MATHEMATICAL COMPUTING IN STATISTICS

FRED TINSLEY (WITH STEVEN JANKE)

JANUARY 18, 2008

2:40 PM

Personal Background

- MS in Statistics; PhD in Mathematics
- 30 years teaching at liberal arts college
  - Service course: Intro
  - Math major courses (statistics track)
    - Linear models; math statistics
- 30 years assisting others inside and outside (mostly legal) of the academy

Colorado College Block Plan

- All intensive study: Course $\Leftrightarrow$ 3.5 Weeks
- Fall and spring semesters: 4 blocks each
- Winter half block
- Students take and faculty teach one block at a time
- **Pedagogical note for upper division courses in math**: student ‘projects’ (broadly defined) are nearly essential.

Backdrop

- **Math Probability → Math Statistics** satisfies the longitudinal sequence requirement for the math major at CC. As such, it enjoys a central role, especially for those following the statistics track.
- **Applied Statistics Degree Programs:** Students view Math Stats as a 'nasty hurdle'.
- **Trends in Teaching Statistics:** Remove all mathematics including probability from lower- and mid-level courses. ***Remove mathematicians from the statistics classroom.***

(Being somewhat of a hybrid, I have refused to leave!)

Natural Question: What does this debate have to say about the content and role of mathematical statistics in the statistics track of the mathematics major at a liberal arts college?

Traditional Role at CC

- **Wishful thinking**: unifying framework
- Prepare students for graduate school/work in statistics
- Application of Probability
- Need for mathematical machinery
- ‘Culminating experience’ for majors
  - Calc, linear alg, comb, prob, complex

Role of Statistical Computing

- Traditional
  - Data Analysis
  - Monte Carlo, Bootstrap, …
- More recent
  - Spreadsheets
- Modern Mathematical ✞
  - Symbolic parametric analysis
  - Density and distribution functions
  - Order statistics: min, median, max, etc
  - Moment generating and characteristic functions

Mathematical Computing

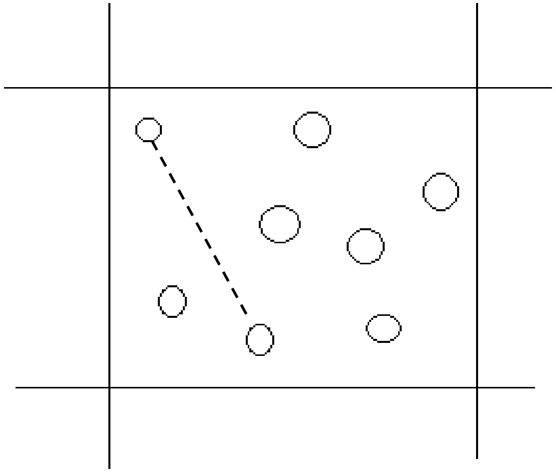
- Derive, Maple, Matlab, Mathematica, etc
  - Helpful to learning and teaching?
  - Applications to data analysis?✞
- Software written in these
  - $$$ (e.g., Mathstatica)
  - Usual plethora on web

Examples

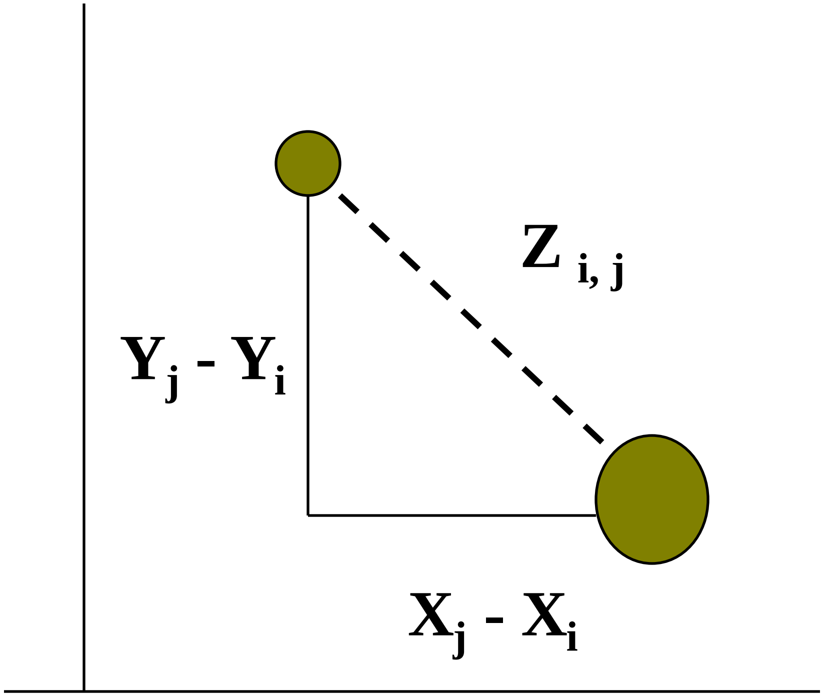
1. Modeling spread of forest fires
   – Student independent study project (advanced)
2. The binomial distribution and mutations
   – Student project for introductory course

