Computational Statistics Workshop

Vesna Lužar-Stiffler, Ph.D.
CAIR-Centar d.o.o. - „The House of Statistics"
and the University of Zagreb, Dept. Of Mathematics

vesna.luzar-stiffler@cair-center.hr
www.cair-center.hr

4th Int. Summer School on Data Science,
Split, 12/9/2019

Computer simulation has become, alongside experimentation and abstract reasoning, the third major tool of science.

James E. Gentle, Professor of Computational Statistics, George Mason University, Department of Computational and Data Sciences

Introduction: Computational Statistics vs. Statistical Computing

- Statistical computing (according to J.E. Gentle)
  - Computational methods, including numerical analysis for statisticians,
  - Database methodology,
  - Computer graphics,
  - Software engineering,
  - Computer/human interface
- Computational Statistics is grounded in mathematical statistics, statistical computing, and applied statistics and includes:
  - statistical computing
  - visualization,
  - other computationally-intensive methods of statistics (including computational inference and Monte Carlo (MC) methods).

Key Characteristics of Computational Statistics (CS)

- Computation is an instrument of discovery.
  - Computers' role is not just to store data, perform computation, create tables and graphs, but also to suggest new models and theories.
- Computational intensity
  - Need for powerful computer systems/software,
  - Graphs and visualization methods are usually integral features of Computational Statistics.

Typical topics covered in CS courses

- Monte Carlo studies in statistics ✓
- Numerical methods in statistics ("statistical computing")
- Computational inference ✓
- Data partitioning and resampling ✓
- Nonparametric probability density estimation
- Statistical models and data fitting

Motivating Examples

- Predicting my grandson's enrolment in MIOC high school
  - Simulating bivariate data from truncated normal distribution
- Estimating lower limit of the 90% CI for f_2 dissolution statistic
  - Nonparametric boostrap (resampling)
- Evaluating empirical power of the PBE statistic for equivalence testing
  - MC simulation/bootstrapping, visualization, computational inference
- Robustness of 1-sample t-test to departures from assumptions
  - MC experiment

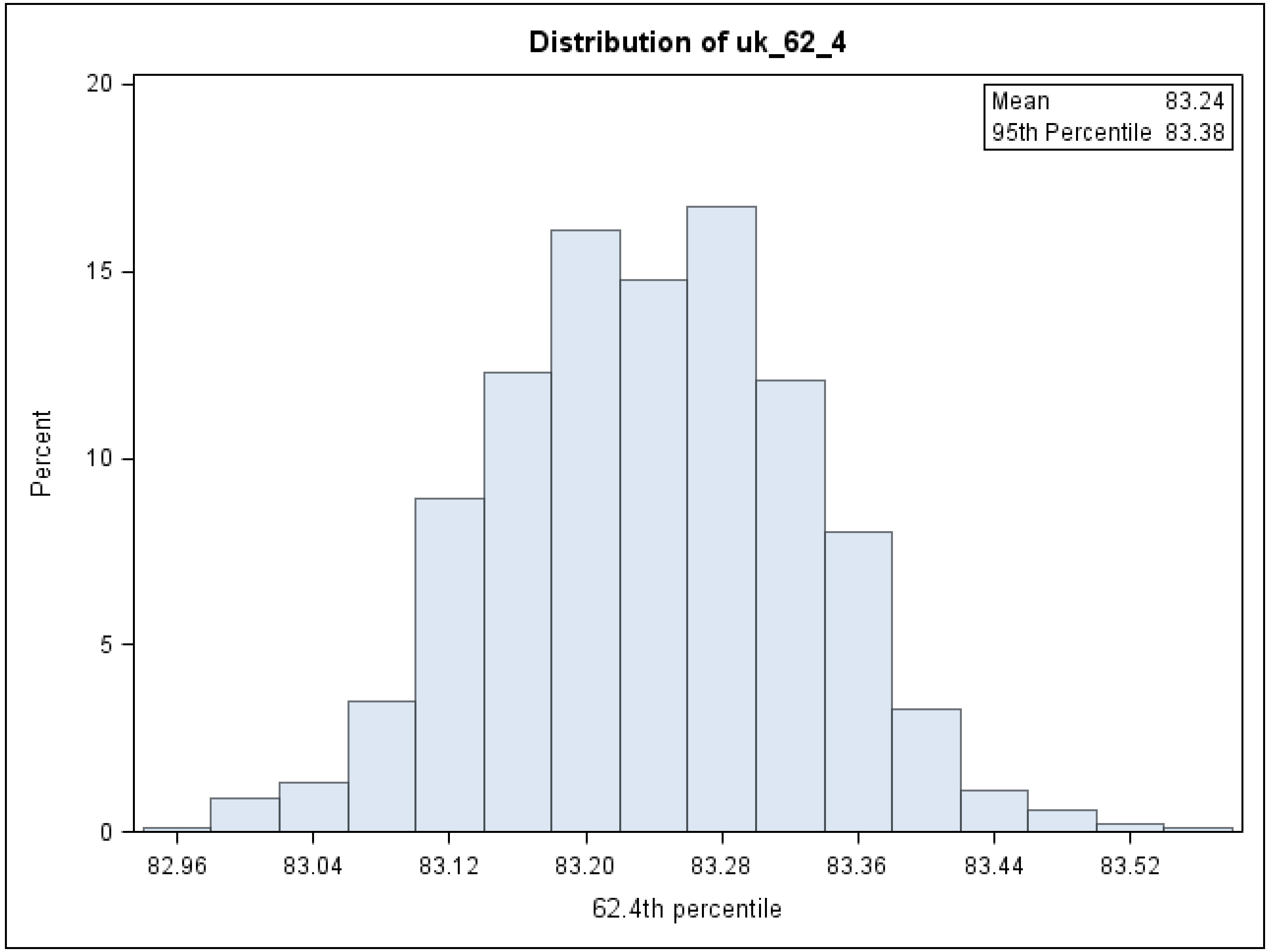
CAIR CENTER Analytic Services
THE HOUSE OF STATISTICS

Ex.1: Predicting the pass/fail under uncertainty

- Problem:
  - Pass/fail depends on a cutoff based on a 62.4th percentile (188 out of 500) of total score *tot*, defined as
    $tot = t1 + t2$,
    (t1=primary school total score, t2=entrance exam score)
  - where only *t2* is known,
  - *t1* can be assumed to follow a truncated normal distribution and has a correlation coefficient of *0.5* with *t2*.
- Question: Is my grandson's score of 83.6 above the cutoff?
- Solution (using CS approach):
  - 1. generate 500 *t1* scores from a truncated normal distribution
  - 2. transform *t1* so as to have a correlation coeff. of *0.5* with *t2*
  - 3. calculate tot and the 62.4th percentile
  - Repeat 1-3 1000 times

CAIR CENTER Analytic Services
THE HOUSE OF STATISTICS

Approximate Sampling Distribution of the 62.4th percentile (the cutoff)

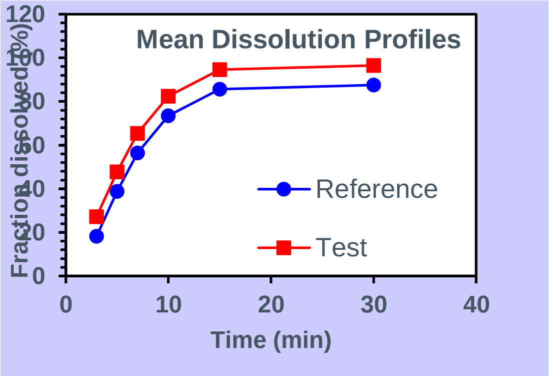


Actual cutoff was 83.2

Ex.2: f_2 statistic for dissolution profile comparison

› f_2 statistic for dissolution profile comparison (T vs. R)

$$f_2 = 50\log\left\{100\left[1+\frac{1}{n}\sum_{t=1}^{n}(\bar{R}_t-\bar{T}_t)^2\right]^{-0.5}\right\}$$



– Sampling distribution - unknown

– FDA Guidance recommends bootstrap CI (5th bootstrap percentile must be >= 50)

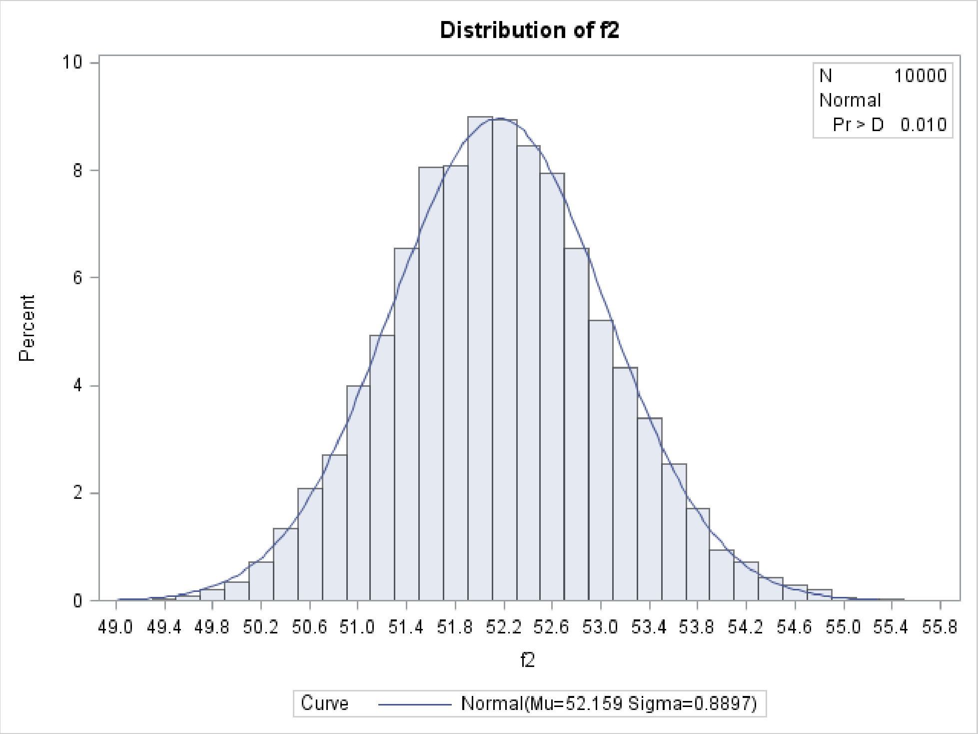
– Questions:

› Is 5th bootstrap percentile ≥ 50?

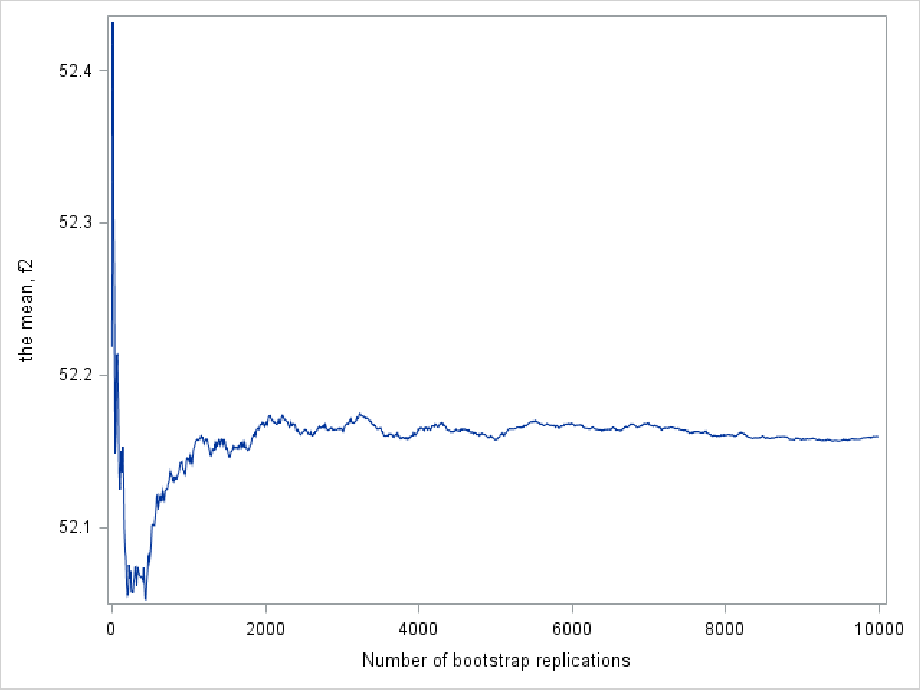
› How many bootstrap replicates B?

CAIR CENTER Analytic Services
THE HOUSE OF STATISTICS

f_2 statistic



Bootstrap distributon of f_2

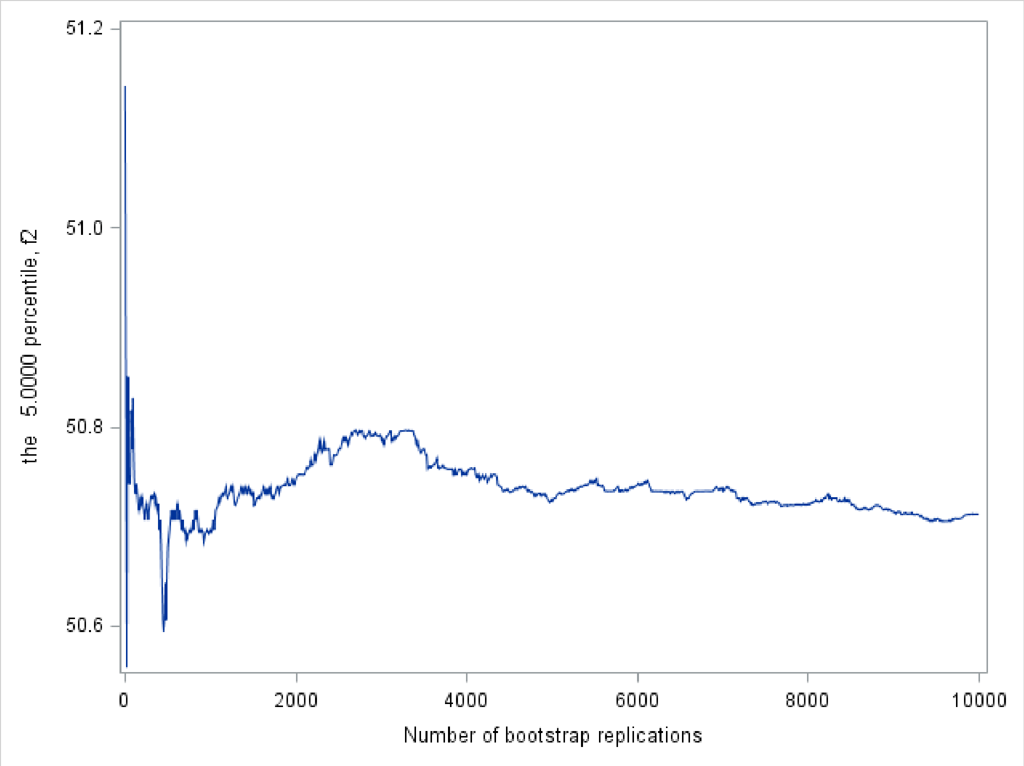


Convergence of bootstrap f_2 mean

Name	Observed Statistic	Approximate Lower Confidence Limit	Approximate Upper Confidence Limit	Confidence Level (%)	Method for Confidence Interval	Number of Resamples
f2	52.1547	**50.7120**	53.6622	90	Bootstrap percentile	10000

5. Centil > 50 ➔ T and R are bioequivalent

f_2 statistic



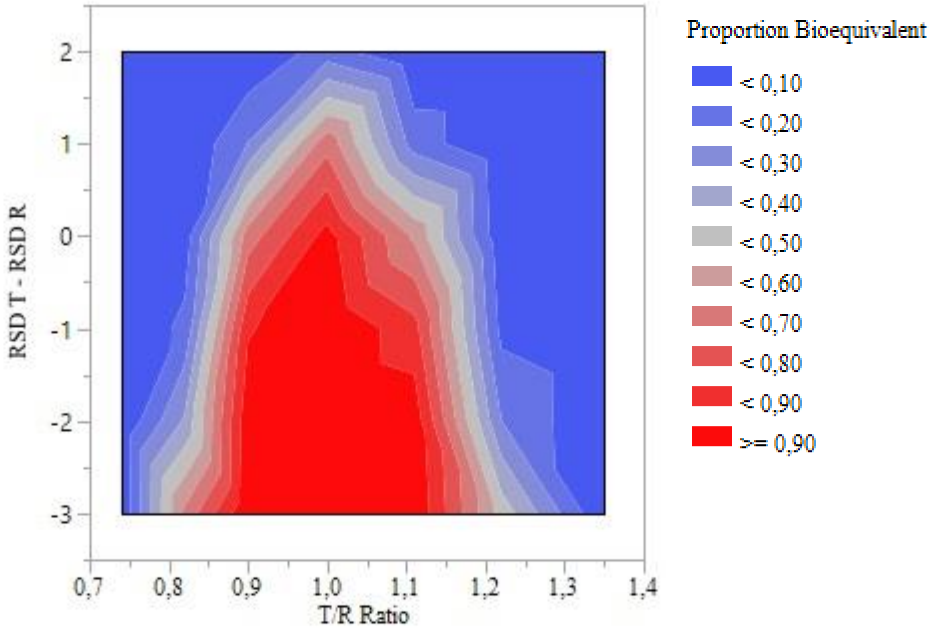
Convergence of bootstrap f_2 5th percentile (>50)

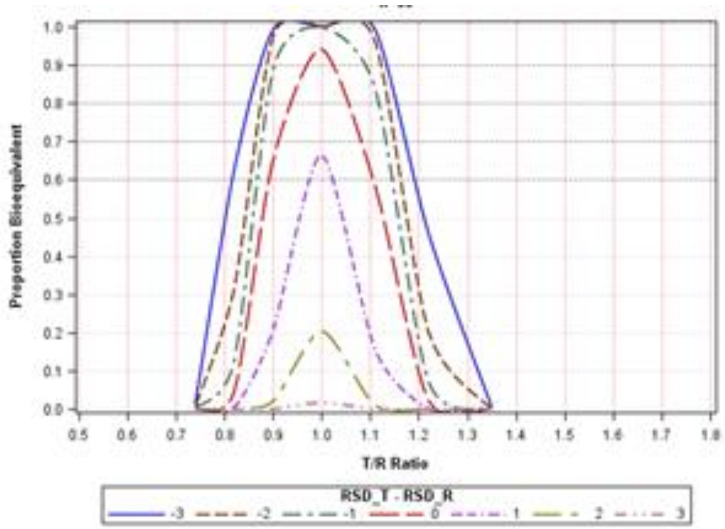
CAIR CENTER Analytic Services
THE HOUSE OF STATISTICS

Ex.3: Empirical power of the PBE statistic for bioequivalence testing (T vs. R)

$$PBE = \frac{(\mu_T - \mu_R)^2 + (Var_T - Var_R)}{max\{\sigma_0^2, Var_R\}}$$

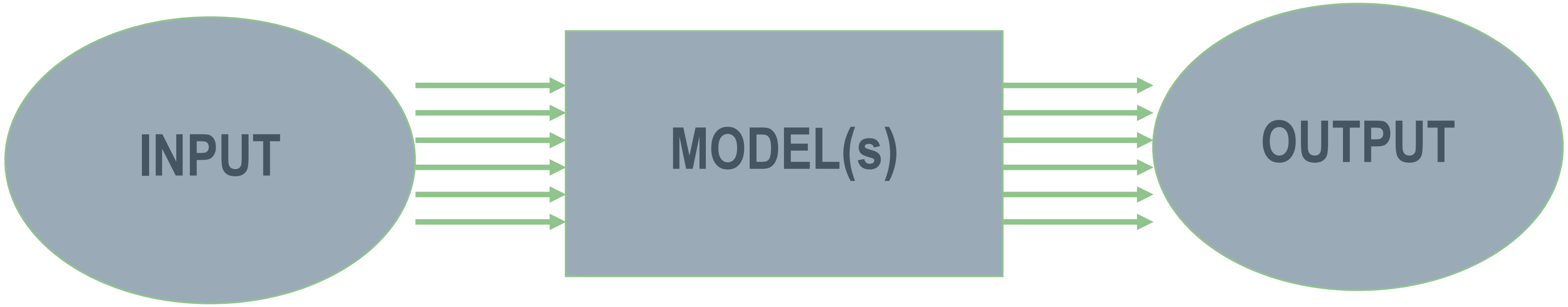
- Exact power formula – unknown
- Empirical power was developed by nonparametric-parametric bootstrap approach using 6 available R batch data





Ex.4: How robust is t statistic?

