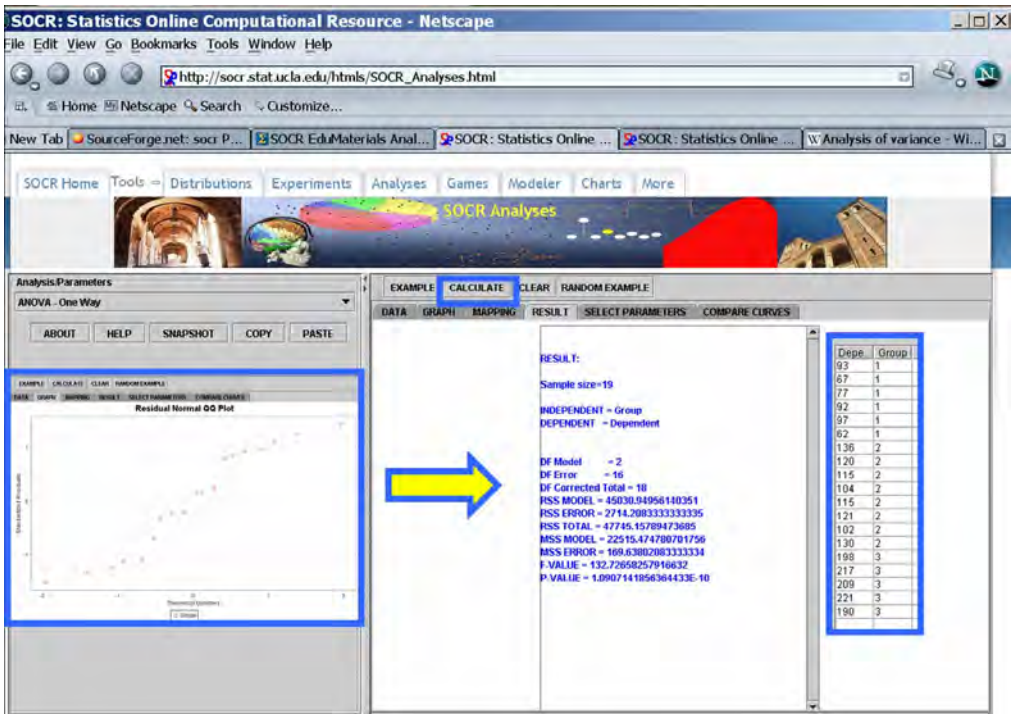
  - **SOCR ANOVA**: Go to SOCR Analyses (www.socr.ucla.edu/htmls/SOCR_Analyses.html) and select **One-way ANOVA** from the drop-down list of SOCR analyses, in the left panel. There are three ways to enter data in the SOCR ANOVA applet:
    - Click on the **Example** button on the top of the right panel.
    - Generate random data by clicking on the **Random Example** button.
    - Pasting your own data from a spreadsheet into SOCR ANOVA data table.



  - Now, map the dependent and independent variables by going to the **Mapping** tab, selecting columns from the available list and sending them to the corresponding bins on the right (see figure). Then press **Calculate** button to carry out the ANOVA analysis.



- The quantitative results will be in the tab labeled **Results**. The **Graphs** tab contains the QQ Normal plot for the residuals. In this case, we have a very significant grouping effect, indicated by the *p-value* $< 10^{-4}$.



- Now try the ANOVA applet on some of the SOCR Datasets (http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data) – e.g., CPI index, Hot-dogs sodium/calorie dataset, Allometric Relations in Plants, etc.

- Simple Linear Regression

This SOCR Activity demonstrates the utilization of the SOCR Analyses package for statistical Computing. In particular, it shows how to use Simple Linear Regression and how to interpret the results. Simple Linear Regression (SLR) is a class of statistical analysis models and procedures which takes one independent variable and one dependent variable, both being quantitative, and models the relationship between them. The model form is:

$$y = intercept + slope * x + error,$$

where x denotes the independent variable and y denotes the dependent variable. So it is linear. The error is assumed to follow the Normal distribution.

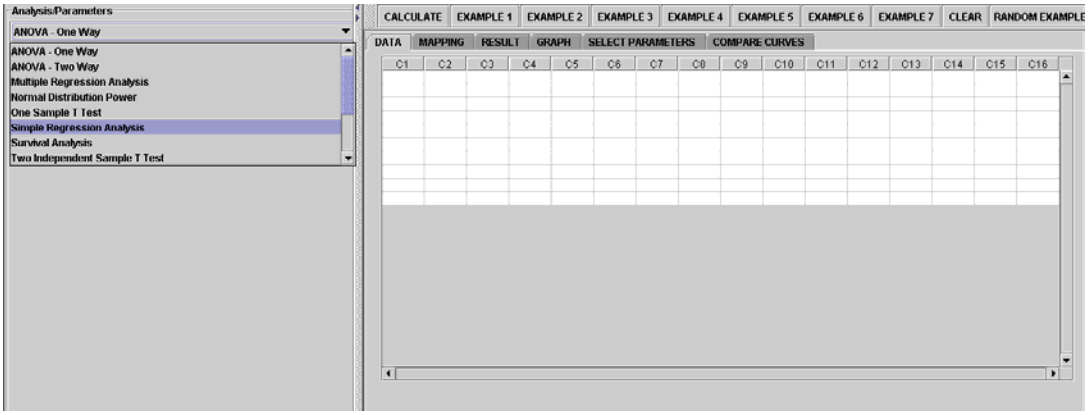
The goal of the Simple Linear Regression computing procedure is to estimate the intercept and the slope, based on the data. Least Squares Fitting is used.

In this activity, the students can learn about:
- Reading results of Simple Linear Regression.
- Interpreting the slope and the intercept.
- Observing and interpreting various data and resulting plots.
  - Scatter plots of the dependent vs. independent variables.
  - Diagnostic plots such as the Residual on Fit plot.
  - Normal QQ plot, etc.

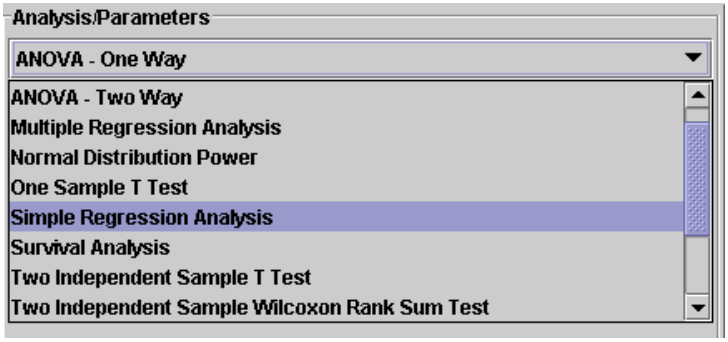
Go to SOCR Analyses (www.socr.ucla.edu/htmls/SOCR_Analyses.html) and select **Simple Linear Regression** from the drop-down list of SOCR analyses, in the left panel. There are three ways to enter data in the SOCR Simple Linear Regression applet:
- Click on the **Example** button on the top of the right panel.
- Generate random data by clicking on the **Random Example** button.
- Paste your own data from a spreadsheet into SOCR Simple Linear Regression data table.

We will demonstrate SLR with some SOCR built-in examples. The first example (EXAMPLE 1) is based on the data taken from "An Introduction to Computational Statistics: Regression Analyses" by Robert Jennrich, Prentice Hall, 1995 (Page 4). The data describe exam and homework scores of a class of students, where **M** stands for midterm, **F** for final, and **H** for homework.

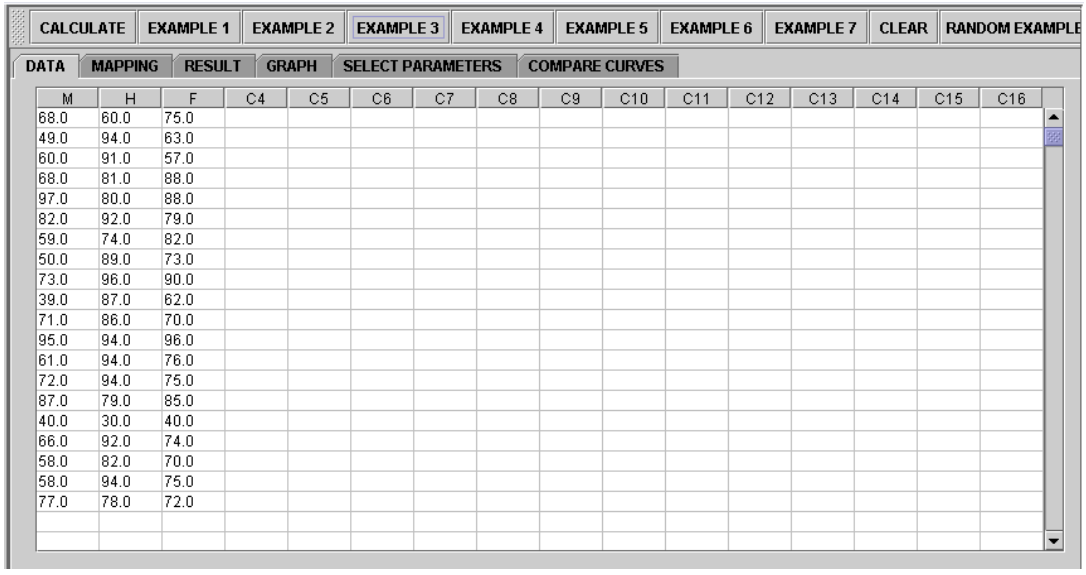
- As you start the SOCR Analyses Applet (www.socr.ucla.edu/htmls/SOCR_Analyses.html), click on **Simple Linear Regression** from the combo box in the left panel. Here's what the screen should look like.



- The left part of the panel looks like this (make sure that the "*Simple Linear Regression*" is showing in the drop-down list of analyses, otherwise you won't be able to find the correct dataset and will not be able to reproduce the results!)

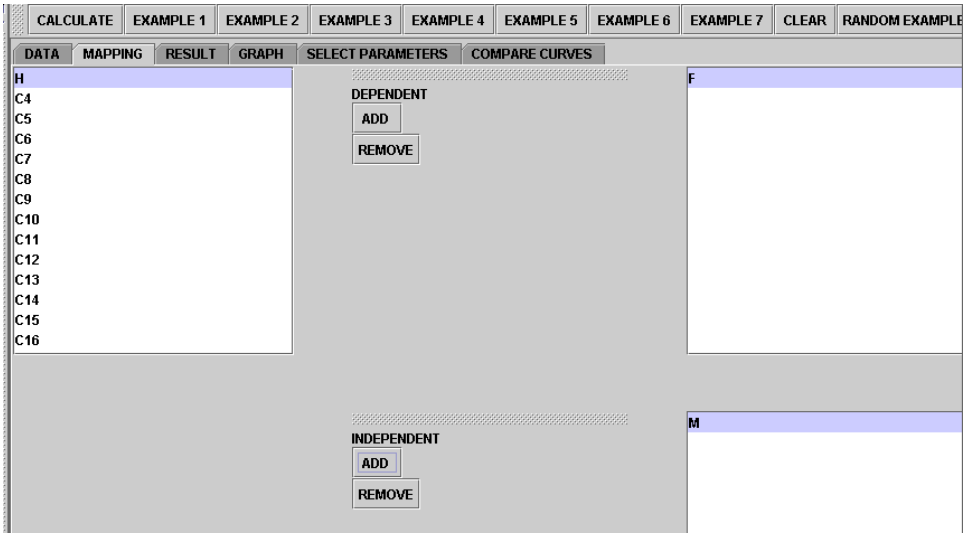


- In the SOCR SLR analysis, there are several SOCR built-in examples. In this activity, we'll be using **Example 3**. Click on the "**Example 3**" button and next, click on the "**Data**" button in the right panel. You should see the data displayed in two columns. There are three columns here, **M**, **F** and **H**.



M	H	F	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16
68.0	60.0	75.0													
49.0	94.0	63.0													
60.0	91.0	57.0													
68.0	81.0	88.0													
97.0	80.0	88.0													
82.0	92.0	79.0													
59.0	74.0	82.0													
50.0	89.0	73.0													
73.0	96.0	90.0													
39.0	87.0	62.0													
71.0	86.0	70.0													
95.0	94.0	96.0													
61.0	94.0	76.0													
72.0	94.0	75.0													
87.0	79.0	85.0													
40.0	30.0	40.0													
66.0	92.0	74.0													
58.0	82.0	70.0													
58.0	94.0	75.0													
77.0	78.0	72.0													

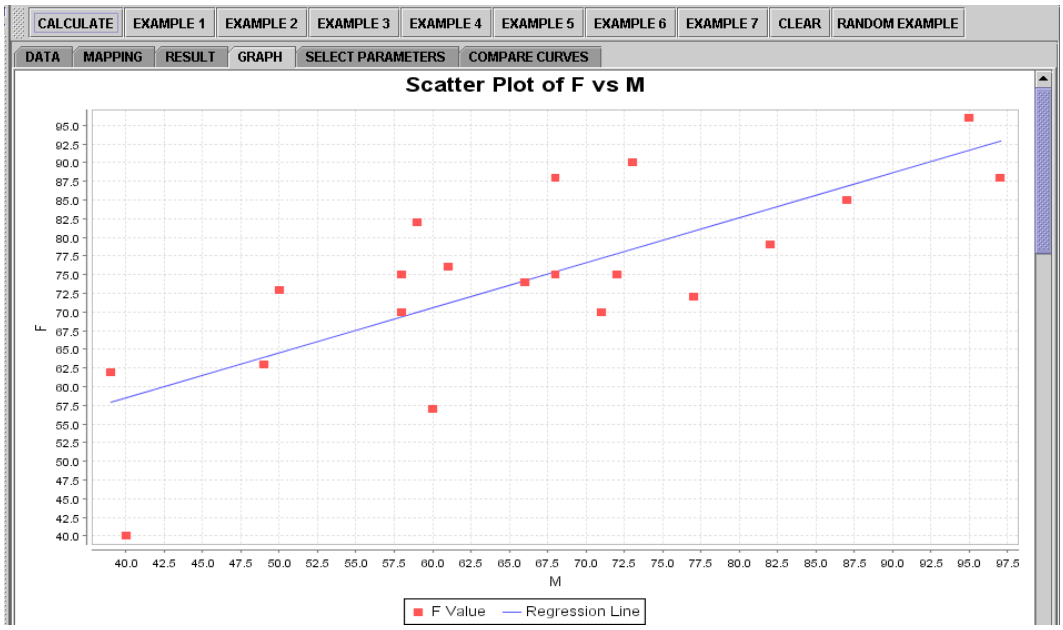
- Use column **M** as the regressor (that is, 'x', the independent variable) and column **F** as the response (that is, 'y', the dependent variable). So you can simply ignore the column **H** for this activity. To tell the computer which variables are assigned to be the regressor and response, we have to do a "Mapping." This is done by clicking on the "**Mapping**" button first to get to the Mapping Panel, and then mapping the variables. For this Simple Linear Regression activity, there are two places the variables can be mapped. The top part says **DEPENDENT** that you'll need to **map** the dependent variable you want here. Just click on **ADD** under **DEPENDENT** and that will do it. If you change your mind, you can click on **REMOVE**. Similar for the **INDEPENDENT** variable. Once you get the screen to look like the screenshot below, you're done with the **Mapping** step. (Note that, since the columns C3 through C16 do not have data and they are not used, just ignore them.)

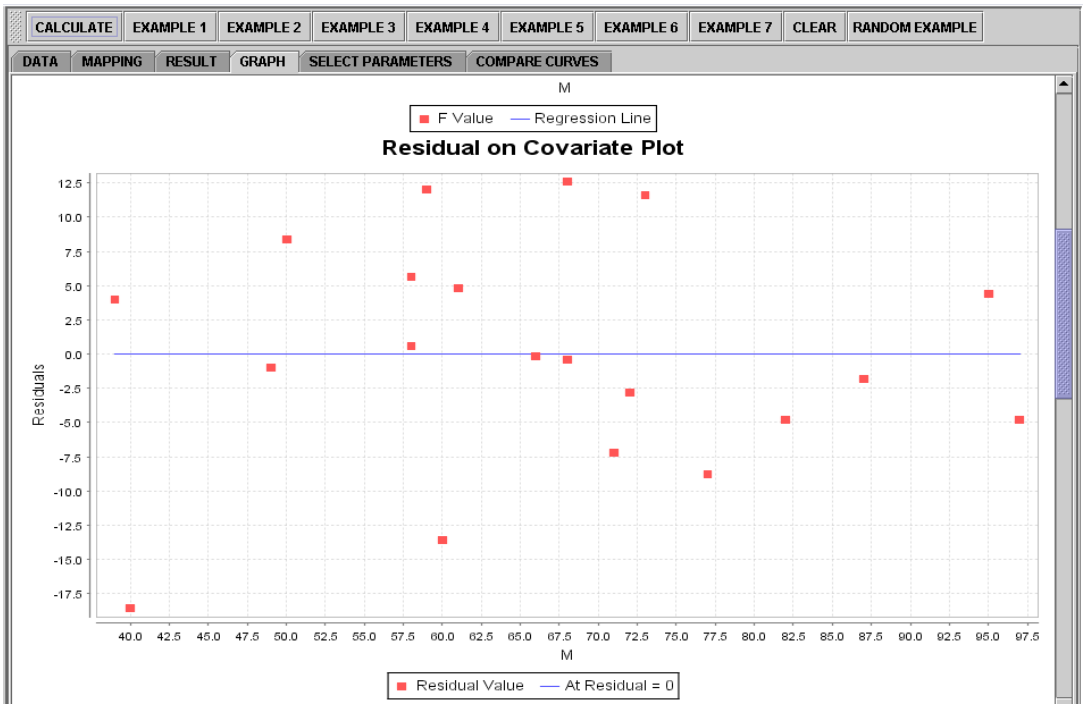


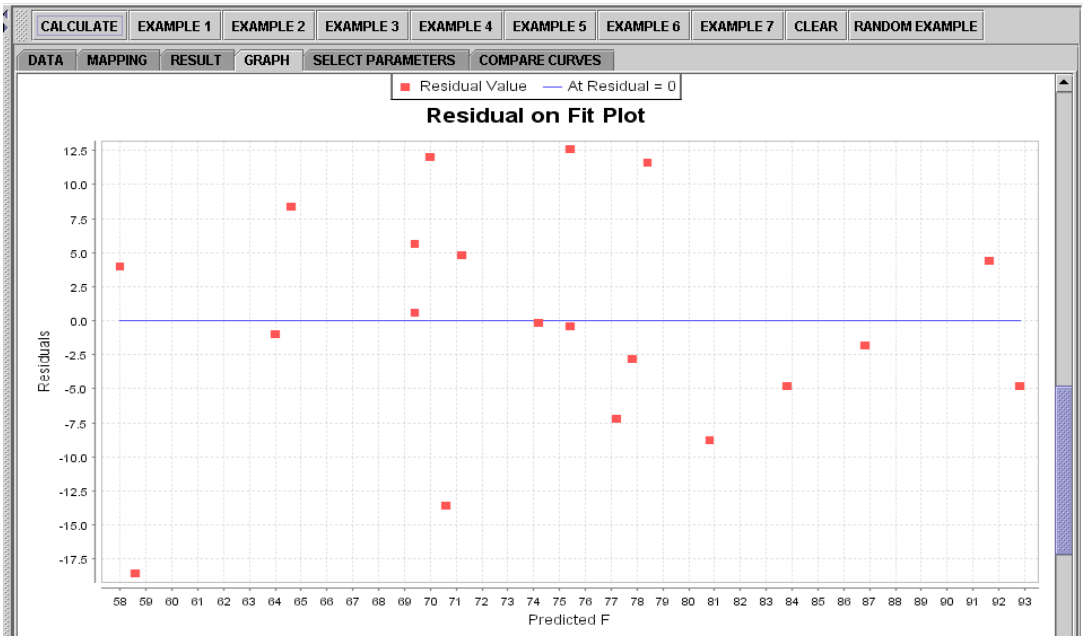
- After we do the "Mapping" to assign variables, now we use the computer to calculate the regression results – click on the "**Calculate**" button. Then select the "**Result**" panel to see the output.

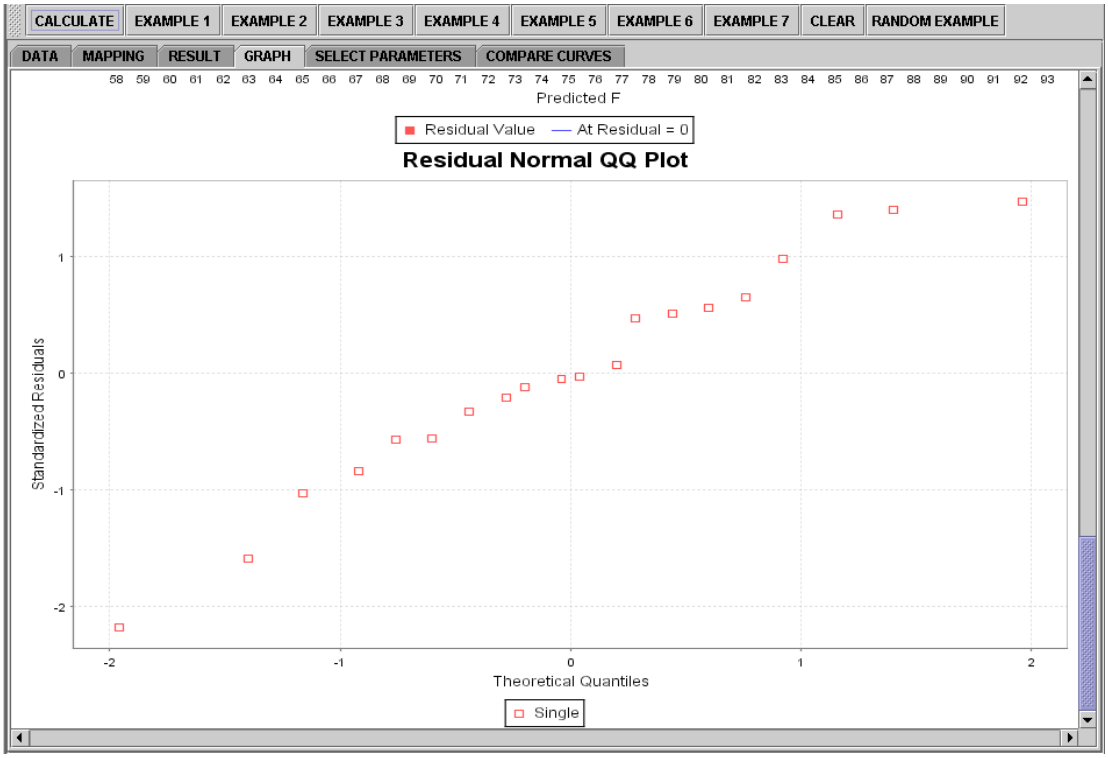
The text in the Result Panel summarizes the results of this simple linear regression analysis. The regression line is displayed. At this point you can think about how the **dependent variable** changes, on average, in response to changes of the **independent variable**.

- If you'd like to see the graphical component of this analysis, click on the "**Graph**" panel. You'll then see the graph panel that displays the scatter plot, as well as diagnostic plots of "residual on fit", "Normal QQ" plots, etc. The plot titles indicate plot types.




CALCULATE
EXAMPLE 1
EXAMPLE 2
EXAMPLE 3
EXAMPLE 4
EXAMPLE 5
EXAMPLE 6
EXAMPLE 7
CLEAR
RANDOM EXAMPLE
DATA
MAPPING
RESULT
GRAPH
SELECT PARAMETERS
COMPARE CURVES
M
F Value
Regression Line
Residual on Covariate Plot
Residuals
M
Residual Value
At Residual = 0


CALCULATE
EXAMPLE 1
EXAMPLE 2
EXAMPLE 3
EXAMPLE 4
EXAMPLE 5
EXAMPLE 6
EXAMPLE 7
CLEAR
RANDOM EXAMPLE
DATA
MAPPING
RESULT
GRAPH
SELECT PARAMETERS
COMPARE CURVES
Residual Value
At Residual = 0
Residual on Fit Plot
Residuals
Predicted F



Note: If you happen to click on the "**Clear**" button in the middle of the procedure, **all the data will be cleared out**. Simply start over from step 1.

Modeler Activities

- SOCR Normal & Beta Distribution Model Fitting.

This activity describes the process of SOCR model fitting in the case of using Normal or Beta distribution models. *Model fitting* is the process of determining the parameters for an analytical model in such a way that we obtain optimal parameter estimates according to some criterion. There are many strategies for parameter estimation. The differences between most of these are the underlying cost-functions and the optimization strategies applied to maximize/minimize the cost-function.

The aims of this activity are to:

- motivate the need for (analytical) modeling of natural processes.
- illustrate how to use the SOCR Modeler to fit models to real data.
- present applications of model fitting.

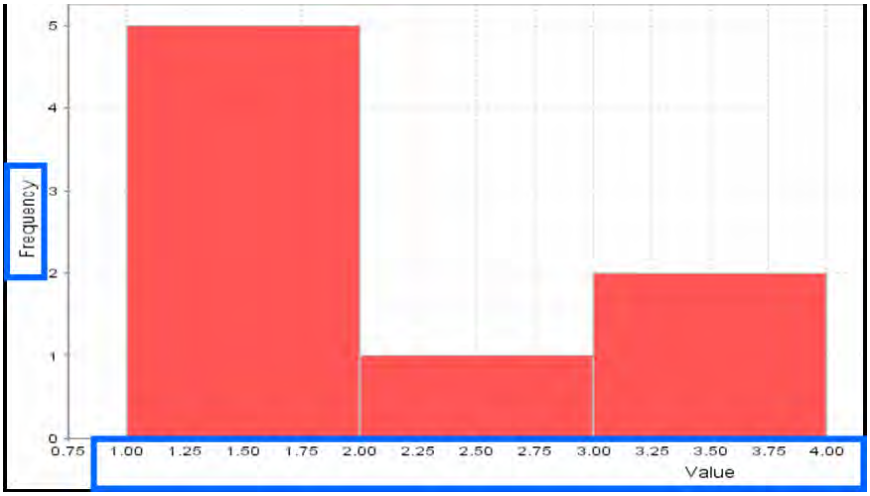
Background & Motivation

Suppose we are given the sequence of numbers {1, 2, 3, 4, 5, 6, 7, 8, 9, 10} and asked to find the best (Continuous) Uniform Distribution that fits that data. In this case, there are two parameters that need to be estimated - the minimum (m) and the maximum (M) of the data. These parameters determine exactly the support (domain) of the continuous distribution and we can explicitly write the density for the (best fit) continuous uniform distribution as:

$$f(x) = \begin{cases} \dfrac{1}{M-m}, & m \le x \le M \\ 0, & x < m \text{ or } x > M \end{cases}$$

Having this model distribution, we can use its analytical form, $f(x)$, to compute probabilities of events, critical functional values and, in general, do inference on the native process without acquiring additional data. Hence, a good strategy for model fitting is extremely useful in data analysis and statistical inference. Of course, any inference based on models is only going to be as good as the data and the optimization strategy used to generate the model.

Let's look at another motivational example. This time, suppose we have recorded the following (sample) measurements from some process {1.2, 1.4, 1.7, 3.4, 1.5, 1.1, 1.7, 3.5, 2.5}. Taking bin-size of 1, we can easily calculate the frequency histogram for this sample, {6, 1, 2}, as there are 6 observations in the interval [1:2), 1 measurement in the interval [2:3) and 2 measurements in the interval [3:4).



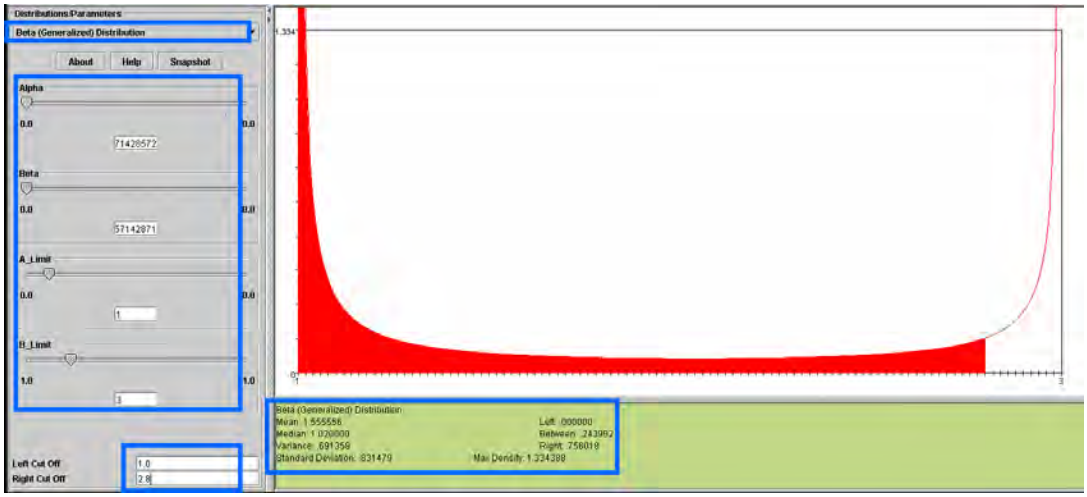
We can now ask about the *best Beta distribution model fit to the histogram of the data*!

Most of the time when we study natural processes using probability distributions, it makes sense to fit distribution models to the frequency histogram of a sample, not the actual sample. This is because our general goals are to model the behavior of the native process, understand its distribution and quantify likelihoods of various events of interest (e.g., in terms of the example above, we may be interested in the probability of observing an outcome in the interval [1.50:2.15) or the chance that an observation exceeds 2.8).

Exercise 1

Let's first solve the challenge we presented in the background section, where we calculated the frequency histogram for a sample to be {6, 1, 2}. Go to the SOCR Modeler (www.socr.ucla.edu/htmls/SOCR_Modeler.html) and click on the **Data** tab. Paste in the two columns of data. Column 1 {1, 2, 3} - these are the ranges of the sample values and correspond to measurements in the intervals [1:2), [2:3) and [3:4). The second column represents the actual frequency counts of measurements within each of these 3 histogram bins - these are the values {6, 1, 2}. Now press the **Graphs** tab. You should see an image like the one below. Then choose **Beta_Fit_Modeler** from the drop-down list of models in the top-left and click the estimate parameters check-box, also on the top-left. The graph now shows you the best Beta distribution model fit to the frequency histogram {6, 1, 2}. Click the **Results** tab to see the actual estimates of the two parameters of the corresponding Beta distribution (*Left Parameter (α) = 0.0446428571428572; Right Parameter (β)= 0.11607142857142871; Left Limit = 1.0; Right Limit = 3.0*).

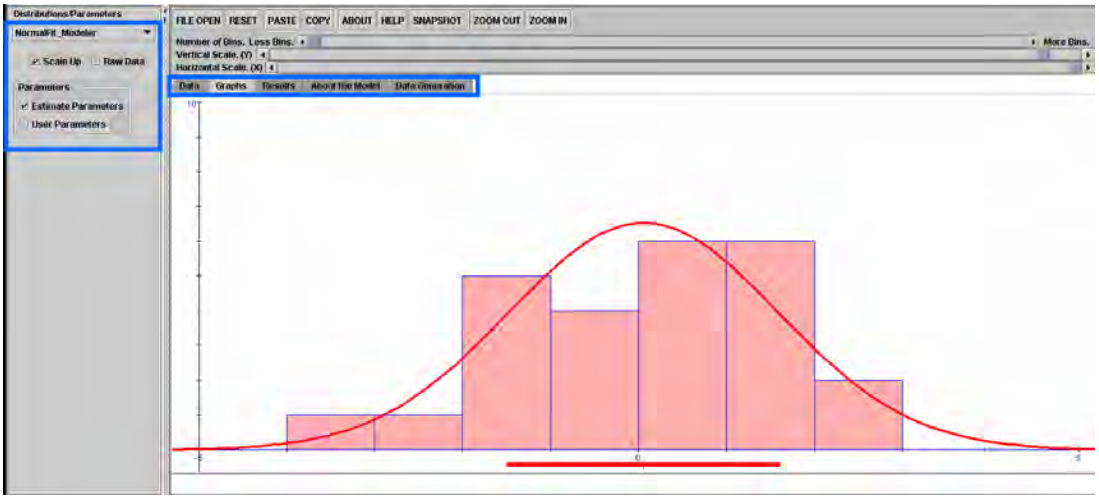
You can also see how the (general) Beta distribution degenerates to this shape by going to SOCR Distributions (www.socr.ucla.edu/htmls/SOCR_Distributions.html), selecting the **(Generalized) Beta Distribution** from the top-left and setting the 4 parameters to the 4 values we computed above. Notice how the shape of the Beta distribution changes with each change of the parameters. This is also a good demonstration of why we did the distribution model fitting to the frequency histogram in the first place - precisely to obtain an analytic model for studying the general process without acquiring mode data. Notice how we can compute the odds (probability) of any event of interest, once we have an analytical model for the distribution of the process. For example, this figure depicts the probabilities that a random observation from this process exceeds 2.8 (the right-limit). This probability is computed to be 0.756.



Exercise 2

Go to the SOCR Modeler (www.socr.ucla.edu/htmls/SOCR_Modeler.html) and select the **Graphs** tab and click the "Scale Up" check-box. Then select **NormalFit_Modeler** from the drop-down list of models and begin clicking inside the graph panel. The latter allows you to construct manually a histogram of interest. Notice that these are not random measurements, but rather frequency counts from which you are manually constructing the histogram of. Try to make the histogram

bins form a unimodal, bell-shaped and symmetric graph. Observe that as you click, new histogram bins will appear and the model fit will update. Now click the Estimate Parameters check-box on the top-left and see the best-fit Normal curve appear superimposed on the manually constructed histogram. Under the **Results** tab you can find the maximum likelihood estimates for the mean and the standard deviation for the best Normal distribution fit to this specific frequency histogram.