1. Modeling Forest Fires

Uniformly distributed trees



Distribution of Distances Between Trees

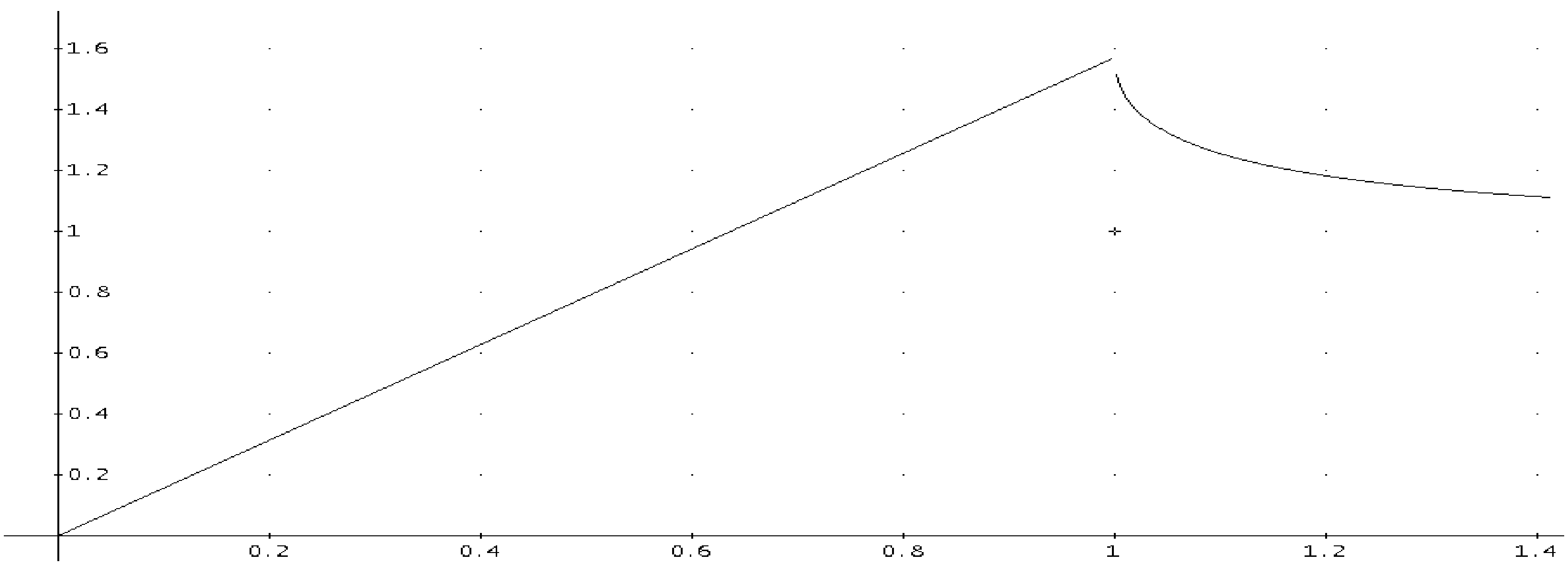


Distributions of $Z_{i, j}$

- X_i′s and Y_i′s are independent (uniform)
- Dist of $Z_{i,j} = ((X_i-X_j)^2 + (Y_i-Y_j)^2)^{½}$ can be computed by distribution function technique
- These data are one component of a model for the spread of forest fires
- Compute distribution of $Z_{i,j}$ from those of X and Y

Technique of Distribution Functions

$$\frac{d}{dz}\int_0^z \int_0^{\sqrt{z^2 - t^2}} f(s)\cdot g(t)\, ds\, dt$$



2. Everything is Binomial!

- Each measurement falls into one and only one category (2 categories)
- Sequence of n such measurements
- X = number of measurements within the preferred category.
- All is well if the **observations are independent** and **each observation has the same chance falling into a preferred category**