- Approach:
  - A Monte Carlo experiment



Sampling from:
-normal, gamma, etc.
-n= 10,20,50, 100, 200
-replications = 10,000

Calculation of t for each:
-distribution
-n
-replication

Graphical and tabular presentation of summary results

Contents

CAIR CENTER Analytic Services
THE HOUSE OF STATISTICS

Brief History of SC/CS

- 1951 – Von Neumann, random number generation and MC
- 1951 – Dwyer, Linear Computations
- 1963 – Wilkinson, rounding errors
- 1964 – Hammersley & Handscomb, MC Methods
- 1967 – Hemmerle, statistical computations
- Conferences, Societies:
  - Interface of Computer Science and Statistics
  - COMPSTAT
  - IASC
  - ITI
- Surveys:
  - 1991, 1999 – Grier, SW and stat.applications
  - 1993 – Billard & Gentle, Interface
- Journals

Computational Statistics I: Simulating Univariate and Multivariate Data

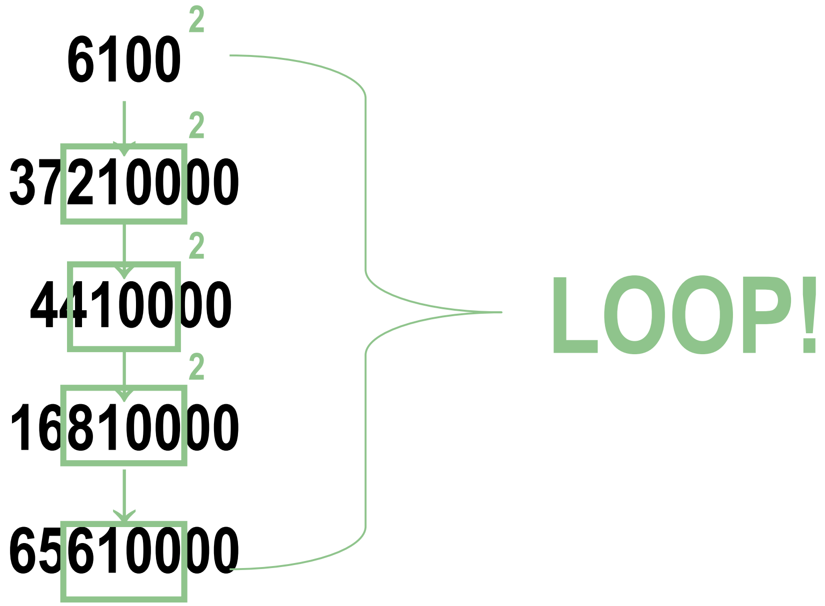
Uniform random numbers
Simulating univariate non-uniform data (continuous and descrete)
Simulating univariate non-uniform data with given first four moments
Simulating bi-variate & multi-variate normal data
Simulating bi-variate & multi-variate nonnormal data
Simulating random matrices
Simulating permutations
Resampling and Bootstrap

Pseudo-random numbers and generators

- Random number generation
  - pseudo-random numbers are generated by (computer) algorithms
- Middle square algorithm (J. Von Neumann, 1949):
  - To generate a sequence of 10 digit integers:
    - start with a 10 digit integer,
    - square it, and then
    - take the middle 10 digits as the next number in the sequence:
  - e.g. 3690295441^2
  - $13618\boxed{2804418653}84481$

Pseudo-random numbers and generators

- The sequence is not random (each number is completely determined from previous)
- It appears random, but it can get into short loops:



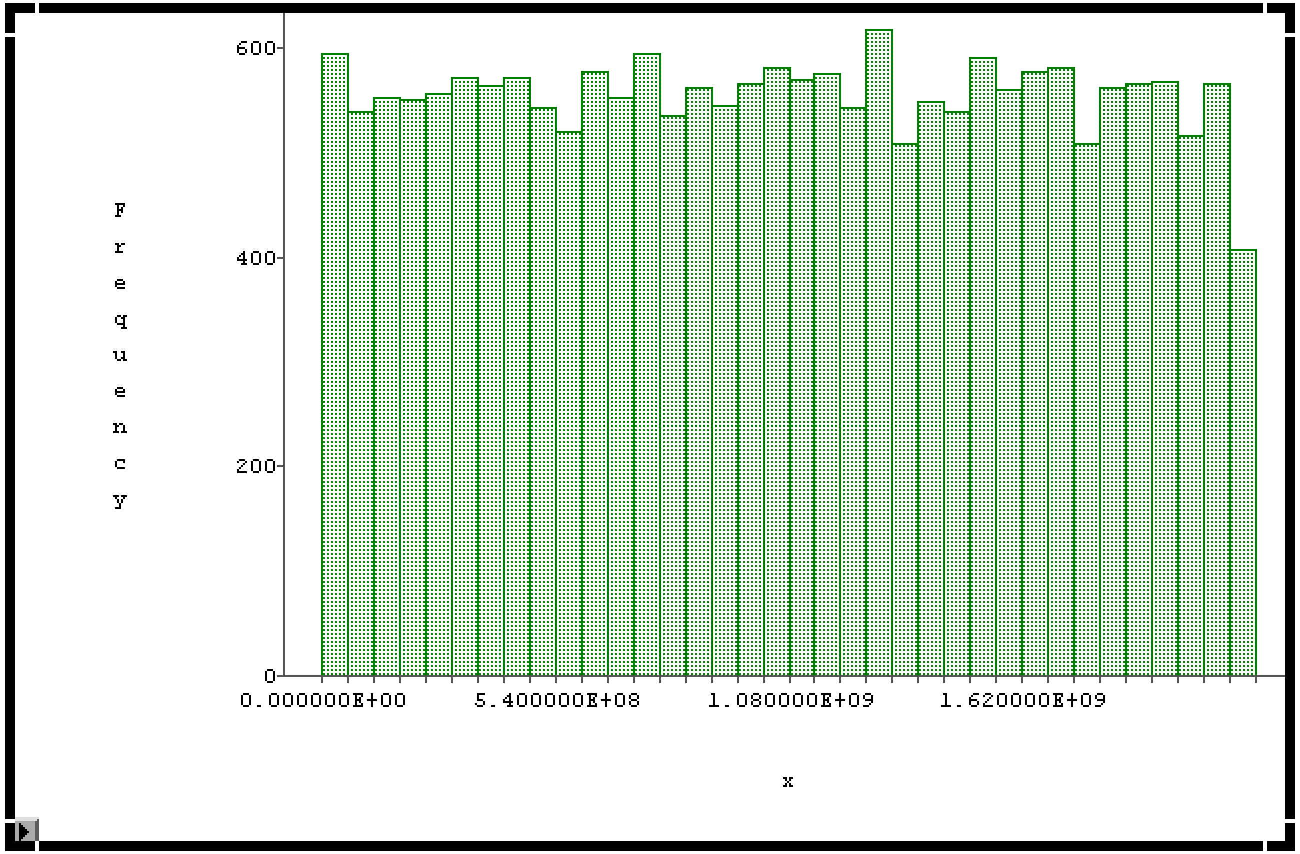
Linear congruential method (Lehmer, 1948)

- $I_{n+1} = (a*I_n + c) \bmod m$
  - I_0 = starting value (seed)
  - $a,c \geq 0$,
  - $m > I_0$, a,c
- Poor choice of constants can lead to poor sequences:
  - e.g., $a=c=I_0=7$, $m=10$ ➔
  - $\underbrace{7,6,9,0,}_{\text{loop}}7,6,9,0,\ldots$

RANDU

- There are well developed rules for selecting constants a,m,c, (if c=0 ➔ multiplicative congruential)
- Nevertheless,
- In the 1960's IBM distributed a popular generator RANDU:
  - $I_{n+1} = (65539 * I_n) \bmod 2^{31}$
  - Later found to have serious problem

RANDU (n=20000, seed=45813)

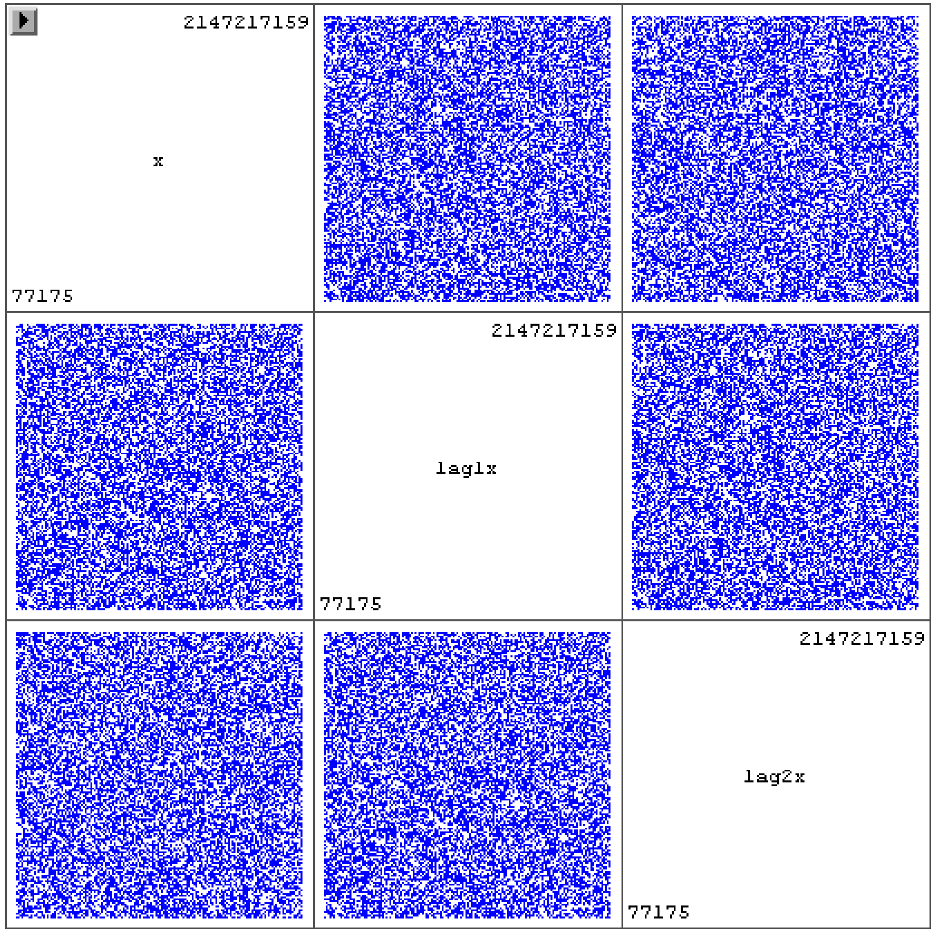


Looks OK

CHAPTER1_1_RANDU.SAS

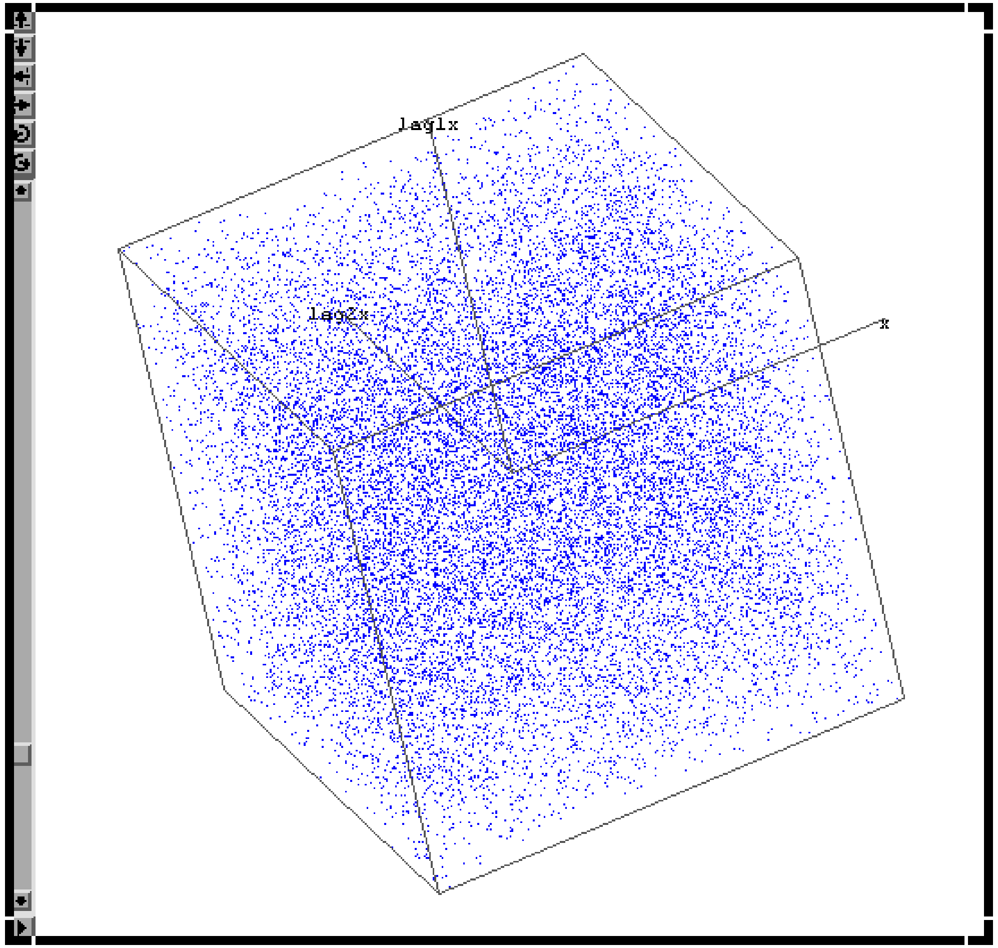
RANDU (n=20000, seed=45813)

x vs lag1(x) vs lag2(x)
($lag1(x_i)=x_i-x_{i-1}$)
($lag2(x_i)=x_i-x_{i-2}$)



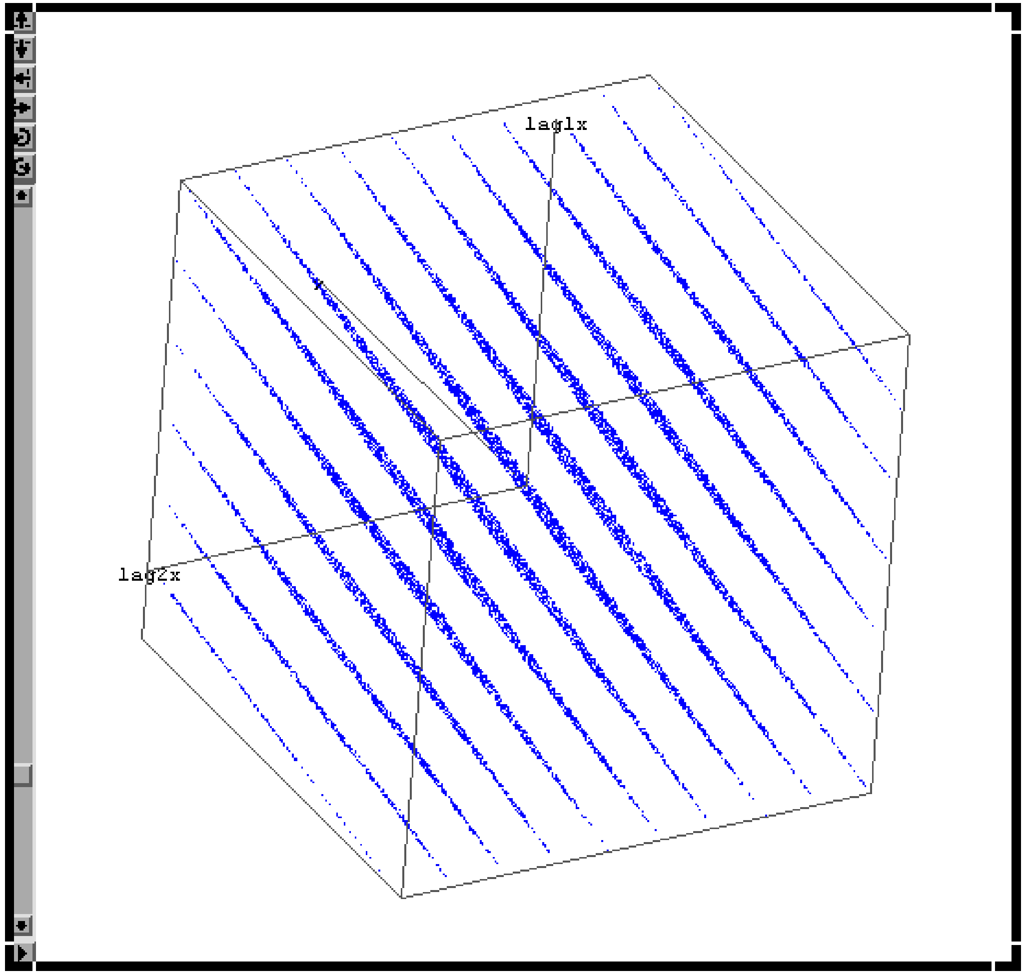
Looks OK

RANDU (n=20000, seed=45813)



Looks OK

RANDU (n=20000, seed=45813)



Problem when 0.5< lag1(x)<=0.51

Requirements for a “good” uniform RN generator

1. A uniform marginal distribution
2. Independence of the uniform variates
3. Reproducibility and portability
4. Computational speed

There are many statistical tests for testing 1. and 2.

CAIR CENTER Analytic Services
THE HOUSE OF STATISTICS

Recommendations on the use of RN generators

- Testing of RN generators is unnecessary in the sense that very good RN generators are available
- Testing is necessary in the sense that bad RN generators still EXIST on many computer systems.
- Use good RN generators with documented properties.
- For this workshop we use SAS and SAS RN generators (and R for hands-on session)

Methods for simulating non-uniform data

› <u>The inverse probability method</u>

THEOREM:

Let x has a continuous CDF F(x) so that $F^{-1}(u)$ exist for $0<u<1$ (and be computable), where

$F^{-1}(u) = \inf \{x: F(x) \geq u\}$

Then the random variable $F^{-1}(U)$ has CDF F(x), if U is uniformly distributed on [0,1] ($U \sim U(0,1)$).

CAIR CENTER Analytic Services
THE HOUSE OF STATISTICS

Methods for simulating non-uniform data

› Example1: exponential distribution

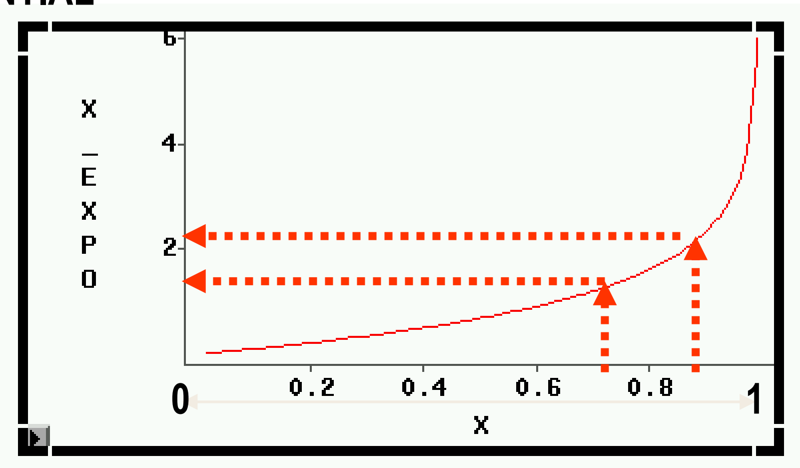
CDF: $F(x) = P(X \le x) = 1 - e^{-\lambda x}, x \ge 0, \lambda > 0$

$0 \quad , x < 0$

➔ $F^{-1}(u) = -\ln(1-u) / \lambda$

➔ If $U \sim U(0,1)$ ➔ $x = F^{-1}(U) \sim \exp(\lambda)$

EXPONENTIAL



UNIFORM

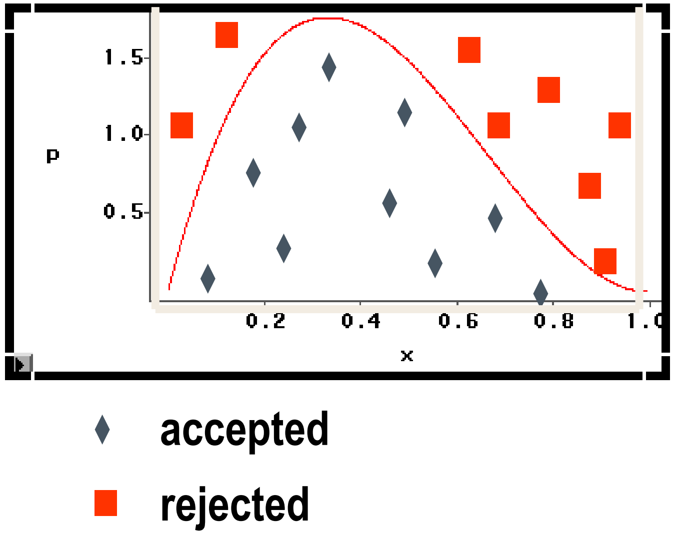
The inverse probability method for simulating exponential data

Example: Generate pseudo-random numbers from exponential distribution using the inverse prob. method

- Run the program: **CHAPTER1_1_EXPO_INVERSION.SAS**
- In Jmp open the dataset **EXPONENTIAL** and examine the distributions of the variables x and x_expo.