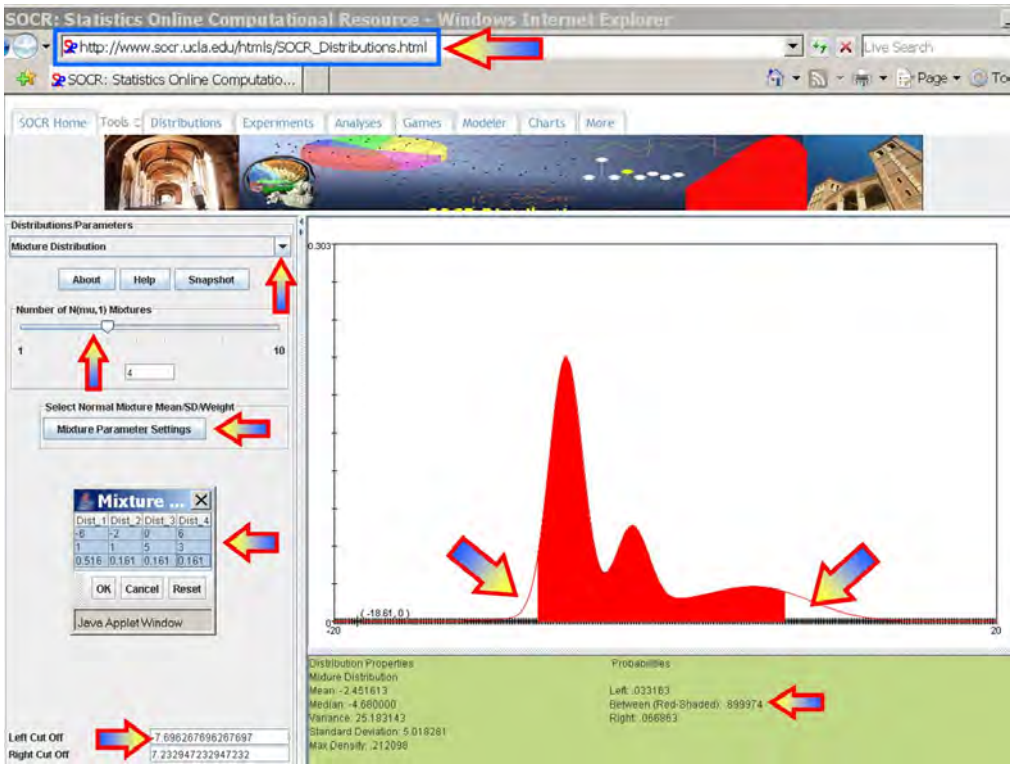


Applications

- Here you can see more instances of using the SOCR Modeler to fit distribution models to real data (http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_Activities_RNG).
- SOCR Modeler allows one to fit distribution, polynomial or spectral models to real data - more information about these is available at the SOCR Modeler Activities (http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_ModelerActivities).

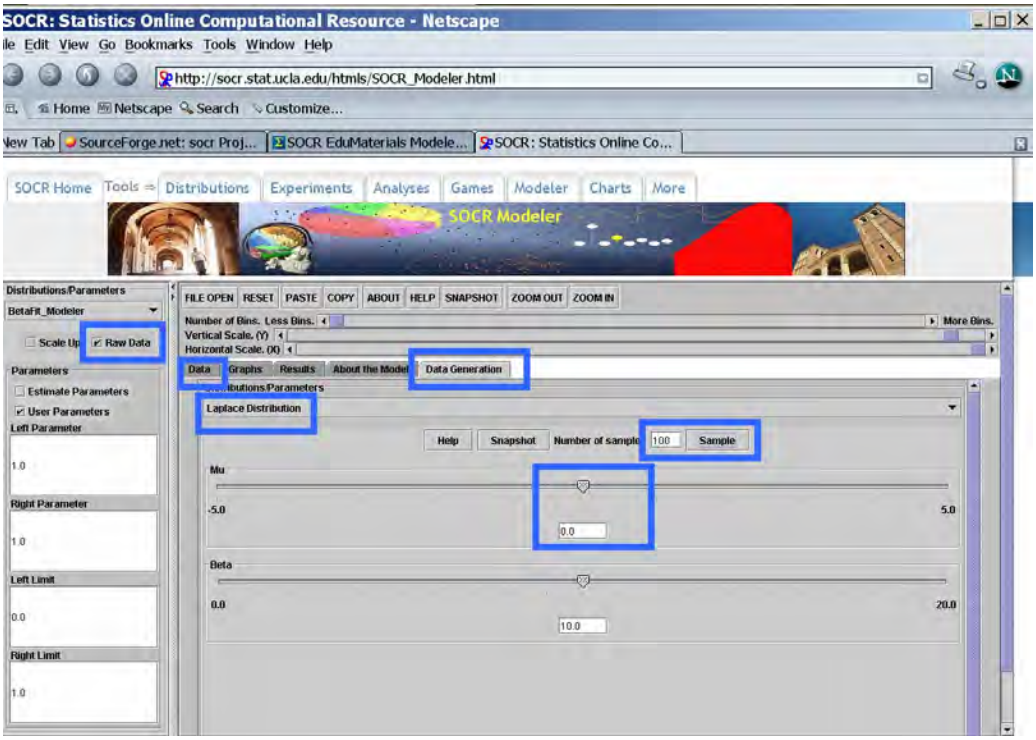
- SOCR Mixture Model Fitting Activity

This is a SOCR Activity that demonstrates random sampling and fitting of mixture models to data. The 1D SOCR mixture-model distribution enables the user to specify the *number of mixture Normal distributions* and their parameters (*means* and *standard deviations*), www.socr.ucla.edu/htmls/dist/Mixture_Distribution.html. This applet demonstrates how unimodal-distributions come together as **building-blocks** to form the backbone of many complex processes. In addition, this applet allows computing probability and critical values for these mixture distributions, and enables inference on such complicated processes. Extensive demonstrations of mixture modeling in 1D, 2D and 3D are available on the SOCR EM Mixture Modeling page (http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_Activities_2D_PointSegmentation_EM_Mixture). The figure below shows one such example of a tri-modal mixture of 4 Normal distributions.



Data Generation

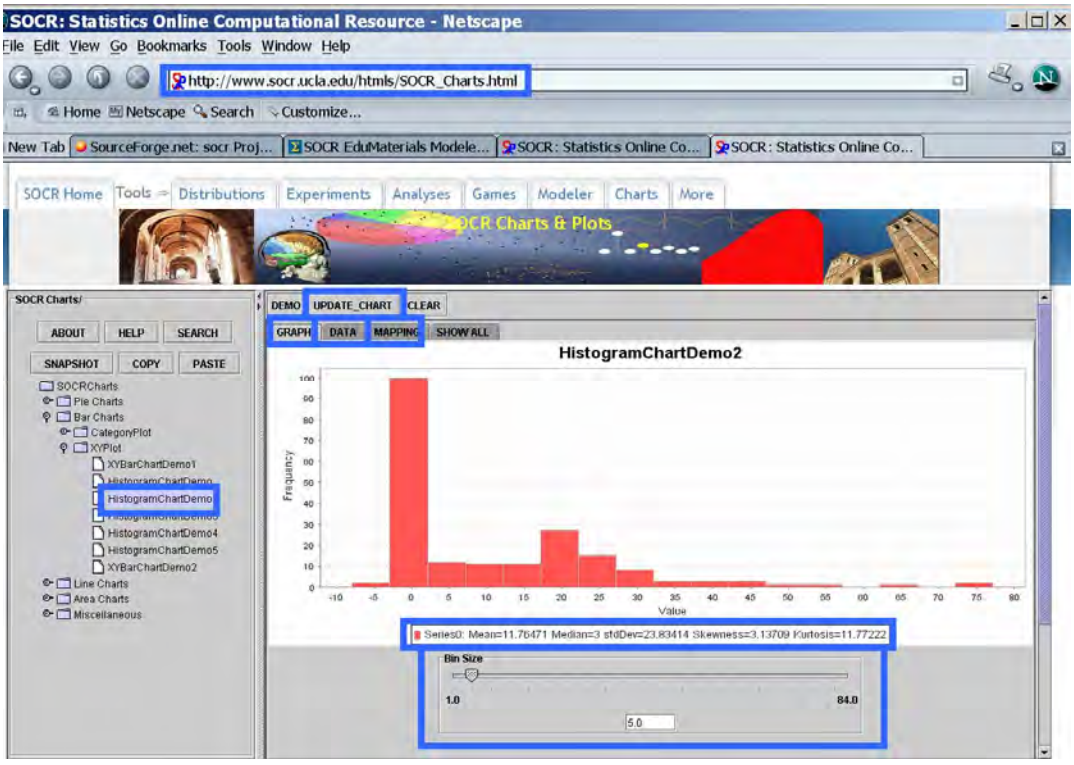
You typically have investigator-acquired data to which you need to fit a model. In this case we will generate the data by randomly sampling using the SOCR resource. Go to the SOCR Modeler (www.socr.ucla.edu/htmls/SOCR_Modeler.html) and select the **Data Generation** tab from the right panel.



- Now, click the **Raw Data** check-box in the left panel, select **Laplace Distribution** (or any other distribution you want to sample from), choose the **sample-size** to be 100, keep the center, Mu ($\mu = 0$), and click **Sample**. Then go to the **Data** tab, in the right panel. There you should see the 100 random Laplace observations stored as a column vector.
- Next, go back to the **Data Generation** tab from the right panel and change the center of the Laplace distribution (set $\mu=20$, say). Click **Sample** again and you will see the list of randomly generated data in the **Data** tab expand to 200 (as you just sampled another set of 100 random Laplace observations).

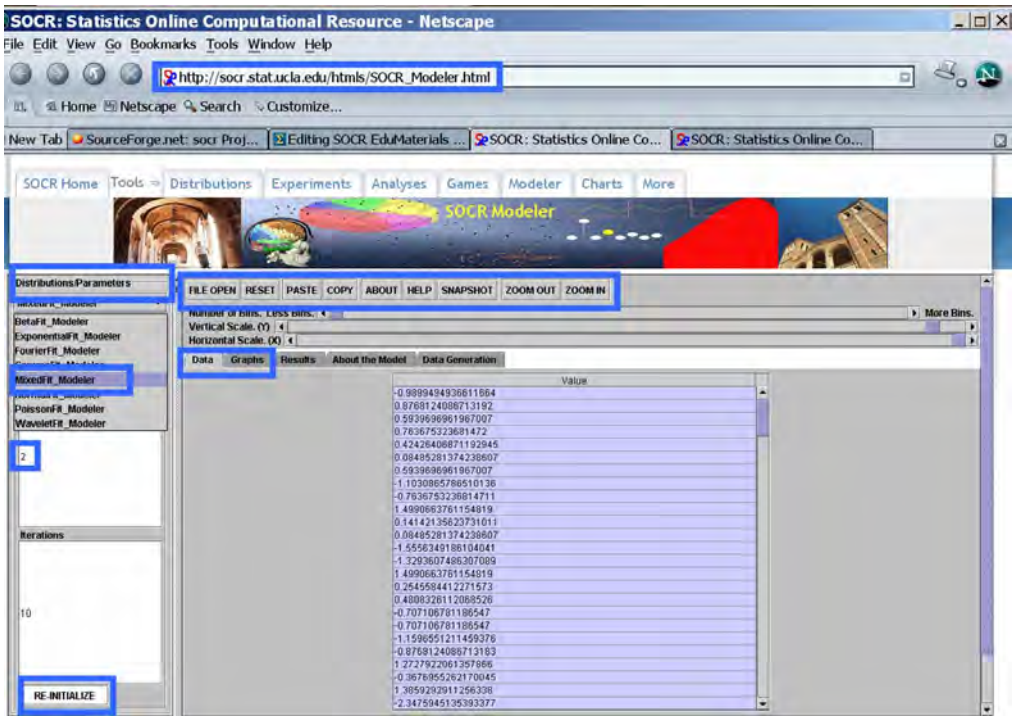
Exploratory Data Analysis (EDA)

Go to the **Data** tab and select all observations in the data column (use CTR-A, or mouse-copy). Then open another web browser and go to SOCR Charts (www.socr.ucla.edu/htmls/chart). Choose **Frequency-Data Histogram Applet**, say, clear the default data (**Data** tab) and paste (CTR-V or mouse paste-in) in the first column the 200 observations that you sampled in the SOCR Modeler Data Generator (above). Then you need to **map** the values - go to the **Mapping** tab, select the first column, where you pasted the data (C1), and click **XValue**. This will move the C1 column label from the left bin to the bottom-right bin. Finally, click **Update Chart**, on the top, and go to the **Graph** tab to see your histogram of the 200 (bimodal) Laplace observations. Notice that you can change the width of the histogram bin to clearly see the bi-modality of the distribution of these 200 measurements. Of course, this is due to the fact that we sampled from two distinct Laplace distributions, one with mean of zero and the second with mean of 20.0.

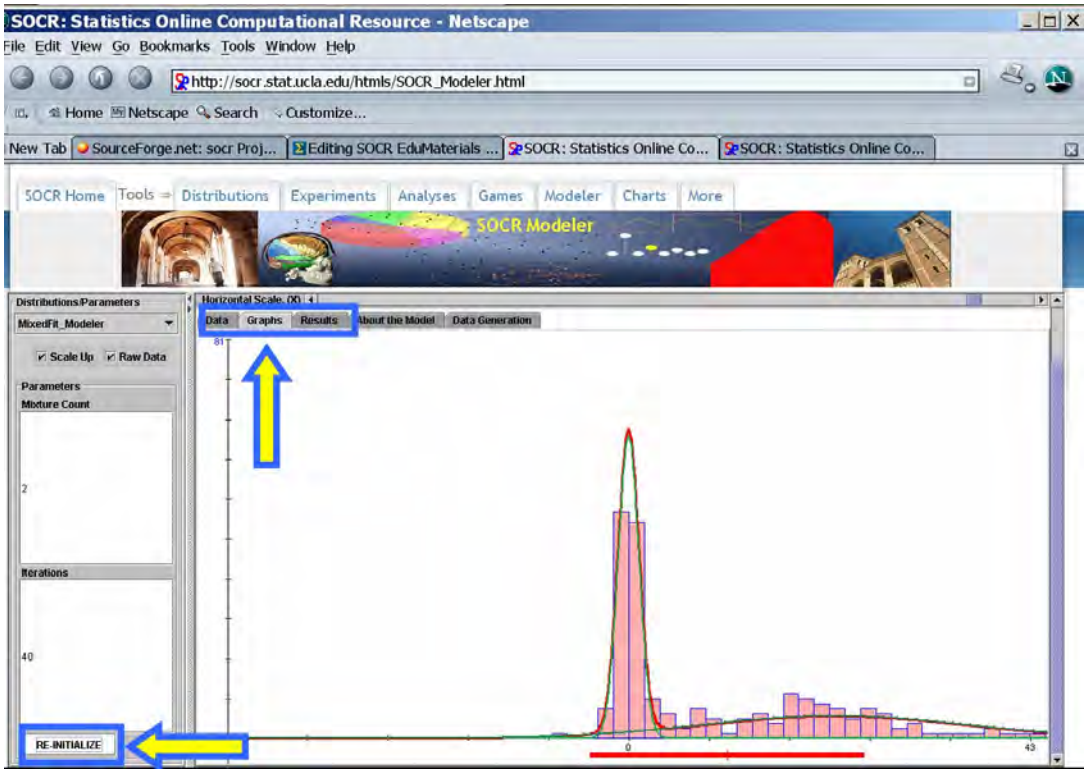


Model Fitting

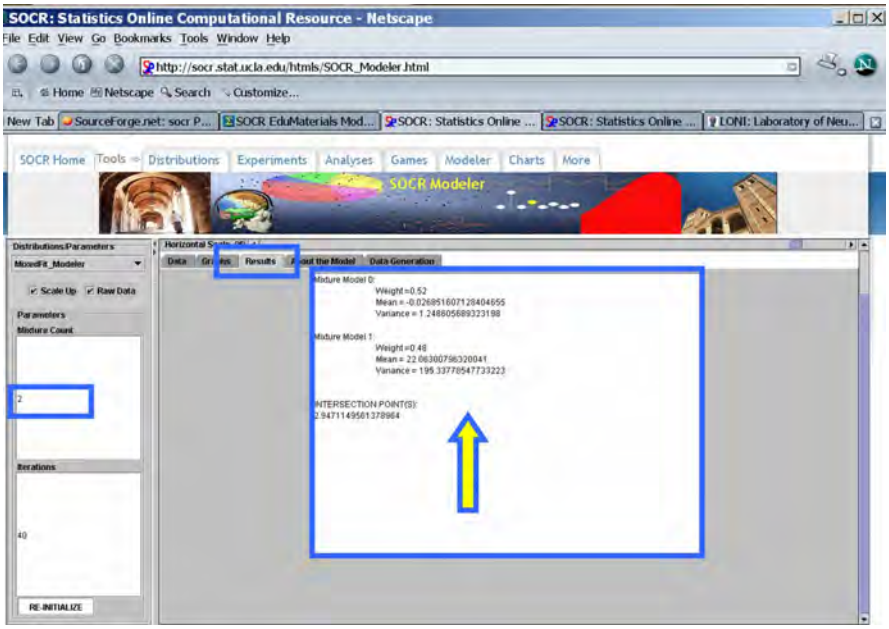
Now go back to the SOCR Modeler (www.socr.ucla.edu/htmls/SOCR_Modeler.html) browser (where you did the data sampling). Choose Mixed-Model-Fit from the drop-down list in the left panel.



- We will now try to fit a 2-component mixture of Gaussian (Normal) distributions to this Bimodal Laplace distribution (of the generated sample). You may need to click the **Re-Initialize** button a few times. The Expectation-Maximization algorithm (http://repositories.cdlib.org/socr/EM_MM) used to estimate the mixture distribution parameters is unstable and will produce somewhat different results for different initial conditions. Hence, you may need to re-initialize the algorithm a few times until a visually satisfactory result is obtained.



- Notice the quantitative results of this mixture model fitting protocol (in the **Results** panel). Recall that we sampled 100 observations from Laplace distribution with mean of zero (not Gaussian, which we could also have done and the fit would have been much better, of course) and then another 100 observations from Laplace distribution with mean = 20.0. In this case, the reported estimates of the means of the two Gaussian mixtures are 0 and 22 (pretty close to the original/theoretical means). We could have also fit in a mixture of 3 (or more) Gaussian mixture components, if we had a reason to believe that the data distribution is tri-modal, or higher, and requires a multi-modal mixture fit.



Caution!

You may need to properly set the values of the sliders on the top of your **Graph** tab, in the right panel, so that you can see the entire graph of the histogram and the models fit to the data. Also, the random data you generate and the EM algorithm are stochastic and you can not expect to get exactly the same results and charts as reported in this SOCR activity (http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_ModelerActivities_MixtureModel_1). Everyone that tries to replicate these steps will obtain different results; however, the principles we demonstrate here are indeed robust.

Footnotes:

- SOCR Mixture-Distribution applet (www.socr.ucla.edu/htmls/dist/Mixture_Distribution.html).
- SOCR 2D Mixture Modeling Activity (http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_Activities_2D_Point Segmentation_EM_Mixture).
- Reference: Dinov, ID. (2008) Expectation Maximization and Mixture Modeling Tutorial (December 9, 2008). Statistics Online Computational Resource. Paper EM_MM, http://repositories.cdlib.org/socr/EM_MM.

Distribution Activities

- Normal Distribution Activity

This is an activity to explore the Normal Probability Distribution and the Normal approximation to the Binomial. You should first review the complete details about the Standard Normal and the General Normal distributions (http://wiki.stat.ucla.edu/socr/index.php/EBook). You can access the applets for the above distributions at www.socr.ucla.edu/htmls/SOCR_Distributions.html.

- **Exercise 1:** Use SOCR to graph and print the distribution of $X \sim N(20,3)$. Show on the graph the following points: $\mu \pm \sigma$; $\mu \pm 2\sigma$; $\mu \pm 3\sigma$. How many standard deviations from the mean is the value $x = 27.5$?
- **Exercise 2:** Graph the distribution of $X \sim N(40,10)$:

1. Find $P(X > 49)$. Submit a printout.

2. Find $P(X < 22)$. Submit a printout.

3. Find $P(X < 58)$. Submit a printout.

4. Find $P(X > 13)$. Submit a printout.

5. Find $P(12 < X < 37)$. Submit a printout.

6. Find $P(33 < X < 60)$. Submit a printout.

7. Find $P(52 < X < 65)$. Submit a printout.

8. Use the mouse or the left cut off or right cut off points to find the 8^{th}, 20^{th}, 45^{th}, 55^{th}, 70^{th}, and 95^{th} percentiles. After you find these percentiles you can place by hand all of them in one printout (or you can submit a printout for each one of them if you want).

9. Make sure you know how to answer the above questions using the *z* score $z = \frac{x - \mu}{\sigma}$ and your *z* table (www.socr.ucla.edu/Applets.dir/Z-table.html)!

- **Exercise 3:** The lifetime of tires of brand *A* follows the Normal distribution with mean 40,000 miles and standard deviation 4,000 miles.

1. Use *SOCR* to find the probability that a tire will last between 40,000 and 46,000 miles.

2. Given that a tire will last more than 46,000 miles what is the probability that it will last more than 50,000 miles? Submit a printout and explain how you get the answer.

3. Given that a tire will last more than 46,000 miles what is the probability that it will last less than 50,000 miles? Submit a printout and explain how you get the answer.

- **Exercise 4:**

1. The probability that a student is admitted to the Math Department Major at a college is 45%. Suppose that this year 100 students will apply for admission into the Math major.

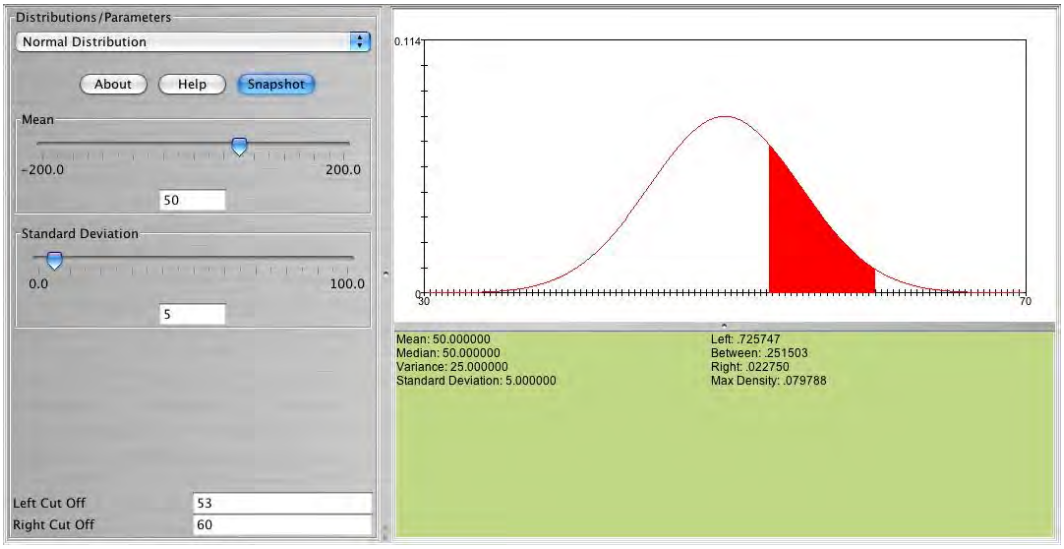
2. What is the distribution of the number of students admitted? Use *SOCR* to graph and print this distribution. What is the shape of this distribution? What is the mean and standard deviation of this distribution?

3. Write an expression for the exact probability that among the 100 students at least 55 will be admitted.

4. Use SOCR to compute the probability of part (3).

5. Use the Normal distribution applet in SOCR to approximate the probability of part (3) (do not forget the continuity correction). What is the error of the approximation?

Below you can see the distribution of a Normal random variable X with $\mu = 50, \sigma = 5$. In this graph you can also see the probability that X is between 53 and 60.



- Relations between distributions

This is a SOCR activity (http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_Activities_Explore_Distributions) to explore the relationships among some of the commonly used probability distributions. This is a complementary SOCR activity to the SOCR Distributions Activity.

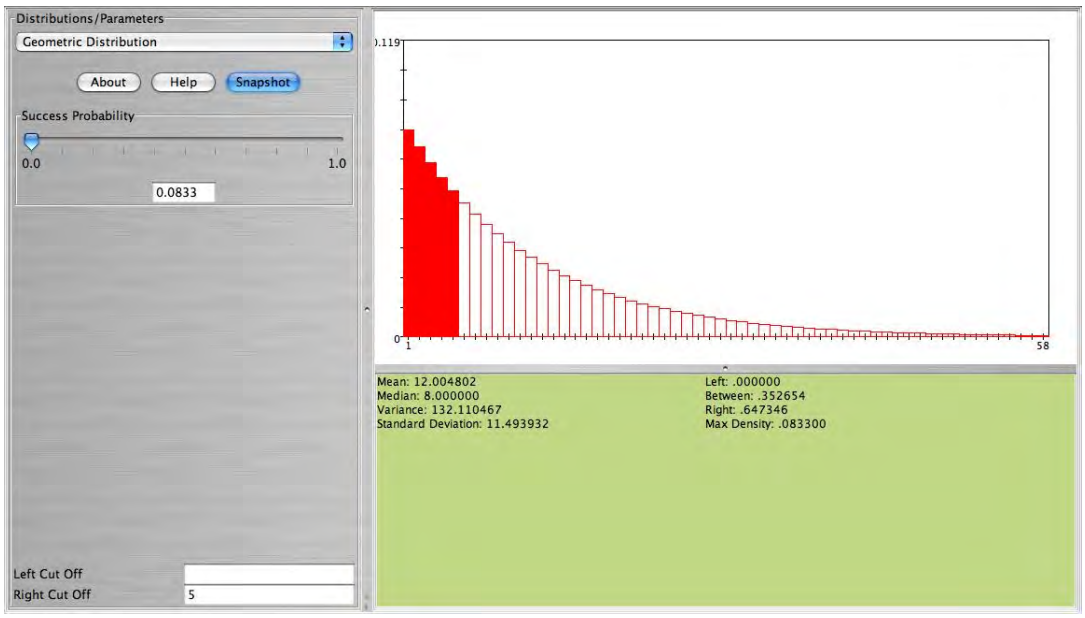
You can access the SOCR Distribution applets for these activities by going to this URL page (use a Java-enabled browser) www.socr.ucla.edu/htmls/dist.

Geometric probability distribution

Suppose we roll two dice until a sum of 10 is obtained. What is the probability that the first sum of 10 will occur after the 5th trial? The answer to this question is

$$P(X > 5) = \left(1 - \frac{3}{36}\right)^5 = 0.6472 .$$

This is equivalent to the event that no sum of 10 is observed on the first 5 trials (5 failures). Now, using SOCR we can obtain this probability easily by entering in the SOCR geometric distribution applet $p = \frac{3}{36} = 0.0833$ and in the Right Cut-Off box 5. We can find the desired probability on the right corner of the applet. The figure below clearly displays this probability.

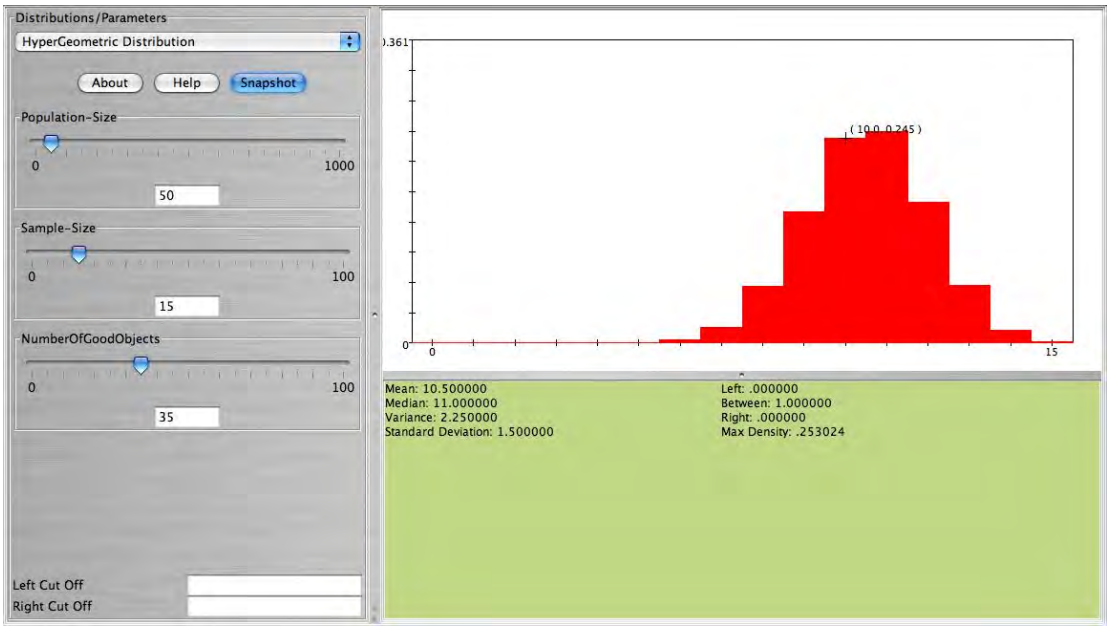


Binomial approximation to Hypergeometric

An urn contains 50 marbles (35 green and 15 white). Fifteen marbles are selected without replacement. Find the probability that exactly 10 out of the 15 selected are green marbles. The answer to this question can be found using the formula:

$$P(X=10)=\frac{\binom{35}{10}\binom{15}{5}}{\binom{50}{15}}=0.2449\,.$$

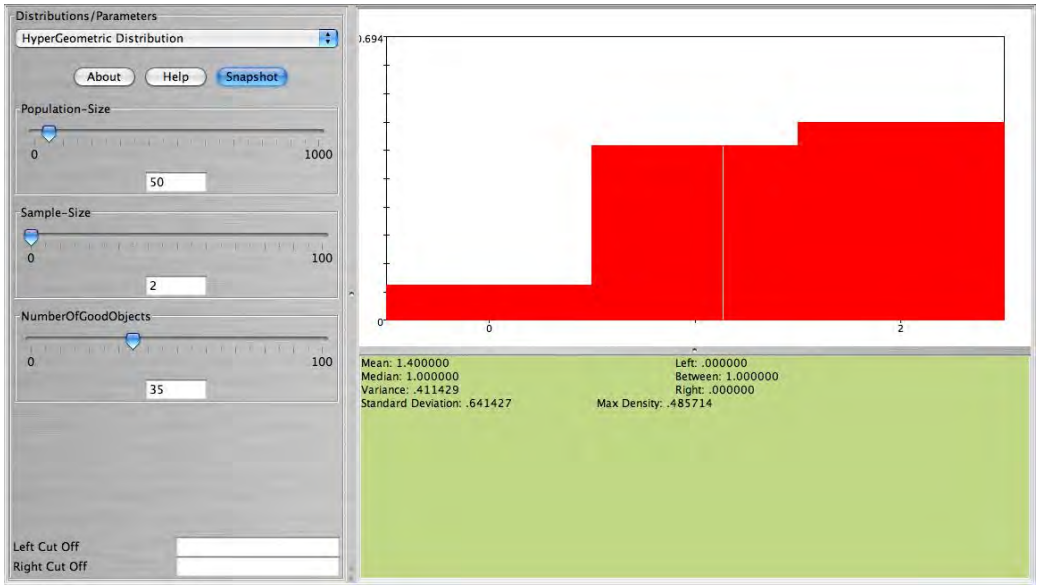
Using SOCR simply enter population size 50, sample size 15, and number of good objects 35, to get the figure below.



Now, select without replacement only 2 marbles. Compute the exact probability that 1 green marble is obtained. This is equal to

$$P(X=1)=\frac{\binom{35}{1}\binom{15}{1}}{\binom{50}{2}}=0.4286\,.$$

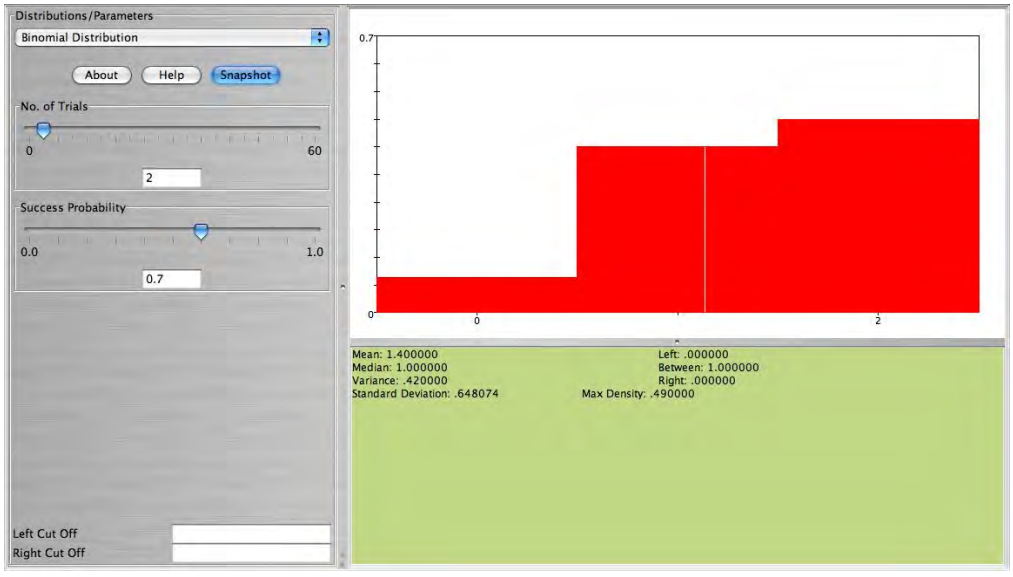
This is also shown on the figure below.