Intro to Stat Project

- Goal: Compare panel of CC students to a panel of experts rating motion pictures
- Questionnaire: Adapted actual questions to a multiple choice format
  - 17 questions about films
    - 16 had four choices for answers
    - 1 had three choices
  - Students: treat each questionnaire as a binomial with success a ‘match’

Two Mutations of the Binomial

1. Probability of a success (match) changes
   – Number of choices changes
2. Trials may be correlated
   – Same student answers questions in the same order.

Mutation 1: Sum of Two Binomials

- $X = Y + Z$
  1. $Y \sim \text{Binomial}(r,p)$
  2. $Z \sim \text{Binomial}(s,q)$
  3. $n = r + s$ is the sample size
- Generalization: Can have a sum of up to n distinct binomial random variables

Independence
(a priori equals conditional)

- X and Y with densities are ***independent*** random variables if $f(x,y) = f_X(x)f_Y(y)$.
  - $f(x,y)$ is the ***joint*** density with domain $\mathbb{R}^2$.
  - Probability is represented by volume.
- If X_1 and X_2 have the same density, then we say they are ***identically distributed***.
- $X_1, \ldots, X_n$ is a ***random sample*** if they are independent (joint density factors into n factors) and identically distributed.

Moment Generating Functions

- Moments

  $\mu_r = E(X^r) = r^{th}$ moment about 0.

  $\mu_r' = E(X - \mu)^r = r^{th}$ moment about μ.

  $\mu_1 = E(X); \mu_2' = E(X - \mu)^2 = Var(X)$

- Moment generating function (introduced in part to 'prove' the Central Limit Theorem

  $M_X(t) = E(e^{tX}) = \int e^{tX} f(x)dx$ (continuous)

  $M_X(t) = E(e^{tX}) = \sum e^{tX} f(x)dx$ (discrete)

  $M_X^{(r)}(0) = ?$

Moment Generating Functions of Linear Combinations

- $M_{X+Y}(t) = E(e^{t(X+Y)}) = E(e^{tX}e^{tY})$
- So, $M_{X+Y}(t) = \int e^{tX}e^{tY}f(x,y)dxdy$

  $= \int e^{tX}e^{tY}f(x)f(y)dxdy$ by independence☺

  $= (\int e^{tX}f(x)\,dx)(\int e^{tY}\,f(y)\,dy) = M_X(t)\,M_Y(t)$
- Thus, if $X_1, \ldots, X_n$ are iid, then

  $M_{\sum Xi}(t) = (M_X(t))^n$
- $M_{cX}(t) = E(e^{t(cX)}) = E(e^{(ct)X}) = M_X(ct)$

Binomial(n,p) ~ X

- Bernoulli trial: X = 0 or 1
- n independent trials
- $P(X_i = 1) = p$; $P(X_i = 0) = 1\text{-}p$
- $X = \sum X_i$ = # of 1's
- $M_{X_i}(t) = (1\text{-}p)\cdot e^{t\cdot 0} + p\cdot e^{t\cdot 1}$
- $M_X(t) = ((1\text{-}p) + p\cdot e^{t\cdot 1})^n$

Characteristic Function

- Problem: Most rv's do not have a MGF.
  - Difficulty: e^{tx} must be integrated from $-\infty$ to ∞ with respect to x.
- Question: Can the density be recovered from the MGF?
- Solution: $\chi_X(t) = E(e^{itx}) = M_X(i \cdot t)$ (complex Fourier transform)

$$(F(b) - F(a)) = \frac{\lim_{n \to \infty} RE\left(\int_{-n}^{n} \frac{e^{-i \cdot a \cdot s} - e^{-i \cdot b \cdot s}}{i \cdot s} \cdot \chi \, ds\right)}{2 \cdot \pi}$$

Mixture of Binomials

- Y~Binomial(m,p); Z ~ Binomial(n,q)
- Y and Z independent
- What is the distribution of Y+Z?
- What is the Characteristic function for Y+Z?

$$\left(p\cdot e^{i\cdot t}+1-p\right)^{m}\cdot\left(q\cdot e^{i\cdot t}+1-q\right)^{n}$$