Methods for simulating non-uniform data

- Inverse probability method can be easily applied for generating random numbers according to various discrete distributions
- For continuous, such as normal and gamma distributions
  - No simple functional form for the inverse
  - Approximations for inverse functions available
- Other methods
  - Simple acceptance-rejection
  - General acceptance-rejection
  - Decomposition (method of mixtures)
  - General decomposition
  - Methods for specific distributions
    - e.g., Box-Muller technique for normal



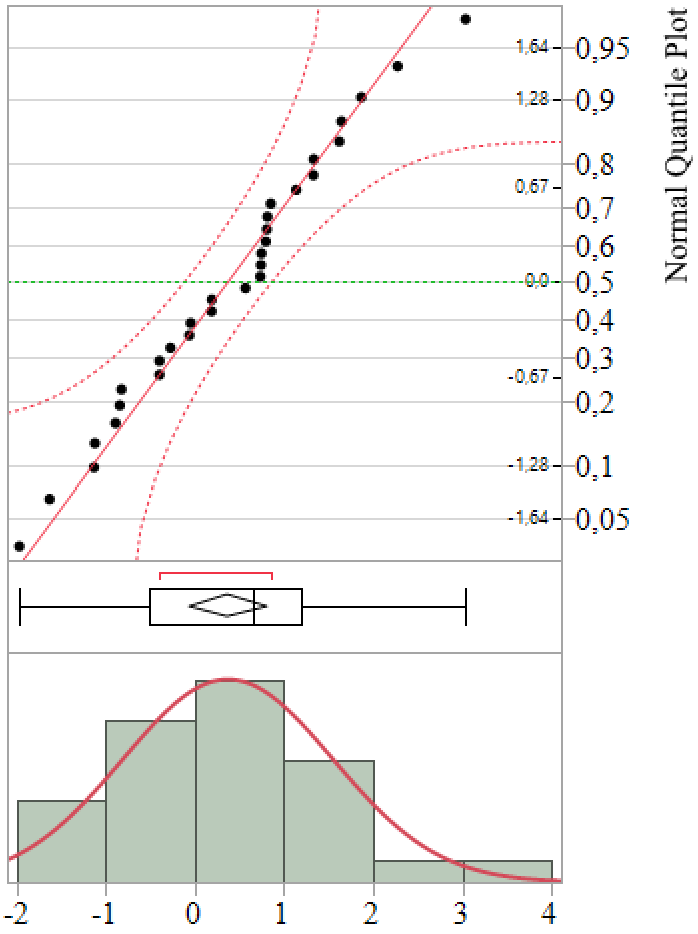
RN Generators in SAS

Distribution	SAS statement	Result
Bernoullijeva	x=rand('BERN',.75);	0
Beta	x=rand('BETA',3,0.1);	.99920
Binomial	x=rand('BINOM',10,0.75);	10
Cauchy	x=rand('CAUCHY');	-1.41525
χ^2	x=rand('CHISQ',22);	25.8526
Erlang	x=rand('ERLANG', 7);	7.67039
exponential	x=rand('EXPO');	1.48847
F	x=rand('F',12,322);	1.99647
Gamma	x=rand('GAMMA',7.25);	6.59588
geometric	x=rand('GEOM',0.02);	43
hypergemoetric	x=rand('HYPER',10,3,5);	1
lognormal	x=rand('LOGN');	0.66522
neg.binomial	x=rand('NEGB',5,0.8);	33
normal	x=rand('NORMAL');	1.03507
Poisson	x=rand('POISSON',6.1);	6
t	x=rand('T',4);	2.44646
table	x=rand('TABLE',.2,.5,.3);	2
triangular	x=rand('TRIANGLE',0.7);	.63811
uniform	x=rand('UNIFORM');	.96234
Weibul	x=rand('WEIB',0.25,2.1);	6.55778

simulating from N(0,1)

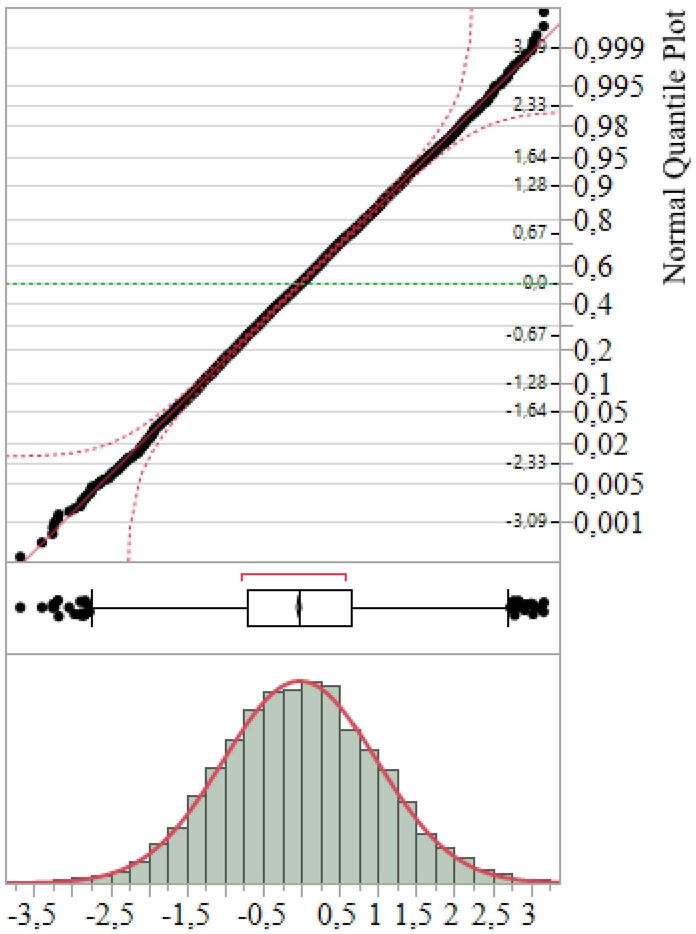
$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad \text{for all } x$$

n = 30



Normal(0,36804,1,17922)

n = 5000

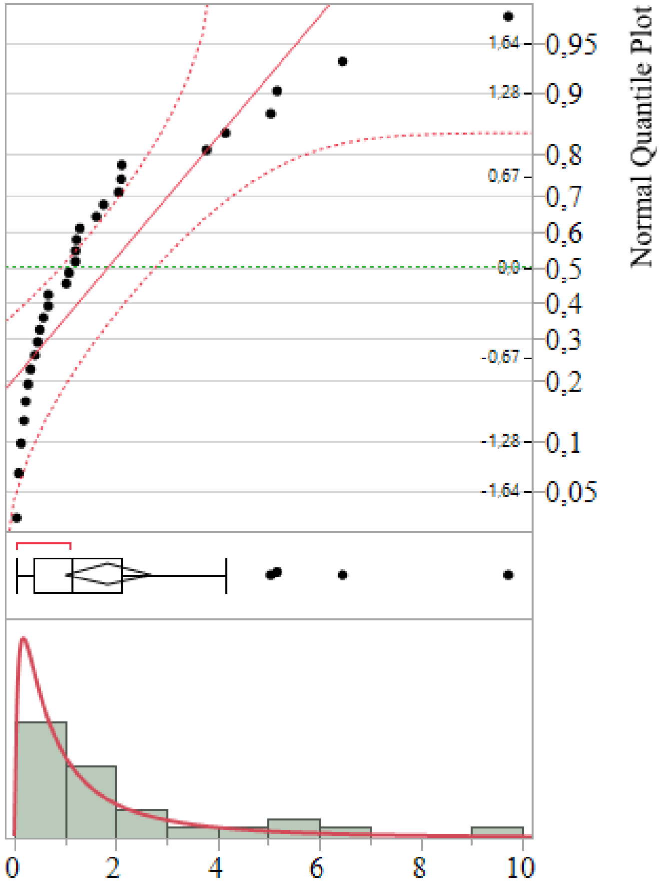


Normal(-0,0169,1,00382)

simulating from lognormal

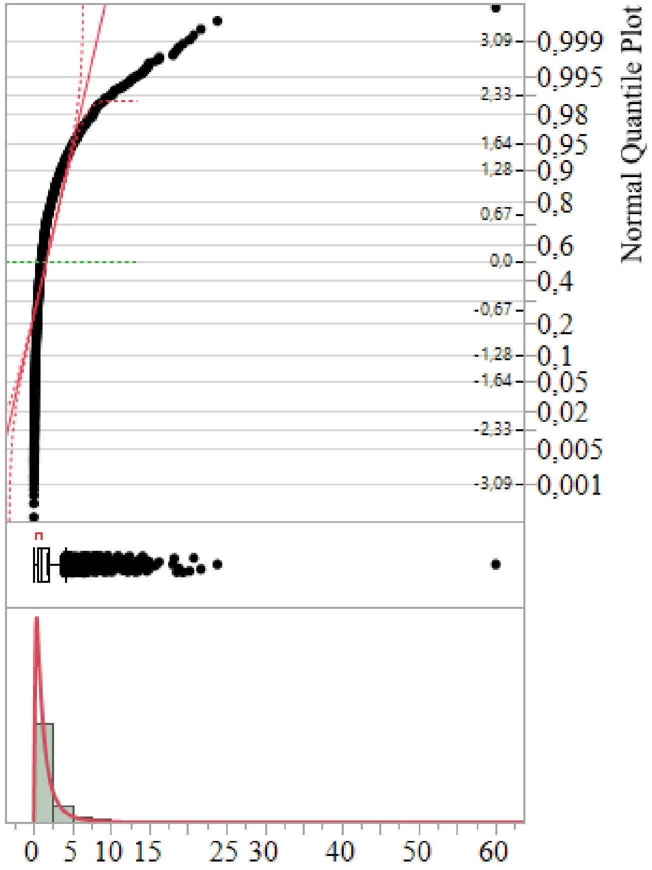
$$p(x) = \begin{cases} \frac{1}{\sigma\sqrt{2\pi}(x-\theta)} \exp\left(-\frac{(\log(x-\theta)-\zeta)^2}{2\sigma^2}\right) & \text{for } x > \theta \\ 0 & \text{for } x \leq \theta \end{cases}$$

n = 30



LogNormal(-0,0747,1,27745)

n = 5000



LogNormal(-0,0042,0,98372)

Using simulated data

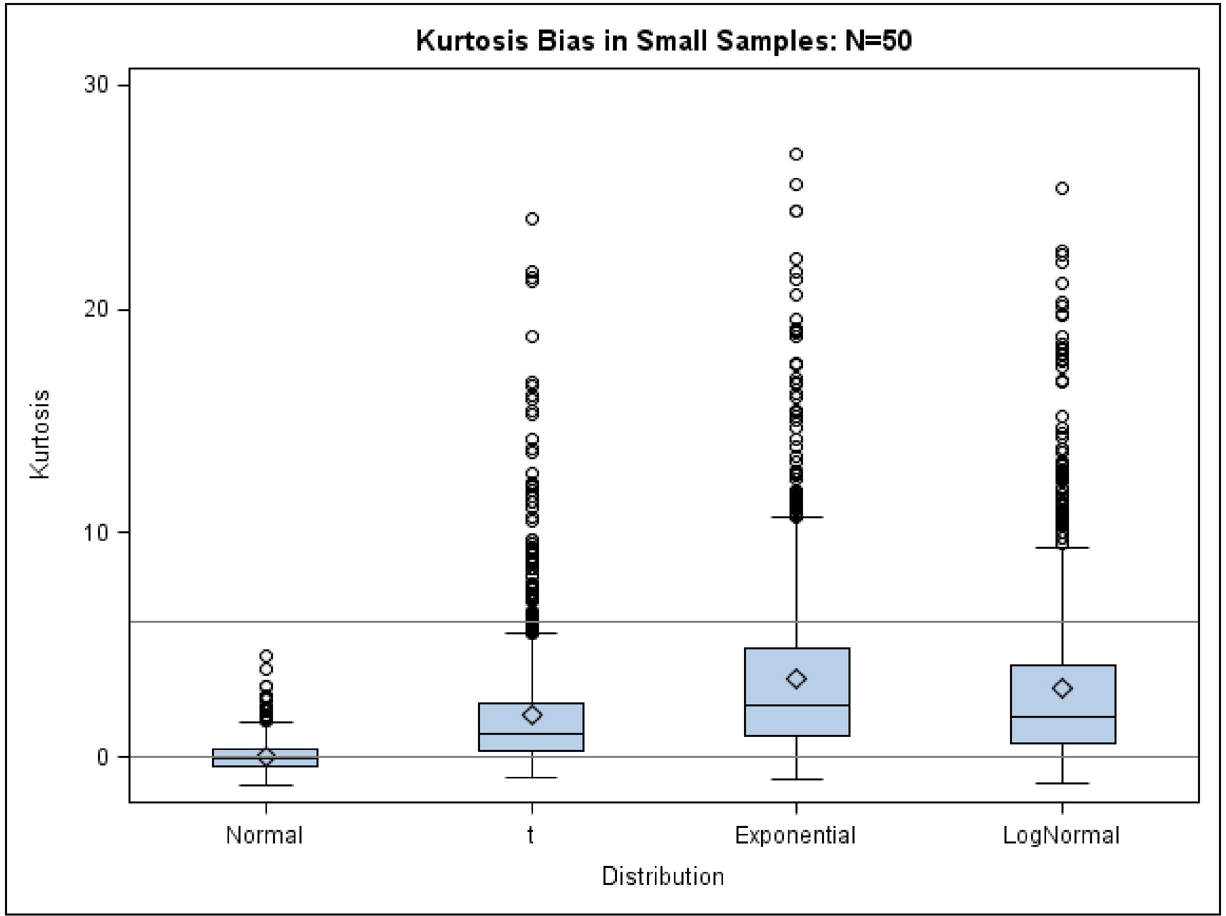
- **Approximate sampling distribution**
  - Mean: CLT, t, other statistics (with intractable/unknown sampling disn) for normal data and nonnormal data
- Evaluating statistical techniques
  - Robustness of statistical tests under various conditions (varying distributions, n, variance, etc.)
  - Comparing models/methods/algorithms
- Empirical power/ simulated power analysis
- CI: Coverage for normal and nonnormal data
- Comparing actual probability of Type I error to nominal α
- Computational inference/ MC tests/using simulation to compute p-values
- Accuracy of estimates
- Improving prediction

CAIR CENTER Analytic Services THE HOUSE OF STATISTICS

Sampling distribution of sample kurtosis γ_2 (for normal, t, exponential and lognormal data)

Example: 1. Generate pseudo-random numbers from a distribution (n=50), 2. Compute $\boldsymbol{\gamma_2}$, 3. Repeat 1-2 1000 times for each disn, 4. Summarize

Distribution	γ_2
Normal	0
t_5	6
Exponential	6
Lognormal (0,0.503)	6



Note: γ_2 stands for excess kurtosis

Simulating data from non-normal distribution with a desired level of skewness and kurtosis

- In MC studies it is often required to generate data with various degrees and types of non-normality defined by
  - Coefficient of skewness (γ_1), and
  - Coefficient of kurtosis (γ_2)
- Algorithms
  - Generalized lambda distribution
  - Fleishman method of polynomial transformation:
    - $Y = a + bZ + cZ^2 + dZ^3$
    - where
    - Y = non-normal varijable with desired γ_1 and γ_2 ,
    - Z = standard normal variable (N(0,1))
    - a,b,c,d (a=-c) coeficients for transformation (available from the table given in Fleishman (1978) or from SAS macro program)

Fleishman A.I. A Method for Simulating Non-Normal Distributions. Psychometrika 43: 521-531. 1978

CAIR CENTER Analytic Services THE HOUSE OF STATISTICS

Methods for simulating bivariate normal data

› simulating **bivariate normal random variables** (with a desired correlation coefficient ρ)

There is a simple **algorithm**:

1. Generate independently x_1, x_2 from N(0,1) (mean=0, std=1)
2. Set

   $x = x_1$

   $y = \rho \; x_1 + (1 - \rho^2)^{1/2} \; x_2$

CAIR CENTER Analytic Services
THE HOUSE OF STATISTICS

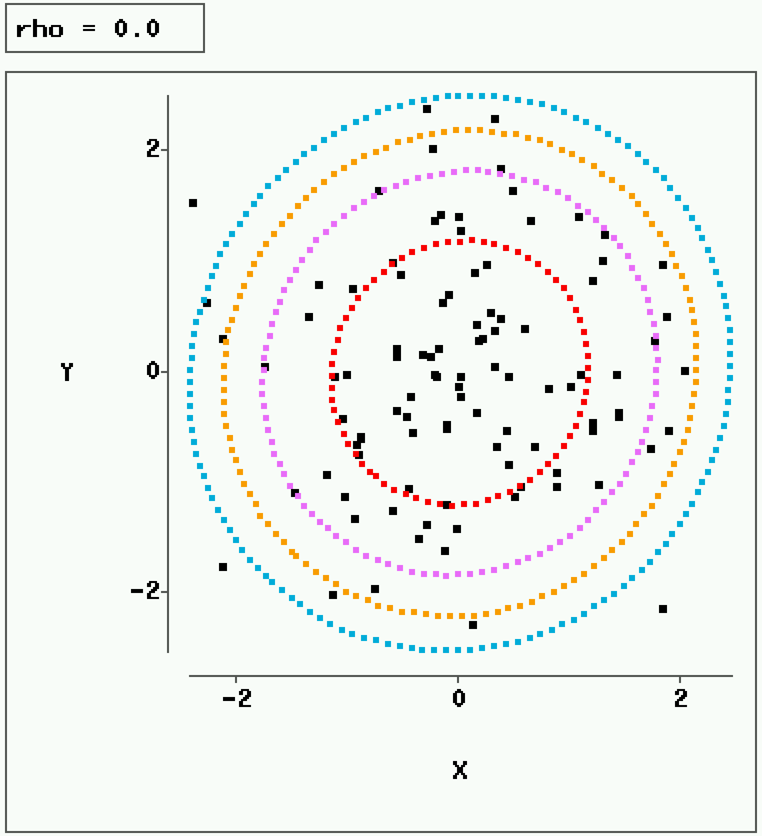
Simulating bivariate normal data

```
/** Generating a REPRODUCIBLE sequence of 100 random numbers from **/
/** bivariate NORMAL  NORMAL distribution **/

%LET SEED =1235;
%LET NREP=100;
%let rho=0.7;

DATA NORMAL2;
 CALL STREAMINIT(&SEED);
 DO REP = 1 TO &NREP;
  X1 =RAND( NORMAL );
  X2 = RAND( NORMAL );
  X = X1;
  Y = &RHO * X1 + SQRT(1 - &RHO**2)* X2;
  OUTPUT;
 END;
RUN;
```

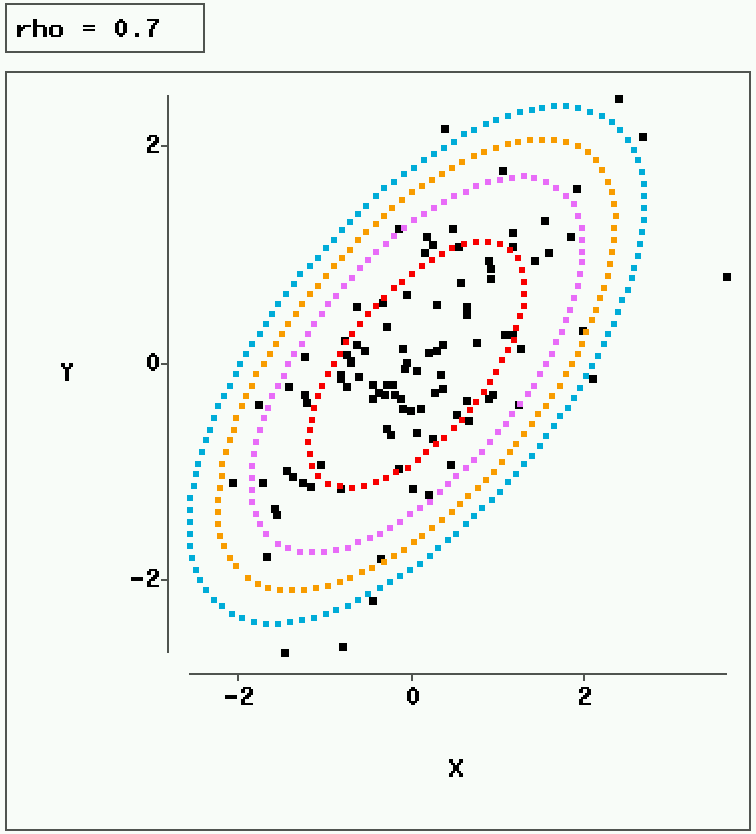
CAIR CENTER Analytic Services
THE HOUSE OF STATISTICS

CHAPTER1_1_NORMAL2.SAS

Simulating bivariate normal data



Confidence Ellipses	
Type	Coefficient
Prediction	0.5000
Prediction	0.8000
Prediction	0.9000
Prediction	0.9500



Confidence Ellipses	
Type	Coefficient
Prediction	0.5000
Prediction	0.8000
Prediction	0.9000
Prediction	0.9500

Intro ex.1
Slide 7

Simulating multivariate normal data

- Problem: Simulate a random nxp matrix X from multivariate normal distribution $N(\mu,\Sigma)$, with a given mean vector μ and a variance-covariance matrix Σ.
- For simplicity we can reformulate the problem to the one of simulating from a mult.normal distribution $N(0,R)$, where R is a pre-specified correlation matrix.

CAIR CENTER Analytic Services
THE HOUSE OF STATISTICS

Simulating multivariate normal data

- Assuming R is a pos.def. symmetric correlation matrix, it can be decomposed as follows:
- $R = Y\Lambda Y^T$, where

  Y is pxp orthogonal matrix of the eigenvectors of R, and

  Λ is a diagonal pxp matrix of eigenvalues of R.

  Then it is easy to demonstrate that
- $X = Z\,\Lambda^{1/2}\,Y^T \sim N(0,R)$, if $Z \sim N(0,I)$.

Note: The algorithm can be easily implemented using the IML matrix language (in SAS), or using SPlus, R

CAIR CENTER Analytic Services
THE HOUSE OF STATISTICS

Simulating random matrices (from Wishart disn. with Σ=I)

The algoritm is based on the Cholesky decomposition of a pxp matrix V with Wishart distribution W(I,n,p):

$V=TT^T$,

where $T=(t_{ij})$ is a pxp lower triangular matrix.

t_{ij} are independently distributed:

$t_{ii}^2 \sim$ chi-square with (n-i) df (i=1,p),

t_{ij} $(i \neq j) \sim N(0,1)$.