We will approximate the probability of obtaining 1 green marble using the Binomial distribution as follows. Select the SOCR Binomial distribution and choose number of trials 2 and probability of success $p = \frac{35}{50} = 0.7$. Compare the figure below with the figure above. They are almost the same! Why? Using the Binomial formula we can compute the approximate probability of observing 1 green marble as

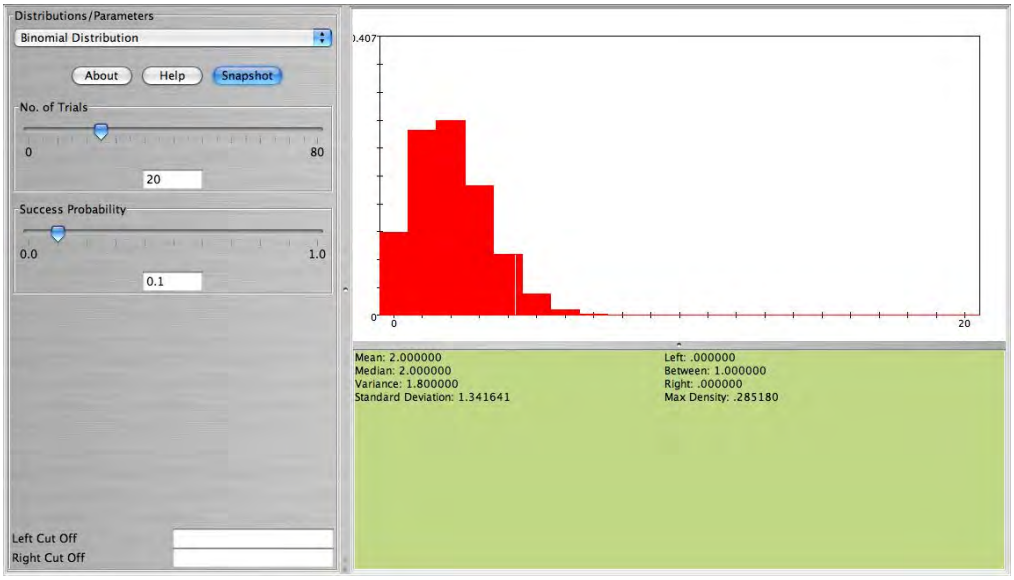
$$P(X = 1) = \binom{2}{1} 0.70^1 0.30^1 = 0.42,$$

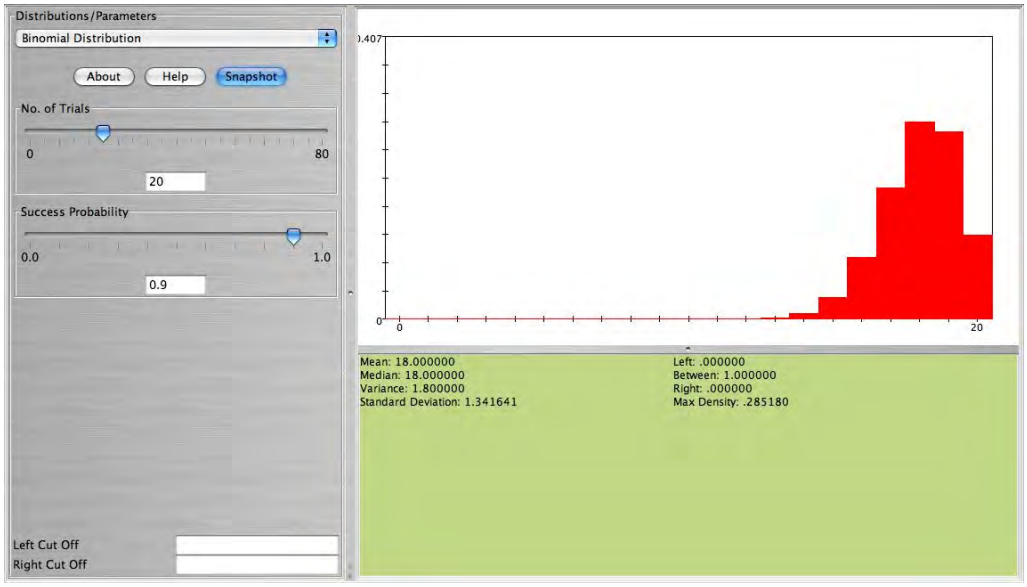
which is very close to the exact probability, **0.4286**.

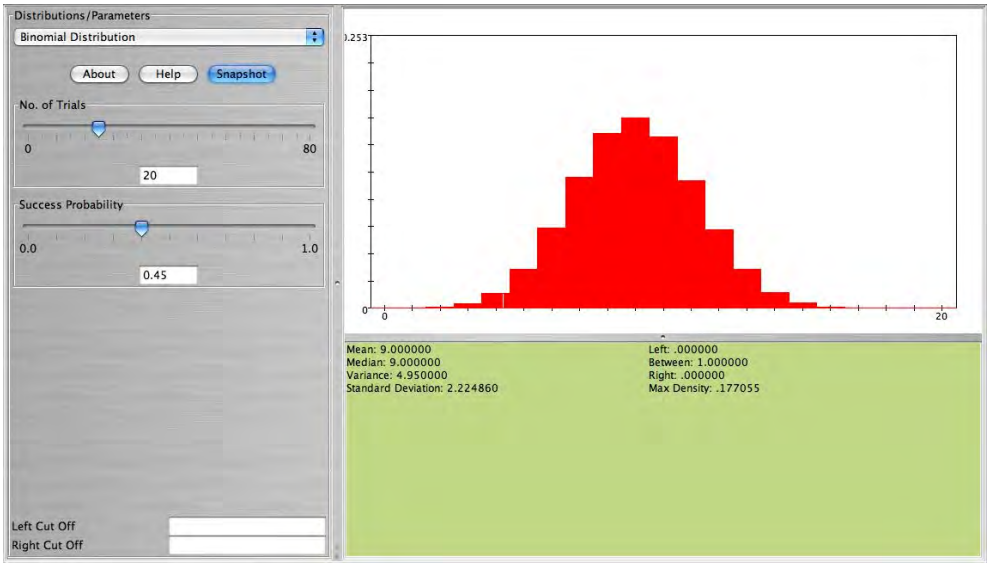


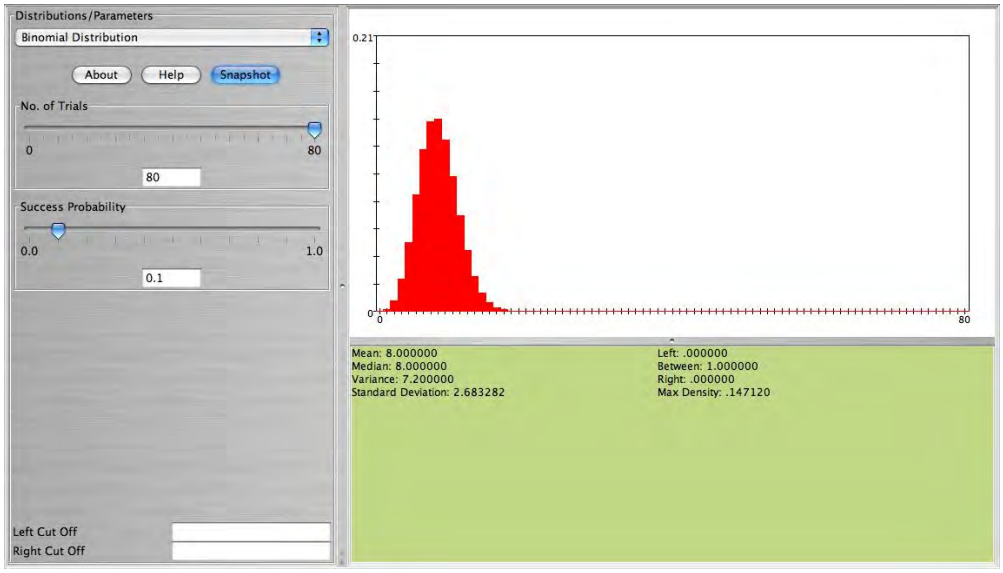
Normal approximation to Binomial

Graph and comment on the shape of the Binomial distribution with $n = 20$, $p = 0.1$ and $n = 20$, $p = 0.9$. Now, keep $n = 20$ but change $p = 0.45$. What do you observe now? How about when $n = 80$, $p = 0.1$. See the four figures below.





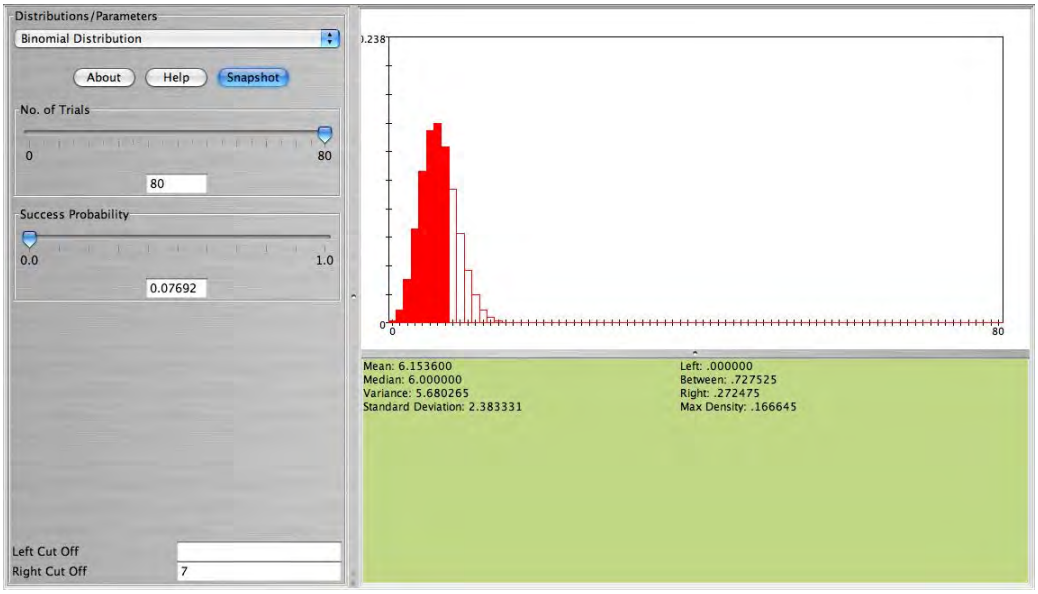




What is your conclusion on the shape of the Binomial distribution in relation to its parameters *n*, *p*? Clearly when *n* is large and *p* small or large the result is a bell-shaped distribution. When *n* is small (10-20) we still get approximately a bell-shaped distribution as long as $p \approx 0.5$. Because of this feature of the Binomial distribution we can approximate Binomial distributions using the Normal distribution when the above requirements hold. Here is one example: Eighty cards are drawn with replacement from the standard 52-card deck. Find the exact probability that at least 8 aces are obtained. This can be computed using the formula

$$P(X \geq 8) = \sum_{x=8}^{80} \left(\frac{4}{52}\right)^{x} \left(\frac{48}{52}\right)^{80-x} = 0.2725.$$

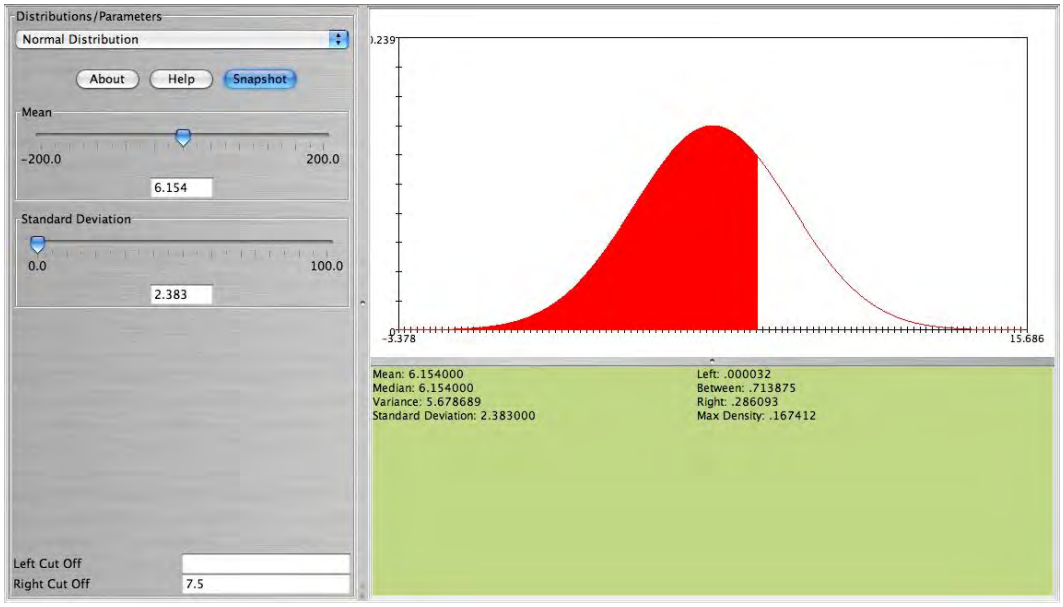
We can easily use SOCR to compute this probability (see figure below).



But we can also *approximate* this probability using the Normal distribution. We will need the mean and the standard deviation of this Normal distribution. These are

$$\mu = np = 80\frac{4}{52} = 6.154 \quad \text{and} \quad \sigma = \sqrt{np(1-p)} = \sqrt{80\frac{4}{52}\frac{48}{52}} = 2.383.$$

Of course this can be obtained directly from the SOCR Binomial applet. Now, all you need to do is to select the SOCR Normal distribution applet and enter for the mean *6.154*, and for the standard deviation *2.383*. To obtain the desired probability in the right cut-off box enter 7.5 (using the continuity correction for better approximation). The approximate probability is $P(X \geq 8) \approx 0.2861$ (see figure below).

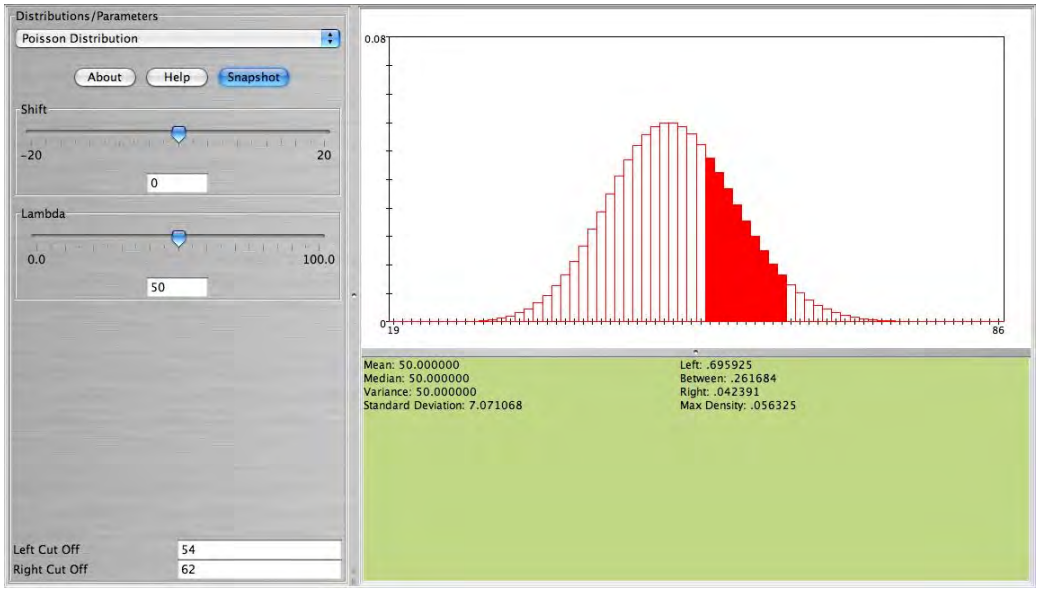


Normal approximation to Poisson

The Poisson distribution with parameter λ can be approximated with the Normal when λ is large. Here is one example. Suppose cars arrive at a parking lot at a rate of 50 per hour. Let's assume that the process is a Poisson random variable with $\lambda = 50$. Compute the probability that in the next hour the number of cars that arrive at this parking lot will be between *54* and *62*. We can compute this as follows:

$$P(54 \le X \le 62) = \sum_{x=54}^{62} \frac{50^x e^{-50}}{x!} = 0.2617 .$$

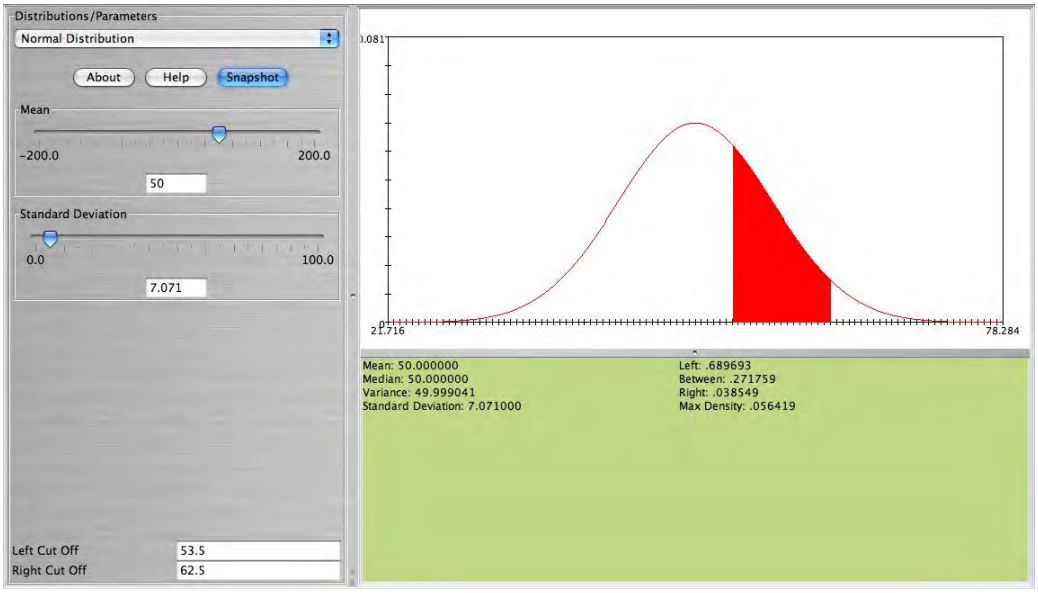
The figure below from SOCR shows this probability.



- **Note**: We observe that this distribution is bell-shaped. We can use the Normal distribution to approximate this probability. Using

$$N\left(\mu = 50, \sigma = \sqrt{50} = 7.071\right),$$

together with the continuity correction for better approximation we obtain $P(54 \le X \le 62) = 0.2718$, which is close to the exact probability that was found earlier. The figure below shows this probability.

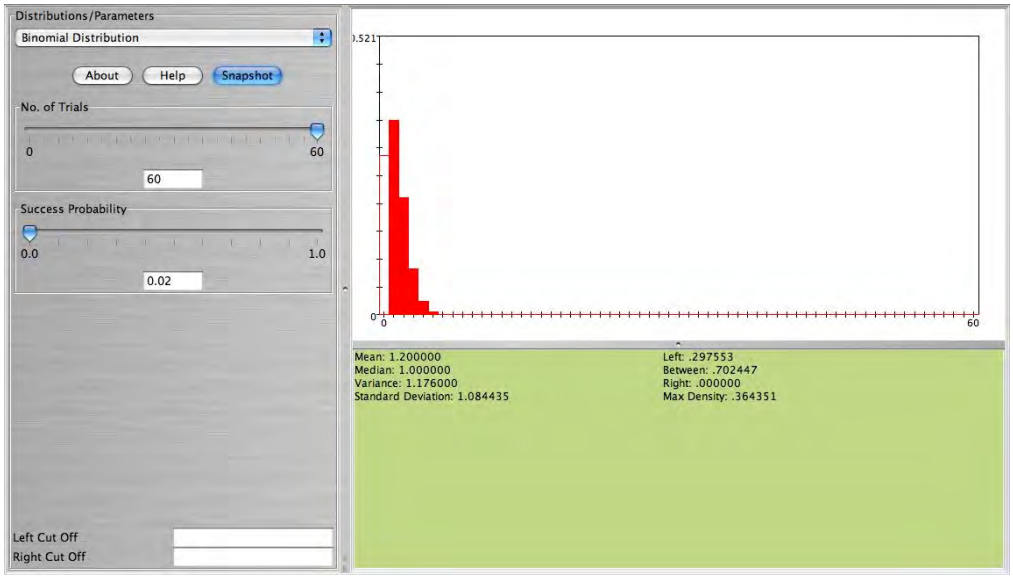


Poisson approximation to Binomial

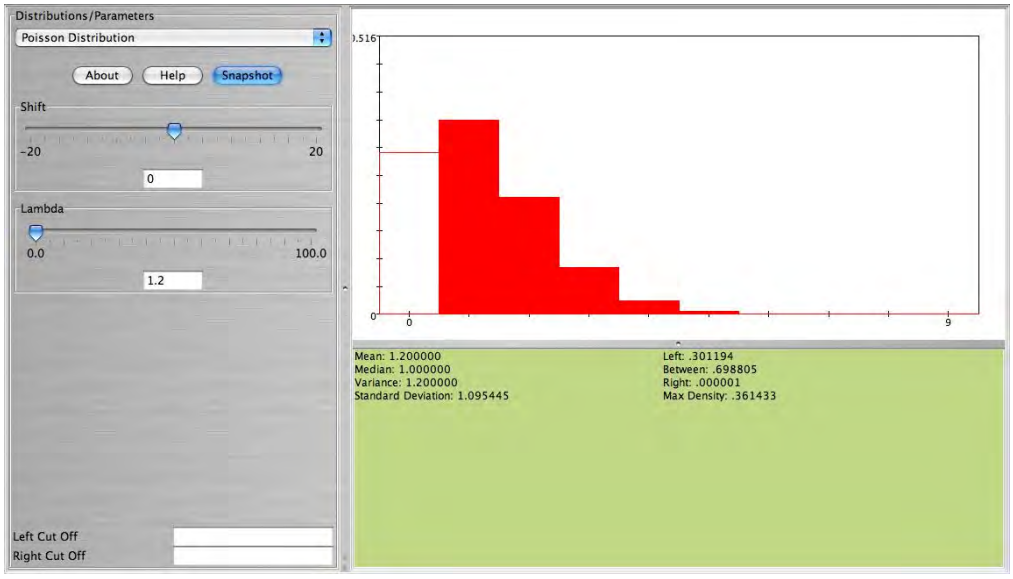
The Binomial distribution can be approximated well by the Poisson when n is large and p is small with $np < 7$. This is true because

$$\lim_{n \to \infty} \left\{ \binom{n}{x} p^x (1-p)^{n-x} \right\} = \frac{\lambda^x e^{-\lambda}}{x!},$$

where $\lambda = np$. Here is an example. Suppose 2% of a certain population have Type AB blood. Suppose 60 people from this population are randomly selected. The number of people X among the 60 that have Type AB blood follows the Binomial distribution with $n = 60$, $p = 0.02$. The figure below represents the distribution of X. This figure also shows $P(X = 0)$.



- **Note**: This distribution can be approximated well with Poisson with $\lambda = np = 60(0.02) = 1.2$. The figure below is approximately the same as the figure above (the width of the bars is not important here. The height of each bar represents the probability for each value of X which is about the same for both distributions).



Afternoon Session: SOCR Activities (cont.)

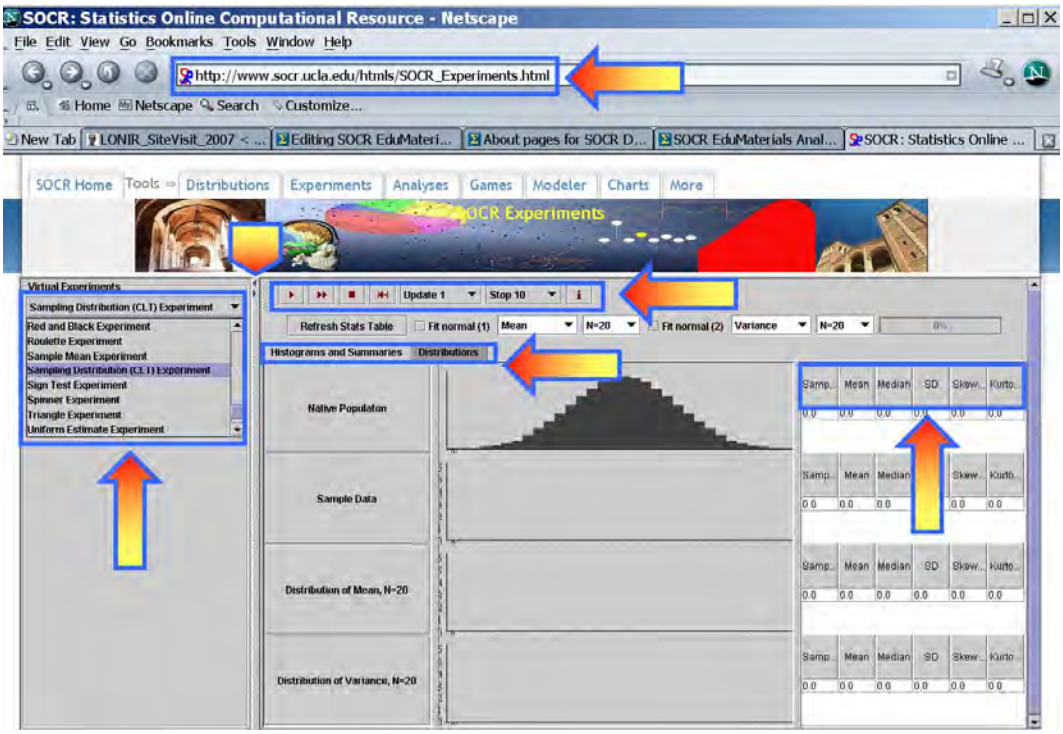
Central Limit Theorem Activity

This activity represents a very general demonstration 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.

The aims of this activity are to:

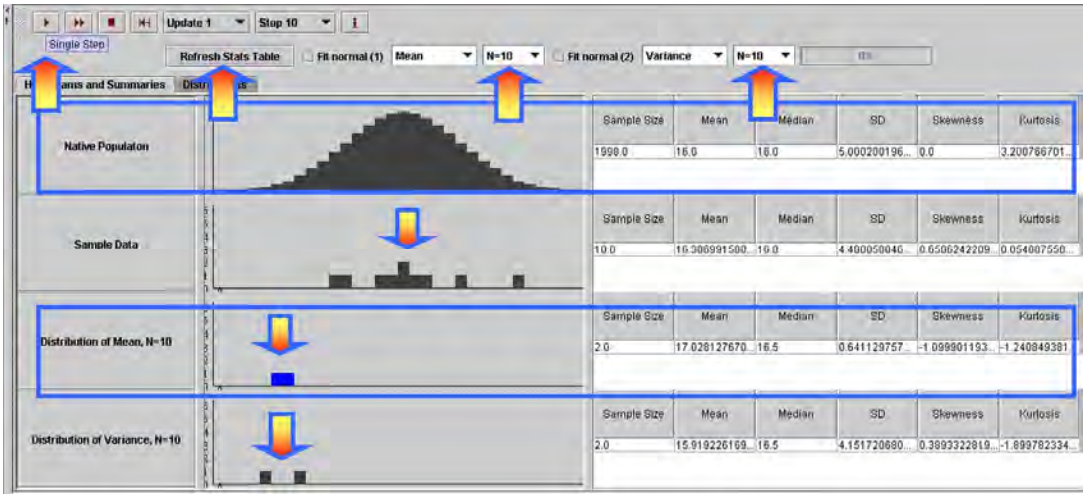
- Provide an intuitive notion of sampling from any process with a well-defined distribution.
- Motivate and facilitate learning of the central limit theorem.
- Empirically validate that sample-averages of random observations (most processes) follow approximately the Normal distribution.
- Empirically 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, at the rate of $\frac{1}{\sqrt{n}}$, as the sample size increases.
- Reinforce the concepts of a native distribution, sample, sample distribution, sampling distribution, parameter estimator and data-driven numerical parameter estimate.

To start the this Experiment, go to SOCR Experiments (www.socr.ucla.edu/htmls/SOCR_Experiments.html) 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).



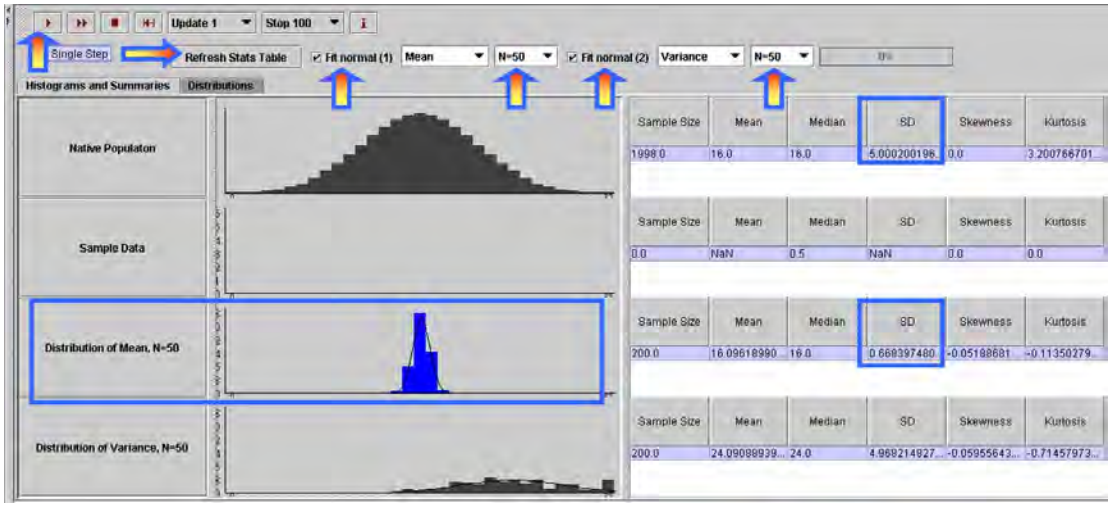
Experiment 1

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 (2-5) times 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 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/histograms become good approximations to the corresponding sampling distributions. 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).



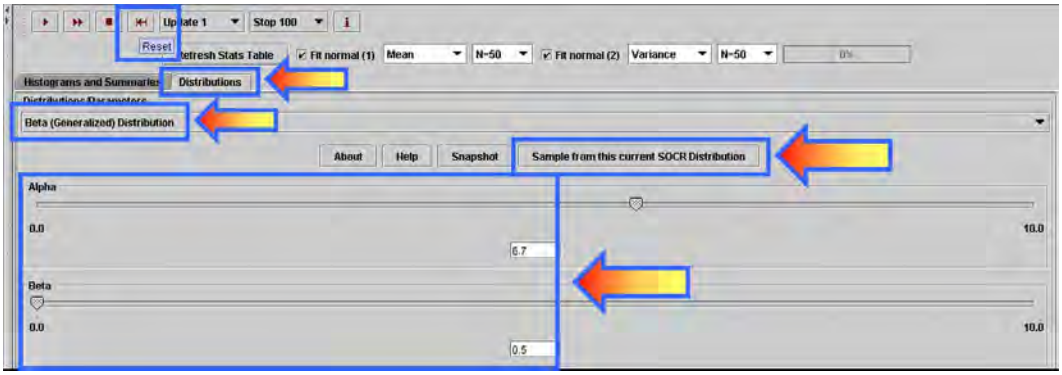
Experiment 2

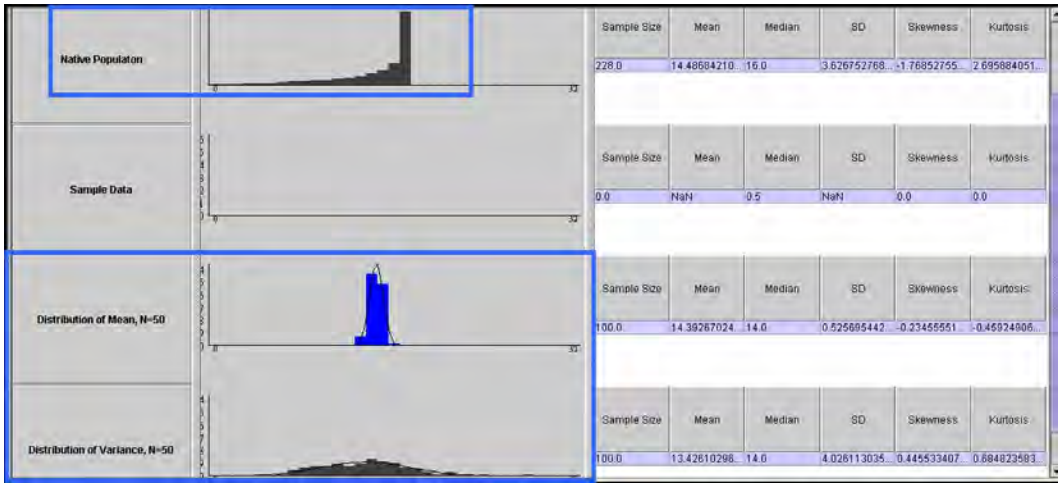
For this experiment we'll look at the mean, standard deviation, skewness and kurtosis of the sample-average and the sample-variance (these are two data-driven statistical estimates). Choose sample-sizes of 50, for both estimates (mean and variance). Select the **Fit Normal Curve** check-boxes for both 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 ($\alpha = 6.7, \beta = 0.5$).

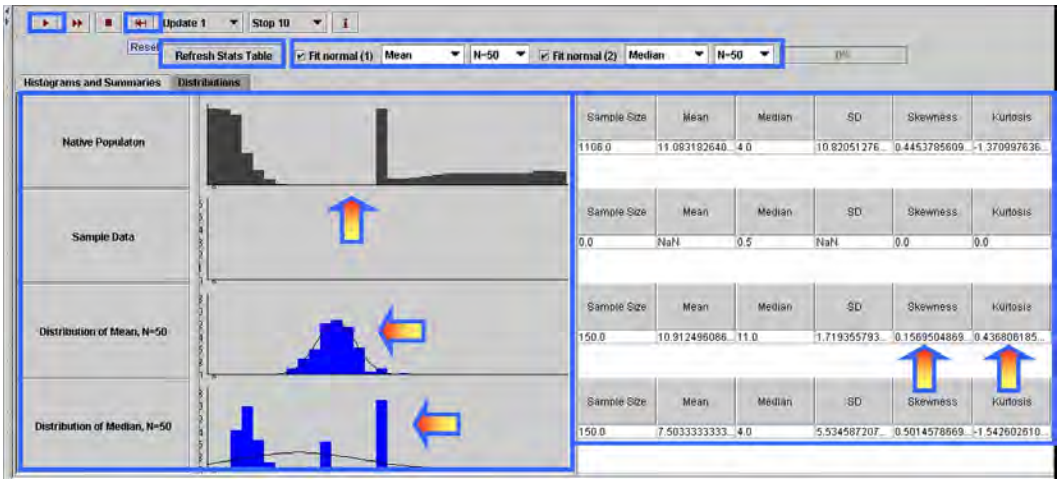




Experiment 4

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 away 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, where as the sampling distribution of the sample median is not.



Questions

- What effects will asymmetry, gaps and continuity of the native distribution have on the applicability of the CLT, or on the asymptotic distribution of various sample statistics?
- 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 $\sigma_{\bar{X}} = \frac{2}{3}\sigma_X$?

CLT Applications

The second part of this SOCR CLT activity demonstrates the applications of the Central Limit Theorem. The aims of this activity are to demonstrate several practical applications of the general CLT. There are many practical examples of using the CLT to solve real-life problems. Here are some examples which may be solved using the SOCR CLT resources.

Application 1 (Poisson)

Suppose a call service center expects to get 20 calls a minute for questions regarding each of 17 different vendors that rely on this call center for handling their calls. What is the probability that in a 1-minute interval they receive less than 300 calls in total?

Let X_i be the random variable representing the number of calls received about the i^{th} vendor within a minute, then $X_i \sim Poisson(20)$, as X_i is the number of arrivals within a unit interval and the mean arrival count is given to be *20*. The distribution of the total number of calls

$$T = \sum_{i=1}^{17} X_i \sim Poisson(17 \times 20 = 340).$$

By the CLT, $T \sim Normal\left(\mu = 340, \sigma^2 = 340\right)$, as an approximation of the exact distribution of the total sum. Using the SOCR Distribution applet one can compute exactly the $P\left(T < 300 \mid T \sim Poisson(340)\right) = 0.014021$. On the other hand, one may use the CLT to compute a Normal approximation probability of the same event,

$$P\left(T < 300 \mid T \sim Normal\left(\mu = 340, \sigma^2 = 340\right)\right) = 0.014896.$$

The last quantity is obtained again using the SOCR Distributions applet, without using continuity correction. Using continuity correction the approximation improves,

$$P\left(T < 300 \mid T \sim Normal\left(\mu = 340, \sigma^2 = 340\right)\right) = 0.0140309.$$

Arguably, the CLT-based calculation is less intense and more appealing to students and trainees, compared to computing the exact probability.

Application 2 (Exponential)

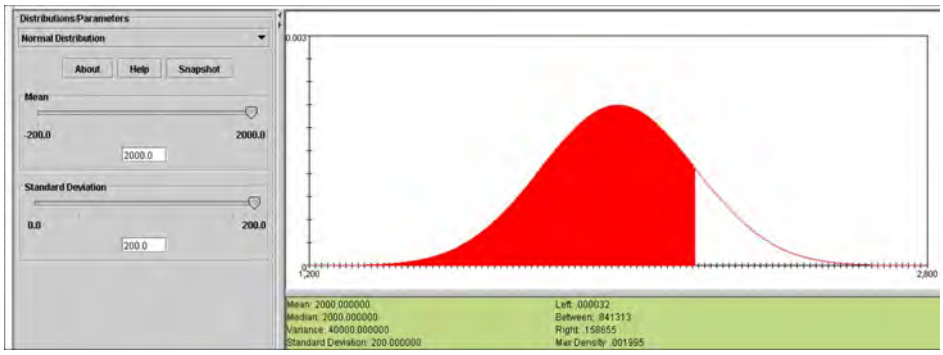
It is believed that life-times, in hours, of light-bulbs are Exponentially distributed, say $Exp\left(\frac{1}{2{,}000}\right)$, mean expected life of 2,000 hours. Recall that the Exponential distribution is called the Mean-Time-To-Failure distribution. You can find out more about it from the SOCR Distributions applet. Suppose a University wants to estimate the average life-span of some light bulbs based on the life-span of 100 such new light-bulbs. What is a CLT-based estimate of the probability that the average life-span exceeds 2,200 hrs? Let $X_i \sim Exp\left(\frac{1}{2{,}000}\right)$ and $\overline{X} = \frac{1}{100}\sum_{i=1}^{100} X_i$.

Notice that in this case, the exact distribution of $\overline{X}$ is (generally) not Exponential, even though the density may be computed in closed form (Khuong & Kong, 2006). If we use the CLT, however, we can approximate the probability of interest

$$P\left(\overline{X} > 2{,}200\right) \cong P\left(\overline{X} > 2{,}200 \mid \overline{X} \sim N\left(\mu_{\overline{X}} = 2{,}000, \sigma^2_{\overline{X}} = \frac{2{,}000^2}{100}\right)\right),$$

as we know that the mean and the standard deviation of X_i are $\frac{1}{\lambda} = 2{,}000$ and the standard deviation of $\overline{X}$ is $\frac{1}{\lambda\sqrt{100}} = 200$.

Therefore, $P\left(\overline{X} > 2{,}200\right) \approx 0.158655$, using the CLT approximation and the SOCR Distributions calculator, see figure below.



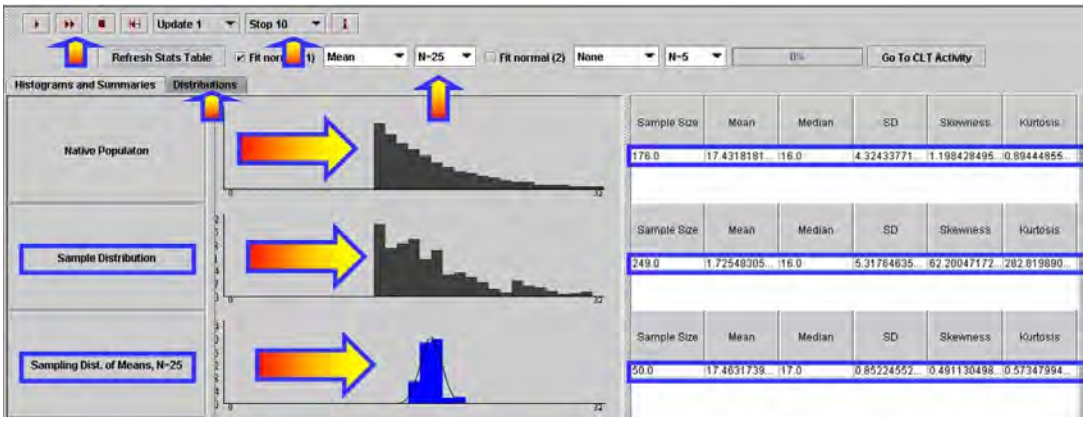
Application 3 (Exponential)

A weekly TV talk show from broadcaster U invites viewers to call to express their opinions about the program. Many people call, which sometimes results in quite a long wait time until the host replies. The time it takes the host to respond tends to follow an *exponential distribution* with mean of 50 seconds. A competing TV network W has another similar talk show and would like to respond to callers faster than broadcaster U. To do that W executives need to know how long U takes to respond. So, W personnel make 25 calls per week to the U show (for 50 weeks) and measure how long it took the U host to respond. Then W executives compute the average length for their weekly samples of size 25. At the end of the year they plot the distribution of the sample means. What do you think are the center, spread and shape of this distribution? Find out using the SOCR CLT applet. Approximately what proportion of time is the average weekly wait time for the U broadcaster exceeding 45 seconds?

The Figure below shows the corresponding sampling simulation (SOCR CLT Applet using $Exp\left(\lambda = \frac{1}{50} = 0.02\right)$ and drawing 50 samples, each of size 25). Notice the differences in the summary statistics between the native distribution, the sample distribution and the sampling distribution for the mean. The answer to this application may then be computed using the SOCR Distributions,

$$P\left(\overline{X} > 45\right) \cong P\left(\overline{X} > 45 \,|\, \overline{X} \sim N\left(\mu_{\overline{X}} = 50, \sigma^2_{\overline{X}} = \frac{50^2}{25}\right)\right) = 0.691463.$$

This chance may also be computed empirically by counting the number of weekly samples that generate an average wait time over 45 seconds and dividing this number by 50 (the total number of weeks in this survey).



Application 4 (Binomial)

Suppose a player plays a standard Roulette game (SOCR Roulette Experiment) and bets $1 on a single number. Find the probability the casino will make at least $28 in 100 games.

- **Solution 1**: One way to solve this problem is to find the distribution of the casino's payoff first: If Y is the random variable representing the payoff for the casino in a single game, the probability mass function for Y is given by $P(Y=1)=\frac{37}{38}$ and $P(Y=-35)=\frac{1}{38}$, as there are 38 numbers in total (0, 00, 1, 2, …, 36). The player may place a bet on any of these numbers, with a player success payoff of $35 (casino loss of $35) and a player loss of $1 (casino win of $1). Therefore, the casino expected return of the game is $\mu_Y = E(Y) = \frac{2}{38} = 0.05263$ and the variance of the casino return is $\sigma_Y^2 = Var(Y) = 33$ ($\sigma_Y = SD(Y) = 5.8$) and the range of the return is [-35 : 1], for one game. The exact probability of interest may be computed by using Binomial Distribution. If the total casino return in 100 games is denoted by $T=\sum_{i=1}^{100} Y_i$, then the expected casino return in 100 games is $5.26 and $P(T>28)=P(X>k)$, where

$$X \sim Binomial\left(n=100,\ p=\frac{37}{38}=0.97368\right)$$

and k is the integer solution of the following dollar amount equation:

$$k\times\$1-(100-k)\times\$35=\$28$$
$$\Rightarrow \quad k=98.$$

Therefore, $P(T \geq 28) = P(X \geq 98) = 0.508326$. The last probability represents the exact solution and is computed using the SOCR Binomial Distribution applet. This exact calculation is numerically intractable for large sample-sizes (*n>200*), albeit approximations exist.

- **Solution 2**: One could use the CLT to find a very good approximation to this type of probability. For example, in the case above (n=100), we can estimate the probability of interest by

$$P(T \geq 28) \approx P\left(T > 28 \mid T \sim Normal\left(\mu_T = 100 \times 0.05263, \sigma^2 = 5.8^2 \times 100\right)\right) = 0.347.$$

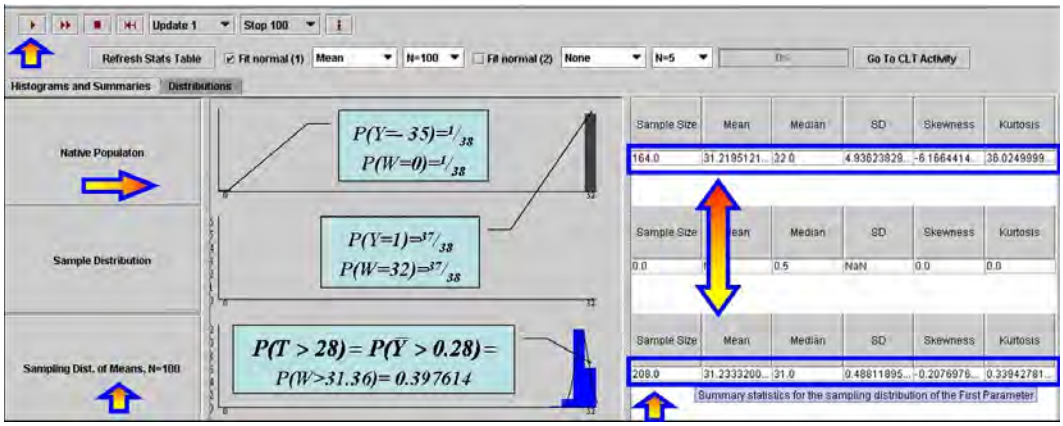
Notice that this calculation is sample-size independent, and hence widely applicable, whereas the former exact probability calculation (Solution 1) is limited for small n. What caused the large discrepancy between the exact $\left(P(T \geq 28) = 0.508326\right)$ and approximate $\left(P(T \geq 28) = 0.347\right)$ values of the probability of interest? This is an example where the usual rule of **30 measurements** breaks down because of the heavily skewed payout distribution. Such limitations of the CLT even for large sample-sizes have been previously observed and reported for severely skewed distributions (Freedman, Pisani, & Purves, 1998). Here one would need a much larger sample to get a reasonably good approximation using CLT. For example, if n=1,000, and we are looking for $P(T > 100)$, then k=975, the exact probability is $P(T > 100) = P(X > 975) = 0.4493287$, and the CLT approximation is much closer:

$$P(T > 28) \cong P\left(T > 28 \mid T \sim N\left(\mu_T = n \times \mu_Y = 1{,}000 \times 0.05263 = 52.63, \sigma_T^2 = 5.8^2 \times 1{,}000\right)\right)$$
$$P(T > 28) \cong 0.3979.$$

- **Solution 3**: Finally, we show how one can use the SOCR CLT applet alone to completely empirically estimate the probability of interest, *P(T>28)*, for n=100. The figure below demonstrates how we can manually construct the native probability mass function for the random variable *Y* (casino payoff of one roulette game). A simple linear transformation is needed to convert the values of *Y* to $W\left(W = \frac{32}{36}(Y + 35)\right)$, so that the range of the Y variable [-35 : 1] may be mapped to the default range of W, the native distribution [0 : 32]. Now, recalling the definitions above (for the n=100 case) we have that

$$P(T > 28) \cong P(Y > 0.28) = P(W > 31.36) \cong 0.397614.$$

The last equality is obtained by noticing that *W* will have approximately *Normal*(μ *= 31.23332;* σ^2 *= 0.48811895*^{2}) distribution, with empirical mean and standard deviation obtained from row 3 in the table.


Update 1
Stop 100
Refresh Stats Table
Fit normal (1)
Mean
N=100
Fit normal (2)
None
N=5
Go To CLT Activity
Histograms and Summaries
Distributions
Native Population
Sample Distribution
Sampling Dist. of Means, N=100
P(Y=35)=1/38
P(W=0)=1/38
P(Y=1)=37/38
P(W=32)=37/38
P(T > 28) = P(Ȳ > 0.28) =
P(W>31.36)= 0.397614
Sample Size
Mean
Median
SD
Skewness
Kurtosis
164.0
31.2195121...
32.0
4.93623829...
-6.1664414...
38.0249999...
0.0
0.5
NaN
0.0
0.0
208.0
31.2333200...
31.0
0.48811895...
-0.2076976...
0.33942781...
Summary statistics for the sampling distribution of the First Parameter

Confidence Intervals Activity

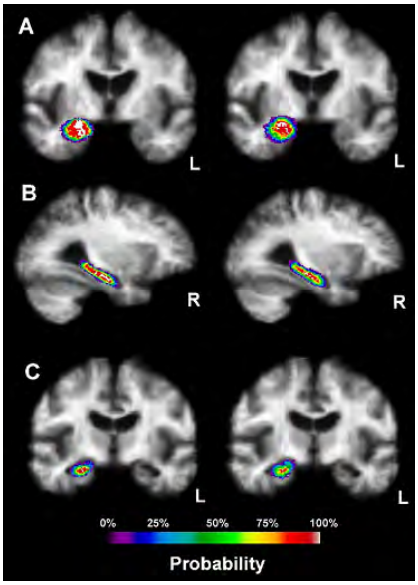
There are two types of parameter estimates – *point-based* and *interval-based* estimates. Point-estimates refer to unique quantitative estimates of various parameters. Interval-estimates represent ranges of plausible values for the parameters of interest. There are different algorithmic approaches, prior assumptions and principals for computing data-driven parameter estimates. Both point and interval estimates depend on the distribution of the process of interest, the available computational resources and other criteria that may be desirable (Stewarty 1999) – e.g., biasness and robustness of the estimates. Accurate, robust and efficient parameter estimation is critical in making inference about observable experiments, summarizing process characteristics and prediction of experimental behaviors.

This activity demonstrates the usage and functionality of SOCR General Confidence Interval Applet. This applet is complementary to the SOCR Simple Confidence Interval Applet and its corresponding activity.

The **aims** of this activity are to:

- Demonstrate the theory behind the use of interval-based estimates of parameters;
- Illustrate various confidence intervals construction recipes;
- Draw parallels between the construction algorithms and intuitive meaning of confidence intervals;
- Present a new technology-enhanced approach for understanding and utilizing confidence intervals for various applications.

Motivational example: Alzheimer's Disease Dataset



A 2005 study proposing a new computational brain atlas for Alzheimer's disease (Mega et al., 2005) investigated the mean volumetric characteristics and the spectra of shapes and sizes of different cortical and subcortical brain regions for Alzheimer's patients, individuals with minor cognitive impairment and asymptomatic subjects. This study estimated a number of centrality and variability parameters for these three populations. Based on these point- and interval-estimates, the study analyzed a number of digital scans to derive criteria for imaging-based classification of subjects based on the intensities of their 3D brain scans. Their results enabled a number of subsequent inference studies that quantified the effects of subject demographics (e.g., education level, familial history, APOE allele, etc.), stage of the disease and the efficacy of new drug treatments targeting Alzheimer's disease. The Figure to the right illustrates the *shape, center* and *distribution parameters* for the 3D geometric structure of the right hippocampus in the Alzheimer's disease brain atlas. New imaging data can then be coregistered and compared relative to the amount of anatomical variability encoded in this atlas. This enables automated, efficient and quantitative inference on

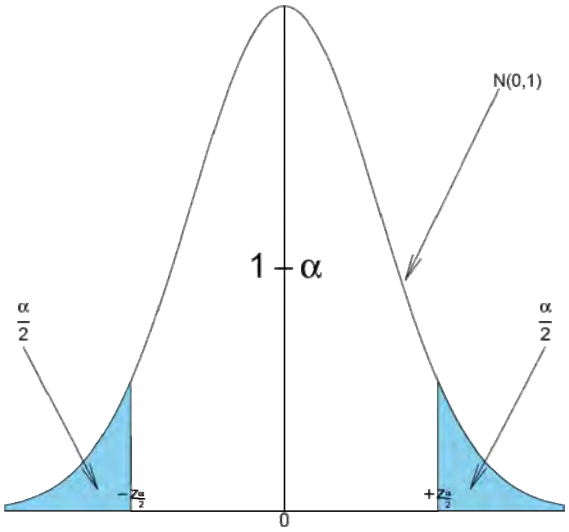
large number of brain volumes. Examples of point and interval estimates computed in this atlas framework include the mean-intensity and mean shape location, and the standard deviation of intensities and the mean deviation of shape, respectively.

Confidence intervals (CI) for the population mean μ of Normal population with known population variance σ^2

Let $X_1, X_2, X_3, ..., X_n$ be a random sample from $N(\mu,\sigma^2)$. We know that $\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. Therefore,

$$P\left(-z_{\frac{\alpha}{2}} \leq \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\frac{\alpha}{2}}\right) = 1-\alpha,$$

where $-z_{\frac{\alpha}{2}}$ and $z_{\frac{\alpha}{2}}$ are defined as shown in the figure below:



The area 1 − α is called *confidence level*. Usually, the choices for confidence levels are the following:

1 − α	$z_{\frac{\alpha}{2}}$
0.90	1.645
0.95	1.960
0.98	2.325
0.99	2.575

The expression above can be written as:

$$P\left(\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha.$$

We say that we are 1 − α confident that the mean μ falls in the interval

$$\bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}.$$

Example 1:

Suppose that the length of iron rods from a certain factory follows the Normal distribution with known standard deviations $\sigma = 0.2\ m$ but unknown mean μ. Construct a *95% confidence interval* for the population mean μ if a random sample of n=16 of these iron rods has sample mean $\bar{x} = 6m$.

We solve this problem by using our CI recipe:

$$\begin{gathered} 6 \pm 1.96 \frac{0.2}{\sqrt{16}} \\ 6 \pm 0.098 \\ 5.902 \leq \mu \leq 6.098. \end{gathered}$$

Sample size determination for a given length of the confidence interval

Find the sample size *n* needed when we want the width of the confidence interval to be ±E with confidence level 1 − α.

Solution: In the expression

$$\bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

the width of the confidence interval is given by $z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ (also called *margin of error*). We want this width to be equal to *E*. Therefore,

$$E = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \Rightarrow n = \left(\frac{z_{\frac{\alpha}{2}} \sigma}{E} \right)^2.$$

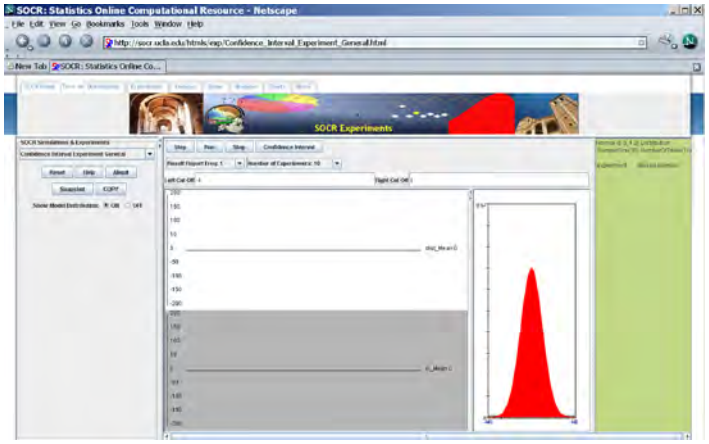
Example 2:

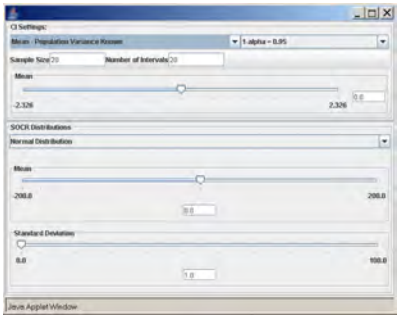
Following our first example above, suppose that we want the entire width of the confidence interval to be equal to *0.05 m*. Find the sample size *n* needed.

$$n = \left(\frac{1.96 \times 0.2}{0.025} \right)^2 = 245.9 \Rightarrow n \approx 246.$$

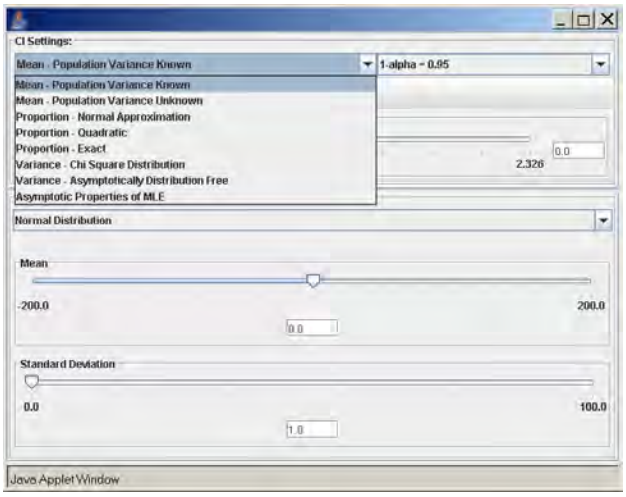
The SOCR Confidence Interval Applet

To access the SOCR applet on confidence intervals go to http://socr.ucla.edu/htmls/exp/Confidence_Interval_Experiment_General.html. To select the type and parameters of the specific confidence interval of interest click on the **Confidence Interval** button on the top – this will open a new pop-up window as shown below:

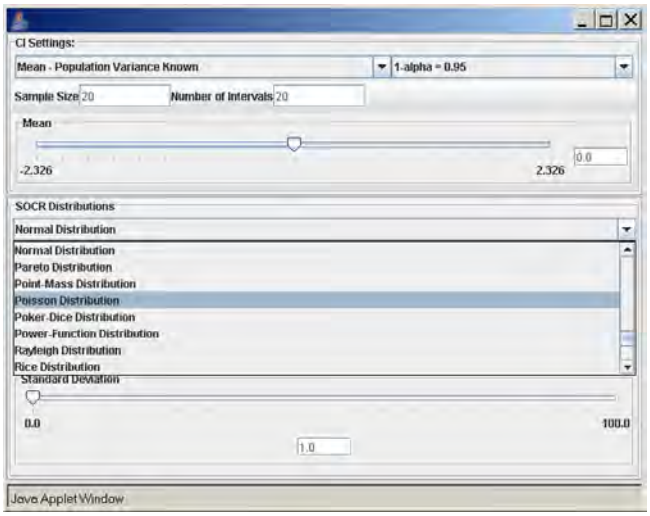




A confidence interval of interest can be selected from the drop-down list under *CI Settings*. In this case, we selected *Mean - Population Variance Known*.



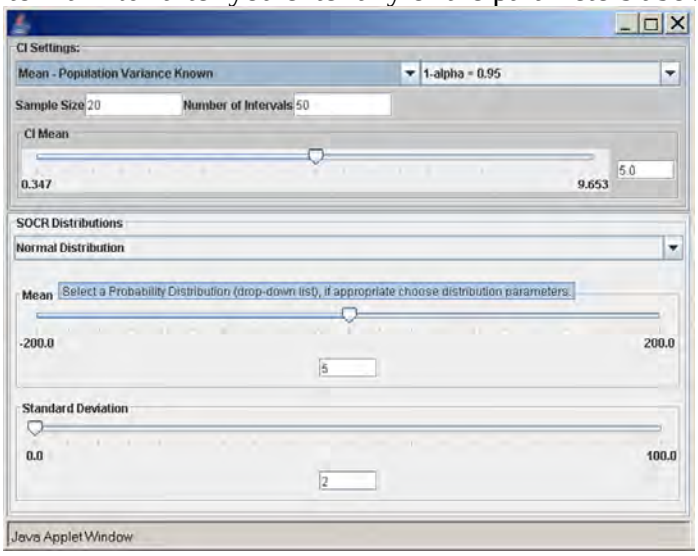
In the same pop-up window, under *SOCR Distributions*, the drop-down menu offers a list of all the available distributions of SOCR. These distributions are the same as the ones included in the SOCR Distributions applet (http://socr.ucla.edu/htmls/dist/).



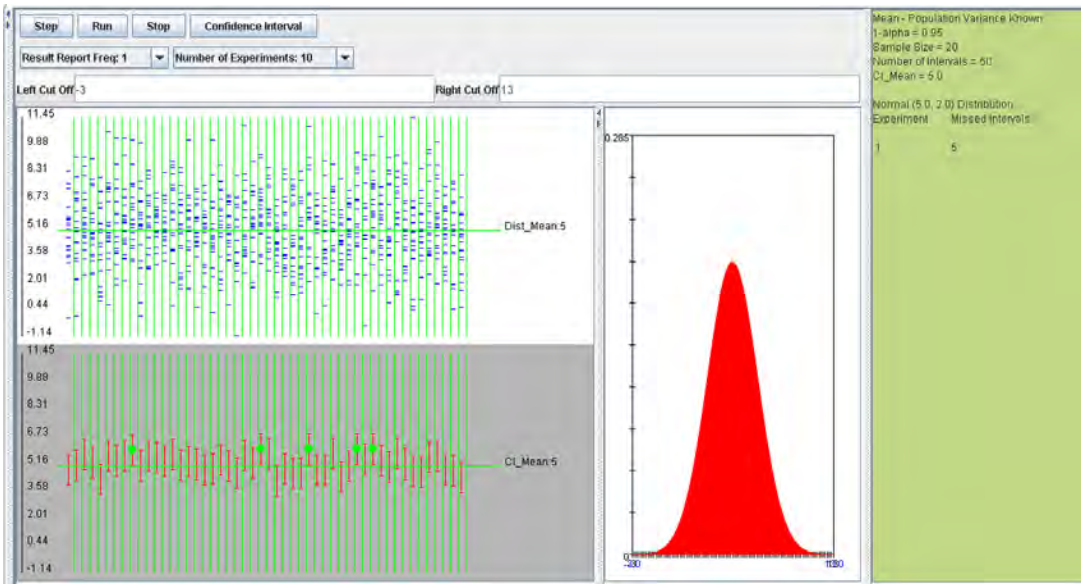
Once the desired distribution is selected, its parameters can be chosen numerically or via the sliders. In this example we select:

- *Normal distribution* with *mean 5* and *standard deviation 2*,
- sample size (number of observations selected from the distribution) is *20*,
- the confidence level ($1-\alpha = 0.95$), and
- the number of intervals to be constructed is 50 (see screenshot below).

Note: Make sure to hit *Enter* after you enter any of the parameters above.



To run the SOCR CI simulation, go back to the applet in the main browser window. We can run the experiment once, by clicking on the *Step* button, or many times by clicking on the *Run* button. The number of experiments can be controlled by the value of the *Number of Experiments* variable (10, 100, 1,000, 10,000, or continuously).



In the screenshot above we observe the following:

- The shape of the distribution that was selected (in this case Normal).
- The observations selected from the distribution for the construction of each of the 50 intervals shown in blue on the top-left graph panel.
- The confidence intervals shown as red line segments on the bottom-left panel.
- The green dots represent instances of confidence intervals that do not include the parameter (in this case population mean of 5).
- All the parameters and simulation results are summarized on the right panel of the applet.

Practice:

Run the same experiment using sample sizes of 20, 30, 40, 50 with the same confidence level (1 − *alpha* = 0.95). What are your observations and conclusions?

Confidence intervals for the population mean μ with known variance σ^2

From the central limit theorem we know that when the sample size is large (usually $n \geq 30$) the distribution of the sample mean $\overline{X}$ approximately follows $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

Therefore, the confidence interval for the population mean μ is approximately given by the expression we previously discussed:

$$P\left(\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) \approx 1 - \alpha.$$

The mean μ falls in the interval $\bar{x} \pm z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$.

Also, the sample size determination is given by the same formula:

$$E = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \Rightarrow n = \left(\frac{z_{\frac{\alpha}{2}} \sigma}{E} \right)^2 .$$

Example 3:

A sample of size *n=50* is taken from the production of light bulbs at a certain factory. The sample mean of the lifetime of these 50 light bulbs is found to be $\bar{x} = 1{,}570$ hours. Assume that the population standard deviation is σ = 120 hours.

- Construct a *95% confidence interval* for μ.
- Construct a *99% confidence interval* for μ.
- What sample size is needed so that the length of the interval is 30 hours with *95% confidence*?