(Olkin I. (1985))

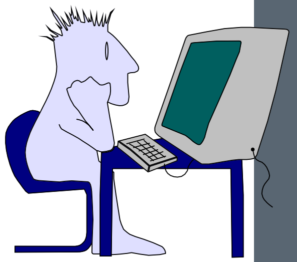
Simulating permutations

- Example: All permutations of 1,2,3 (total: 1*2*3=3!=6):
  123
  132
  213
  231
  312
  321
- In general:
  - No. of permutations of n elements = n!
- Applications in statistics:
  - Experimental plans (e.g., randomization lists)
  - Randomization (or "permutation") tests (i.e., procedures for determining statistical significance directly from data (using permutations), without applying a sampling distribution)
    - Exact r. tests (using ALL permutations)
    - Approximate (MC estimates of p-values) (using a random sample from ALL permutations)
  - For adjusting p-values in multiple testing problems

CAIR CENTER Analytic Services
THE HOUSE OF STATISTICS

Simulating permutations: An example using the exact permutation test

- Experiment to examine if a self-proclaimed vodka “expert” can recognize vodka brands (in a blind experiment with 4 vodka brands tasted in a randomized order).
- H_0: Expert’s opinion of the contents of the glasses is independent of the actual contents of the glasses.
- ➔all permutations are equally possible
- Results of the taste test:



Question: What is the probability of 2 or more correct, if, in fact, the “expert” can not discriminate among the brands (H_0)

<u>Example</u>: Is the “expert” really an expert?

	GLASS 1	GLASS 2	GLASS 3	GLASS 4
Actual contents	Pollish	Premium US	Russian	Budget US
“Expert’s” opinion	Pollish	Premium US	Budget US	Russian

Outcome: 2 correct

CAIR CENTER Analytic Services THE HOUSE OF STATISTICS

CHAPTER_1_1_PERMUTATIONS_VODKA_EX.SAS

Expert's opinion

rep	glass 1: Pollish	2: Premium US	3: Russian	4: Budget US	No. correct
1	Polish	Russian	Premium US	Budget US	2
2	Polish	Russian	Budget US	Premium US	1
3	Polish	Premium US	Russian	Budget US	4
4	Polish	Premium US	Budget US	Russian	2
5	Polish	Budget US	Russian	Premium US	2
6	Polish	Budget US	Premium US	Russian	1
7	Russian	Polish	Premium US	Budget US	1
8	Russian	Polish	Budget US	Premium US	0
9	Russian	Premium US	Polish	Budget US	2
10	Russian	Premium US	Budget US	Polish	1
11	Russian	Budget US	Polish	Premium US	0
12	Russian	Budget US	Premium US	Polish	0
13	Premium US	Polish	Russian	Budget US	2
14	Premium US	Polish	Budget US	Russian	0
15	Premium US	Russian	Polish	Budget US	1
16	Premium US	Russian	Budget US	Polish	0
17	Premium US	Budget US	Polish	Russian	0
18	Premium US	Budget US	Russian	Polish	1
19	Budget US	Polish	Russian	Premium US	1
20	Budget US	Polish	Premium US	Russian	0
21	Budget US	Russian	Polish	Premium US	0
22	Budget US	Russian	Premium US	Polish	0
23	Budget US	Premium US	Polish	Russian	1
24	Budget US	Premium US	Russian	Polish	2

CHAPTER_1_1_PERMUTATIONS_VODKA_EX.SAS

p=7/24=0.29

Note:
This experiment is inspired by the famous Fisher's "Tea lady" experiment (1935) ("Fisher's exact test") - from E.W. Noreen p.12

Resampling and Bootstrap

› According to Efron: Statistics operates on 2 levels:
  - Algorithms
  - Accuracy of these algorithms

  „Bootstrap is a way of using computer power to answer the question of accuracy (because problems are becoming more difficult, data sets enormous and questions are much more intricate)"

Nonparametric Bootstrap

Random sampling with replacement from data:

x_1	x_2	x_3	$\rightarrow \hat{\theta}$
x_2	x_2	x_1	$\rightarrow \hat{\theta}^*(1)$
x_2	x_1	x_3	$\rightarrow \hat{\theta}^*(2)$
x_3	x_1	x_3	$\rightarrow \hat{\theta}^*(3)$
x_1	x_2	x_2	$\rightarrow \hat{\theta}^*(4)$
	. . .		
x_3	x_2	x_1	$\rightarrow \hat{\theta}^*(B)$

Bootstrap estimate $\hat{\theta}^* = \sum_{b=1}^{B} \hat{\theta}*(b)/B$, and $\widehat{\sigma_B}$

- Suppose $x_1, x_2, \ \ldots \ x_n \sim F$ on R^1
- Denote by $\hat{\theta}$ the estimate of unknown parameter θ
- $se(F;n,\theta)$ is unknown
- Empirical distribution function $\hat{F}$ puts probability $1/n$ on each of the n observed points x_i , i=1,..n.
- We estimate F by $\hat{F}$, and $se(F;n,\theta)$ by $se(\hat{F};n,\hat{\theta})$

Basic Bootstrap Methods

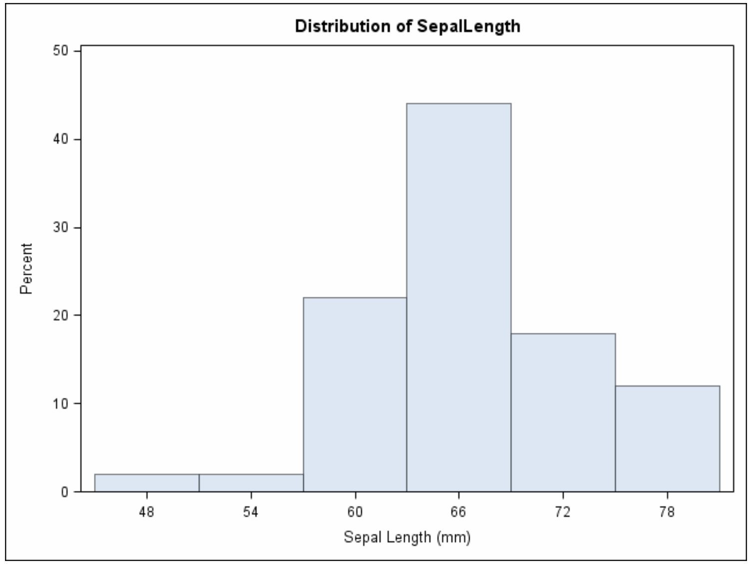
- In many cases there is no simple expression for the function se(F;n,θ) , but it is easy to numerically evaluate se($\hat{F}$;n,$\hat{\theta}$), by taking "bootstrap samples" from:
  - actual sample (nonparametric bootstrap), or from
  - fitted distribution $\hat{F}$ (parametric bootstrap).

Intro ex.2
Slide 9

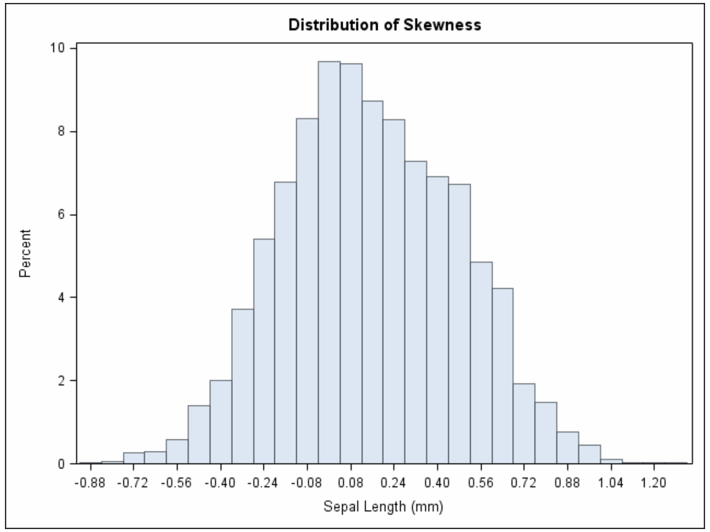
Nonparametric Bootstrap
An example: bootstrap estimates of skewness and kurtosis (using Fisher's Iris data)

<u>Example</u>: 1. Take a bootstrap sample from data (n=50),
2. Compute γ_1 and γ_2, 3. Repeat 1-2 5000 times, 4. Summarize (90% CI)

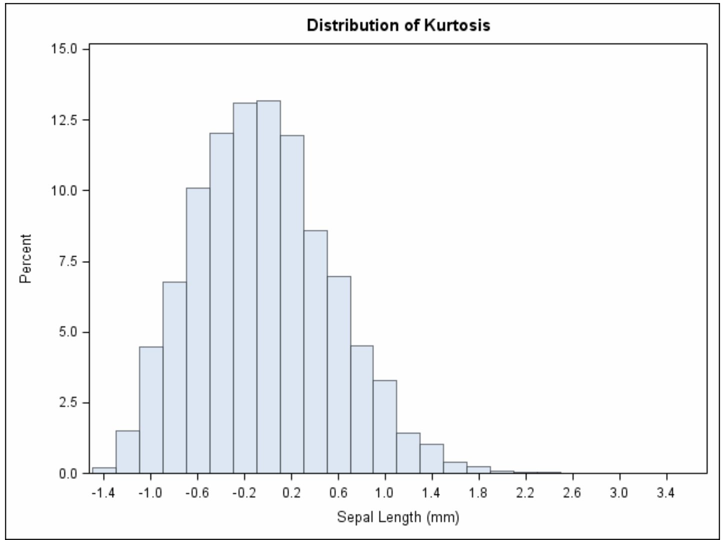
- Virginica species from Iris data (n=50)
- Sepal length
- Skewness = 0.118, kurtosis = 0.0329
- Accuracy?

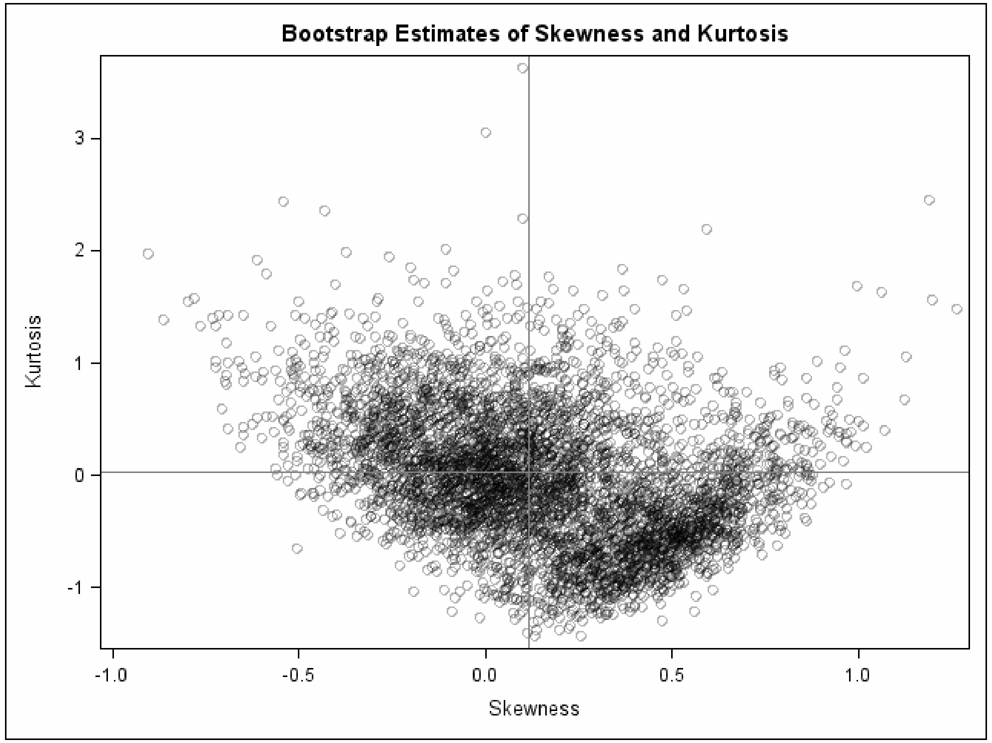


Bootstrap estimates of skewness and kurtosis (using Fisher's Iris data)



Variable	N	Mean	Std Dev	5th Pctl	95th Pctl
Skewness	5000	0.153	0.321	-0.348	0.677
Kurtosis	5000	-0.038	0.590	-0.942	0.986



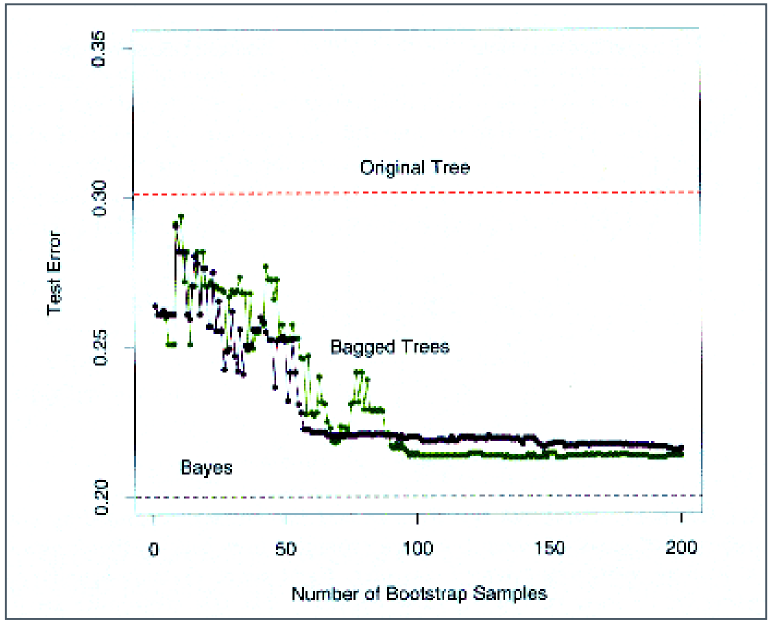


Bootstrap and Big Data

- Decision trees
- Low predictive power (and unstable)
- Leo Breiman (1996): „Bagging” – ensemble of trees on bootstrap samples
- In “bagging” bootstrap is not used for estimating accuracy, but for improving (reducing) the prediction error

- Complex algorithms using enormous data sets
- Models are compared and a model with the lowest prediction error is selected
- For complex models there is no simple way to estimate prediction error
- Bootstrap for estimating prediction error (e.g. 632+ rule)

An example of Bagging trees



(from T. Hasti, R. Tibshirani, J. Friedman (2001) The Elements of Statistical Learning, p.247)

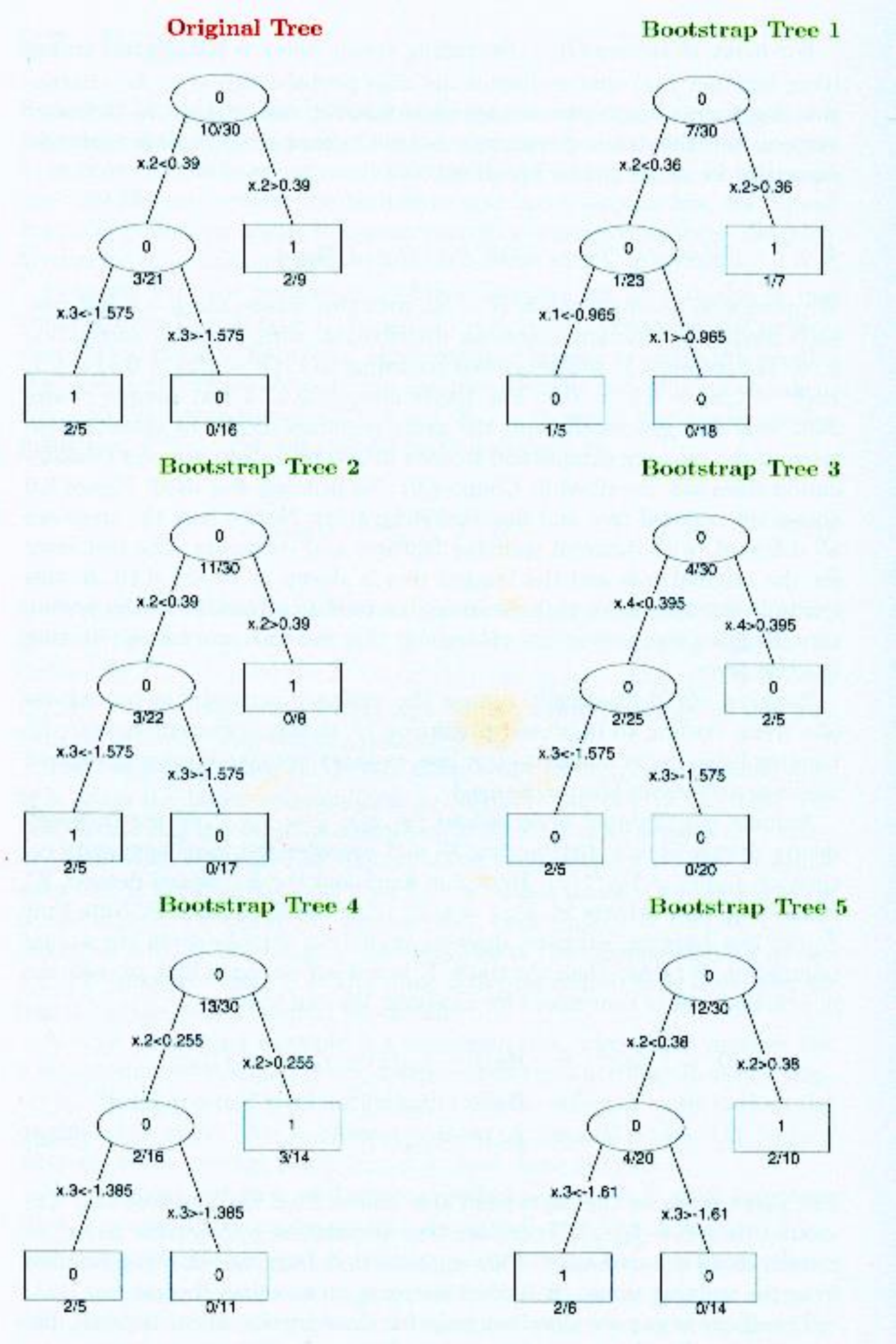


FIGURE 8.9. *Bagging trees on simulated dataset. Top left panel shows original tree. Five trees grown on bootstrap samples are shown.*

CAIR CENTER Analytic Services
THE HOUSE OF STATISTICS

Computational Statistics II: Using Simulation to Evaluate Statistical Techniques and Models

MC studies
MC study of the robustness of a 1-sample t-test
The power of a regression test
Coverage probability of 90% and 95% CI for the mean

Using simulated data

- Approximate sampling distribution ✓
  - Mean: CLT, t, other statistics (with intractable/unknown sampling disn) for normal data and nonnormal data
- Evaluating statistical techniques ✓
  - Robustness of statistical tests under various conditions (varying distributions, n, variance, etc.)
  - Comparing models/methods/algorithms
- Empirical power/ simulated power analysis ✓
- CI: Coverage for normal and nonnormal data ✓
- Comparing actual probability of Type I error to nominal α ✓
- Computational inference/ MC tests/using simulation to compute p-values ✓
- Accuracy of estimates ✓
- Improving prediction ✓

CAIR CENTER Analytic Services
THE HOUSE OF STATISTICS

When are MC studies needed?

- When the theoretical assumptions of the underlying statistical theory are not fulfilled
  - e.g., studies that examine the consequences of departures from theoretical conditions, (such as normality)

and/or

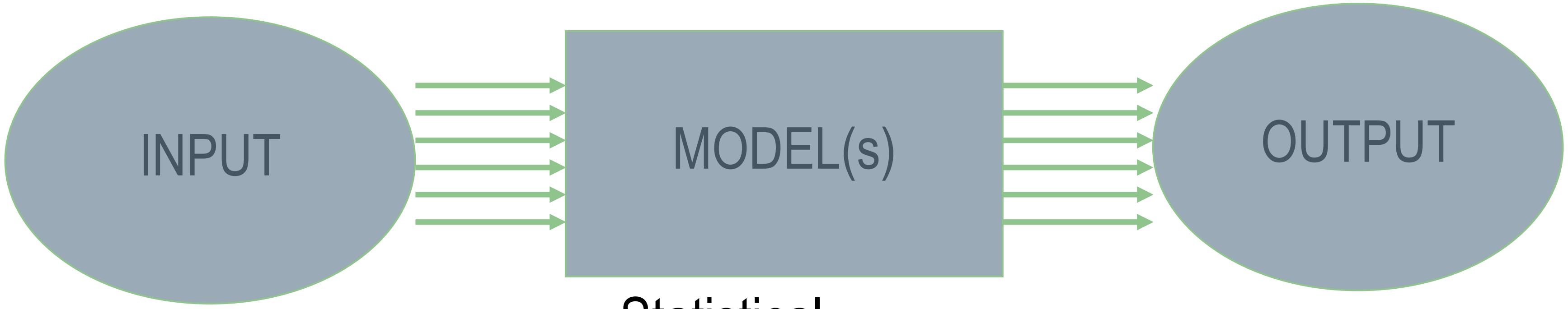
- When the underlying statistical theory is not completely developed, doesn't exist or is not tractable
  - e.g., determining (MC) sampling distribution of a statistics without theoretical distribution.

Steps in a MC study

- Ask questions that can be examined through a MC study.
- Design a MC study to provide answers to the questions.
- Generate data.
- Implement the technique/model/algorithm you want to study (e.g., using functions, procedures, macro, IML (matrix language) code)
- Obtain and accumulate the statistic of interest from each iteration.
- Analyze the accumulated statistic of interest.
- Draw conclusions based on empirical results.

Monte Carlo experiments

- Controlled statistical experiments performed on a computer
- Should be used only when analytical and numerical techniques cannot supply answers.