Empirical Investigation

Two dice are rolled and the sum X of the two numbers that occurred is recorded. The probability distribution of X is as follows:

X	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
P(X)	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

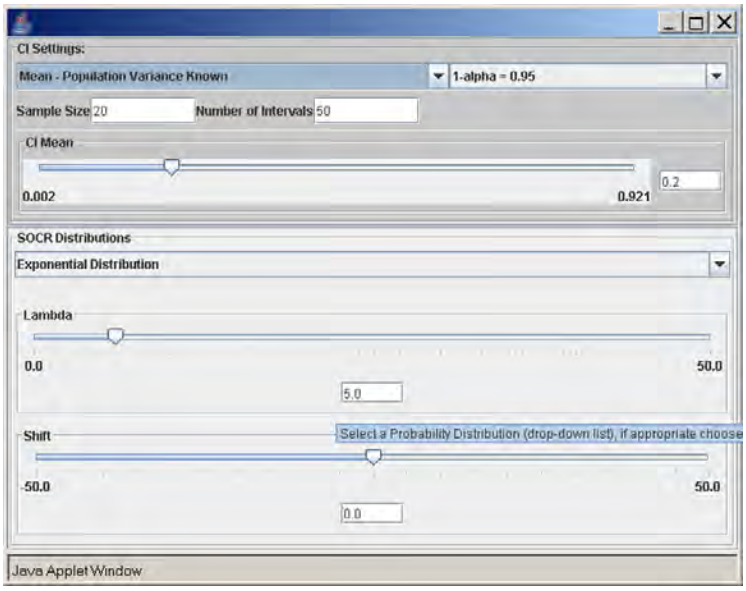
This distribution has mean μ = 7 and standard deviation σ = 2.42. We take 100 samples of size n=50 each from this distribution and compute for each sample the sample mean $\bar{x}$. Pretend now that we only know that σ = 2.42, and that μ is unknown. We are going to use these 100 sample means to construct 100 confidence intervals, each one with *95% confidence level* for the true population mean μ. Here are the results:

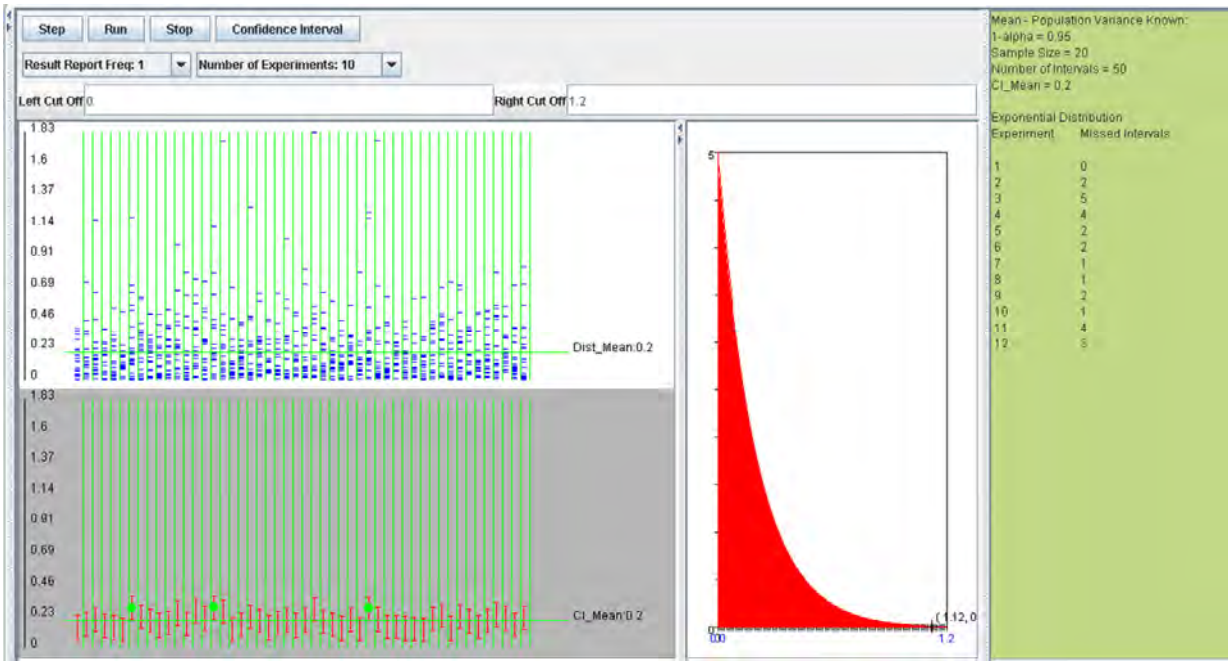
Sample	$\bar{x}$	**95% CI for μ:** $\bar{x} - 1.96\frac{2.42}{\sqrt{50}} \le \mu \le \bar{x} + 1.96\frac{2.42}{\sqrt{50}}$	**Is μ = 7 included?**
1	6.9	$6.23 \le \mu \le 7.57$	YES
2	6.3	$5.63 \le \mu \le 6.97$	NO
3	6.58	$5.91 \le \mu \le 7.25$	YES
4	6.54	$5.87 \le \mu \le 7.21$	YES
5	6.7	$6.03 \le \mu \le 7.37$	YES
6	6.58	$5.91 \le \mu \le 7.25$	YES
7	7.2	$6.53 \le \mu \le 7.87$	YES
		…	
100	7.3	$6.63 \le \mu \le 7.97$	YES

We observe that four confidence intervals among the 100 that we constructed fail to include the true population mean $\mu = 7$ (about 5%).

Example 4:

For this example, we will select the Exponential distribution with $\lambda = 5$ (mean of 1/5 = 0.2), sample size 60, confidence level 0.95, and number of intervals 50. These settings, along with the results of the simulations are shown below





Confidence intervals for the population mean of Normal distribution when the population variance σ^2 is unknown

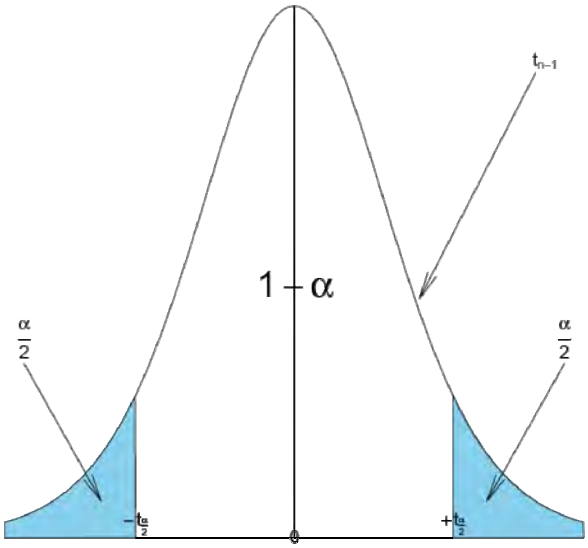
Let $X_1, X_2, X_3, ..., X_n$ be a random sample from $N(\mu,\sigma^2)$. It is known that

$$\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}.$$

Therefore,

$$P\left(-t_{\frac{\alpha}{2};n-1} \leq \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}} \leq t_{\frac{\alpha}{2};n-1}\right) = 1-\alpha,$$

where $-t_{\frac{\alpha}{2};(n-1)}$ and $t_{\frac{\alpha}{2};(n-1)}$ are defined as follows:



As before, the area $1-\alpha$ is called the *confidence level*. The values of $t_{\frac{\alpha}{2};(n-1)}$ can be found from:

- SOCR Student's T-distribution applet (http://socr.ucla.edu/htmls/dist/StudentT_Distribution.html), or
- SOCR T-table (http://socr.ucla.edu/Applets.dir/T-table.html). Below are some examples:

1 − α	**n**	$t_{\frac{\alpha}{2};n-1}$
0.90	13	1.782
0.95	21	2.086
0.98	31	2.457
0.99	61	2.660

Note: The sample standard deviation is computed as follows:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

or using the shortcut formula.

$$s = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right]}$$

After some rearranging the expression above can be written as:

$$P\left(\bar{x} - t_{\frac{\alpha}{2};n-1}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\frac{\alpha}{2};n-1}\frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

We say that we are 1 − α confident that μ falls in the interval:

$$\bar{x} \pm t_{\frac{\alpha}{2};n-1}\frac{s}{\sqrt{n}}.$$

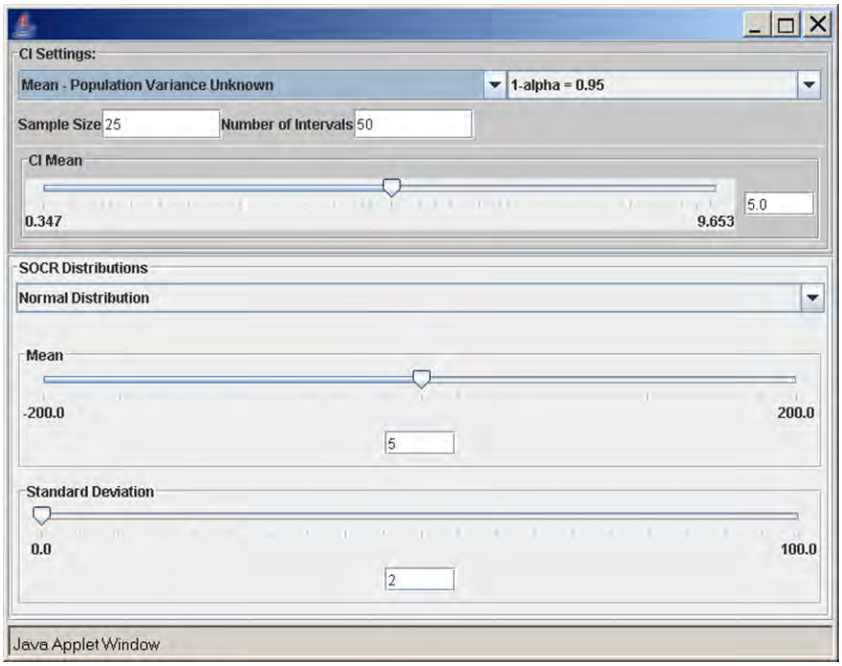
Example 3:

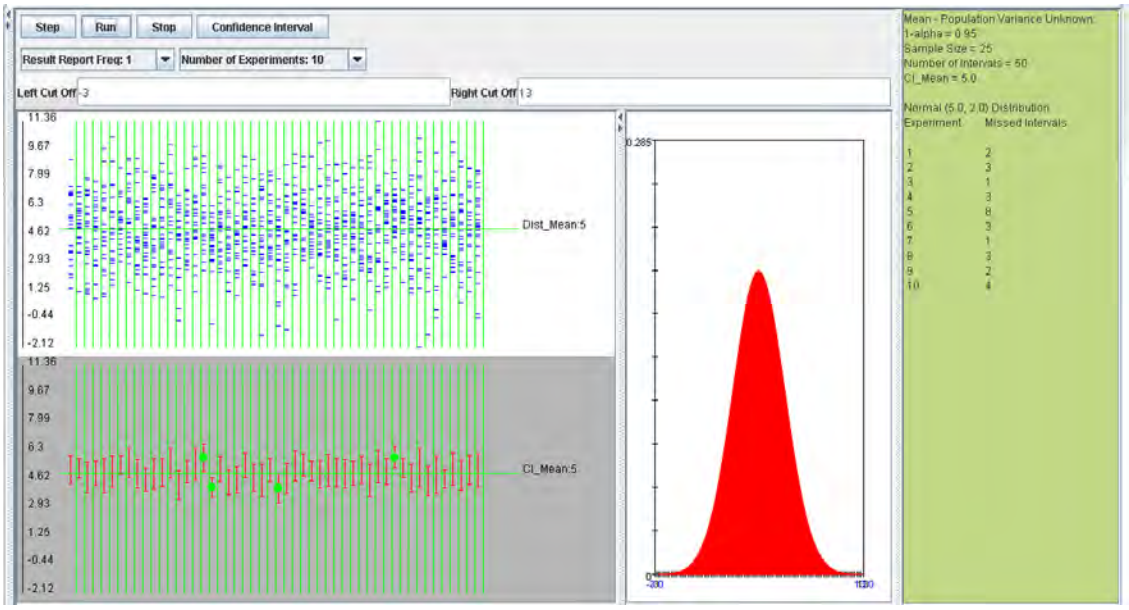
The daily production of a chemical product last week in tons was: 785, 805, 790, 793, and 802.

- Construct a *95% confidence interval* for the population mean μ.
- What assumptions are necessary?

SOCR Investigation

For this case, we will select the Normal distribution with mean 5 and standard deviation 2, sample size of 25, number of intervals 50, and confidence level 0.95. These settings and simulation results are shown below:





We observe that the length of the confidence interval differs for all the intervals because the margin of error is computed using the sample standard deviation.

Confidence interval for the population proportion *p*

Let $Y_1, Y_2, Y_3, ..., Y_n$ be a random sample from the Bernoulli distribution with probability of success *p*. To construct a confidence interval for *p*, we let $X = \sum_{i=1}^{n} Y_i$. The following result is used based on Normal approximation:

$$\frac{X - np}{\sqrt{np(1-p)}} \sim N(0,1).$$

Therefore,

$$P\left(-z_{\frac{\alpha}{2}} \leq \frac{X - np}{\sqrt{np(1-p)}} \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha,$$

where $-z_{\frac{\alpha}{2}}$ and $z_{\frac{\alpha}{2}}$ are defined as above.

After rearranging we get:

$$P\left(\frac{X}{n} - z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \frac{X}{n} + z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right) = 1 - \alpha.$$

The ratio $\frac{x}{n}$ is the *point estimate* of the population *p* and it is denoted with $\hat{p} = \frac{x}{n}$. The problem with this interval is that the unknown *p* appears also at the end points of the interval. As an approximation we can simply replace *p* with its estimate $\hat{p} = \frac{x}{n}$. Finally the confidence interval is given by:

$$P\left(\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 1 - \alpha.$$

We say that we are 1 − α confident that *p* falls in

$$\hat{p} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Calculating Sample Sizes

The basic problem we will address now is how to determine the sample size needed so that the resulting confidence interval will have a fixed margin of error *E* with confidence level 1 − α.

In the expression

$$\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

the width of the confidence interval is given by the *margin of error* $z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

We can simply solve for *n*:

$$E = z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \Rightarrow n = \frac{z_{\frac{\alpha}{2}}^2 \hat{p}(1-\hat{p})}{E^2}.$$

However, the value of $\hat{p}$ is not known because we have not observed our sample yet. If we use $\hat{p} = 0.5$, we will obtain the largest possible sample size. Of course, if we have an idea about its value (from another study, etc.) we can use it.

Example 6

At a survey poll before the elections candidate *A* receives the support of 650 voters in a sample of 1,200 voters.

- Construct a 95% confidence interval for the population proportion *p* that supports candidate *A*.
- Find the sample size needed so that the margin of error will be ± 0.01 with confidence level 95%.

Another formula for the confidence interval for the population proportion *p*

$$P\left(-z_{\frac{\alpha}{2}} \leq \frac{X - np}{\sqrt{np(1-p)}} \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

$$P\left(-z_{\frac{\alpha}{2}} \leq \frac{\frac{X}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

$$P\left(\frac{|\hat{p}-p|}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{\frac{\alpha}{2}}\right) = 1-\alpha$$
$$P\left(\frac{(\hat{p}-p)^2}{\frac{p(1-p)}{n}} \leq z^2_{\frac{\alpha}{2}}\right) = 1-\alpha$$

We obtain a quadratic expression in terms of *p*:

$$(\hat{p}-p)^2 - z^2_{\frac{\alpha}{2}}\frac{p(1-p)}{n} \leq 0$$

$$(1+\frac{z^2_{\frac{\alpha}{2}}}{n})p^2 - (2\hat{p}+\frac{z^2_{\frac{\alpha}{2}}}{n})p + \hat{p}^2 = 0$$

Solving for *p* we get the following confidence interval:

$$\frac{\hat{p}+\frac{z^2_{\frac{\alpha}{2}}}{2n} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}+\frac{z^2_{\frac{\alpha}{2}}}{4n^2}}}{1+\frac{z^2_{\frac{\alpha}{2}}}{n}}.$$

When *n* is large, this interval is the same as before.

Exact Confidence Interval for *p*

The first interval for proportions above (Normal approximation) produces intervals that are too narrow when the sample size is small. The coverage is below 1 − α. The following **exact** method (or Clopper-Pearson) improves the low coverage of the Normal approximation confidence interval. The exact confidence interval however has higher coverage than 1 − α.

$$\left[1+\frac{n-x+1}{xF_{1-\frac{\alpha}{2};2x,2(n-x+1)}}\right]^{-1} < p < \left[1+\frac{n-x}{(x+1)F_{\frac{\alpha}{2};2(x+1),2(n-x)}}\right]^{-1},$$

where, *x* is the number of successes among *n* trials, and $F_{a,b,c}$ is the *a* quantile of the *F* distribution with numerator degrees of freedom *b*, and denominator degrees of freedom *c*.

Confidence interval for the population variance σ^2 of the Normal distribution

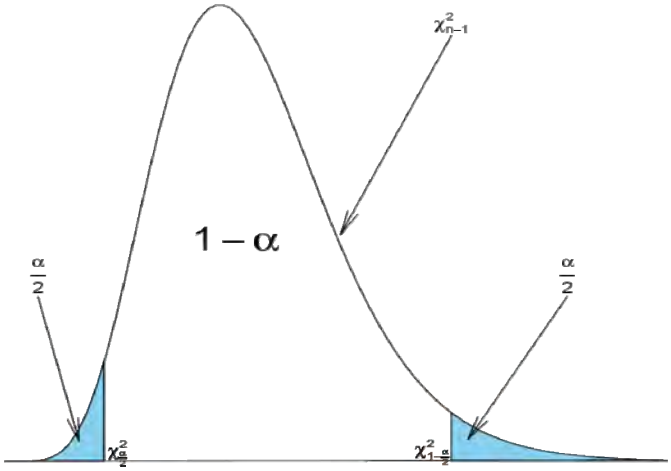
Again, let $X_1, X_2, X_3, ..., X_n$ random sample from $N(\mu,\sigma^2)$. It is known that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Therefore,

$$P\left(\chi^2_{\frac{\alpha}{2};n-1} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{1-\frac{\alpha}{2};n-1}\right) = 1-\alpha,$$

where $\chi^2_{\frac{\alpha}{2};(n-1)}$ and $\chi^2_{1-\frac{\alpha}{2};(n-1)}$ are defined as follows:



As with the T-distribution, the values of $\chi^2_{\frac{\alpha}{2};n-1}$ and $\chi^2_{1-\frac{\alpha}{2};n-1}$ may be found from:

- Interactive SOCR Chi-Square Distribution applet, (http://socr.ucla.edu/htmls/dist/ChiSquare_Distribution.html) or
- The SOCR Chi-Square Table (http://socr.ucla.edu/Applets.dir/ChiSquareTable.html). Some examples are included below:

1 − α	**n**	$\chi^2_{\frac{\alpha}{2};n-1}$	$\chi^2_{1-\frac{\alpha}{2};n-1}$
0.90	4	0.352	7.81
0.95	16	6.26	27.49
0.98	25	10.86	42.98
0.99	41	20.71	66.77

If we rearrange the inequality above we get:

$$P\left(\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2};n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2};n-1}}\right) = 1-\alpha.$$

We say that we are $1 - \alpha$ confident that the population variance σ^2 falls in the interval:

$$\left[\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2};n-1}}, \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2};n-1}}\right]$$

Comment: When the sample size *n* is large, the χ^2_{n-1} distribution can be approximated by $N(n-1,2(n-1))$. Therefore, in such situations, the confidence interval for the variance can be computed as follows:

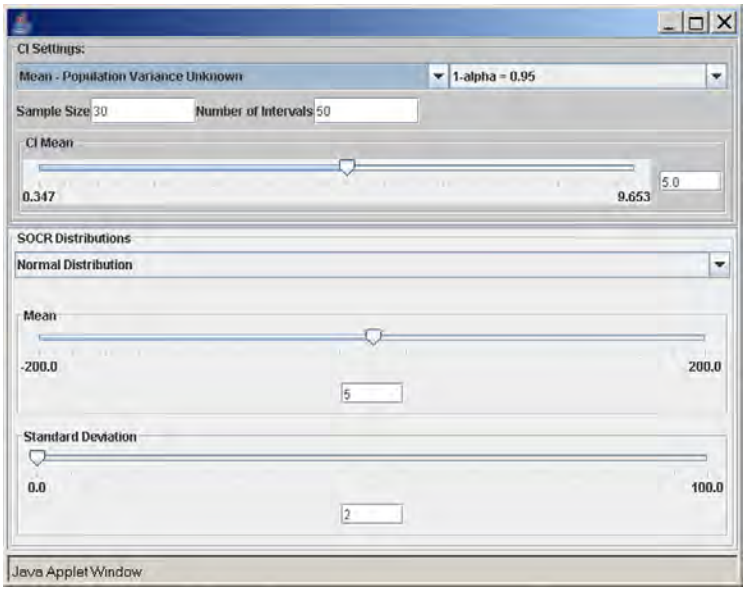
$$\frac{s^2}{1+z_{\frac{\alpha}{2}}\sqrt{\frac{2}{n-2}}} \leq \sigma^2 \leq \frac{s^2}{1-z_{\frac{\alpha}{2}}\sqrt{\frac{2}{n-2}}}.$$

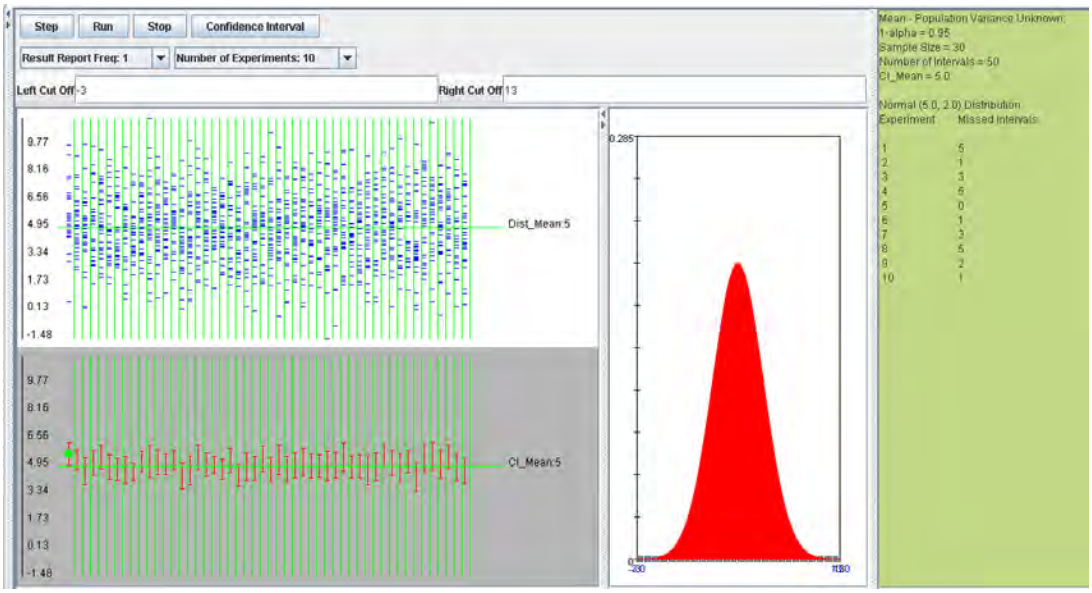
Example 7:

A precision instrument is guaranteed to read accurately to within 2 units. A sample of 4 instrument readings on the same object yielded the measurements 353, 351, 351, and 355. Find a 90% confidence interval for the population variance. Assume that these observations were selected from a population that follows the Normal distribution.

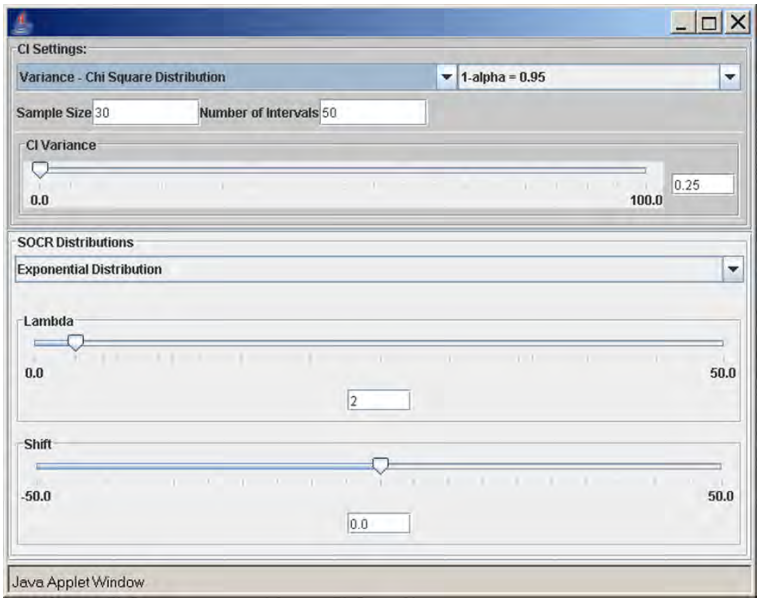
SOCR Investigation

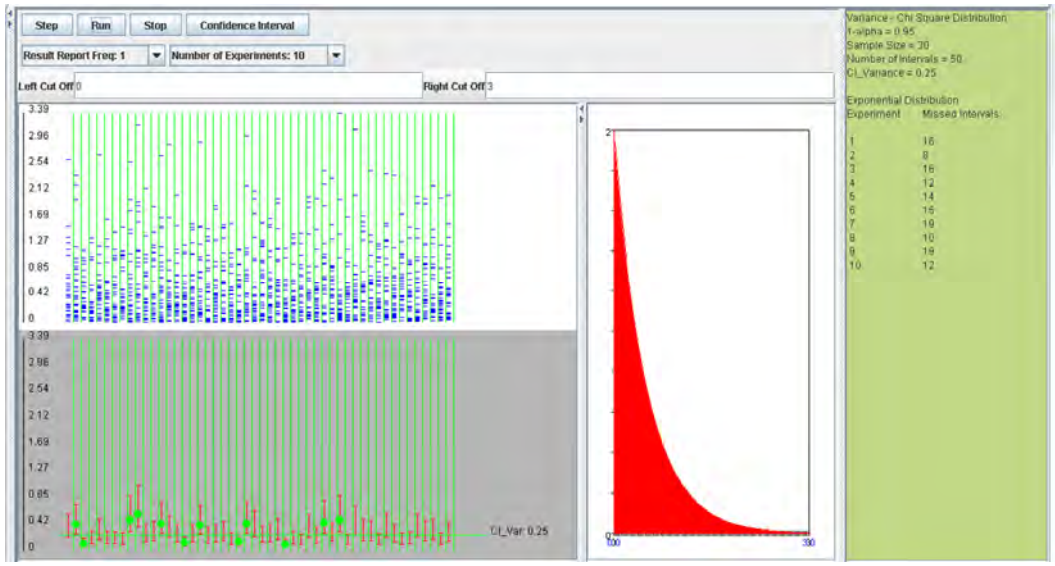
Using the SOCR confidence intervals applet, we run the following simulation experiment: Normal distribution with mean 5 and standard deviation 2, sample size 30, confidence intervals 50, and confidence level 0.95.

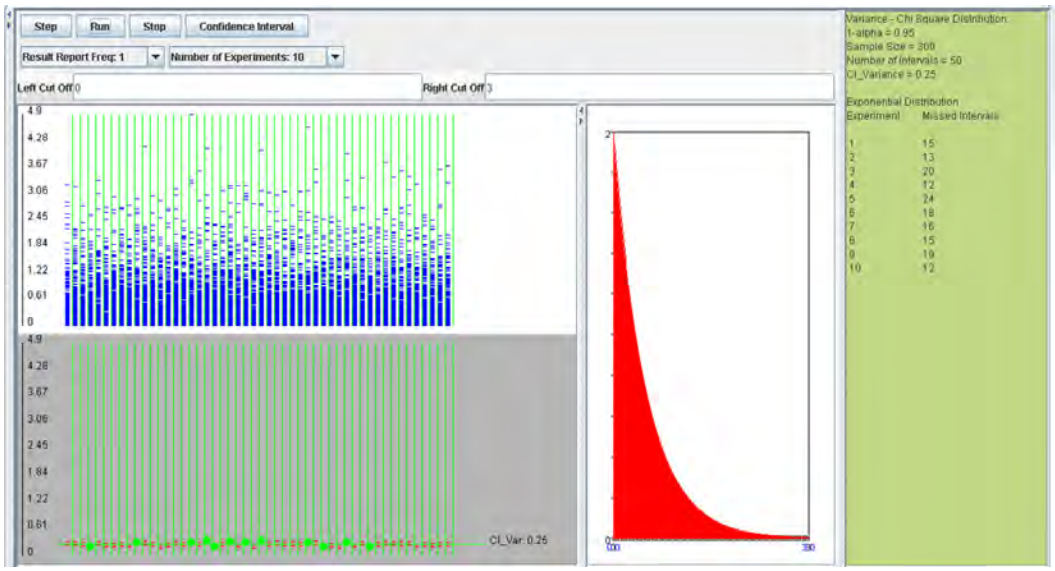




However, if the population is *not Normal*, the coverage is poor and this can be seen with the following SOCR example. Consider the exponential distribution with $\lambda = 2$ (variance is $\sigma^2 = 0.25$). If we use the confidence interval based on the χ^2 distribution, as described above, we obtain the following results (first with sample size 30 and then sample size 300).


CI Settings:
Variance - Chi Square Distribution
1-alpha = 0.95
Sample Size 30
Number of Intervals 50
CI Variance
0.0
100.0
0.25
SOCR Distributions
Exponential Distribution
Lambda
0.0
50.0
2
Shift
-50.0
50.0
0.0
Java Applet Window


Step
Run
Stop
Confidence Interval
Result Report Freq: 1
Number of Experiments: 10
Left Cut Off 0
Right Cut Off 3
CI_Var: 0.25
Variance - Chi Square Distribution
1-alpha = 0.95
Sample Size = 30
Number of Intervals = 50
CI_Variance = 0.25
Exponential Distribution
Experiment Missed Intervals
1 16
2 8
3 16
4 12
5 14
6 16
7 19
8 10
9 19
10 12


Step
Run
Stop
Confidence Interval
Result Report Freq: 1
Number of Experiments: 10
Left Cut Off 0
Right Cut Off 3
CI_Var: 0.25
Variance - Chi Square Distribution
1-alpha = 0.95
Sample Size = 300
Number of Intervals = 50
CI_Variance = 0.25
Exponential Distribution
Experiment Missed Intervals
1 15
2 13
3 20
4 12
5 24
6 18
7 16
8 15
9 19
10 12

We observe that regardless of the sample size the 95% CI(σ^2) coverage is poor. In these situations, (sampling from non-Normal populations) an asymptotically distribution-free confidence interval for the variance can be obtained using the following large sample theory result:

$$\sqrt{n}\left(s^2 - \sigma^2\right) \longrightarrow N\left(m_4 - \sigma^4\right).$$

That is,

$$\frac{\sqrt{n}(s^2 - \sigma^2)}{\sqrt{\mu_4 - \sigma^4}} \rightarrow N(0,1),$$

where, $\mu_4 = E(X - \mu)^4$ is the fourth moment of the distribution. Of course, μ_4 is unknown and will be estimated by the fourth sample moment

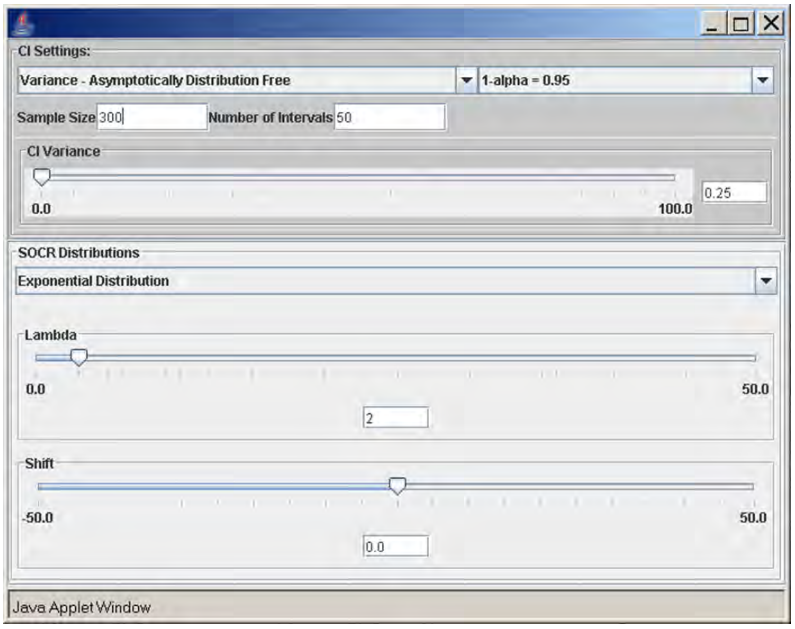
$$\mu_4 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^4$$

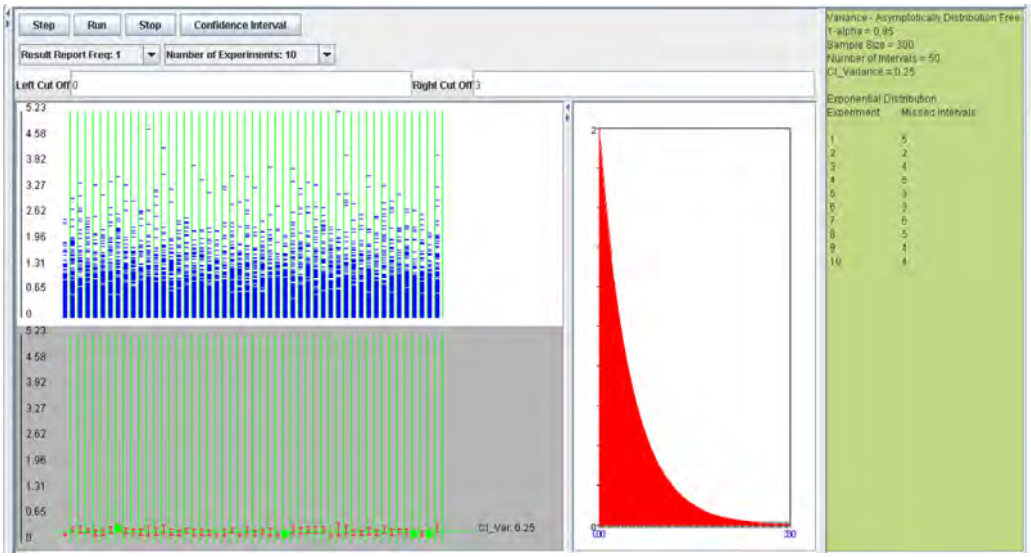
.

The confidence interval for the population variance is computed as follows:

$$s^2 - z_{\frac{\alpha}{2}}\frac{\sqrt{m_4 - s^4}}{\sqrt{n}} \leq \sigma^2 \leq s^2 + z_{\frac{\alpha}{2}}\frac{\sqrt{m_4 - s^4}}{\sqrt{n}}.$$

Using the SOCR CI Applet with exponential distribution ($\lambda = 2$), sample size 300, number of intervals 50, and confidence level 0.95, we see that the coverage of *this interval* is approximately 95%.





The 95% CI(σ^2) coverage for the intervals constructed using the method of asymptotic distribution-free intervals is much closer to 95%.

Confidence intervals for the population parameters of a distribution based on the *asymptotic properties of maximum likelihood estimates*

To construct confidence intervals for a parameter of some distribution the following method can be used based on the large sample theory of maximum likelihood estimates. As the sample size n increases it can be shown that the maximum likelihood estimate $\hat{\theta}$ of a parameter θ follows approximately the Normal distribution with mean θ and variance equal to the lower bound of the Cramer-Rao inequality.

$$\hat{\theta} \sim N\left(\theta, \frac{1}{nI(\theta)}\right),$$

where $\sqrt{\frac{1}{nI(\theta)}}$ is the lower bound of the Cramer-Rao inequality.