•Design
•Simulation of data

•Statistical technique(s) applied to each replication and factor level
•Accumulation

•Analysis of results
•Conclusions

Monte Carlo experiments

› Example (simple):
  – Robustness of t statistic to departures from normality

•Sampling from:
-normal, gamma, etc.
-n= 10,20,50, 100, 200
-# of replications = 10,000

•Calculation of t for each:
-distribution
-n
-replication
•Accumulation

•Graphical and tabular presentation of summary results
•Conclusions

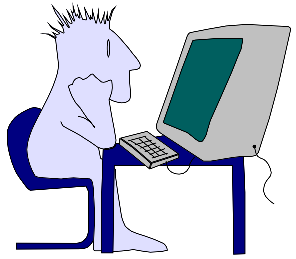
THE HOUSE OF STATISTICS

Assumptions i.e. conditions for applying t-test

- Normality ($x_1, x_2, \ldots\ x_n$ are assumed to be distributed as $N(\mu, \sigma^2)$)
- or, for larger n, we assume normal sampling distribution of the sample mean (central limit theorem)
- iid ($x_1, x_2, \ldots\ x_n$ are assumed to be independently identically distributed)

- Question:
  - How sensitive is the t-test (i.e., t statistic distribution) to the departures from normality assumption (robustness)? What happens when sampling distribution of the mean is not normal?

¸Robustness of the t statistics MC experiment



<u>Example</u>:
1. Generate data from a distribution (n),
2. Compute t (test statistic),
3. Repeat 1-2 10,000 times for various ns and distributions,
4. Summarize

› Examine, using a Monte Carlo experiment (and animation), the samfor pling distribution of the t statistic if the underlying data is from:

- Normal distribution N(0,1)
- Gamma distribution Gamma(0.5,1)

Use sample sizes n=10,20,30,50,100,200.

Compare simulation moments to the asymptotic moments of the t distribution.

Discuss the consequences of applying t-test in the case of very skewed parent distribution.

(Pearson (1929), Geary (1936), Gayen (1949), Pearson and Please(1975), Efron (1969), ...)

CAIR CENTER Analytic Services
THE HOUSE OF STATISTICS

Design of a MC study: Robustness of t statistic

- Based on the identified questions (**e.g., robustness of t statistic**) we consider the major factors that may affect the behavior of the statistic of interest (e.g., sampling distribution of t statistic):
  - Sample size (➔1st factor)
    - e.g.,. 10, 20, 50,
  - Distribution (➔2nd factor)
    - e.g., normal, gamma, uniform
  - Number of samples to be drawn for each combination of factor level (number of replications)
    - e.g., 10000

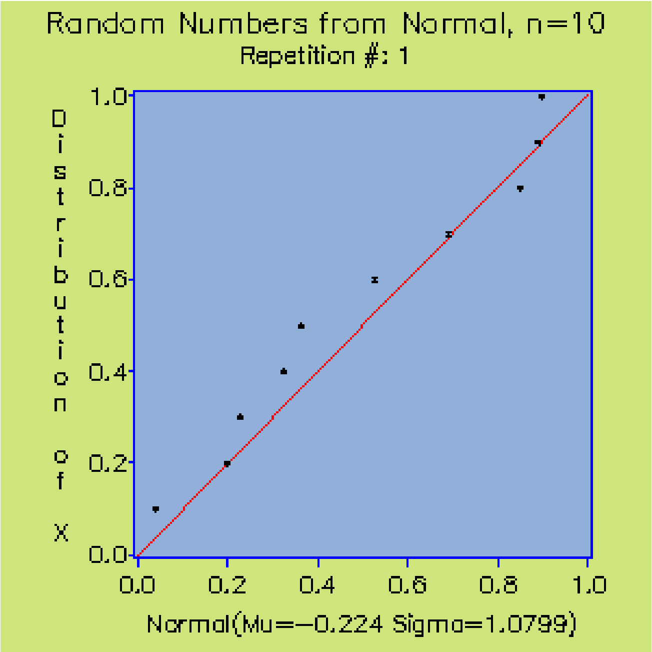
Total:
3 x 10 x 10000 +
3 x 20 x 10000 +
3 x 50 x 10000 +
3 x 100 x 10000 +
3 x 200 x 10000 =
11,400,000 random numbers

	Distribution		
Sample size	**Normal**	**Gamma**	**Uniform**
10	10000	10000	10000
20	10000	10000	10000
50	10000	10000	10000
100	10000	10000	10000
200	10000	10000	10000

CAIR CENTER Analytic Services
THE HOUSE OF STATISTICS

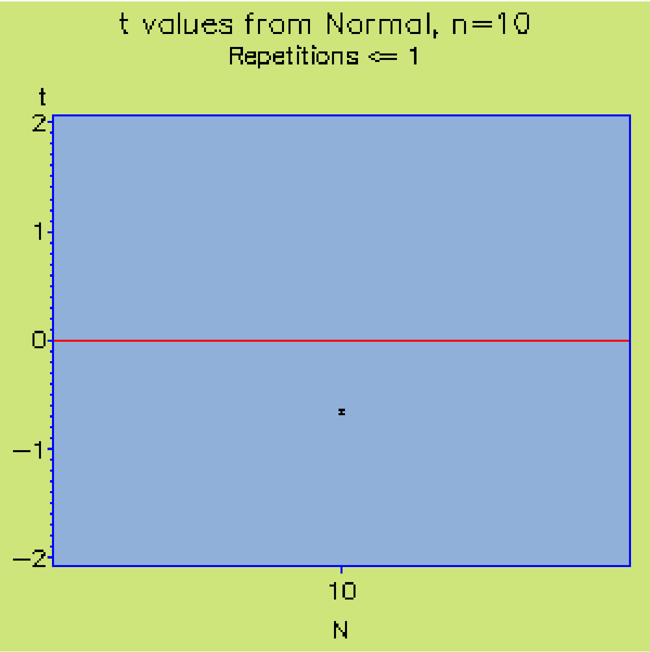
10 random samples of size 10 from normal distribution and the resulting 10 t values (animation)

probability-probability plot



Caution: Natural randomness results in departures from straightness even for N(0,1) data

scatter plot



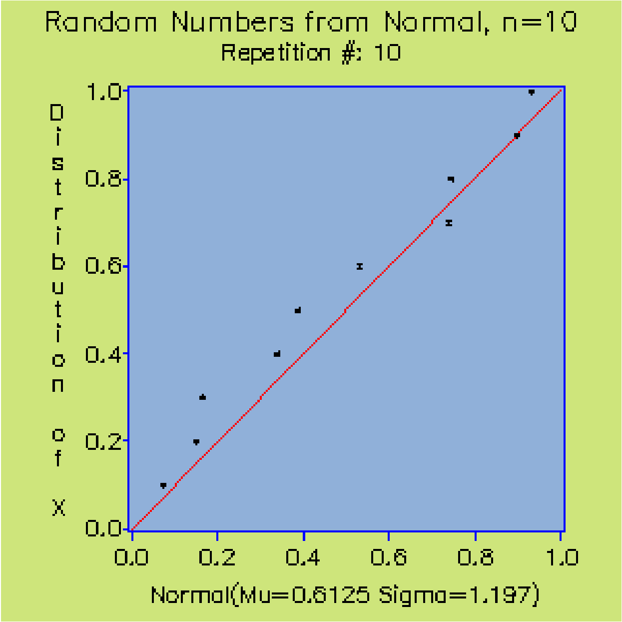
10 values of the t statistic, each computed from a sample of 10 N(0,1) data (t=xbar/se)

CHAPTER1_2_T_NORMAL1.SAS

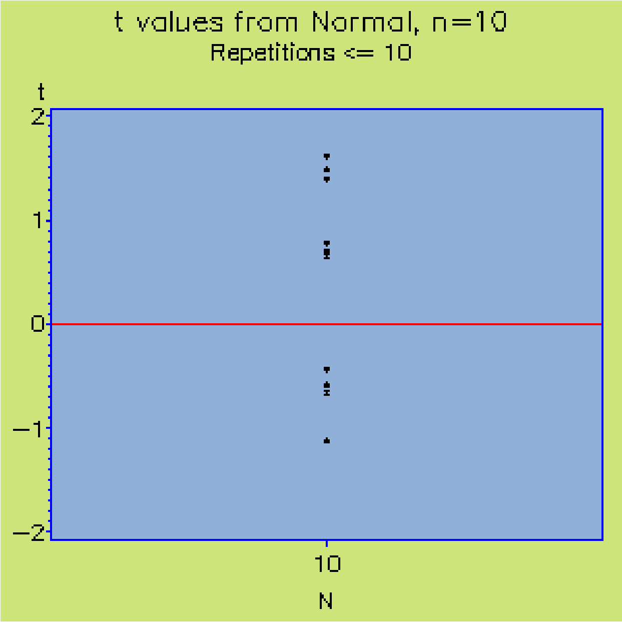
CAIR CENTER Analytic Services
THE HOUSE OF STATISTICS

10 random samples of size 10 from normal distribution and 10 resulting t values

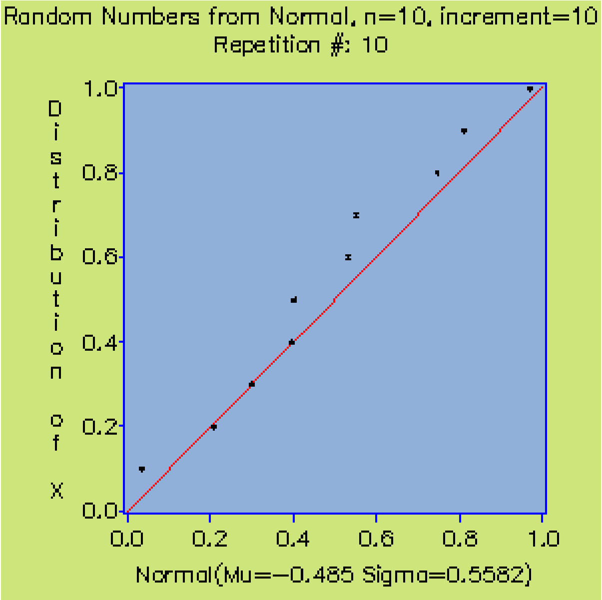
probability-probability plot

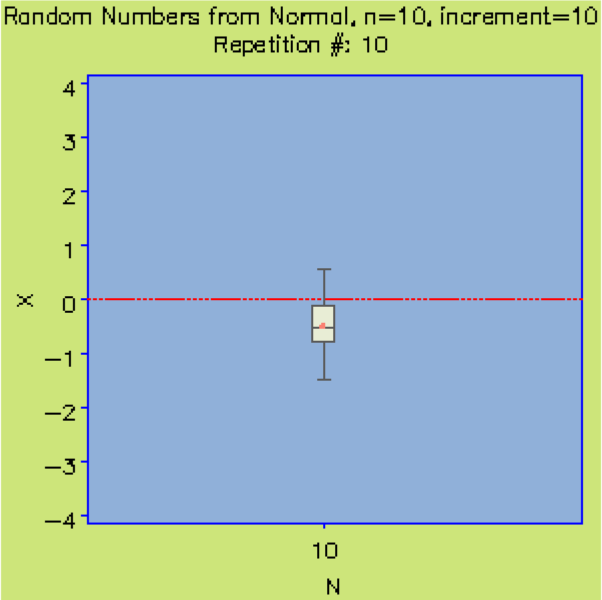


scatter plot

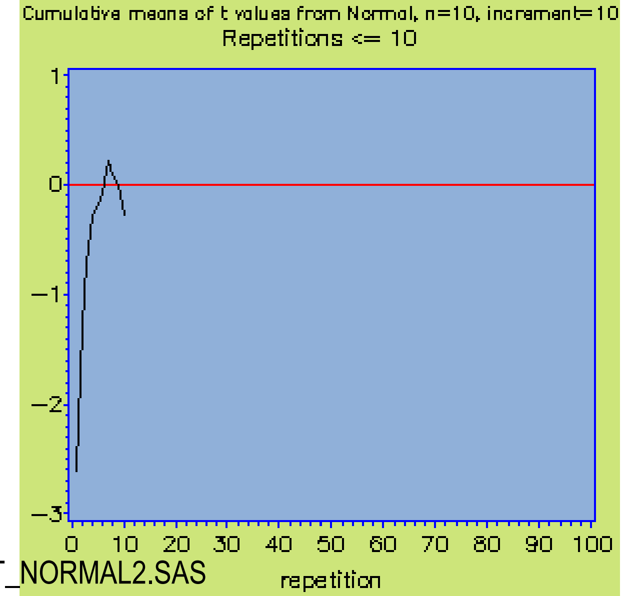


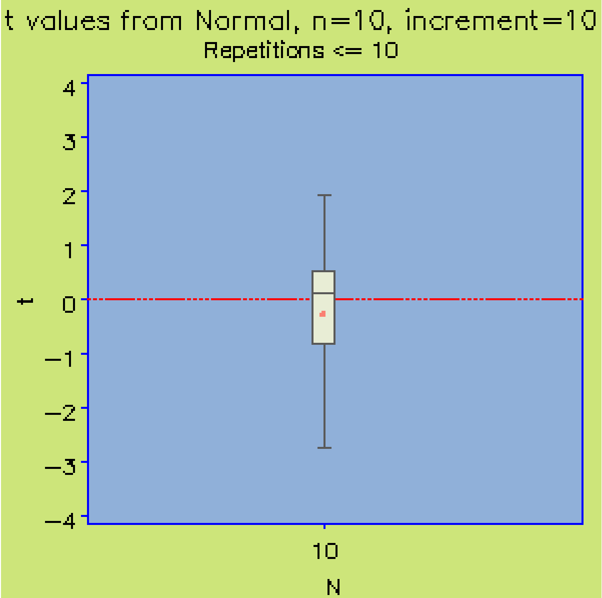
100 random samples of size 10 from N(0,1) (anim)





normal samples, at every 10^{th} repetition

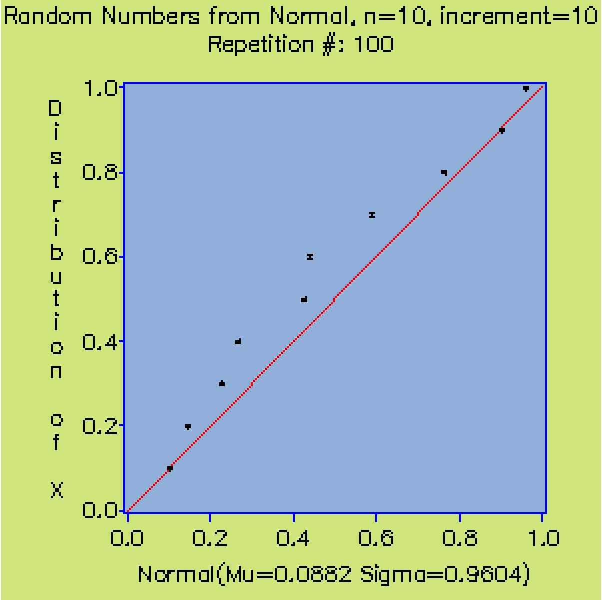


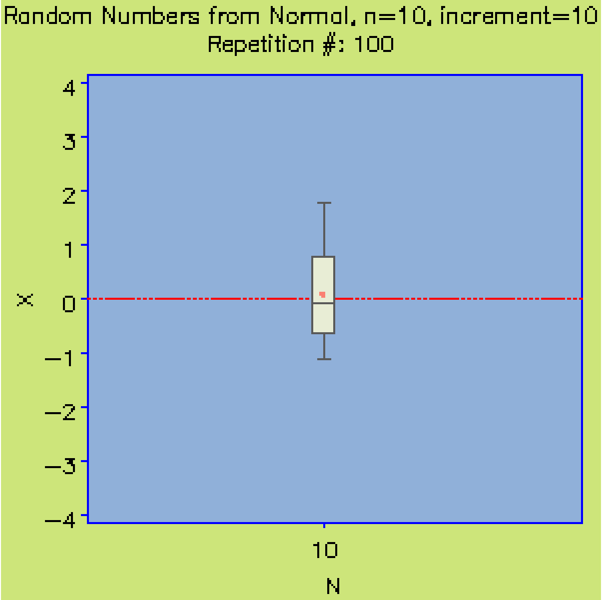


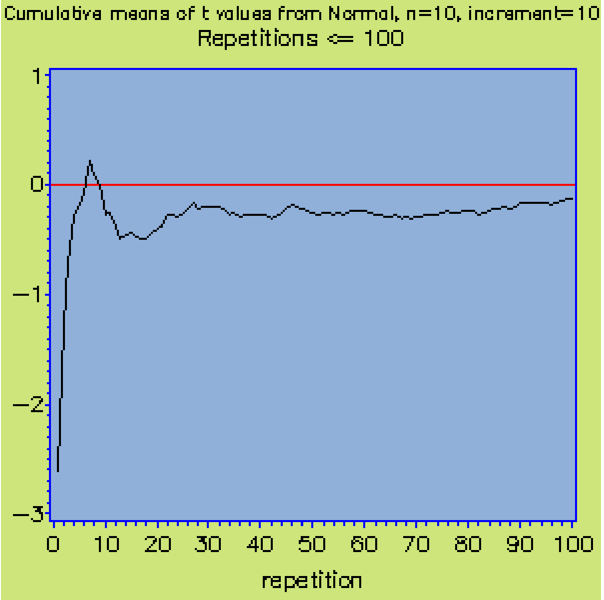
cumulative means and distribution of 100 resulting t values

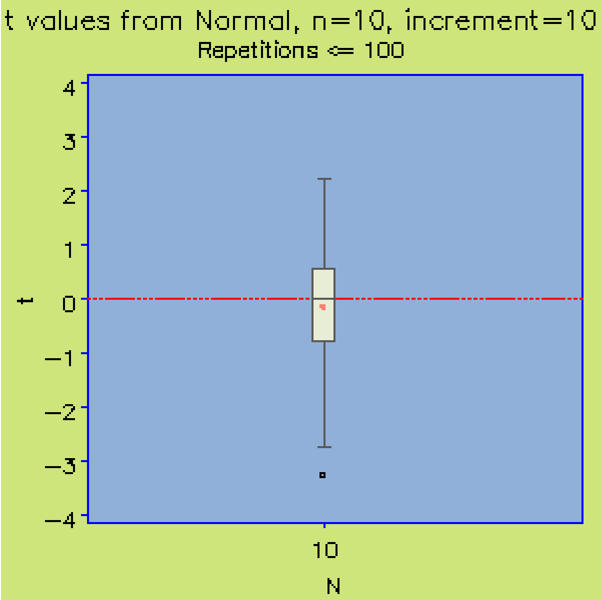
CHAPTER1_2_T_NORMAL2.SAS

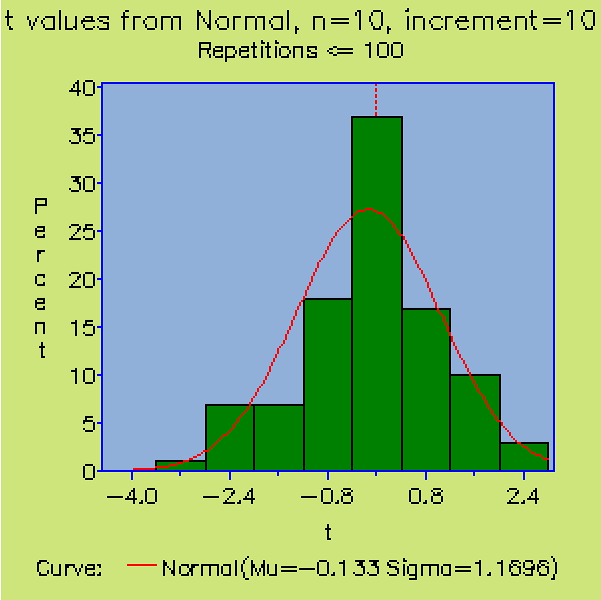
100 random samples of size 10 from N(0,1)



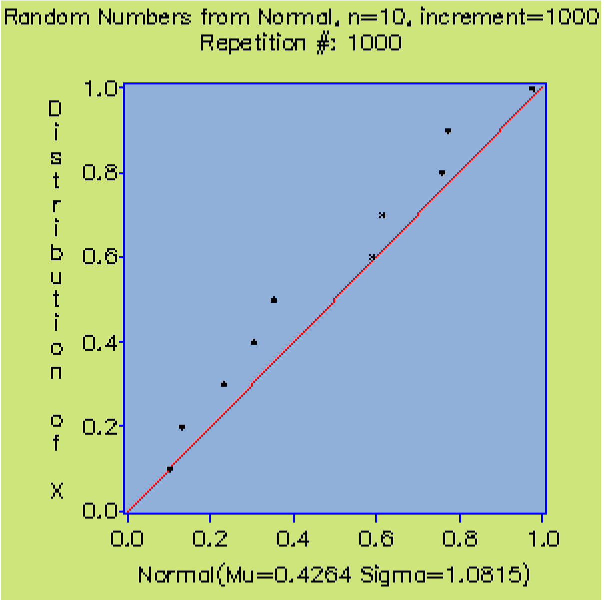


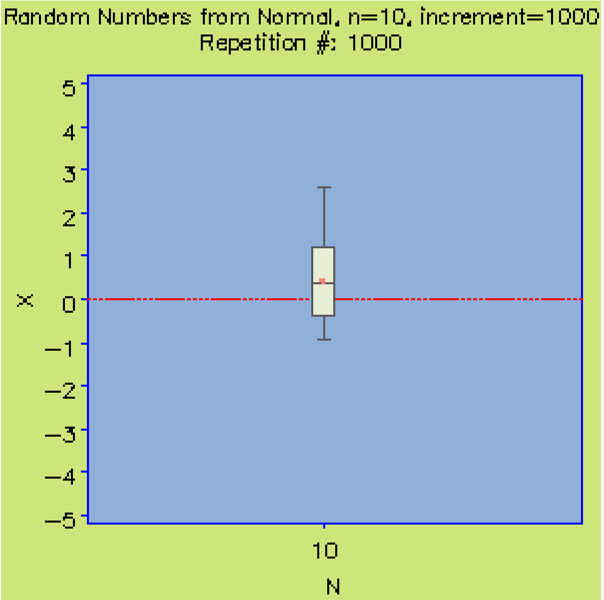




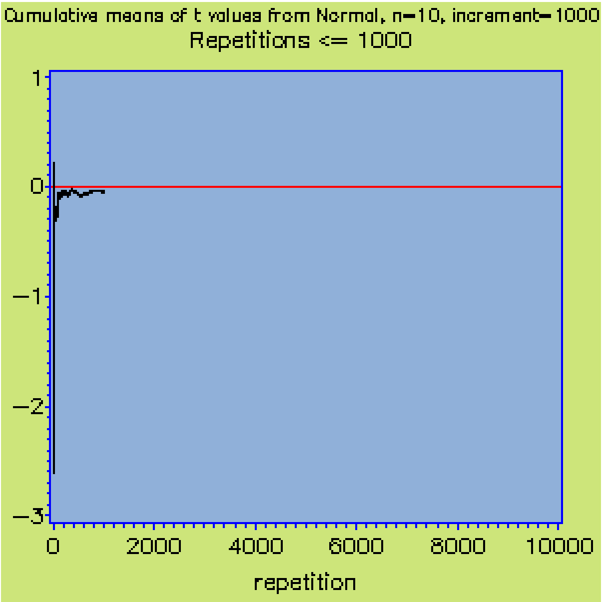


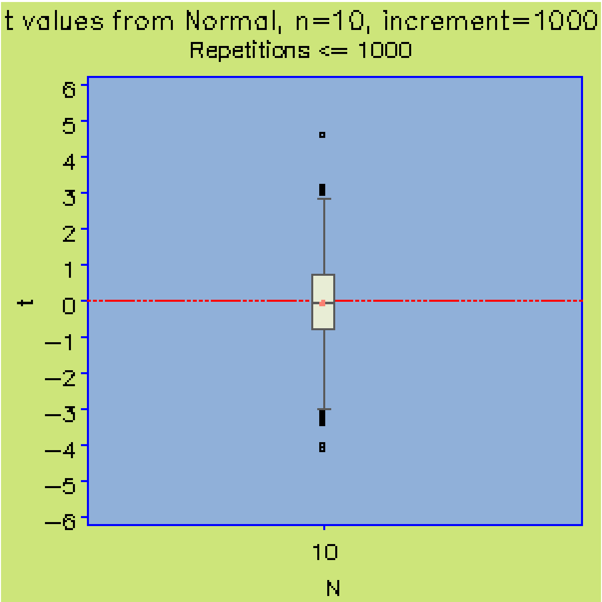
10,000 random samples of size 10 from N(0,1)





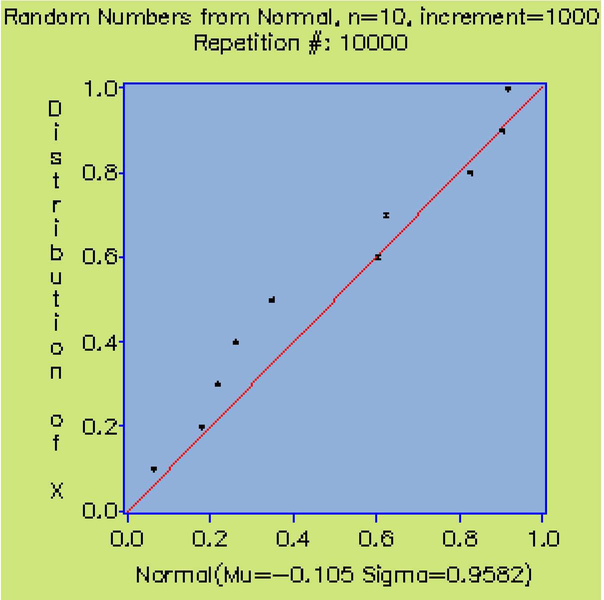
normal samples, at every 1,000th repetition

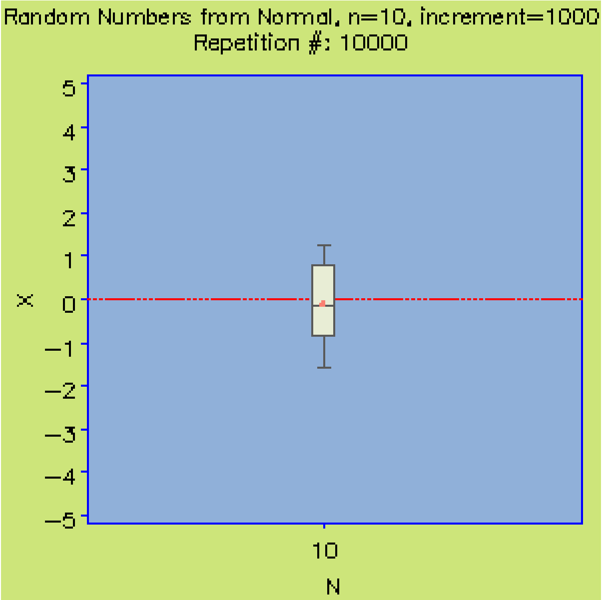


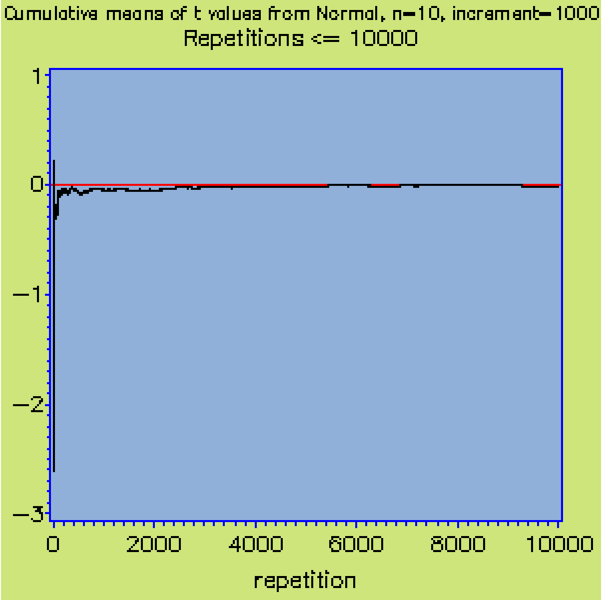


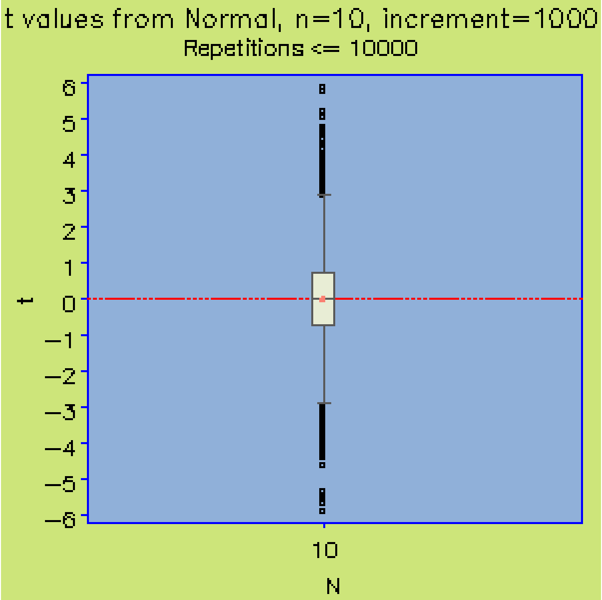
cumulative means and distribution of 10,000 resulting t values

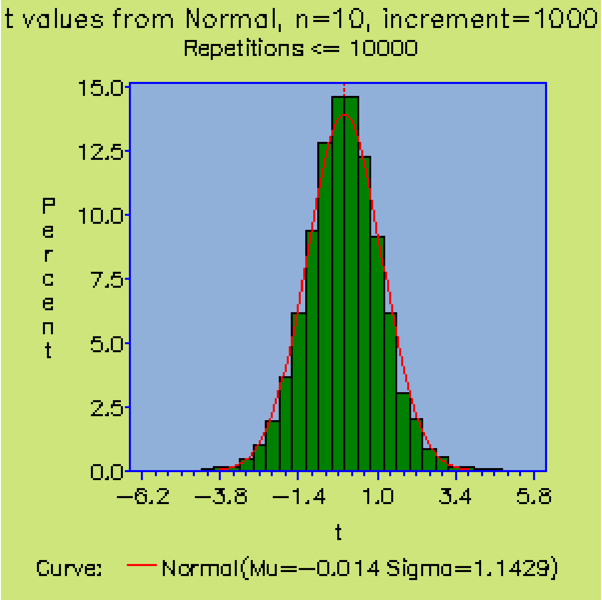
10,000 random samples of size 10 from N(0,1)



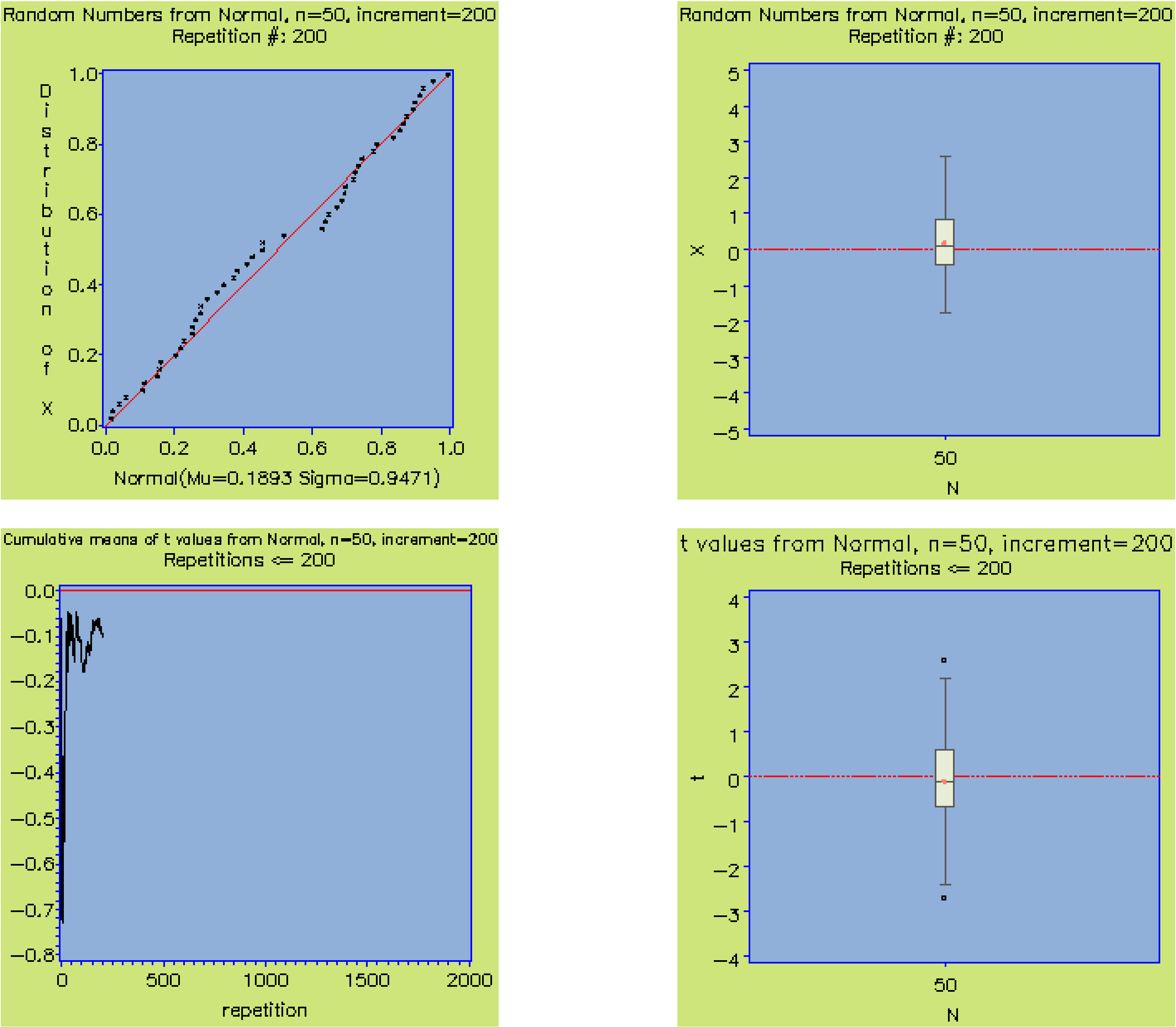




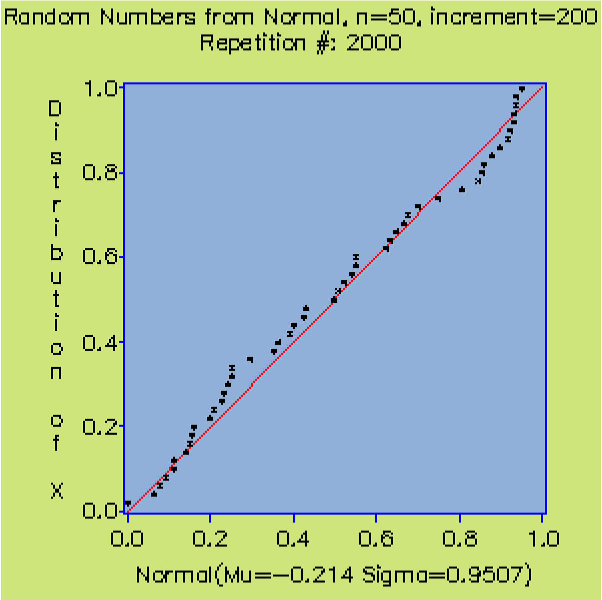


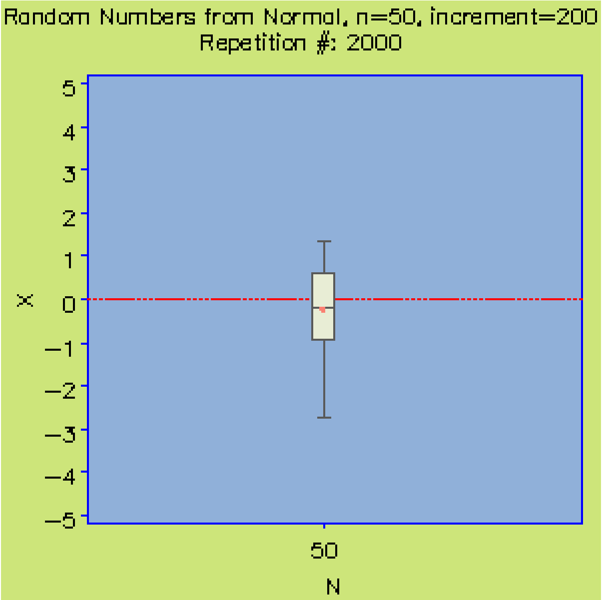


2,000 random samples of size 50 from N(0,1)

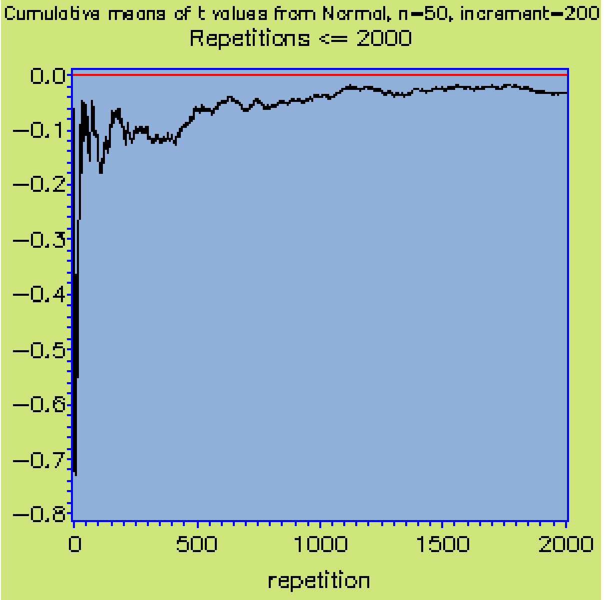


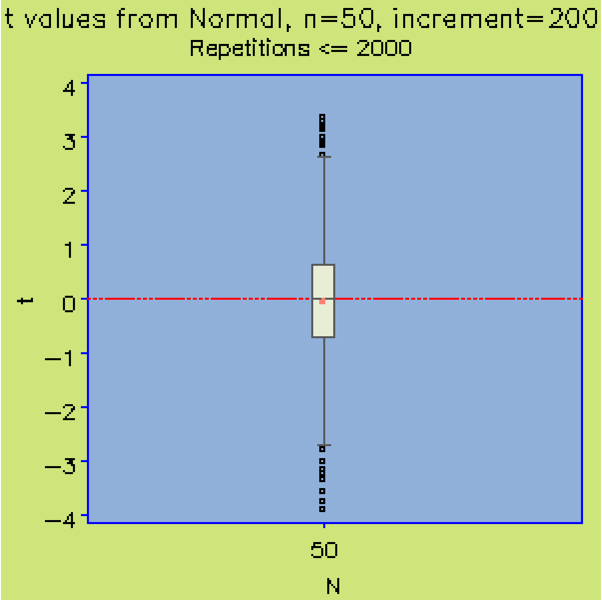
2,000 random samples of size 50 from N(0,1)

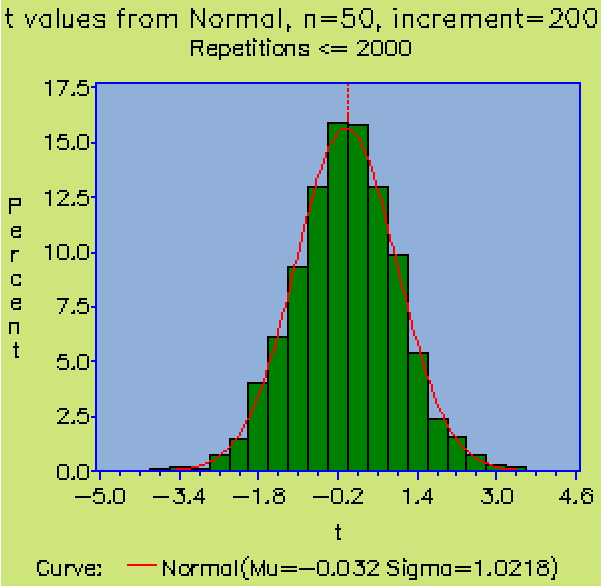




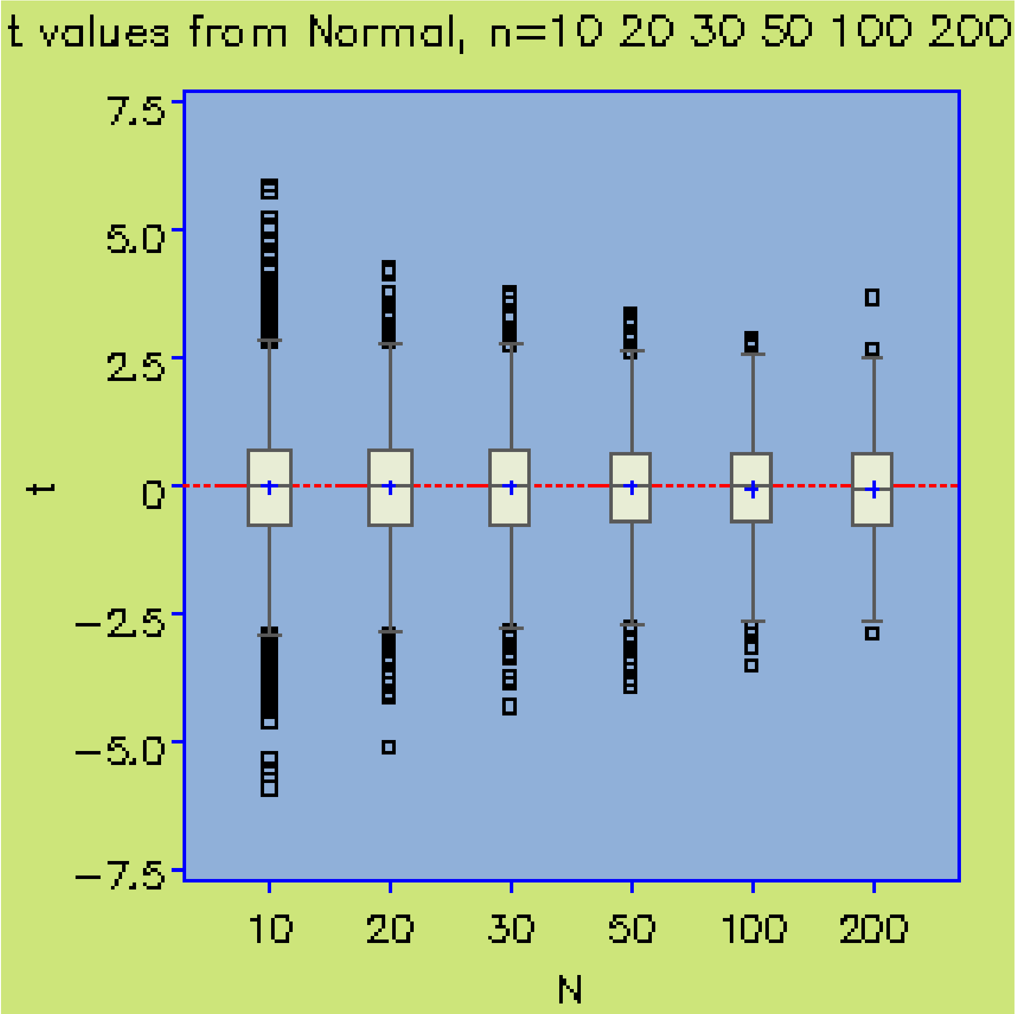
For samples of size n>=40 the distribution of t statistic is close to normal