Because $I(\theta)$ (Fisher's information) is a function of the unknown parameter θ we replace θ with its maximum likelihood estimate $\hat{\theta}$ to get $I\left(\hat{\theta}\right)$.

Since,

$$Z = \frac{\hat{\theta} - \theta}{\sqrt{\frac{1}{nI(\hat{\theta})}}},$$

we can write $P\left(-z_{\frac{\alpha}{2}} \le Z \le z_{\frac{\alpha}{2}}\right)$ as

$$P\left(-z_{\frac{\alpha}{2}} \le \frac{\hat{\theta}-\theta}{\sqrt{\frac{1}{nI(\hat{\theta})}}} \le z_{\frac{\alpha}{2}}\right).$$

Therefore,

$$P\left(\hat{\theta} - z_{\frac{\alpha}{2}}\sqrt{\frac{1}{nI(\hat{\theta})}} \le \theta \le \hat{\theta} + z_{\frac{\alpha}{2}}\sqrt{\frac{1}{nI(\hat{\theta})}}\right).$$

Thus, we are 1 − α confident that θ falls in the interval

$$\hat{\theta} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{1}{nI(\hat{\theta})}}.$$

Example 8:

Use the result above to construct a confidence interval for the Poisson parameter λ. Let $X_1, X_2, X_3, ..., X_n$ be independent and identically distributed random variables from a Poisson distribution with parameter λ.

We know that the maximum likelihood estimate of λ is $\hat{\lambda} = \bar{x}$. We need to find the lower bound of the Cramer-Rao inequality:

$$f(x) = \frac{\lambda e^{-\lambda x}}{x!} \Rightarrow lnf(x) = xln\lambda - \lambda - lnx!.$$

Let's find the first and second derivatives w.r.t. λ.

$$\frac{\partial lnf(x)}{\partial \lambda} = \frac{x}{\lambda} - 1,$$
$$\frac{\partial^2 lnf(x)}{\partial \lambda^2} = -\frac{x}{\lambda^2}.$$

Therefore,

$$\frac{1}{-nE\left(\frac{\partial^2 lnf(x)}{\partial \lambda^2}\right)} = \frac{1}{-nE(-\frac{X}{\lambda^2})} = \frac{\lambda^2}{\lambda n} = \frac{\lambda}{n}.$$

When *n* is large, $\hat{\lambda}$ follows approximately $\hat{\lambda} \sim N\left(\lambda, \frac{\lambda}{n}\right)$. Because λ is unknown, we replace it with its MLE estimate $\hat{\lambda}$:

$$\hat{\lambda} \sim N\left(\overline{X}, \frac{\overline{X}}{n}\right).$$

Therefore, the confidence interval for λ is:

$$\overline{X} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\overline{X}}{n}}.$$

Application:

The number of pine trees at a certain forest follows the Poisson distribution with unknown parameter λ per acre. A random sample of size *n=50* acres is selected and the number of pine trees in each acre is counted. Here are the results:

```
7 4 5 3 1 5 7 6 4 3 2 6 6 9 2 3 3 7 2 5 5 4 4 8 8

7 2 6 3 5 0 5 8 9 3 4 5 4 6 1 0 5 4 6 3 6 9 5 7 6
```

The sample mean is $\overline{x} = 4.76$. Therefore, a 95% confidence interval for the parameter λ is

$$4.76 \pm 1.96\sqrt{\frac{4.76}{50}}$$

That is, 4.76±0.31.

Therefore,

$$4.15 \leq \lambda \leq 5.34.$$

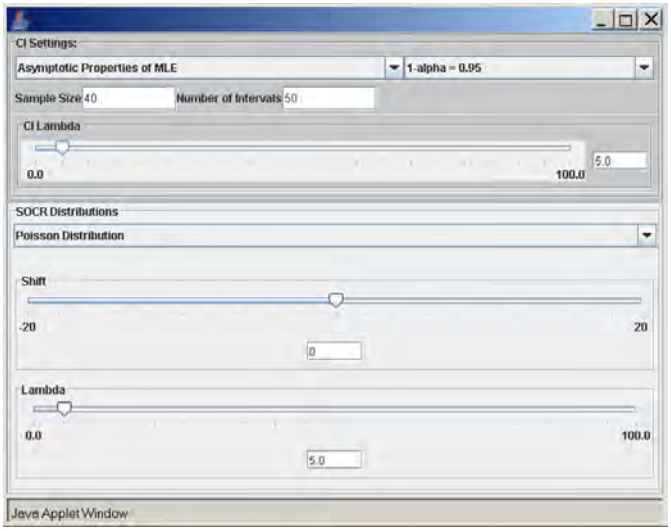
Exponential Distribution

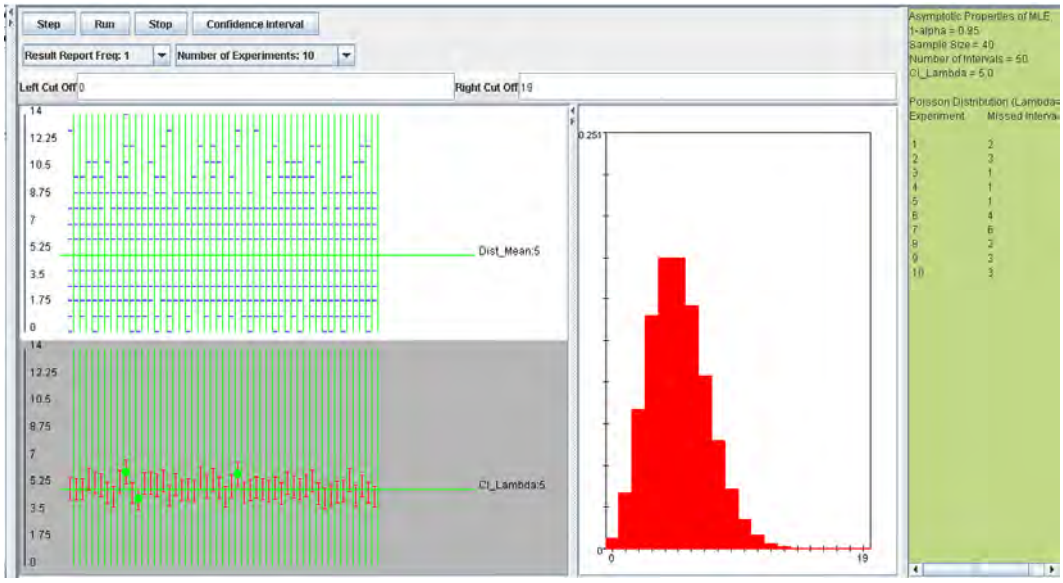
Verify that for the parameter λ of the exponential distribution the confidence interval obtained by this method is given as follows:

$$\frac{1}{\bar{x}} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{1}{n\bar{x}^2}}.$$

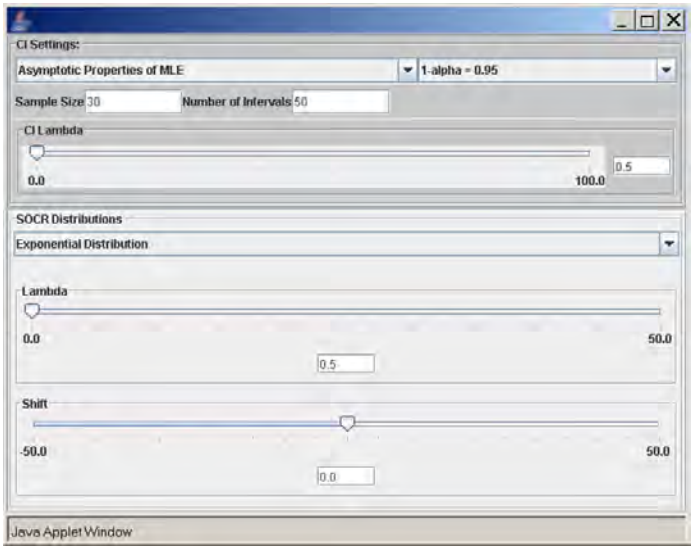
The following SOCR simulations refer to:

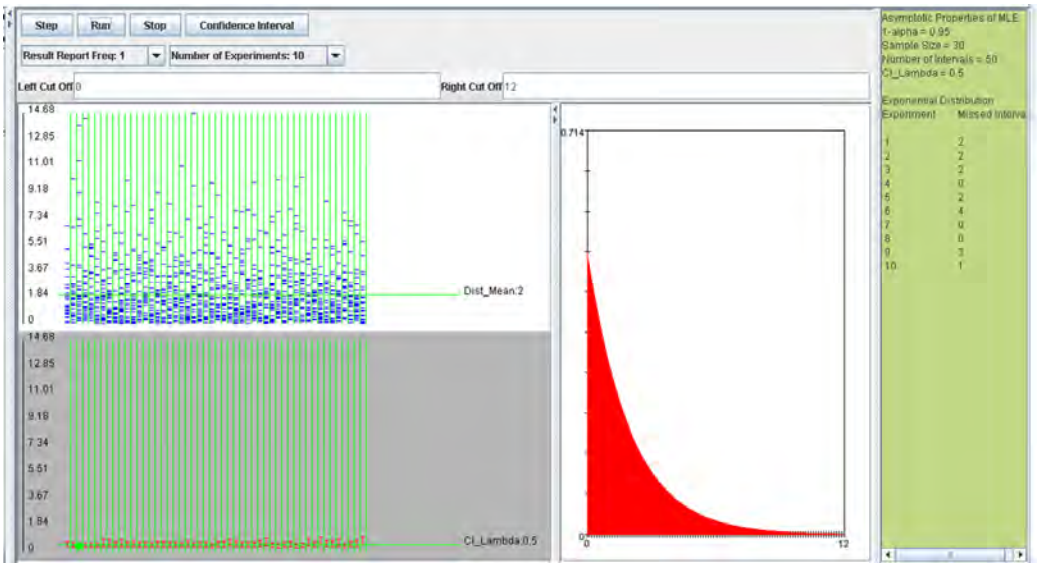
- Poisson distribution, $\lambda = 5$, sample size 40, number of intervals 50, confidence level 0.95.





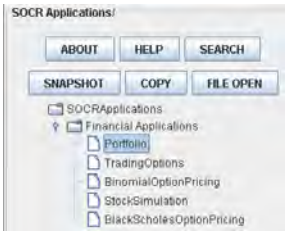
- Exponential distribution, λ = 0.5, sample size 30, number of intervals 50, confidence level 0.95.





SOCR Application Activities

- Portfolio Risk Management



The SOCR portfolio applet is part of the SOCR Applications (www.socr.ucla.edu/htmls/SOCR_Applications.html). Use the hierarchical application navigator to select *Financial Applications* → *Portfolio*.

An investor has a certain amount of dollars to invest into two stocks (*IBM* and *TEXACO*). A portion of the available funds will be invested into IBM (denote this portion of the funds with x_A) and the remaining funds into TEXACO (denote it with x_B). Thus, $x_A + x_B = 1$. The resulting portfolio will be $x_A R_A + x_B R_B$, where R_A is the monthly return of *IBM* and R_B is the monthly return of *TEXACO*. The goal here is to find the most efficient portfolios given a certain amount of risk.

Using market data from January 1980 until February 2001 we compute that $E(R_A) = 0.010$, $E(R_B) = 0.013$, $Var(R_A) = 0.0061$, $Var(R_B) = 0.0046$, and $Cov(R_A,R_B) = 0.00062$. We first want to *minimize the variance of the portfolio*. This will be:

$$\begin{aligned} &\text{Minimize}: && VAR(x_A R_A + x_B R_B) \\ &\text{Subject to restriction}: && x_A + x_B = 1. \end{aligned}$$

In other words,

$$\begin{aligned} &\text{Minimize}: && x_A^2\, VAR(R_A) + x_B^2\, VAR(R_B) + 2x_A x_B\, \text{COV}(R_A, R_B) \\ &\text{Subject to restriction}: && x_A + x_B = 1. \end{aligned}$$

Therefore our goal is to find x_A and x_B, the percentage of the available funds that will be invested in each stock. Substituting $x_B = 1 - x_A$ into the equation of the variance we get

$$\text{Minimize}: \quad x_A^2\, VAR(R_A) + (1-x_A)^2\, VAR(R_B) + 2x_A(1-x_A)\text{COV}(R_A, R_B)$$

To minimize the above expression we take the derivative with respect to x_A, set it equal to zero and solve for x_A. The result is:

$$x_A = \frac{VAR(R_B) - \text{COV}(R_A, R_B)}{VAR(R_A) + VAR(R_B) - 2COV(R_A, R_B)}$$

and therefore

$$x_B = \frac{VAR(R_A) - \text{COV}(R_A, R_B)}{VAR(R_A) + VAR(R_B) - 2COV(R_A, R_B)}$$

Plugging in the given information we get the values of x_A and x_B:

$$x_A = \frac{0.0046 - 0.0062}{0.0061 + 0.0046 - 2 \times 0.0062} = 0.42$$

and

$$x_B = 1 - x_A = 1 - 0.42 = 0.58.$$

Therefore if the investor invests 42% of the available funds into *IBM* and the remaining 58% into *TEXACO*, the *variance of the portfolio will be minimal* and equal to:

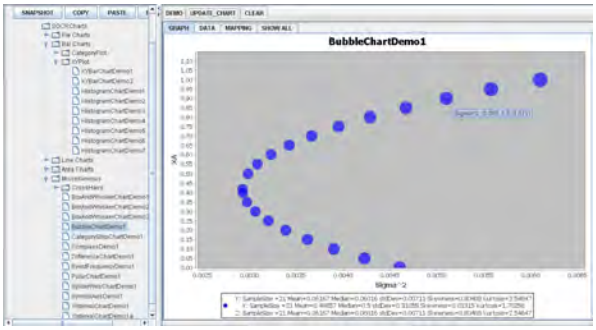
$$VAR(0.42R_A + 0.58R_B) = 0.42^2(0.0061) + 0.58^2(0.0046) + 2(0.42)(0.58)(0.00062) = 0.002926.$$

The corresponding expected return of this portfolio will be:

$$E(0.42R_A + 0.58R_B) = 0.42(0.010) + 0.58(0.013) = 0.01174.$$

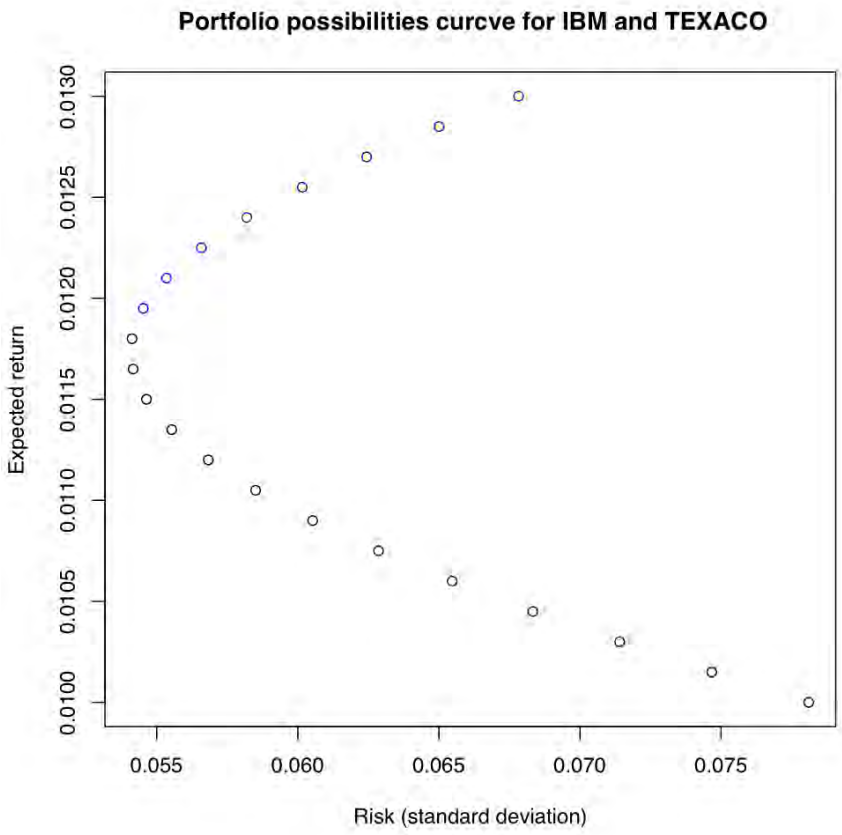
We can try many other combinations of x_A and x_B (but always $x_A + x_B = 1$) and compute the risk and return for each resulting portfolio. This is shown in the table and the graph below.

x_A	x_B	$Risk(\sigma^2)$	*Return*	$Risk(\sigma)$
1.00	0.00	0.006100	0.01000	0.078102
0.95	0.05	0.005576	0.01015	0.074670
0.90	0.10	0.005099	0.01030	0.071404
0.85	0.15	0.004669	0.01045	0.068329
0.80	0.20	0.004286	0.01060	0.065471
0.75	0.25	0.003951	0.01075	0.062859
0.70	0.30	0.003663	0.01090	0.060526
0.65	0.35	0.003423	0.01105	0.058505
0.60	0.40	0.003230	0.01120	0.056830
0.55	0.45	0.003084	0.01135	0.055531
0.50	0.50	0.002985	0.01150	0.054635
0.42	0.58	0.002926	0.01174	0.054088
0.40	0.60	0.002930	0.01180	0.054126
0.35	0.65	0.002973	0.01195	0.054524
0.30	0.70	0.003063	0.01210	0.055348
0.25	0.75	0.003201	0.01225	0.056580
0.20	0.80	0.003386	0.01240	0.058193
0.15	0.85	0.003619	0.01255	0.060157
0.10	0.90	0.003899	0.01270	0.062439
0.05	0.95	0.004226	0.01285	0.065005
0.00	1.00	0.004600	0.01300	0.067823



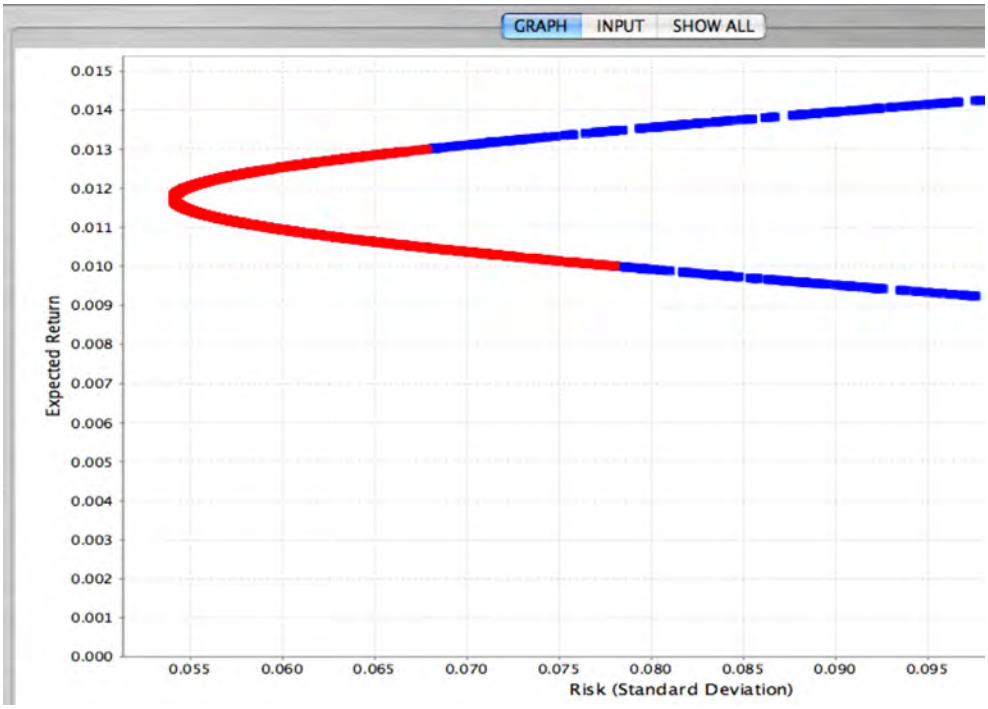
Using SOCR Bubble Chart to display the data using the following mapping:

www.socr.ucla.edu/htmls/SOCR_Charts.html



For the above calculations short selling was not allowed ($0 \le x_A \le 1$ and $0 \le x_B \le 1$, in addition to $x_A + x_B = 1$). The efficient portfolios are located on the top part of the graph between the minimum risk portfolio point and the maximum return portfolio point, which is called the efficient frontier (the blue portion of the graph). Efficient portfolios should provide higher expected return for the same level of risk or lower risk for the same level of expected return.

If short sales are allowed, which means that the investor can sell a stock that he or she does not own the graph has the same shape but now with more possibilities. The investor can have very large expected return but this will be associated with very large risk. The constraint here is only $x_A + x_B = 1$, since either x_A or x_B can be negative. The SOCR applet snapshot below shows the "*short sales scenario*" for the IBM and TEXACO stocks. The blue portion of the portfolio possibilities curve occurs when short sales are allowed, while the red portion corresponds to the case when short sales are not allowed.



When the investor faces the efficient frontier when short sales are allowed and he or she can lend or borrow at the risk-free interest rate the efficient frontier will change in the following way: Let x be the portion of the investor's wealth invested in portfolio A that lies on the efficient frontier, and $1 - x$ the portion invested in a risk-free asset. This combination is a new portfolio and has

$$\overline{R}_p = x\overline{R}_A + (1-x)R_f ,$$

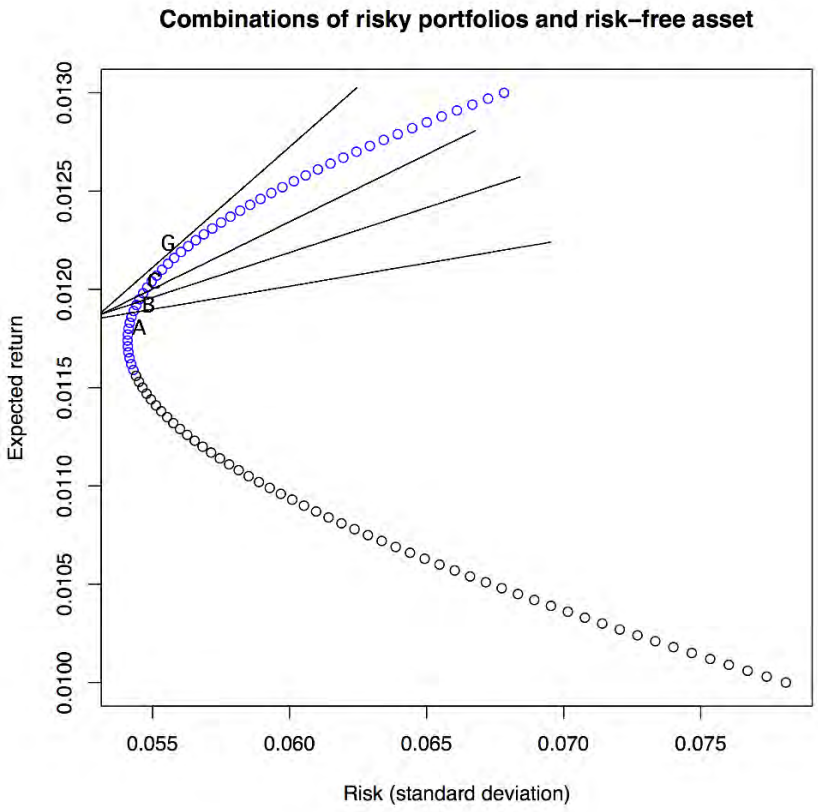
where R_f is the return of the risk-free asset. The variance of this combination is simply

$$\sigma_p^2 = x^2\sigma_A^2 \quad \Rightarrow \quad x = \frac{\sigma_p}{\sigma_A}$$

From the last two equations we get

$$\overline{R}_p = R_f + \left(\frac{\overline{R}_A - R_f}{\sigma_A}\right)\sigma_p$$

This equation describes a straight line. On this line we find all the possible combinations of portfolio A and the risk-free rate. Another investor can choose to combine the risk-free rate with portfolio B or portfolio C. Clearly, for the same level risk the combinations that lie on the $R_f - B$ line have higher expected returns than those on the line $R_f - A$ (see figure below). And $R_f - C$ will produce combinations that have higher return than those on $R_f - B$ for the same level of risk, etc.



The solution, therefore, is to find the point of tangency of this line to the efficient frontier. Let's call this point *G*. To find this point we want to maximize the slope of the line in (1) as follows:

$$\max\{\theta\} = \frac{\overline{R}_p - R_f}{\sigma_p}.$$

Subject to the restriction

$$\sum_{i=1}^{n} x_i = 1.$$

Since,

$$R_f = R_f \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} R_f x_i .$$

We can write the maximization problem as

$$\max\{\theta\} = \frac{\sum_{i=1}^{n} x_i \left(\overline{R}_i - R_f\right)}{\left(\sum_{i=1}^{n} x_i^2 \sigma_i^2 + \sum_{i,j:\ i\neq j}^{n,n} x_i x_j \sigma_{ij}\right)^{1/2}}.$$

Take now the partial derivative with respect to each x_i, $1 \le i \le n$, set them equal to zero and solve. Let's find the partial derivative with respect to x_i:

$$0 = \frac{\partial\theta}{\partial x_k} = \left(\overline{R}_k - R_f\right)\left(\sum_{i=1}^{n} x_i^2 \sigma_i^2 + \sum_{i,j:\ i\neq j}^{n,n} x_i x_j \sigma_{ij}\right)^{-\frac{1}{2}} -$$

$$\frac{1}{2}\left(\sum_{i=1}^{n} x_i \left(\overline{R}_I - R_f\right)\right)\left(2x_k\sigma_k^2 + 2\sum_{j=1, j\neq k}^{n} x_J \sigma_{kj}\right)\left(\sum_{i=1}^{n} x_i^2 \sigma_i^2 + \sum_{i,j:\ i\neq j}^{n,n} x_i x_j \sigma_{ij}\right)^{-\frac{1}{2}}.$$

Multiply both sides by

$$\left(\sum_{i=1}^{n} x_i^2 \sigma_i^2 + \sum_{i,j:\ i\neq j}^{n,n} x_i x_j \sigma_{ij}\right)^{\frac{1}{2}}$$

To get

$$\left(\overline{R}_k - R_f\right) - \frac{\sum_{i=1}^{n} x_i \left(\overline{R}_I - R_f\right)}{\sum_{i=1}^{n} x_i^2 \sigma_i^2 + \sum_{i,j:\ i\neq j}^{n,n} x_i x_j \sigma_{ij}}\left(x_k\sigma_k^2 + \sum_{j=1, j\neq k}^{n} x_J \sigma_{kj}\right) = 0.$$

Now, if we let

$$\lambda = \frac{\sum_{i=1}^{n} x_i \left(\overline{R}_I - R_f\right)}{\sum_{i=1}^{n} x_i^2 \sigma_i^2 + \sum_{i,j:\ i\neq j}^{n,n} x_i x_j \sigma_{ij}},$$

the previous expression will be

$$\left(\overline{R}_k - R_f\right) - \lambda x_k \sigma_k^2 - \sum_{j=1, j\neq k}^{n} \lambda x_J \sigma_{kj} = 0.$$

Therefore,

$$\overline{R}_k - R_f = \lambda x_k \sigma_k^2 + \sum_{j=1, j\neq k}^{n} \lambda x_J \sigma_{kj}.$$

Let's define now a new variable,

$$z_k = \lambda x_k.$$

Then,

$$\overline{R}_k - R_f = z_k \sigma_k^2 + \sum_{j=1, j\neq k}^{n} z_J \sigma_{k\,j}.$$

We have one such equation for each $1 \leq k \leq n$. Explicitly, these equations are:

$$\overline{R}_1 - R_f = z_1 \sigma_1^2 + \sum_{j=1, j\neq 1}^{n} z_J \sigma_{1\,j}$$

$$\overline{R}_2 - R_f = z_2 \sigma_2^2 + \sum_{j=1, j\neq 2}^{n} z_J \sigma_{2\,j}$$

...

$$\overline{R}_n - R_f = z_n \sigma_n^2 + \sum_{j=1, j\neq n}^{n} z_J \sigma_{n\,j}$$

The solution ($\overline{R}$) involves solving the system of these simultaneous equations, which can be written in matrix form as:

$$\overline{R} = \sum Z,$$

where Σ is the variance-covariance matrix of the returns of the *n* stocks. To solve for *Z*:

$$Z = \sum^{-1} \overline{R}.$$

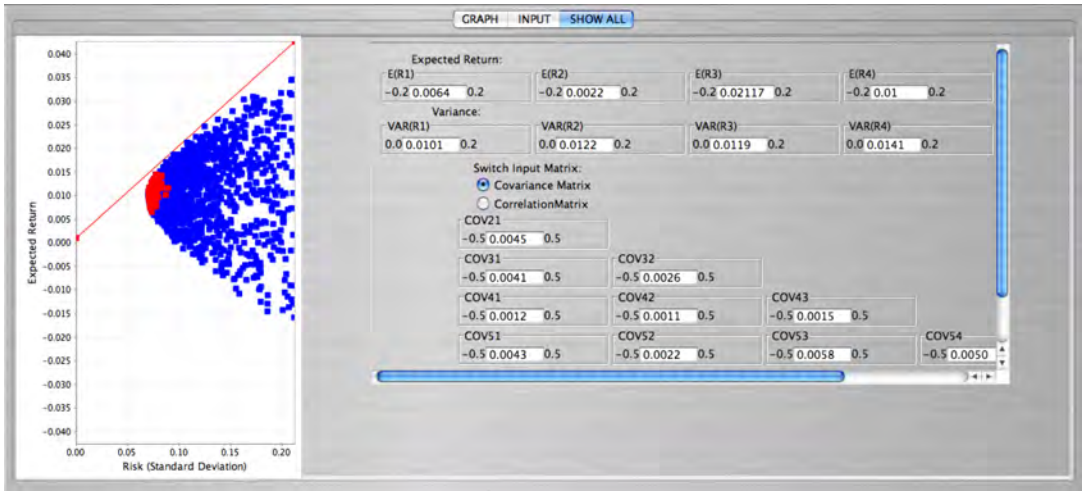
Once we find the z_i's it is easy to find the x_i's (the fraction of funds to be invested in each security). Earlier we defined

$$z_k = \lambda x_k \;\rightarrow\; x_k = \frac{z_k}{\lambda}.$$

As $\sum_{i=1}^{n} x_i = 1$, we can find λ as follows:

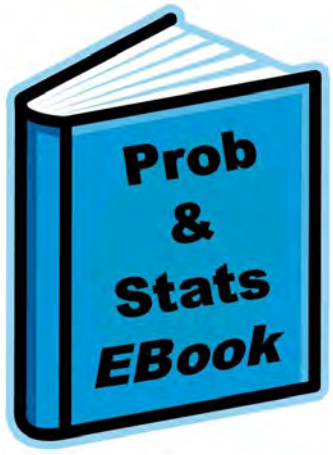
$$\sum_{i=1}^{n} z_i = \lambda \sum_{i=1}^{n} x_i = \lambda.$$

The snapshot from the SOCR portfolio applet shows an example with 5 stocks. Again, the red points in the applet correspond to the case when short sales are not allowed. The point of tangency can be found with a choice of the risk-free rate that can be entered in the input dialog box.



Day 3: Wed 08/12/09

Morning Session: SOCR Activities



EBook and Exploratory Data Analyses (EDA)

- EBook (http://wiki.stat.ucla.edu/socr/index.php/EBook)
  - 1 Preface
    - 1.1 Format
    - 1.2 Learning and Instructional Usage
  - 2 Chapter I: Introduction to Statistics
    - 2.1 The Nature of Data and Variation
    - 2.2 Uses and Abuses of Statistics
    - 2.3 Design of Experiments
    - 2.4 Statistics with Tools (Calculators and Computers)
  - 3 Chapter II: Describing, Exploring, and Comparing Data
    - 3.1 Types of Data
    - 3.2 Summarizing Data with Frequency Tables
    - 3.3 Pictures of Data
    - 3.4 Measures of Central Tendency
    - 3.5 Measures of Variation
    - 3.6 Measures of Shape
    - 3.7 Statistics
    - 3.8 Graphs and Exploratory Data Analysis
  - 4 Chapter III: Probability
    - 4.1 Fundamentals
    - 4.2 Rules for Computing Probabilities
    - 4.3 Probabilities Through Simulations
    - 4.4 Counting
  - 5 Chapter IV: Probability Distributions
    - 5.1 Random Variables
    - 5.2 Expectation (Mean) and Variance
    - 5.3 Bernoulli and Binomial Experiments
    - 5.4 Multinomial Experiments
    - 5.5 Geometric, Hypergeometric and Negative Binomial
    - 5.6 Poisson Distribution
  - 6 Chapter V: Normal Probability Distribution
    - 6.1 The Standard Normal Distribution
    - 6.2 Nonstandard Normal Distribution: Finding Probabilities
    - 6.3 Nonstandard Normal Distribution: Finding Scores (Critical Values)
  - 7 Chapter VI: Relations Between Distributions
    - 7.1 The Central Limit Theorem
    - 7.2 Law of Large Numbers
    - 7.3 Normal Distribution as Approximation to Binomial Distribution
    - 7.4 Poisson Approximation to Binomial Distribution
    - 7.5 Binomial Approximation to Hypergeometric
    - 7.6 Normal Approximation to Poisson

- *SOCR Charts*
  We had previously reviewed the functionality and some of the applications of the SOCR Charts applets (www.socr.ucla.edu/htmls/chart).