Distribution of the t statistic from N(0,1) data for various sample sizes



CHAPTER1_2_T_NORMAL3.SAS

Distribution of the t statistic with N(0,1) data for various sample sizes

Analysis Variable : t						
N	**N Obs**	**Mean**	**Std Dev**	**Std Error**	**Skewness**	**Kurtosis**
10	10000	-0.014	1.143	0.011	0.006	0.956
20	5000	-0.021	1.075	0.015	0.000	0.357
30	3333	-0.026	1.046	0.018	0.014	0.144
50	2000	-0.032	1.022	0.023	-0.007	0.226
100	1000	-0.048	1.026	0.032	-0.066	0.040
200	500	-0.070	1.001	0.045	0.075	0.054

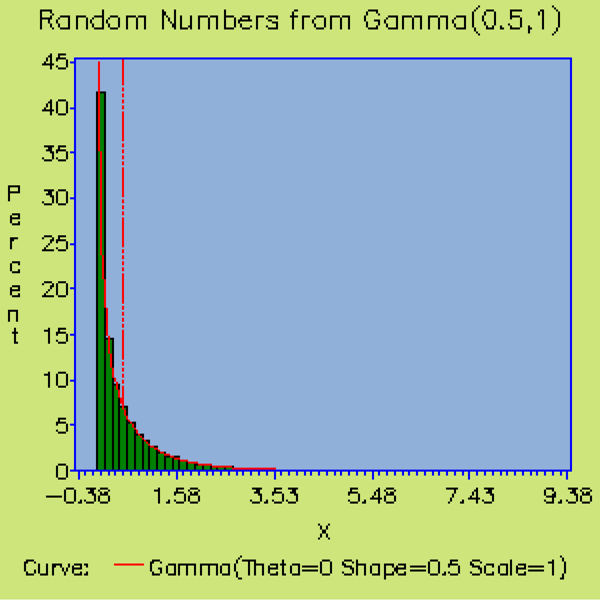
1. super-replication (seed1)

Analysis Variable : t						
N	**N Obs**	**Mean**	**Std Dev**	**Std Error**	**Skewness**	**Kurtosis**
10	10000	0.007	1.134	0.011	-0.019	1.100
20	5000	0.007	1.058	0.015	0.001	0.308
30	3333	0.007	1.043	0.018	0.029	0.335
50	2000	0.009	1.018	0.023	0.001	0.160
100	1000	0.014	1.033	0.033	0.017	-0.183
200	500	0.014	0.997	0.045	0.107	0.013

2. super-replication (seed2)

Sampling from Gamma distribution

Recall some basic properties of the Gamma(k,β) distribution:

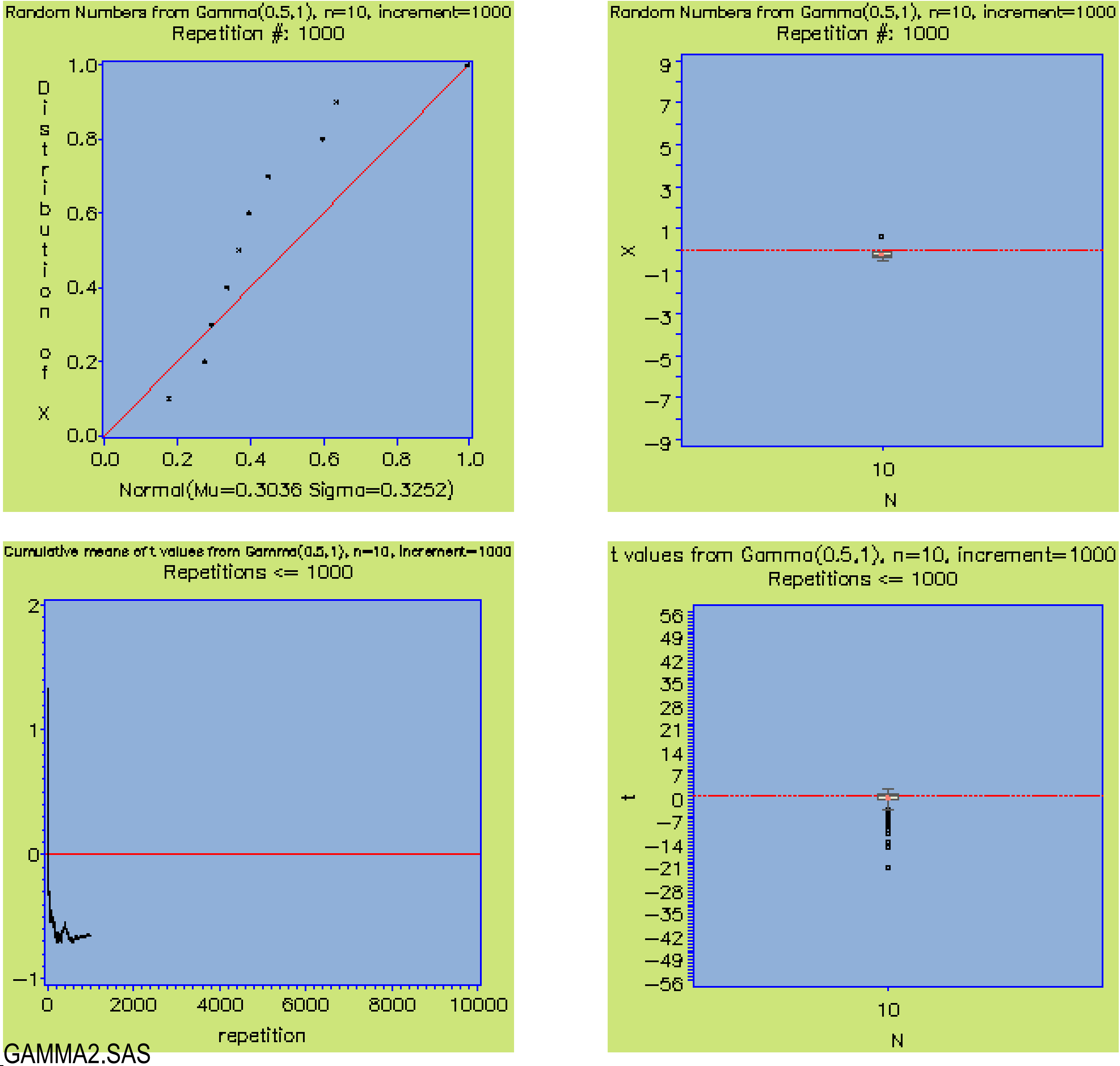


$$f(x) = \frac{\beta^k x^{k-1} e^{-\beta x}}{\Gamma(k)}$$

e.g. k=0.5, β=1:
β=1/τ
τ=scale parameter
k=shape parameter

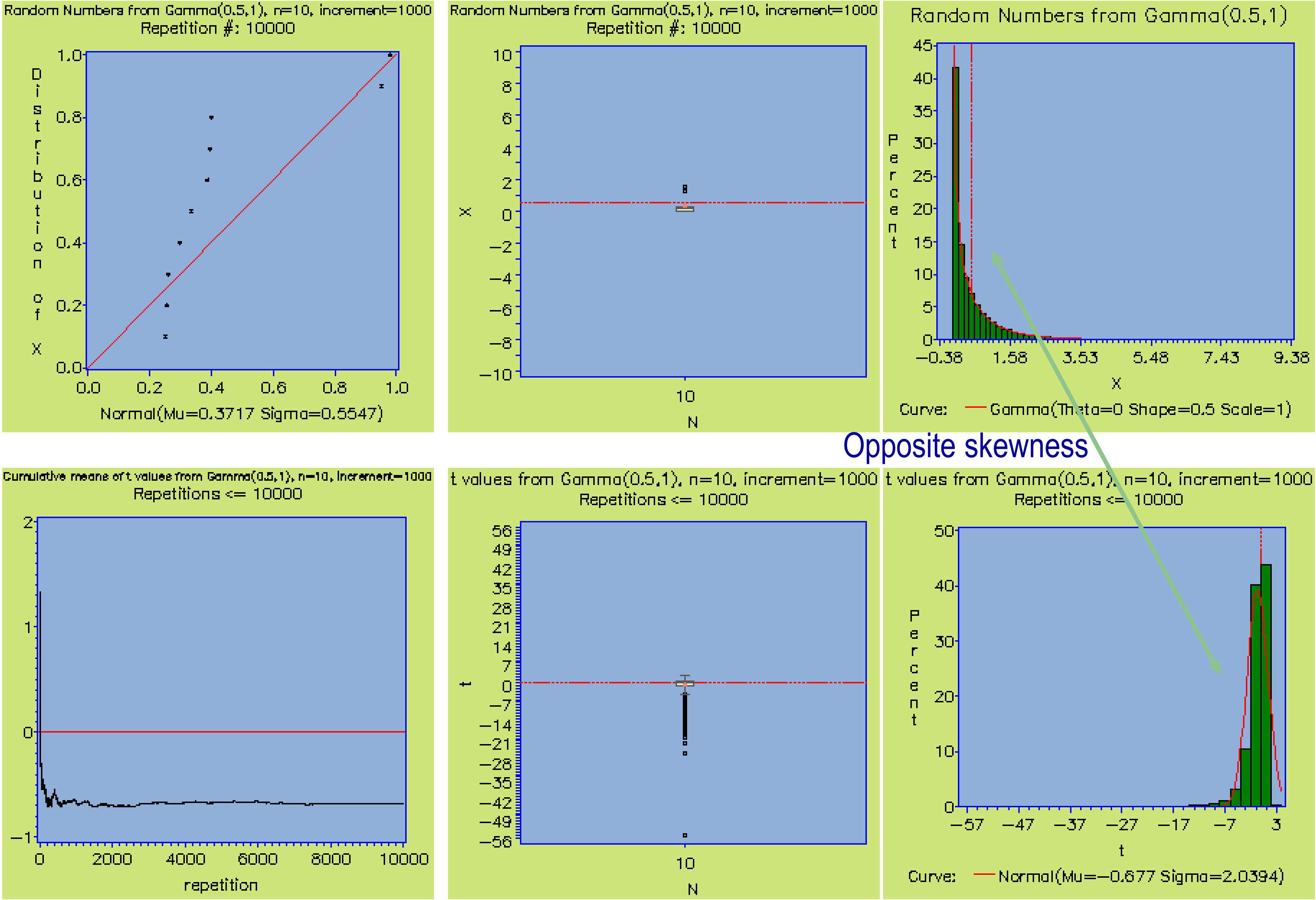
mean	$\mu = k/\beta$	=0.5
st.deviation	$\sigma = k^{1/2}/\beta$	=0.707
skewness	$\gamma_1 = 2/k^{1/2}$	=2.83
kurtosis	$\gamma_2 = 6/k$	=6

10,000 random samples of size 10 from Gamma(0.5,1)



CHAPTER1_2_T_GAMMA2.SAS

10,000 random samples of size 10 from Gamma(0.5,1)



CHAPTER1_2_T_GAMMA2_no_anim.SAS

Some theory: Power series expansions for the moments of t

- $E(t) = -\gamma_1 / (2\,n^{1/2}) + O(n^{-3/2})$,
- $Var(t) = 1 + n^{-1/2}\,(2 + (7/4)\,\gamma_1^{2}) + O(n^{-1/2})$,
- $\gamma_1(t) = -2\,\gamma_1 / n^{1/2} + O(n^{-3/2})$,

where

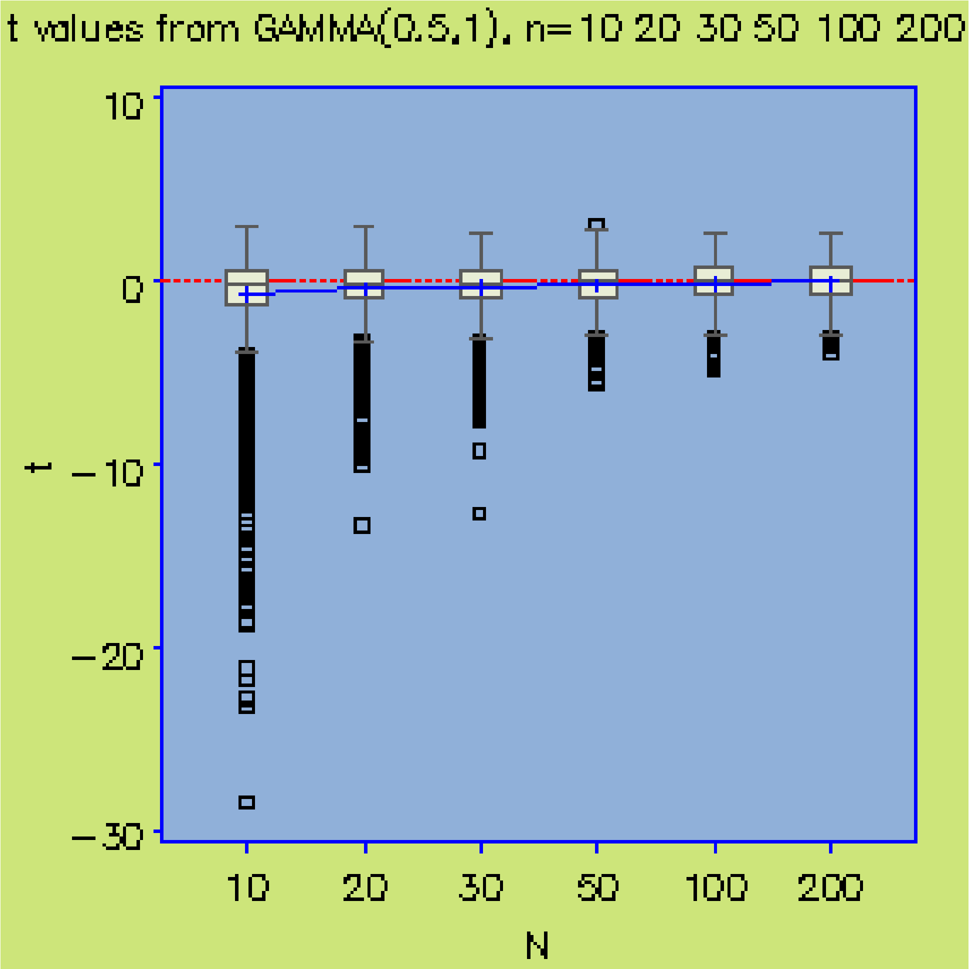
γ_1 is the skewness of the parent population, and

$\gamma_1(t)$ is the skewness of t

The skewness of t is in the opposite direction from the parent population

THE HOUSE OF STATISTICS

Distribution of the t statistic from Gamma(0.5,1) data for various sample sizes



CHAPTER1_2_T_GAMMA3.SAS

Monte Carlo studies of t-test

- MC study by Pearson and Please (1975)
  - Tabulation of **fraction of samples** falling above, below, and outside the appropriate α =0.05 and 0.01 t critical values for various combinations of
  - n = 10, 20, 25,
  - $\gamma_1 = 0$ (0.2) 0.8 (skewness),
  - $\gamma_2 = -1$ to 14 (kurtosis).
- etc.

CAIR CENTER Analytic Services
THE HOUSE OF STATISTICS

Robustness of t statistic (for normal and gamma distributed data)

Example: Tabulation of fraction of samples outside the appropriate critical values

- Calculate the number of samples outside critical values of t for α =0.01, 0.025 and 0.05, i.e., estimate $Pr(t \geq t_{0.01})$, $Pr(t \leq -t_{0.01})$, $Pr(t \geq t_{0.025})$, $Pr(t \leq -t_{0.025})$, $Pr(t \geq t_{0.05})$, $Pr(t \leq -t_{0.05})$.
- How do these fractions (estimated (MC) probabilities of Type I error) compare to the corresponding α values?

CHAPTER1_2_T_GAMMA3_FRACTION_CRITICAL_VALUES.SAS

Fraction of samples outside the appropriate critical values

Normal parent population

N	fraction_crit_01_left	fraction_crit_01_right	fraction_crit_025_left	fraction_crit_025_right	fraction_crit_05_left	fraction_crit_05_right
10	0.009	0.011	0.024	0.025	0.048	0.050
20	0.011	0.009	0.025	0.027	0.047	0.050
30	0.009	0.009	0.024	0.025	0.050	0.050
50	0.010	0.009	0.026	0.024	0.051	0.049
100	0.013	0.009	0.027	0.025	0.053	0.049
200	0.012	0.011	0.026	0.024	0.052	0.051

Gamma parent population

N	fraction_crit_01_left	fraction_crit_01_right	fraction_crit_025_left	fraction_crit_025_right	fraction_crit_05_left	fraction_crit_05_right
10	0.098	0.001	0.137	0.002	0.176	0.009
16	0.083	0.001	0.115	0.003	0.152	0.013
20	0.072	0.001	0.105	0.004	0.143	0.014
30	0.059	0.001	0.088	0.006	0.121	0.017
50	0.047	0.001	0.073	0.007	0.104	0.021
100	0.029	0.003	0.051	0.009	0.083	0.028
200	0.022	0.004	0.041	0.014	0.069	0.035

The Power of a Regression Test H_0:β_i=0

- Linear regression model:
- $y = \beta_0 + \beta_1 x + \beta_2 z + e$

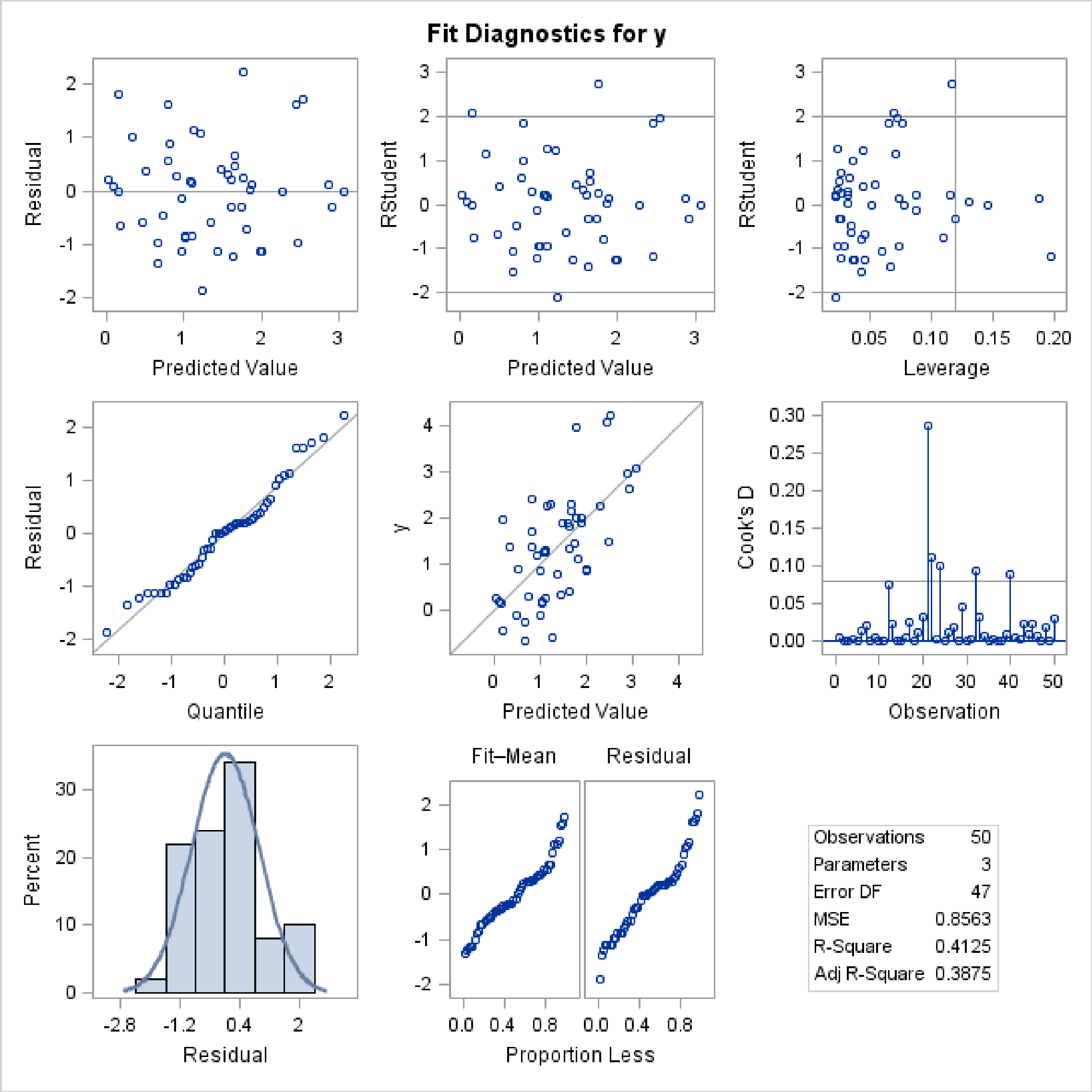
where

- $x \sim N(0,1)$, $z \sim N(0,1)$, corr(x,z)=0,
- Sample size n=50
- $\beta_0 = \beta_1 = 1$, β_2= 0 to 1 by 0.1
- Model 1: $e \sim N(0,1)$
- Model 2: $e \sim t(5)$
- Model 3: $e \sim N(0, \exp(x/5))$ (heteroscedastic)
- Number of samples=1000