- *Chart Activities*
  Many SOCR Charts are paired with hands on activities that demonstrate their utilization as EDA tools (http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_ChartsActivities).

  For example, the Power-Transformation activity (http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_Activities_PowerTransformFamily_Graphs) illustrates the effects of the family of power-transformations on raw data, histogram plots, QQ-Normal Probability plots and XY-scatter plots.

  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 a continuous piece-wise function, for all $y > 0$ by

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

  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. In this situation, one observes paired (X,Y) values which are typically plotted Y (abscissa) vs. X (ordinate) 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, go to SOCR Charts (www.socr.ucla.edu/htmls/SOCR_Charts.html) and select the **PowerTransformXYStatterChart** (Line-Charts → PowerTransformXYStatterChart). You may use the default data provided for this chart, enter your own data (remember to **MAP** the data before you **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 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

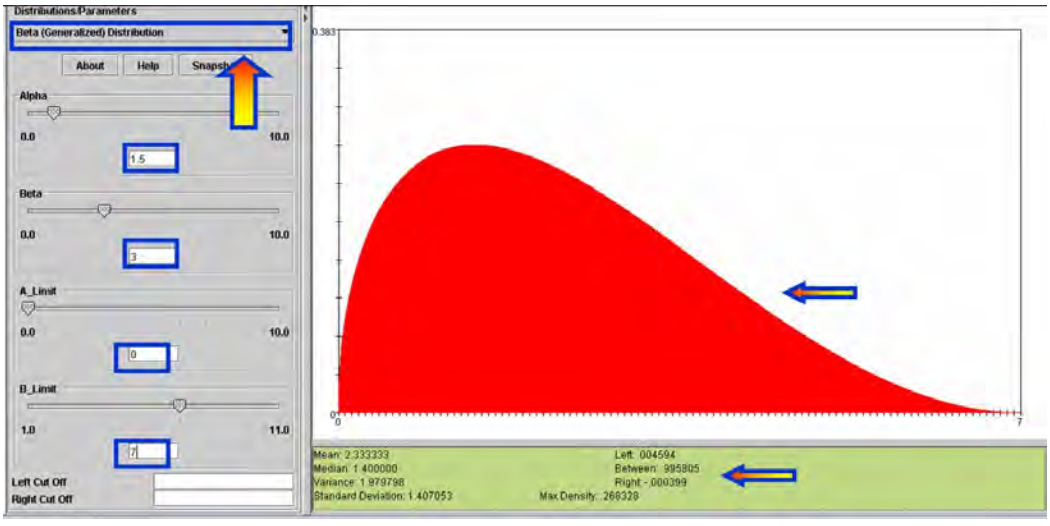
- The second exercise demonstrates the effects of the power-transform applied to data in a QQ-Normal plot setting. We are interested in studying the effects of power transforming the native (original) data on the quantiles, relative the Normal quantiles (i.e., QQ-Normal plot effects). How and why do you expect the QQ-Normal plot to change as we vary the power parameter?

- Again go to SOCR Charts (www.socr.ucla.edu/htmls/SOCR_Charts.html) 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 you **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 in 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 will make the (transformed) data quantiles similar to those of the Normal distribution? Why?

Exercise 3: Power Transformation Family in a Histogram Plot Setting

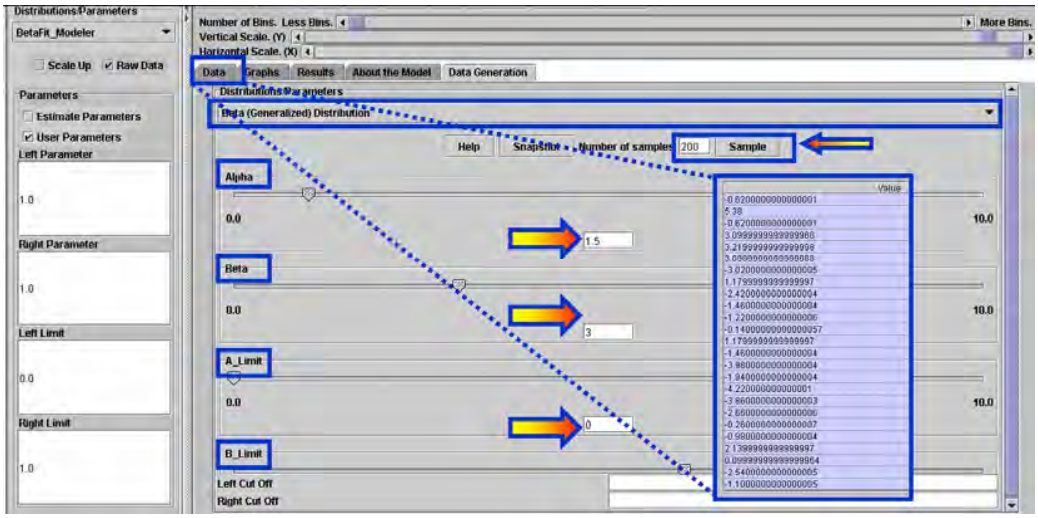
- This exercise demonstrates the effects on the histogram distribution after applying the power-transform to the (observed or simulated) data. In this experiment, we want to see whether we can reduce the variance of a dataset and make its histogram more symmetric, unimodal and bell-shaped.

- 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.

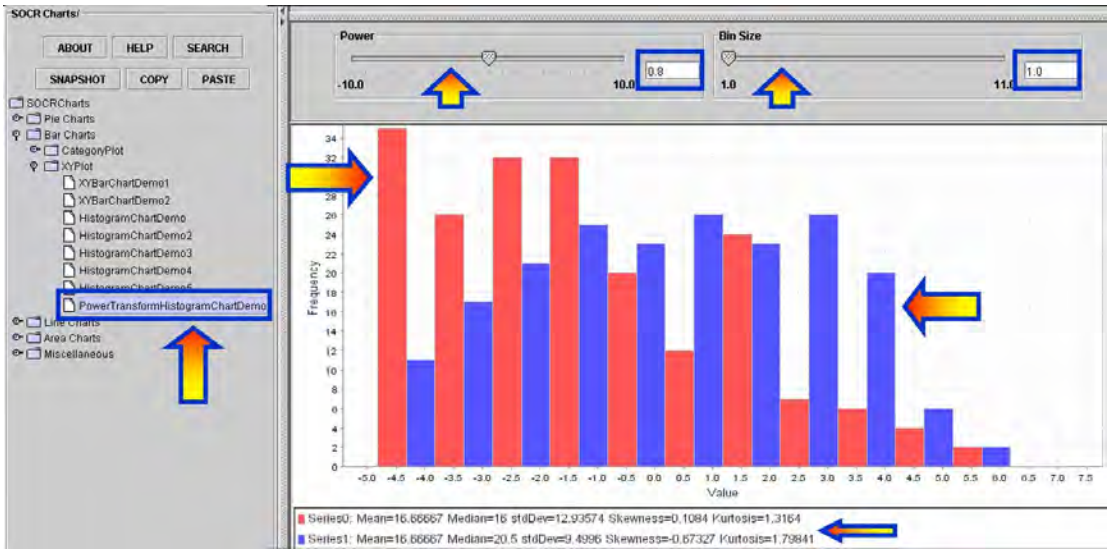


- Go to the SOCR Modeler and select 200 observations from the Generalized Beta Distribution ($\alpha = 1.5; \beta = 3; A = 0; B = 7$), as shown on the image below. Copy these 200 values in your mouse buffer (CNT-C) and paste them in the **Data** tab of the **PowerTransformHistogramChart**. Then *map* 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.

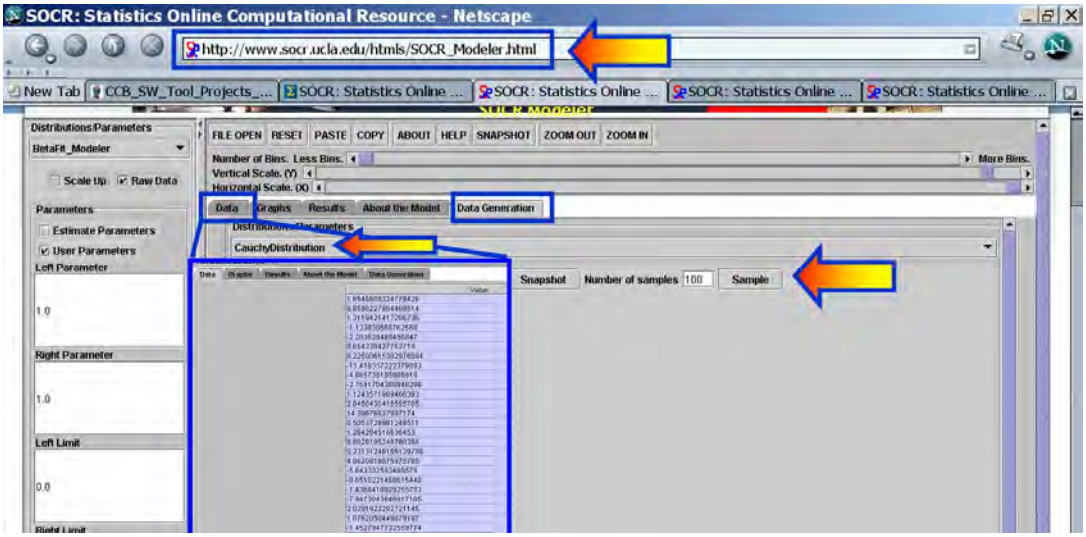


- In the **Graph** tab of the **PowerTransformHistogramChart**, change the power-transform parameter (using the slider on the top). All SOCR Histogram charts allow you to choose the width of the histogram bins, using the second slider on the top. Observe the graphical behavior of the **histogram** of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

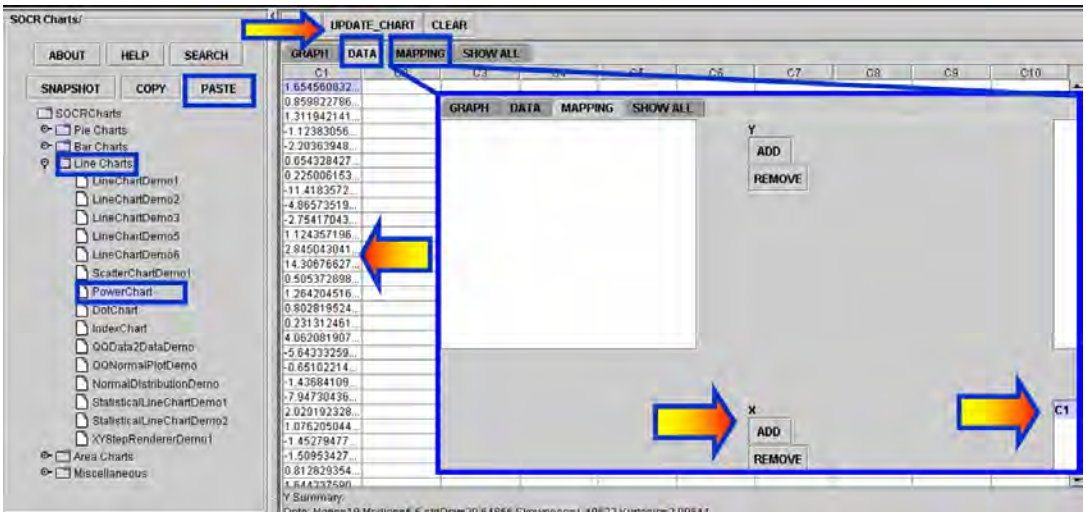


Exercise 4: Power Transformation Family in a Time/Index Plot Setting

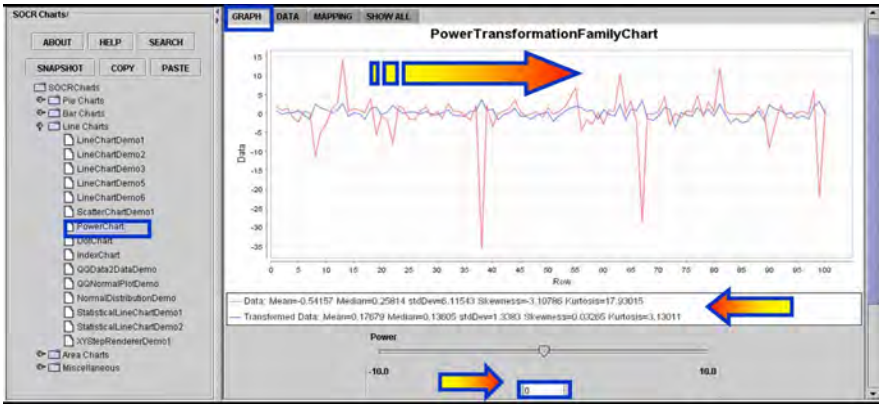
- Let's first get some data: Go to SOCR Modeler and generate 100 Cauchy Distributed variables. Copy these data in your mouse buffer (CNT-C). Of course, you may use your own data throughout. We choose Cauchy data to demonstrate how the Power Transform Family allows us to normalize data that is far from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



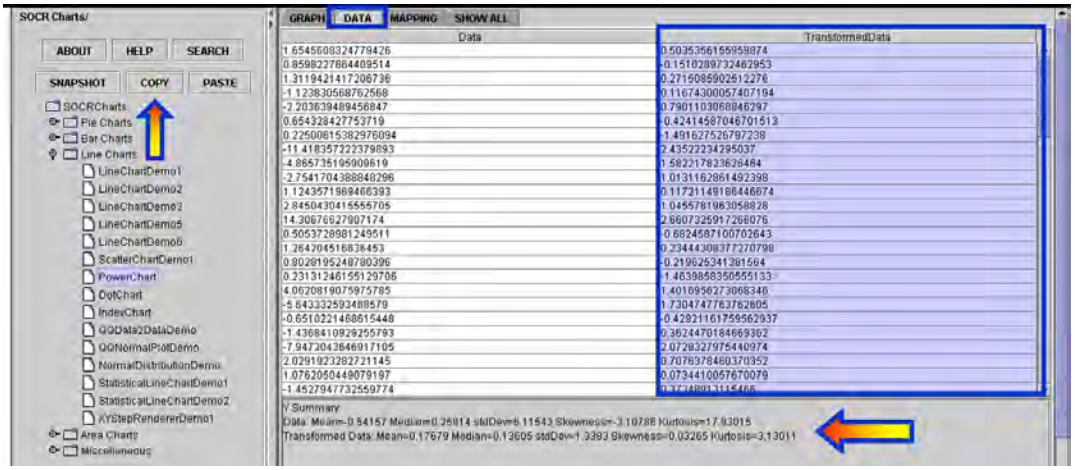
- Next, paste (CNT-V) these 100 observations in SOCR Charts (Line-Charts → Power Transform Chart). Click **Update Chart** to see the index plot of this data in RED!



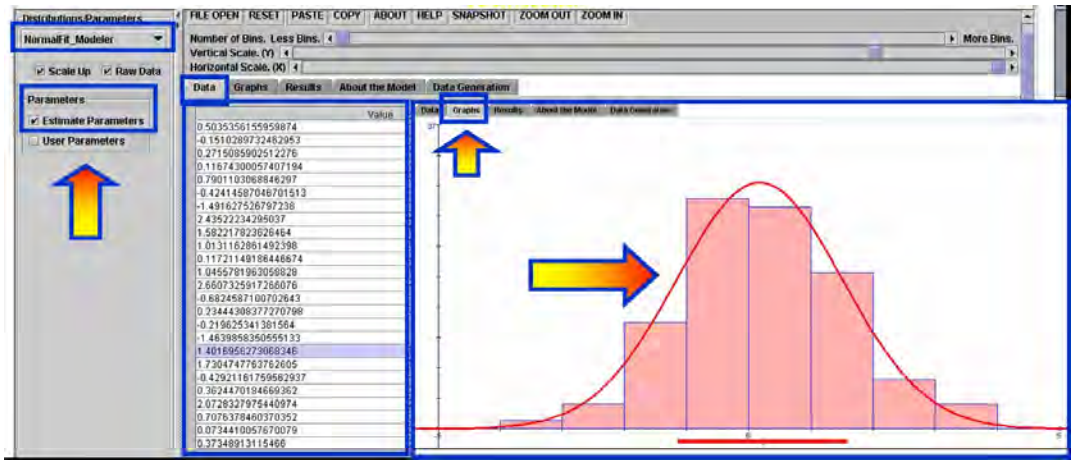
- Now go to the **Graph Tab-Pane** and choose $\lambda = 0$ (the power parameter). Why is $\lambda = 0$ the best choice for this data? Try experimenting with different values of λ. Observe the variability in the Graph of the transformed data in Blue (relative to the variability of the native data in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



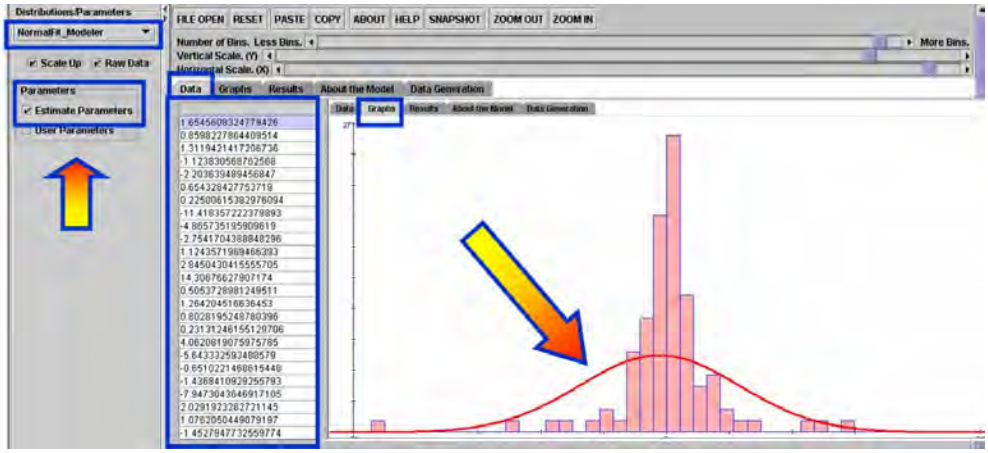
- Then go back to the **Data Tab-Pane** and copy in your mouse buffer the transformed data. We will compare how well the Normal distribution fits the histograms of the raw data (Cauchy distribution) and the transformed data. One can experiment with other values of λ, as well! In the case of $\lambda = 0$, 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 the 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*, though these will, in general, vary.



- 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 a 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 good 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.

SOCR MotionChart Activities

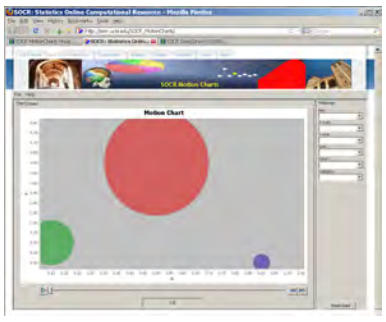
The amount, complexity and provenance of data has dramatically increased in the last few years. Visualization of observed and simulated data is a critical component of any social, environmental, biomedical or scientific quest. Dynamic, exploratory and interactive visualization of multivariate data, without preprocessing by dimensionality reduction, remains an insurmountable challenge. SOCR MotionCharts (www.socr.ucla.edu/SOCR_MotionCharts) provide a new paradigm for discovery-based exploratory analysis of multivariate data. This interactive data visualization tool enables the visualization of high-dimensional longitudinal data. SOCR Motion Charts allows mapping of ordinal, nominal and quantitative variables onto time, axes, size, colors, glyphs and appearance characteristics, which facilitates the interactive display of multidimensional data. SOCR Motion Charts can be used as instructional tool for rendering and interrogating high-dimensional data in the classroom, as well as a research tool for exploratory data analysis. One activity we chose to present here is based on the SOCR Housing Price Index dataset (http://wiki.stat.ucla.edu/socr/index.php/SOCR_MotionCharts_HousingPriceIndex).

The aims of this activity are to:

- demonstrate the SOCR MotionCharts data import, manipulations and graphical interpretation.
- explore interactively the graphical visualization of real-life multidimensional datasets.
- illustrate complex data navigation from different directions (using data mappings).

Open 3 browser tabs pointing to the activity, the SOCR MotionCharts applet and the SOCR Housing Price Index dataset. The images below show the arrangement of these 3 browser tabs.

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

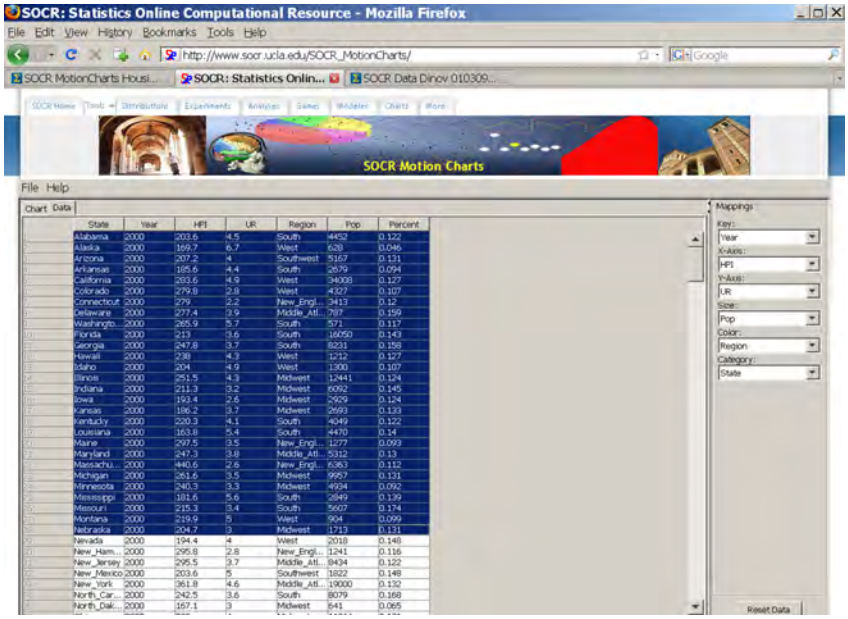


www.socr.ucla.edu/SOCR_MotionCharts

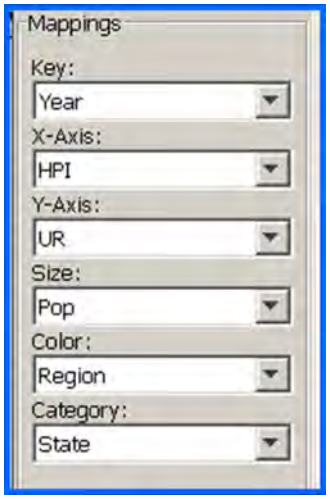
http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_Dinov_010309_HousingPriceIndex

The house price index data was provided by the Office of Federal Housing Enterprise Oversight. The data represent the average housing price for all states between the years 2000 and 2006. The data also include the average unemployment rate, population (in thousands), the percent subprime loans, and the region by state.

- Using the mouse, copy the data from the SOCR data web page, click on the first cell (top-left) in the DATA tab of the SOCR Motion Charts applet, and paste the data in the spreadsheet.

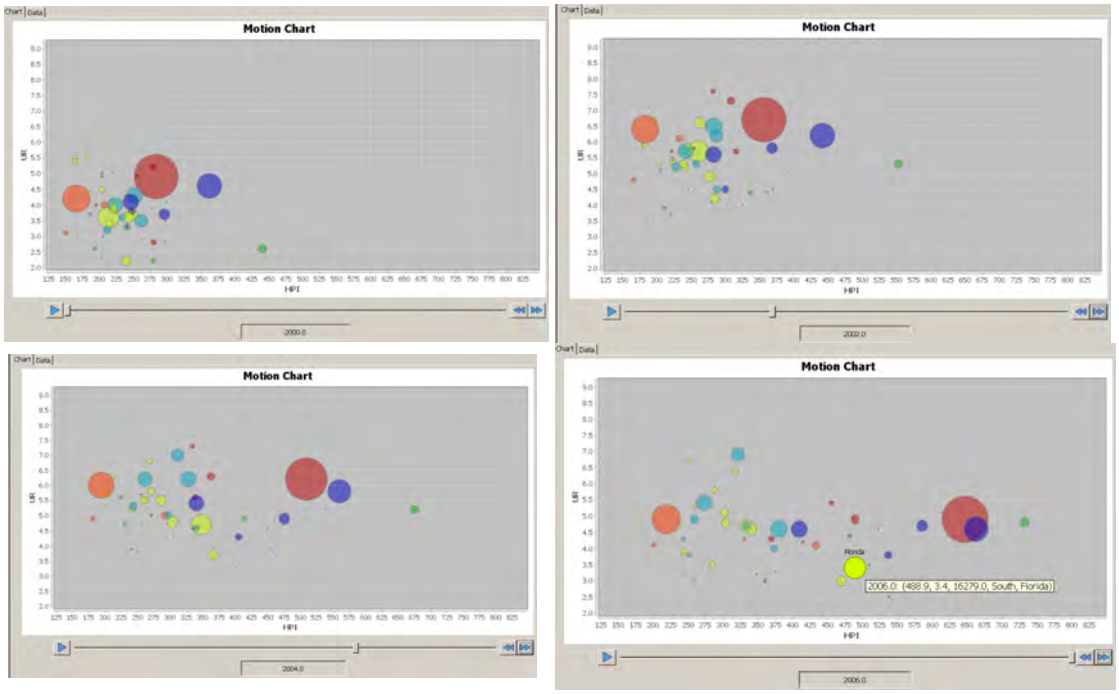


- Next, you need to map the column-variables to different properties in the SOCR MotionChart. For example, you can use the following mapping:



	Variables					
SOCR MotionChart Property	Key	X-Axis	Y-Axis	Size	Color	Category
Data Column Name	Year	HPI	UR	Pop	Region	State

The figures below represent snapshots of the generated dynamic SOCR motion chart. In the real applet, you can *play* (animate) or *scroll* (1-year steps) through the years (2000, ..., 2006). Notice the position change between different snapshots of the time slider on the bottom of these figures. Also, a mouse-over a blob will trigger a dynamic graphical pop-up providing additional information about the data for the specified blob in the chart.



You can also change what variables (data columns) are mapped to the following SOCR MotionCharts properties:

- *Key*, *X-Axis*, *Y-Axis*, *Size*, *Color* and *Category*. Changing the variable-property mapping allows you to explore the data from a different perspective.

Data Type and Format

SOCR MotionCharts currently accept three types of data: numbers, dates/time, and strings. With these data types, the SOCR MotionCharts applet is able to handle the majority of data types. Internally, the applet uses the *natural (lexicological) ordering* of these data as defined by Java primitive data types. While many types of data can be interpreted as strings, in some cases it may not be appropriate to use string-data lexicological ordering. When designing SOCR MotionCharts, we took this into consideration and designed the applet so that it can easily be extended to provide a greater variety of interpreted types. Thus, a developer may easily extend the applet to provide another data type interpretation for specific types of data.

Applications

The SOCR MotionCharts can be used in a variety of applications to visualize dynamic relationships in multidimensional data in up to 5 dimensions, plus a 6^{th} temporal component. The applet's design and implementation support plug-ins to increase dimensionality. The overall purpose of SOCR MotionCharts is to provide users with a way to visualize the relationships between multiple variables over a period of time in a simple, intuitive and animated fashion.

Law of Large Numbers (LLN) Activity

This is a heterogeneous activity that demonstrates the theory and applications of the Law of Large Numbers (LLN). The SOCR LLN applet is available here: www.socr.ucla.edu/htmls/exp/LLN_Simple_Experiment.html.

The goals of this activity are to:

- illustrate the theoretical meaning and practical implications of the LLN.
- present the LLN in a variety of situations.
- provide empirical evidence in support of the LLN-convergence and dispel common LLN misconceptions.

Example

The average weight of 10 students from a class of 100 students is most likely closer to the *real average* weight of all 100 students, compared to the average weight of 3 randomly chosen students from that same class. This is because the sample of 10 is a *larger number* than the sample of only 3 and better represents the entire class. At the extreme, a sample of 99 of the 100 students will produce a sample average almost exactly the same as the average for all 100 students. On the other extreme, sampling a single student will be an extremely variant estimate of the overall class average weight.

Statement of the Law of Large Numbers

If an event with probability p is observed repeatedly during **independent repetitions**, the ratio of the observed frequency of that event to the total number of repetitions converges towards p as the number of repetitions becomes arbitrarily large.

The complete details about the *weak* and *strong* laws of large numbers may be found here: http://en.wikipedia.org/wiki/Law_of_large_numbers.

Exercise 1

This exercise illustrates the statement and validity of the LLN in the situation of repeatedly tossing a fair or a biased coin. Suppose we let H and T denote Heads and Tails, the probabilities of observing a Head or a Tail at each trial are $0 < p < 1$ and $0 < 1 - p < 1$, respectively. The sample space of this experiment consists of sequences of H's and Ts. For example, an outcome may be $\{H,H,T,H,H,T,T,T,....\}$. If we toss a coin n times, the size of the sample-space is 2^n, as the coin tosses are independent. The Binomial Distribution governs the probability of observing k Heads in n experiments ($0 \le k \le n$), which is evaluated by the Binomial density at k.

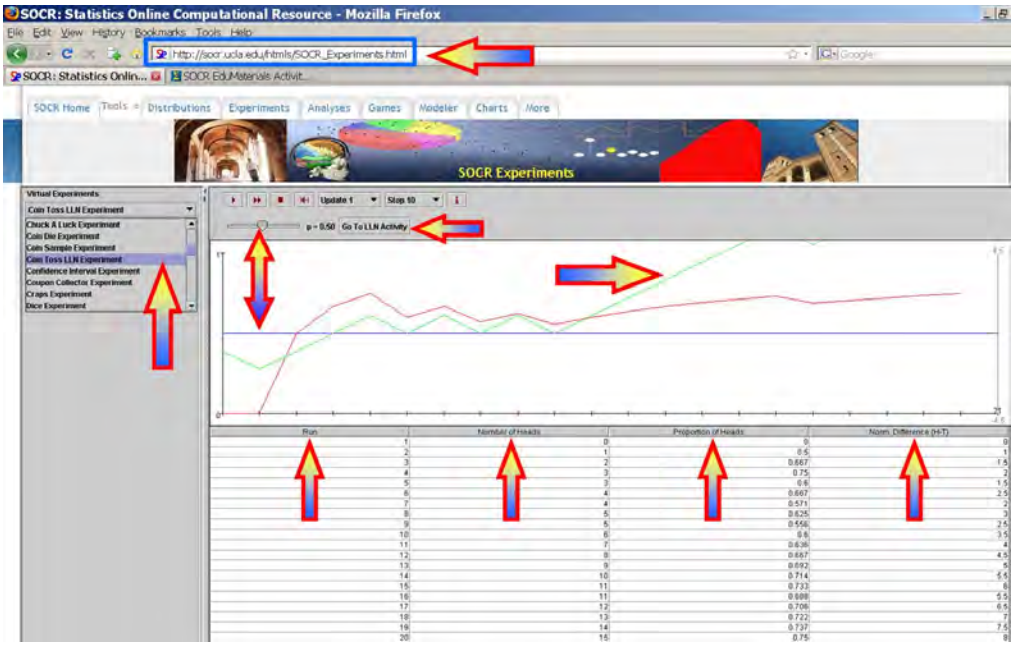
In this case we will be interested in two random variables associated with this process. The first variable will be the *proportion of Heads* and the second will be the *differences of the number of Heads and Tails*. This will empirically demonstrate the LLN and its most common misconceptions (presented below). Point your browser to the SOCR Experiments and select the **Coin Toss LLN Experiment** from the drop-down list of

experiments in the top-left panel. This applet consists of a control toolbar on the top followed by a graph panel in the middle and a results table at the bottom. Use the toolbar to flip coins one at a time, 10, 100, 1,000 at a time or continuously! The toolbar also allows you to stop or reset an experiment and select the probability of Heads (**p**) using the slider. The graph panel in the middle will dynamically plot the values of the two variables of interest (*proportion of heads* and *difference of Heads and Tails*). The outcome table at the bottom presents the summaries of all trials of this experiment. From this table, you can copy and paste the summary for further processing using other computational resources (e.g., SOCR Modeler or MS Excel).

- **Note:** We report the normalized differences of the number of Heads minus the number of Tails in the graph and result table. Let H and T be the number of Heads and Tails, up to the current trial (k), respectively. Then we define the normalized difference

$$|H - T| = p + \frac{(1-p)H - pT}{\frac{2}{3} Max_k},$$

where $Max_k = \max_{1 \le i \le k} |H - T|_i$ and $|H - T|_i$ is the maximum difference of Heads and Tails up to the i^{th} trial. Observe that the expectation of the normalized difference $E(|H - T|) = p$, since $E((1-p)H - pT) = 0$. This ensures that the normalized differences oscillate around the chosen p (the LLN limit of the proportion of Heads) and they are visible within the graph window.



Now, select **n=100** and **p=0.5**. The figure above shows a snapshot of the applet. Remember that each time you run the applet the random samples will be different and the figures and results will generally vary. Click on the **Run** or **Step** buttons to perform the experiment and observe how the *proportion of heads* and *differences* evolve over time.

Choosing **Continuous** from the number of experiments drop-down list in the tool bar will run the experiment in a continuous mode (use the **Stop** button to terminate the experiment in this case). The statement of the LLN in this experiment is simply that **as the number of experiments increases the sample proportion of Heads (red curve) will approach the theoretical (user preset) value of p (in this case *p=0.5*)**. Try to change the value of **p** and run the experiment interactively several times. Notice the behavior of the graphs of the two variables we study. Try to pose and answer questions like these:

- If we set **p=0.4**, how large of a sample-size is needed to ensure that the sample-proportion stays within [0.4; 0.6]?
- What is the behavior of the curve representing the differences of Heads and Tails (red curve)?
- Is the convergence of the sample-proportion to the theoretical proportion (that we preset) dependent on p?
- Remember that the more experiments you run the closer the theoretical and sample proportions will be (by LLN). Go in **Continuous run mode** and watch the convergence of the sample proportion to *p*. Can you explain in words, why we can't expect the second variable of interest (the differences of Heads and Tails) to converge?

Exercise 2

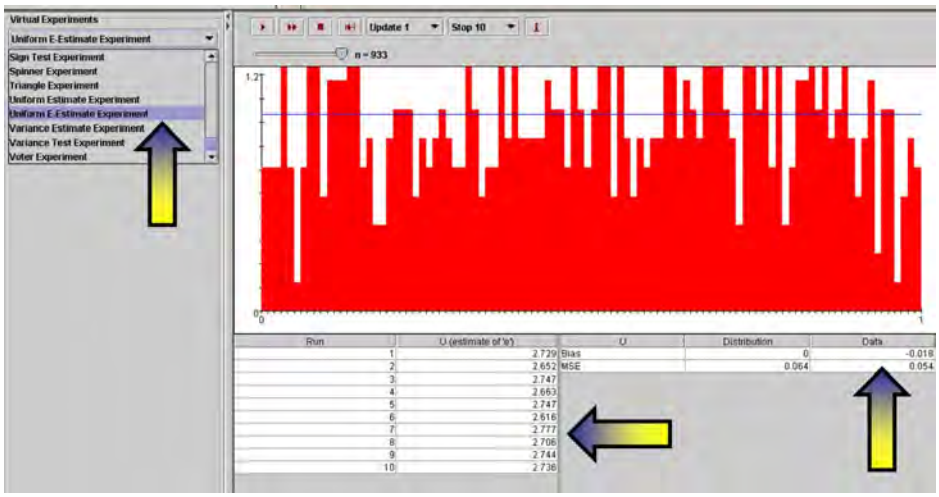
The second SOCR demonstration of the law of large numbers will be quite different and practically useful. Here we show how the LLN implies practical algorithms for estimation of transcendental numbers. The two most popular transcendental numbers are π and e.

Estimating *e* using SOCR simulation

The SOCR E-Estimate Experiment provides the complete details of this simulation. In a nutshell, we can estimate the value of the natural number e using random sampling from a Uniform distribution. Suppose $X_1, X_2, ..., X_n$ are drawn from a uniform distribution on (0, 1) and define $U = \arg\min_n \left(X_1 + X_2 + X_3 + ... + X_n > 1 \right)$, note that all $X_i \geq 0$.

Now, the expected value $E(U) = 3 \cong 2.7182$. Therefore, by LLN, taking averages of $\{U_1, U_2, U_3, ..., U_k\}$ values, each computed from random samples $X_1, X_2, X_3, ..., X_n, \sim U(0,1)$ as described above, will provide a more accurate estimate (as $k \longrightarrow \infty$) of the natural number *e*.

The **Uniform E-Estimate Experiment**, part of SOCR Experiments, provides a hands-on demonstration of how the LLN facilitates stochastic simulation-based estimation of *e*.



Estimating π using SOCR Simulation

Similarly, one may approximate the transcendental number π, using the SOCR Buffon's Needle Experiment. Here, the LLN again provides the foundation for a better approximation of π by virtually dropping needles (many times) on a tiled surface and observing if the needle crosses a tile grid-line. For a tile grid of size 1, the odds of a needle-line intersection are $\frac{2}{\pi} \approx 0.63662$. In practice, to estimate π from a number (N) of needle drops, we take the reciprocal of the sample odds-of-intersection.

Experiment 3

Suppose we roll 10 loaded hexagonal (6-face) dice 8 times and we are interested in the probability of observing the event A={3 ones, 3 twos, 2 threes, and 2 fours}. Assume the dice are loaded to the small outcomes according to the following probabilities of the 6 outcomes (*one* is the most likely and *six* is the least likely outcome).

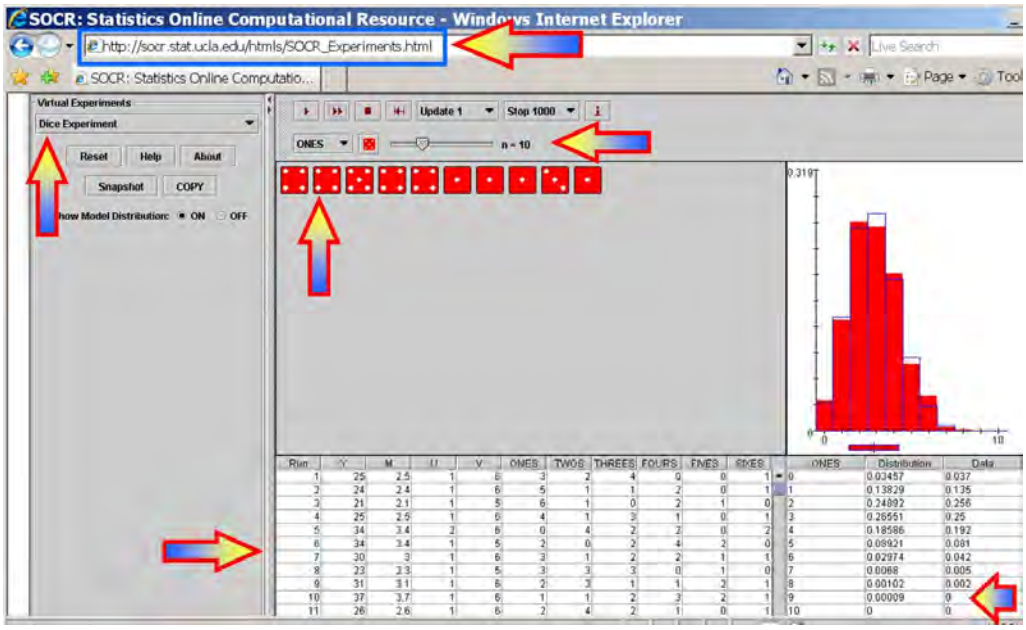
x	1	2	3	4	5	6
P(X=x)	0.286	0.238	0.19	0.143	0.095	0.048

P(A)=? Of course, we can compute this number exactly as:

$$P(A) = \frac{10!}{3!\times 3!\times 2!\times 2!} \times 0.286^3 \times 0.238^3 \times 0.19^2 \times 0.143^2 = 0.00586690138.$$

However, we can also find a pretty close empirically-driven estimate using the SOCR Dice Experiment.

For instance, running the SOCR Dice Experiment 1,000 times with number of dice n=10, and the loading probabilities listed above, we get an output like the one shown below.



Now, we can actually count how many of these 1,000 trials generated the event *A* as an outcome. In one such experiment of 1,000 trials, there were 8 outcomes of the type {3 ones, 3 twos, 2 threes and 2 fours}. Therefore, the relative proportion of these outcomes to 1,000 will give us a fairly accurate estimate of the exact probability we computed above

$$P(A) \approx \frac{8}{1{,}000} = 0.008.$$

Note that this approximation is close to the exact answer above. By the Law of Large Numbers, we know that this SOCR empirical approximation to the exact multinomial probability of interest will significantly improve as we increase the number of trials in this experiment to 10,000.

Hands-on Activities

The following practice problems will help students experiment with the SOCR LLN activity and understand the meaning, ramifications and limitations of the LLN.

- Run the SOCR Coin Toss LLN Experiment twice with stop=100 and *p=0.5*. This corresponds to flipping a fair coin 100 times and observing the behavior of the proportion of heads across (discrete) time.
  - What will be different in the outcomes of the 2 experiments?
  - What properties of the 2 outcomes will be very similar?
  - If we did this 10 times, what is expected to vary and what may be predicted accurately?

- Use the SOCR Uniform *e*-Estimate Experiment to obtain stochastic estimates of the natural number $e \approx 2.7182$.

- Try to explain in words, and support your argument with data/results from this simulation, why the expected value of the variable *U* (defined above) is equal to *e*, $E(U) = e$.
- How does the LLN come into play in this experiment?
- How would you estimate $e^2 \approx 7.38861124$?
- Similarly, try to estimate $\pi \approx 3.141592$ and $\pi^2 \approx 9.8696044$ using the SOCR Buffon's Needle Experiment.

- Run the SOCR Roulette Experiment and bet on 1-18 (out of the 38 possible numbers/outcomes).
  - What is the probability of success (*p*)?
  - What does the LLN imply about p and repeated runs of this experiment?
  - Run this experiment 3 times. What is the sample estimate of p ($\hat{p}$)? What is the difference $p - \hat{p}$? Would this difference change if we ran the experiment 10 or 100 times? How?
  - In 100 Roulette experiments, what can you say about the difference of the number of successes (outcome in 1-18) and the number of failures? How about the proportion of successes?

Additional SOCR LLN Activities

A number of additional examples supplementing this SOCR LLN activity are available online:

- http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_Activities_LawOfLargeNumbers2
- http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_Activities_LawOfLargeNumbersExperiment.

Common Misconceptions regarding the LLN

- **Misconception 1**: If we observe a streak of 10 consecutive heads (when p=0.5, say) the odds of the 11^{th} trial being a Head is > p! This is of course, incorrect, as the coin tosses are independent trials (an example of a *memoryless* process).
- **Misconception 2**: If we run a large number of coin tosses, the **number of heads** and **number of tails** become more and more equal. This is incorrect, as the LLN only guarantees that the sample proportion of heads will converge to the true population proportion (the p parameter that we selected). In fact, the difference |Heads - Tails| diverges!

Group Interactive Discussion on Hands-on Activities
What works, what doesn't, how to extend the collection and how improve teaching of statistical analysis methodologies?

Workshop Evaluation by Participants (see forms below)

Afternoon Session: Visit to the J. Paul Getty Center (www.Getty.edu)

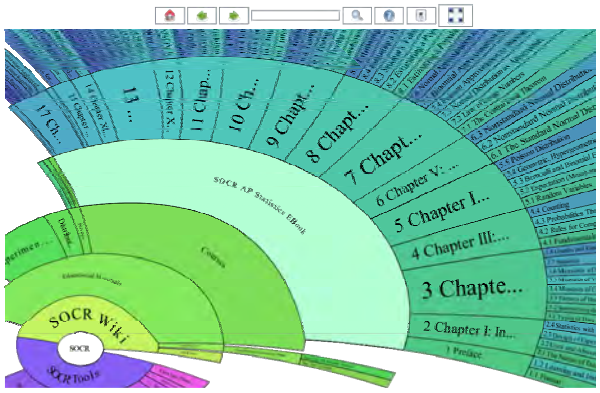
The J. Paul Getty Trust is one of the largest supporters of arts in the world. It is an international cultural and philanthropic institution that focuses on the visual arts in all their dimensions. The Getty serves both the general public and a wide range of professional communities in Los Angeles and throughout the world.

Through the work of the 4 Getty programs—the Museum, Research Institute, Conservation Institute, and Foundation—the Getty aims to further knowledge and nurture critical seeing through the growth and presentation of its collections and by advancing the understanding and preservation of the world's artistic heritage. The Getty pursues this mission with the conviction that cultural awareness, creativity, and aesthetic enjoyment are essential to a vital and civil society. The Getty (www.Getty.edu) is based in Los Angeles, California, and welcomes nearly 1.8 million visitors each year to its two locations, the Getty Center in Los Angeles and the Getty Villa in Malibu.

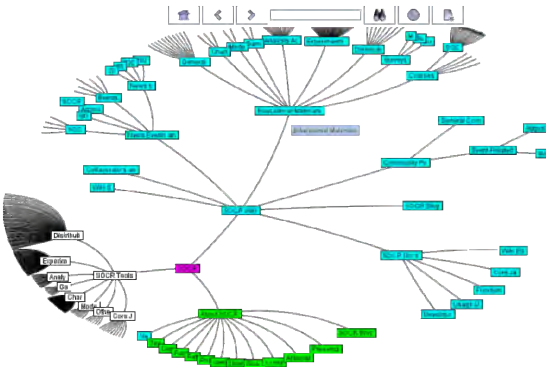
SOCR Resource Navigation

The SOCR resources include a large number of applets, learning materials, datasets, web-services and instructional resources. Interactive discovery, navigation and exploration of these resources are facilitated by the following 4 complementary types of interfaces:

SOCR Hyperbolic Wheel Navigator
www.socr.ucla.edu/SOCR_HW_Viewer.html



SOCR Hyperbolic Tree Viewer
www.socr.ucla.edu/SOCR_HT_ResourceViewer.html



SOCR Carousel Viewer
www.socr.ucla.edu/Carousel

Keyword Searches
http://wiki.stat.ucla.edu/socr/index.php/Socr:Searching

Workshop Evaluation Forms

The next 3 forms contain Workshop evaluation materials that should be returned as per the instructions on each form. Please complete and return these forms to us in due time.

No Need to Return This Information Page!

Workshop Evaluation – Information

Workshop Participants Expectations and Understanding of Workshop Evaluation Expectations

The Statistics Online Computational Resource (www.SOCR.ucla.edu) has received NSF funds to develop educational resources and organize workshops and professional development events for statistics educators across the United States. This requires that an external evaluation be conducted to determine the effectiveness of these events and to identify strengths and limitations. As a workshop participant, you are being asked to participate in the evaluation. All information you provide is strictly confidential. When reporting results, neither your name nor your school will be linked to the data. Only group data will be reported.

BENEFITS AND SERVICES

Participants will have access to all opportunities provided by the workshop, receive a variety of resources for supporting teaching and learning of statistics, and receive financial support as described in recruitment materials.

EXPECTATIONS. Beyond the project expectations for participation and receipt of benefits and services, workshop participants will be expected to participate in the following evaluation activities:

- Complete a pre-workshop survey (at beginning of workshop).
- Complete an end-of-workshop questionnaire (at end of workshop).
- Assist evaluators in administering pre/post statistics content assessments for college students in selected classes (sample of participants only).
- Provide copies of syllabi and other curricula and instructional materials created as a result of participating in the SOCR workshop.
- Provide feedback about classroom usage of SOCR resources in the future.

The evaluation team will coordinate these activities with project staff so as to minimize intrusions and inconveniences. Questions about SOCR evaluations should be addressed to the project PI: Ivo Dinov.

Please Complete This Form on the First Day of the Workshop (08/10/09)!

Workshop Evaluation – Agreement Form

Workshop Participants Expectations and Understanding of Workshop Evaluation Expectations

The Statistics Online Computational Resource (www.SOCR.ucla.edu) has received NSF funds to develop educational resources and organize workshops and professional development events for statistics educators across the United States. This requires that an external evaluation be conducted to determine the effectiveness of these events and to identify strengths and limitations. As a workshop participant, you are being asked to participate in the evaluation. All information you provide is strictly confidential. When reporting results, neither your name nor your school will be linked to the data. Only group data will be reported.

BENEFITS AND SERVICES

Participants will have access to all opportunities provided by the workshop, receive a variety of resources for supporting teaching and learning of statistics, and receive financial support as described in recruitment materials.

EXPECTATIONS. Beyond the project expectations for participation and receipt of benefits and services, workshop participants will be expected to participate in the following evaluation activities:

- Complete a pre-workshop survey (at beginning of workshop).
- Complete an end-of-workshop questionnaire (at end of workshop).
- Assist evaluators in administering pre/post statistics content assessments for college students in selected classes (sample of participants only).
- Provide copies of syllabi and other curricula and instructional materials created as a result of participating in the SOCR workshop.
- Provide feedback about classroom usage of SOCR resources in the future.

The evaluation team will coordinate these activities with project staff so as to minimize intrusions and inconveniences. Questions about SOCR should be directed to the project PIs: Ivo Dinov.

Signature of Workshop Participant: _____ Date: _____

Please print your name: _____

This memo of agreement is not legally binding, but represents a good faith commitment to full participation in the SOCR evaluation by this participant for one year beginning with the workshop.

Please Complete This Form on the First Day of the Workshop (08/10/09)!

Workshop Evaluation – Pre-Program Participant Survey

As part of the required evaluation of this workshop, participants are asked to complete the following *16-question pre-program survey*. The information will be used to help us learn about the effectiveness of the program. You will be asked to complete a similar survey at the end of the next school year. The information is strictly confidential--no one except project evaluators will see individual survey results. Only group data will be reported. The code number is for follow-up purposes and to analyze pre- and post-program data.

When You Have Completed The Survey, Return It To The Workshop Organizers

Thanks for taking time to complete this survey! If you have questions, please ask the facilitator.

PART A: About you.

1. What courses do you teach that include instruction in statistics? _____

_____

2. How many years have you been an instructor? _____

3. What is the subject-area of your highest college degree? _____

4. Are you a member of any national statistics education professional organizations? ___Yes___ No

If yes, which one(s)? _____

5. Why did you choose to participate in this SOCR workshop?

6. What are your expectations for this workshop? What do you want to get out of it?

7. How did you learn about this workshop?

8. Were you familiar with SOCR (the Statistics Online Computational Resource) before enrolling in this workshop?

9. Are you familiar with the www.SOCR.ucla.edu page?

10. How well prepared are you to teach the following statistics topics in the courses you teach? Rate each item on a 4 point scale, with 1 = not adequately prepared and 4 = very well prepared.

	Not Adequately Prepared			Very Well Prepared
a. Data collection (surveys and experiments)	1	2	3	4
b. Summary statistics & graphics (e.g., histograms & boxplots)	1	2	3	4
c. Probability	1	2	3	4
d. Sampling distributions	1	2	3	4
e. Confidence intervals	1	2	3	4
f. Hypothesis testing (one sample for means and proportions)	1	2	3	4
g. Simple linear regression and correlation	1	2	3	4
h. Using open web-resources in probability and statistics	1	2	3	4

11. About how often do you teach probability and statistics classes? List by course title.

_____

_____

_____

12. What is your perception of using open Internet-accessible resources in and out of class for teaching probability and statistics?

13. What are the major issues or concerns for you related to the teaching and learning of statistics at your grade level? Use the back of this page if you need more room.

Please Complete this Form on the Last day of the Workshop (08/12/09)!

End-of-Workshop Evaluation Questionnaire

As part of the required evaluation of SOCR, workshop participants are asked to respond to the following about the 2009 SOCR workshop. Your comments are important in helping us improve our materials, extend our resources and fine-tune our pedagogical approaches. Your responses are anonymous & confidential. They will be compiled and reported only as group data. ***DO NOT WRITE YOUR NAME ON THIS FORM.*** **We appreciate your comments!**

A. ABOUT YOU:

1. What **courses** do you teach (that the material of this workshop may be relevant to)? Please include: title, Upper/Lower/Graduate division, number of students and number of offerings per year.

   _____

   _____

   _____

2. How many years have you been an instructor in these courses?

3. What attracted you to apply and take part in this Workshop?

B. WORKSHOP OUTCOMES. The workshop sessions were designed to address various strategies for teaching introductory statistics. Please rate each one of the associated objectives according to:
1) Your perception of the VALUE (V) of the session objective and
2) Whether you think it was ACCOMPLISHED (A)

Note: "1" represents the lowest score; a "5" represents the highest score. Please make comments.

The workshop sessions helped me to do the following:

		1	2	3	4	5	
1. Learn how to emphasize statistical literacy.	V						Comments:
	A						
2. Learn how to develop statistical thinking.	V						Comments:
	A						
3. Learn to use real data.	V						Comments:
	A						
4. Learn how to focus on conceptual understanding rather than only knowledge of procedures.	V						Comments:
	A						

		1	2	3	4	5	
5. Foster active learning among your students.	V						Comments:
	A						
6. Learn to use technology for developing conceptual understanding.	V						Comments:
	A						
7. Learn to use technology to analyze data.	V						Comments:
	A						

C. WORKSHOP ARRANGEMENTS. Rate the following on a scale of 1-5, with 1 = Disagree, 5 = Agree.

	(Disagree) 1	2	3	4	5 (Agree)	
1. Workshop facilities were satisfactory.						Comments:
2. Workshop facilitators were effective in communicating ideas and issues.						Comments:
3. Workshop facilitators were effective in organizing sessions so that I was actively involved.						Comments:
4. A collaborative and helpful tone was established during the session.						Comments:

5. What 2 or 3 BIG ideas about the teaching and learning of statistics did you learn during this workshop?

_____

_____

_____

6. In what ways could this workshop be improved?

7. In what ways do you plan to use what you have learned in this workshop in your own teaching?

Acknowledgments

The SOCR faculty and the authors of this handbook are deeply indebted to all students, developers, instructors and researchers, including the 2009 SOCR workshop attendees, for their significant contributions, constructive critiques and continued support of SOCR activities, developments of new materials and improvement of existent resources. Specifically we would like to acknowledge the contributions of Robert Gould, Juana Sanchez, Jenny Cui, Jameel Al-Aziz, Annie Chu, Rahul Gidwani, Priscilla Chui, Siu-Ling Teresa Lam, Lei Ricky Jin, Victor Zhu, Beryl Lou, Charles Dang, Stephan Chiu, Jay Zhou, Liao Weien, Agapios Constantinides, Brigid Brett-Esborn, Ariana Anderson, and Jenny Nguyen. David Zes provided invaluable critiques and recommendations for improving the style, syntax and organization of the materials in this book. The dedication and hard work of all these individuals made possible the 2009 SOCR Continuing Statistics Education Workshop and provided the necessary materials to compose and validate this handbook.

The National Science Foundation (NSF) support for the SOCR project (DUE 0716055 & 0442992) provides the fundamental resources necessary to design, build, test, validate and disseminate these integrated materials freely, openly and platform-independently to the entire community in multiple languages and via diverse electronic media formats. Administrative and logistical support from the UCLA Department of Statistics, the College of Letters and Science and the Office of Instructional Development were critical for the organization of this training workshop.

SOCR

UCLA Statistics

UCLA OID

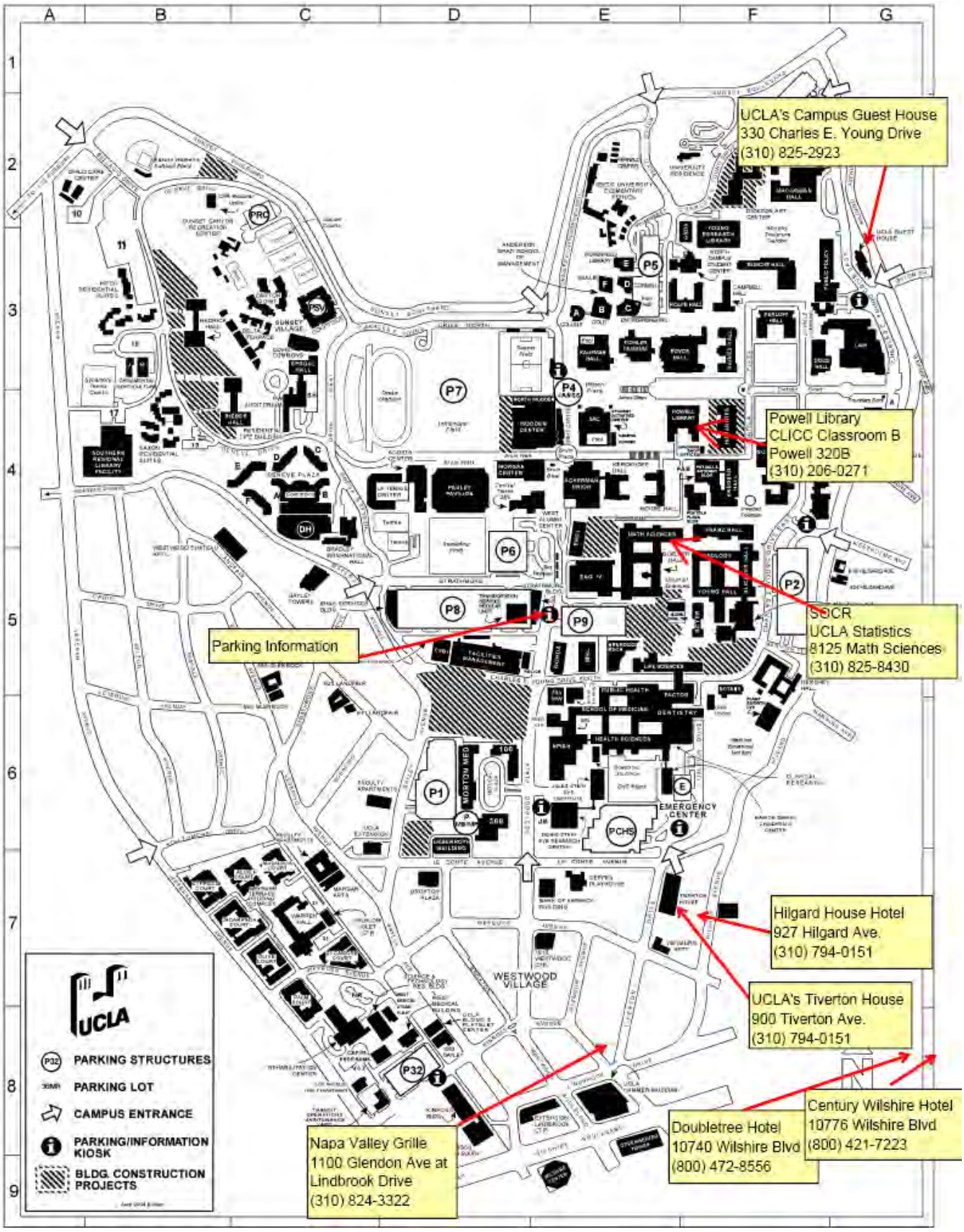
NSF

References

- Agresti A, Coull A. “Approximate is Better than Exact for Interval Estimation of Binomial Proportions” American Statistician (1998).
- Che, Annie, Cui, Jenny, and Dinov, Ivo (2009). SOCR Analyses: Implementation and Demonstration of a New Graphical Statistics Educational Toolkit. JSS, Vol. 30, Issue 3, Apr 2009.
- Dinov, ID, Christou, N. (2009) "Statistics Online Computational Resource for Education". Teaching Statistics, Vol. 31, No. 2, 49-51, 2009.
- Che, A, Cui, J, and Dinov, ID (2009) "SOCR Analyses – an Instructional Java Web-based Statistical Analysis Toolkit", JOLT, 5(1), 1-19, March 2009.
- Dinov, ID, Christou, N, and Gould, R (2009) "Law of Large Numbers: the Theory, Applications and Technology-based Education". JSE, Vol. 17, No. 1, 1-15, 2009.
- Dinov, ID, Christou, N, and Sanchez, J (2008) Central Limit Theorem: New SOCR Applet and Demonstration Activity. Journal of Statistics Education, Volume 16, Number 2, http://www.amstat.org/publications/jse/v16n2/dinov.html.
- Dinov, ID (2008) Integrated Multidisciplinary and Technology-Enhanced Science Education: The Next Frontier, JOLT, 4(1), March 2008, 84-93.
- Dinov, ID, Sanchez, J, and Christou, N (2008) Pedagogical Utilization and Assessment of the Statistic Online Computational Resource in Introductory Probability and Statistics Courses, Journal of Computers & Education, http://doi:10.1016/j.compedu.2006.06.003, 50, 284–300.
- Dinov, ID., Christou, N. and Sanchez, J. Handbook on Continuing Statistics Education with Technology Workshop, August 6, 2007.
- Christou, N., Sanchez, J. and Dinov, ID. Design and Evaluation of SOCR Tools for Simulation in Undergraduate Probability and Statistics Courses. Proceedings of the International Statistics Institute meeting (ISI), Lisbon, Portugal, August 2007.
- Dinov, ID. Grant Review: American Idol or Big Brother? Cell, Vol 127, No 2, 663-664, 17 November 2006.
- Dinov, ID. Statistics Online Computational Resource, Journal of Statistical Software, Vol. 16, No. 1, 1-16, October 2006.
- Dinov, ID. SOCR: Statistics Online Computational Resource: socr.ucla.edu, Statistical Computing & Graphics. Vol. 17, No. 1, 11-15, 2006.
- Dinov, ID and Che, A. Statistics Online Computational Resource for Education, Joint AMS/MAA Meeting, San Antonio, TX, January 12-15, 2006.
- Dinov, ID and Sanchez, J. Assessment of the pedagogical utilization of the statistics online computational resource in introductory probability courses: a quasi-experiment. International Association for Statistical Education, ICOTS7, July 2-7, 2006, Salvador, Brazil (conference proceedings).
- Elton, EJ., Gruber, MJ., Brown, SJ., and Goetzmann, WN. *Modern Portfolio Theory*, Sixth Edition, Wiley, 2003.
- Ferguson, T., S., A Course in Large Sample Theory, Chapman & Hall (1996).
- Freedman, D, Pisani, R, & R. Purves. (2007). Statistics. (Fourth Edition). New York, NY: W.W. Norton & Company.
- Hogg, R. V., Tanis, E. A., Probability and Statistical Inference, 3rd Edition, Macmillan (1988).

- Leslie, M. NetWatch EDUCATION: Statistics Starter Kit, Science Magazine, Volume 302, Number 5651, Issue of 5 December 2003.
- Mega, M., Dinov, I., Thompson, P., Manese, M., Lindshield, C., Moussai, J., Tran, N., Olsen, K., Felix, J., Zoumalan, C., Woods, R., Toga, A., and Mazziotta, J. (2005). Automated brain tissue assessment in the elderly and demented population: Construction and validation of a sub-volume probabilistic brain atlas. NeuroImage, 26(4), 1009-1018.
- Rice, J. Mathematical Statistics and Data Analysis, Third Edition, Duxbury Press (2006).
- Sauro, J., Lewis, J. R., Estimating Completion Rates From Small Samples Using Binomial Confidence Intervals: Comparisons and Recommendations, Proceedings of the Human Factor AND Ergonomics Society 49th Annual Meeting (2005).
- Stewarty, C. (1999). Robust Parameter Estimation in Computer Vision. SIAM Review, 41(3), 513–537.
- Wolfram, S. (2002). A New Kind of Science, Wolfram Media Inc.

Index


UCLA's Campus Guest House
330 Charles E. Young Drive
(310) 825-2923
Powell Library
CLICC Classroom B
Powell 320B
(310) 206-0271
SOCR
UCLA Statistics
8125 Math Sciences
(310) 825-8430
Parking Information
Hilgard House Hotel
927 Hilgard Ave.
(310) 794-0151
UCLA's Tiverton House
900 Tiverton Ave.
(310) 794-0151
Century Wilshire Hotel
10776 Wilshire Blvd
(800) 421-7223
Doubletree Hotel
10740 Wilshire Blvd
(800) 472-8556
Napa Valley Grille
1100 Glendon Ave at
Lindbrook Drive
(310) 824-3322
WESTWOOD VILLAGE
EMERGENCY CENTER
PARKING STRUCTURES
PARKING LOT
CAMPUS ENTRANCE
PARKING/INFORMATION KIOSK
BLDG. CONSTRUCTION PROJECTS

Program

DAY 1 (Mon 8/10/09)

Registration and Coffee (8:00 - 9:00 AM)

Welcome, Ivo Dinov, SOCR Director
(9:00 - 9:40 AM)

Morning Session -
SOCR Motivational Datasets

(9:40 - 10:40 AM)

Morning Break (10:40 - 10:50 AM)

SOCR Open Motivational Datasets
(cont.) (10:50-11:30 AM)

Interactive Discussion
(11:30AM - 12:00 PM)

Lunch Break (12:00 - 1:00 PM)
UCLA Cafeteria

Afternoon Session -
SOCR Applets and Tools
(1PM - 4PM)

Dinner
6:30-8:00 PM UCLA Cafeteria

DAY 2 (Tue 8/11/09)

Morning Session -
SOCR Activities

SOCR Analyses Activities
(9:00 - 10:00 AM)

SOCR Modeler Activities
(10:00 - 10:30 AM)

Break (10:30 - 10:45 AM)

SOCR Distribution Activities
(10:45 - 11:45 AM)

Group Interactive Discussion
(11:45 AM - 12:00 PM)

Lunch Break (12:00 - 1:00 PM)
UCLA Cafeteria

Afternoon Session -
SOCR Activities (cont.)
(1PM - 4PM)

Dinner
6:30-8:00 PM UCLA Cafeteria

DAY 3 (Wed 8/12/09)

Morning Session -
SOCR Activities and EDA

Exploratory Data Analyses
(9:00 - 10:15 AM)

Break (10:15 - 10:30 AM)

Law of Large Numbers LLN
Activity
(10:30-11:30 AM)

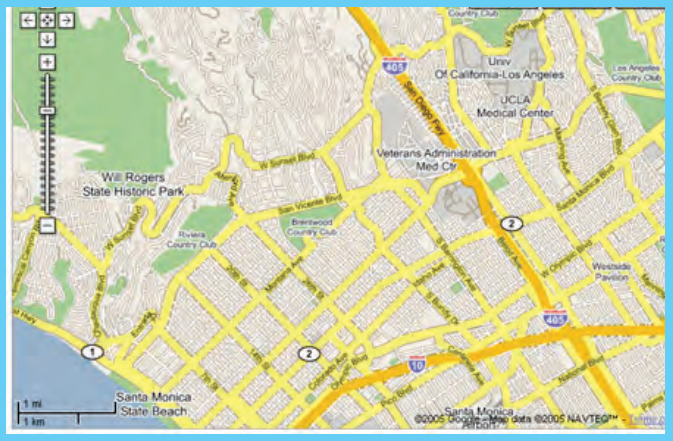
Group Interactive Discussion
(11:30AM - 11:45AM)

Workshop Evaluation
(11:45 AM - 12:00PM)

Lunch Break & Adjourn
(12:00 - 1:00 PM)

Afternoon Session
Visit to J. Paul Getty Museum
Group Tour starts at 2PM
(1PM - 4PM)

www.StatisticsResource.org



$0.00
ISBN-10: 0-615-30464-8
ISBN-13: 978-0-615-30464-9
90000>
9 780615 304649