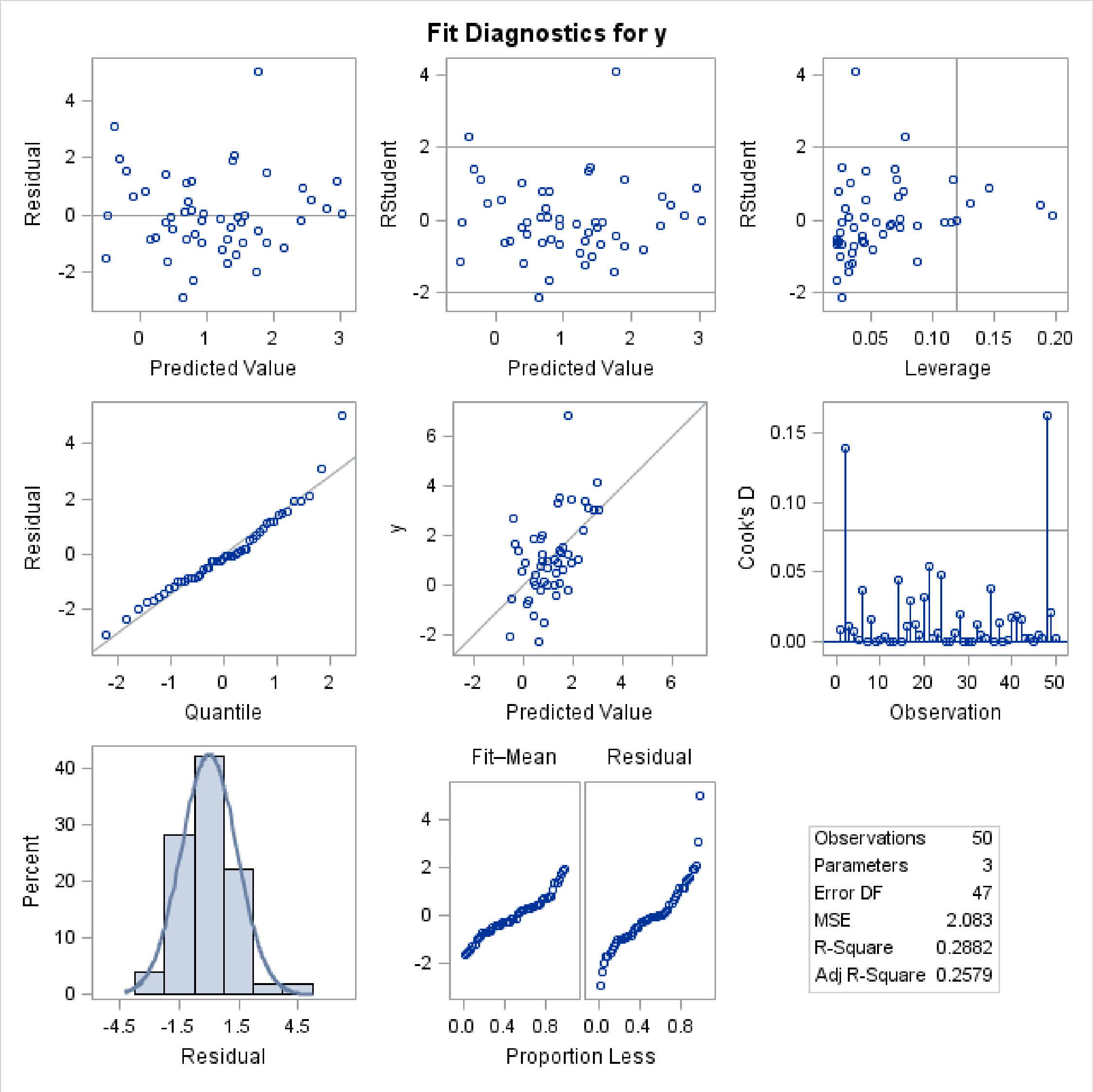
<u>Example</u>:
1. Simulate data from a bivariate lin.reg.model $y = \beta_0 + \beta_1 x + \beta_2 z + e$,
2. Calculate rej.indicator =(ProbF≤0.05) for H0: β_2 =0
3. Repeat 1000 times for β_2 ranging from 0 to 1 and for 3 error disns, summarize – calculate proportion of rejections

CAIR CENTER Analytic Services
THE HOUSE OF STATISTICS

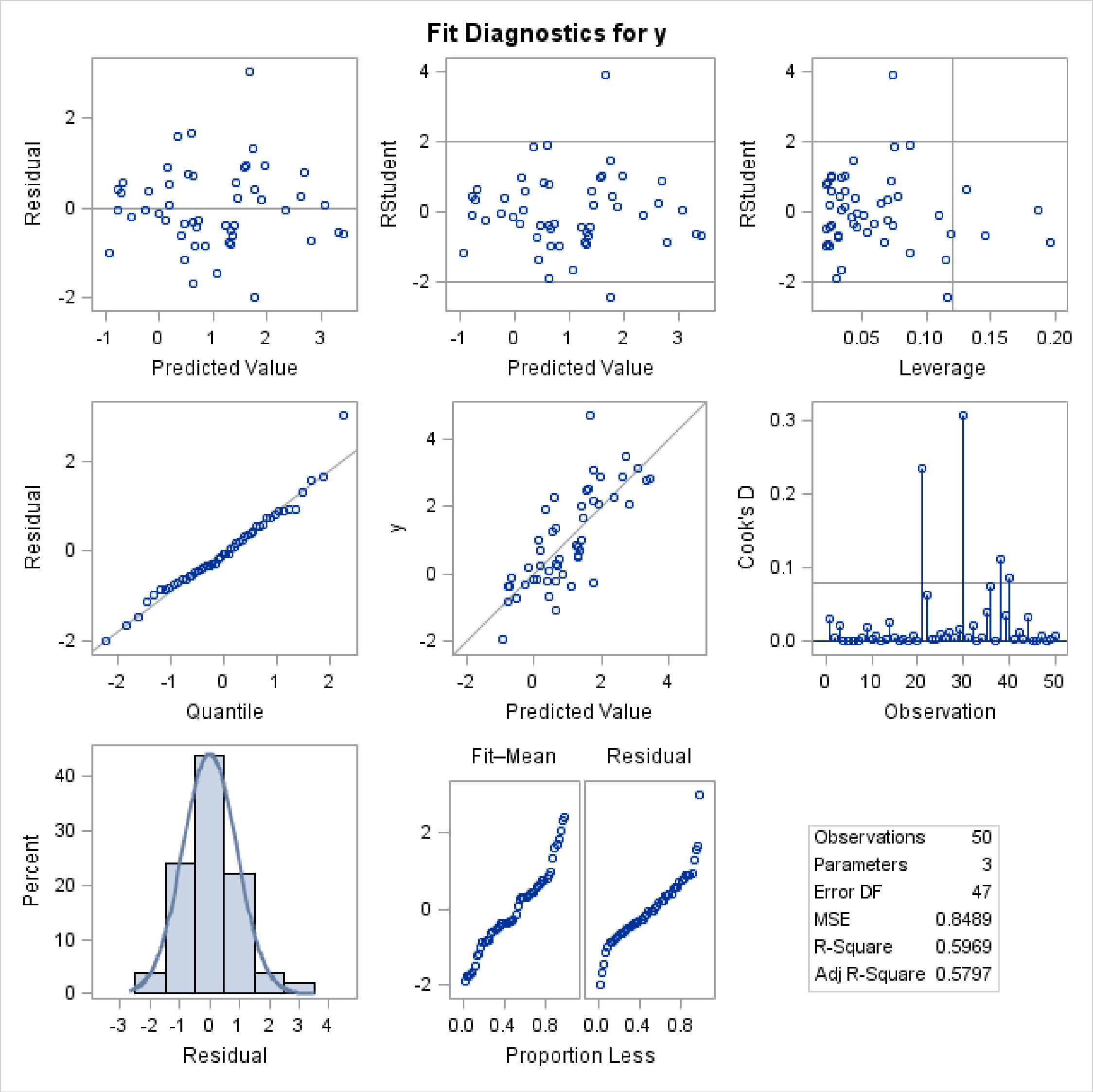
Diagnostic plot for Model1 : e ~ N(0,1)



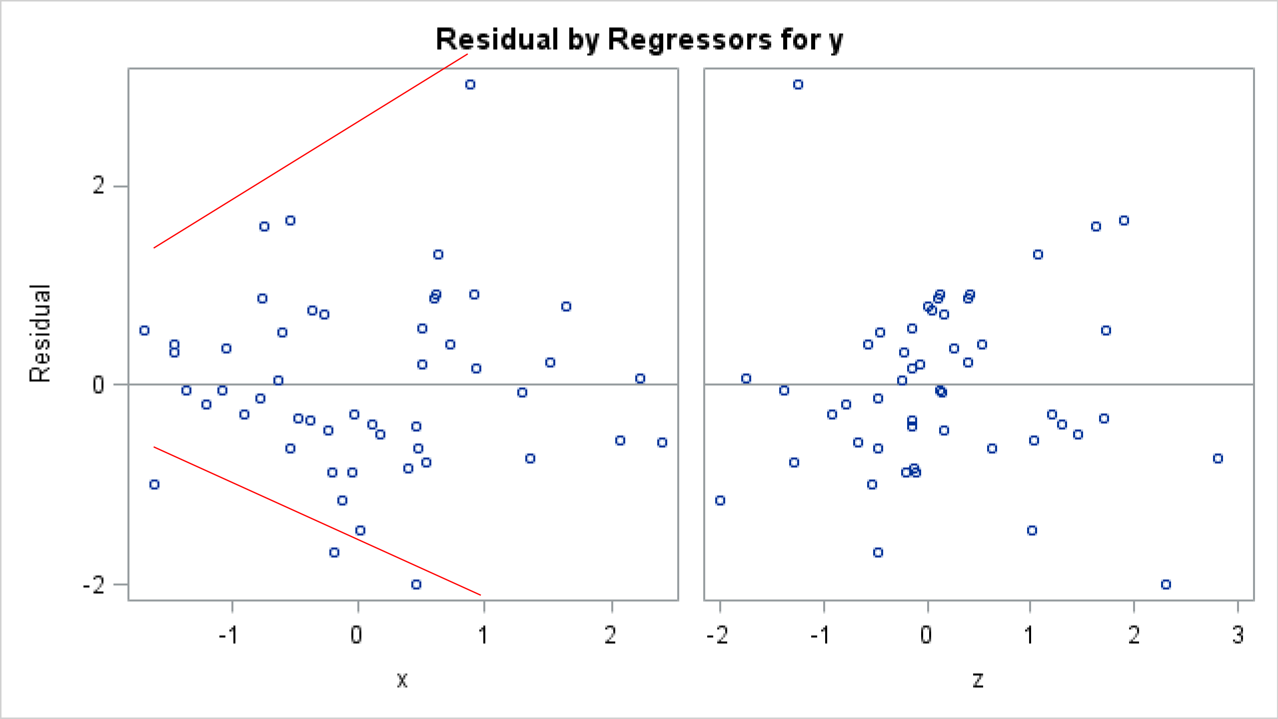
Diagnostic plot for Model1 : e ~ t(5)



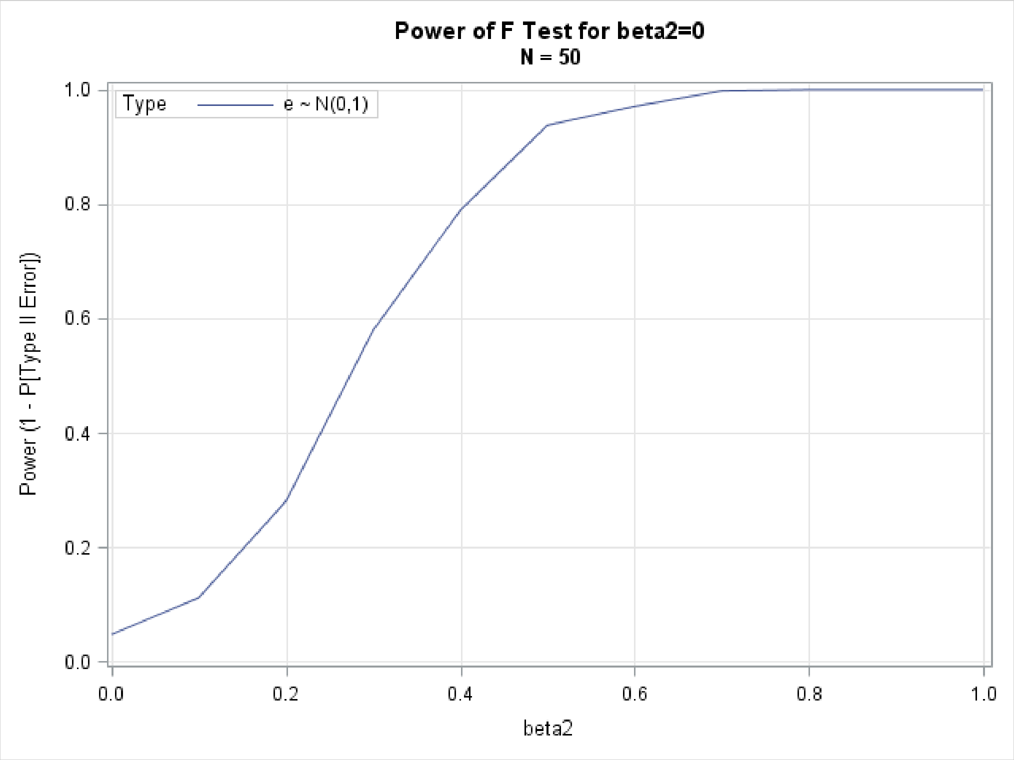
Diagnostic plot for Model1 : e ~ N(0, exp(x/5)) (heteroscedastic)

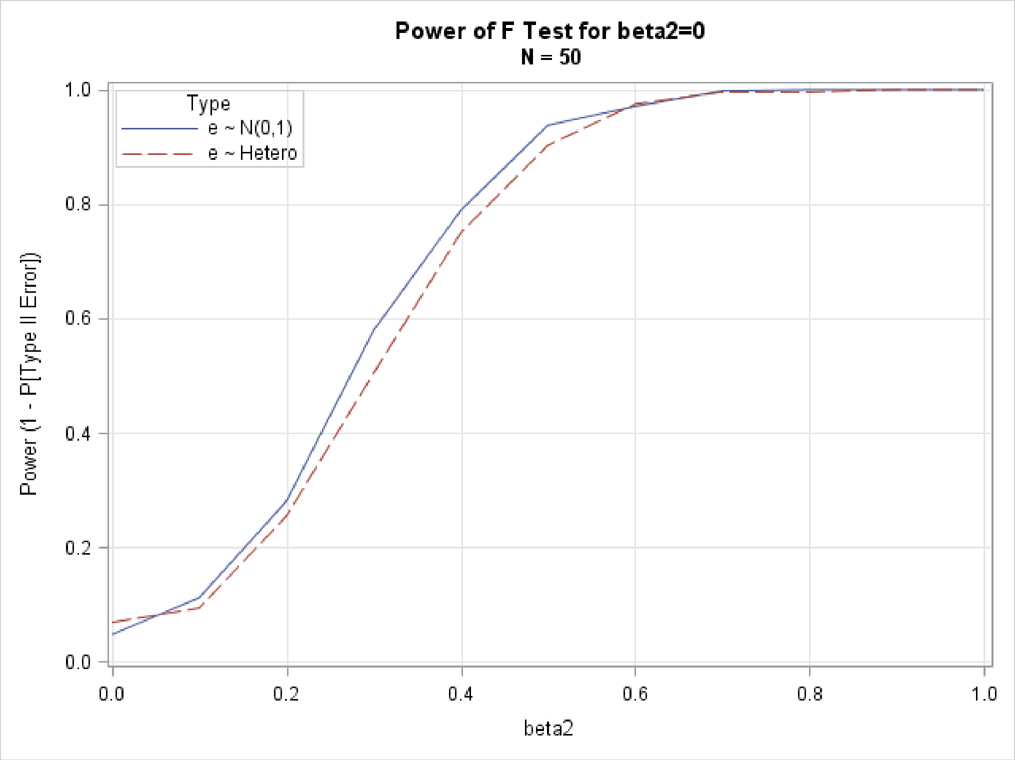


Diagnostic plot for Model1 : e ~ N(0, exp(x/5)) (heteroscedastic)



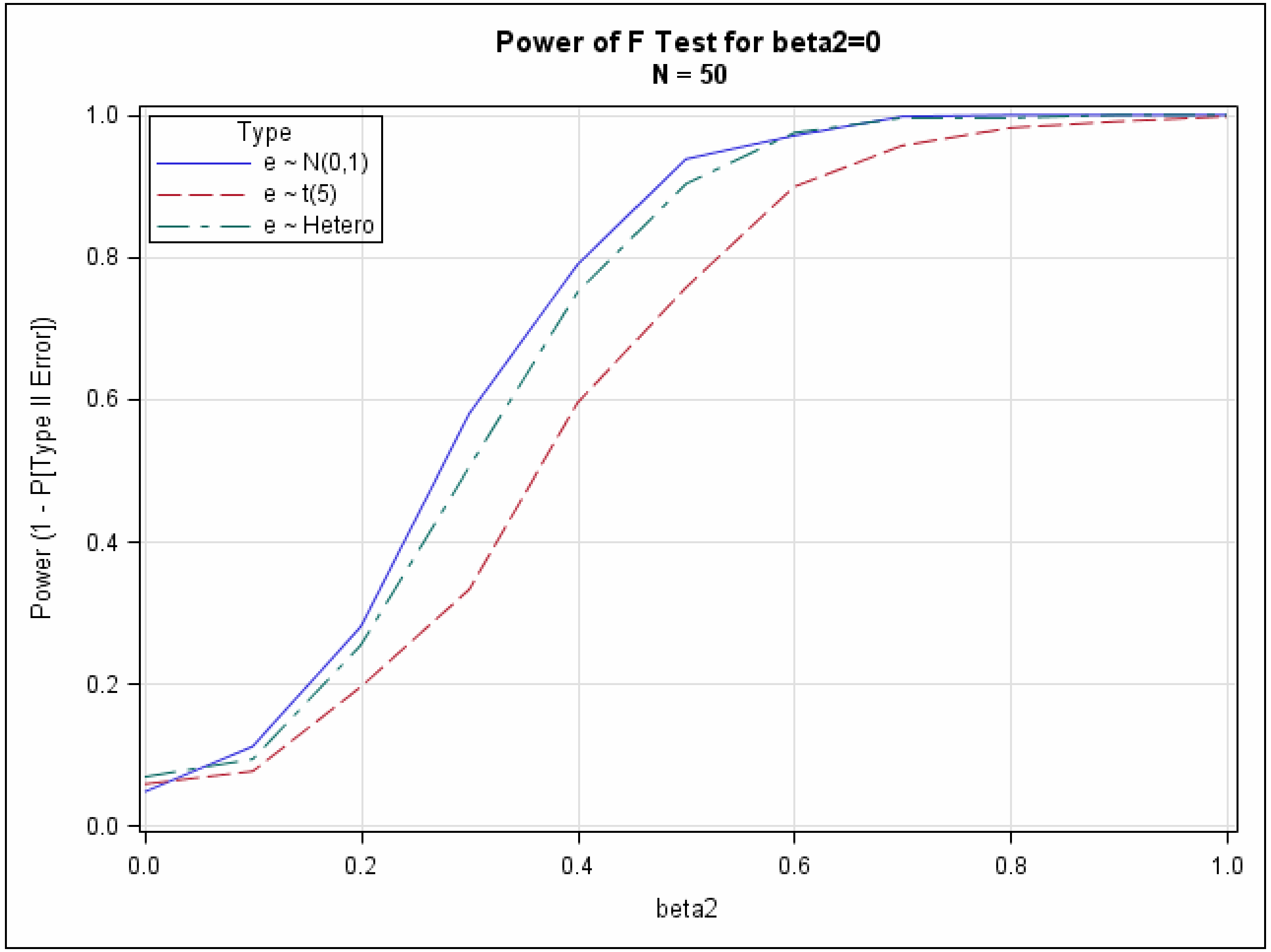
The Power of a Regression Test





CAIR CENTER Analytic Services
THE HOUSE OF STATISTICS

The Power of a Regression Test H_0:β_2=0



Proportion of samples where H_0 is rejected (at α=0.05 i.e., when Pr(F)$\leq$0.05)

Coverage probability – student project

- 90% and 95% confidence interval for the mean
- Data from
  - Normal
  - Laplace
  - Gamma
  - Weibull
  - Uniform
- Sample size n = 5, 10, 15, 20, 40, 80, 100
- Coverage probability = proportion of cases when the estimated confidence interval contains the actual expected value

Projekt_8_prezentacija.pdf

Conclusions

- There are ways of making decisions when the answer is „not in the back of the book".
- There are ways of examining the validity of a statistical test (when conditions for the test are not fulfilled)
- Use MC experiments to evaluate and compare algorithms/ methods/ models under various conditions
- When in doubt, apply computational statistics

Computer simulation has become, alongside experimentation and abstract reasoning, the third major tool of science.