Distribution of Y + Z

- Invert the characteristic function
- RE( ) is necessary because of underflow

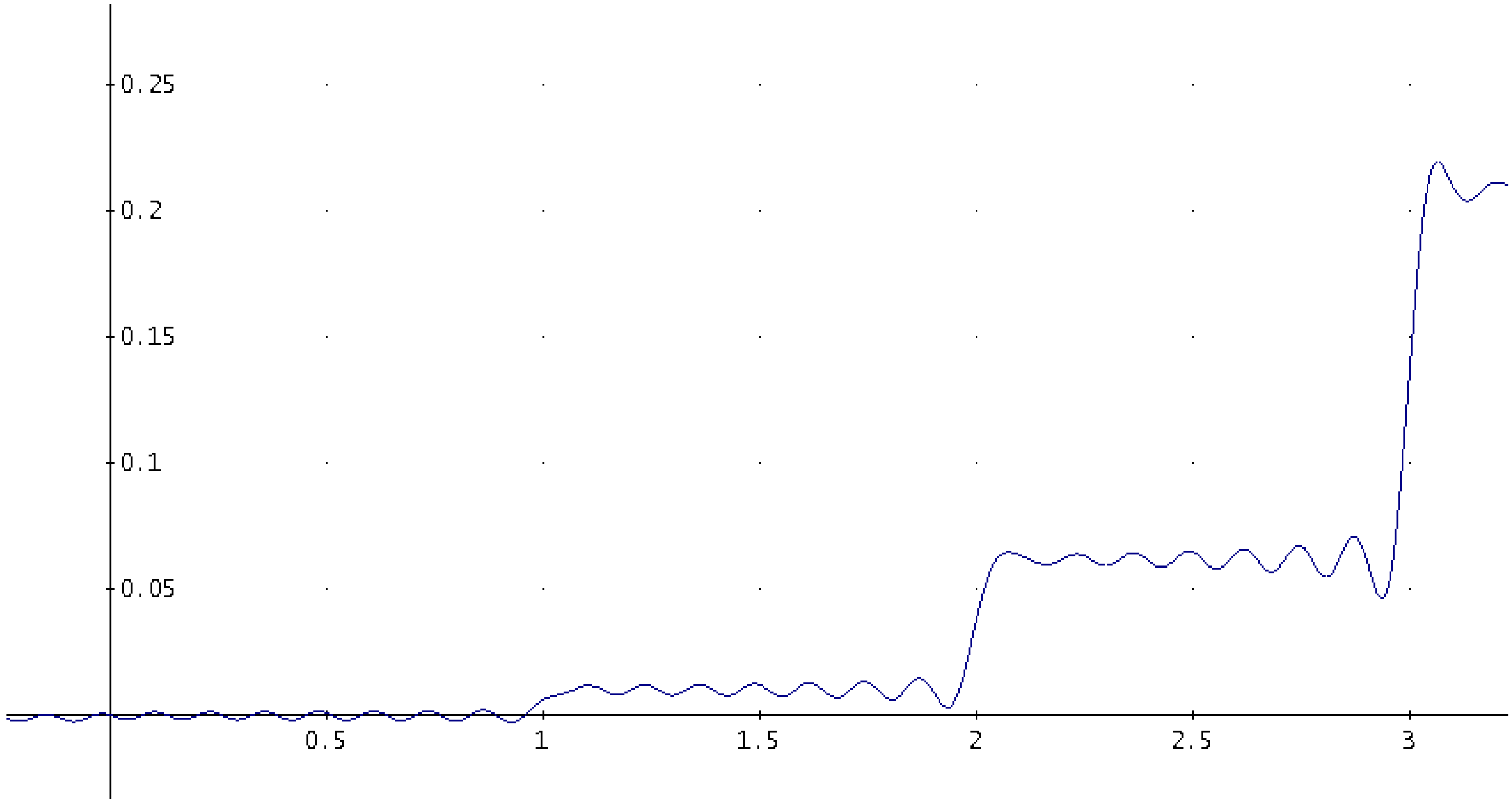
$$\lim_{c \to \infty} \frac{\mathrm{RE}\left(\displaystyle\int_{-c}^{c} \frac{e^{-i \cdot a \cdot s} - e^{-i \cdot b \cdot s}}{i \cdot s} \cdot \chi \, ds\right)}{2 \cdot \pi}$$

$m = 5$, $p=0.5$; $n = 3$, $q = .7$

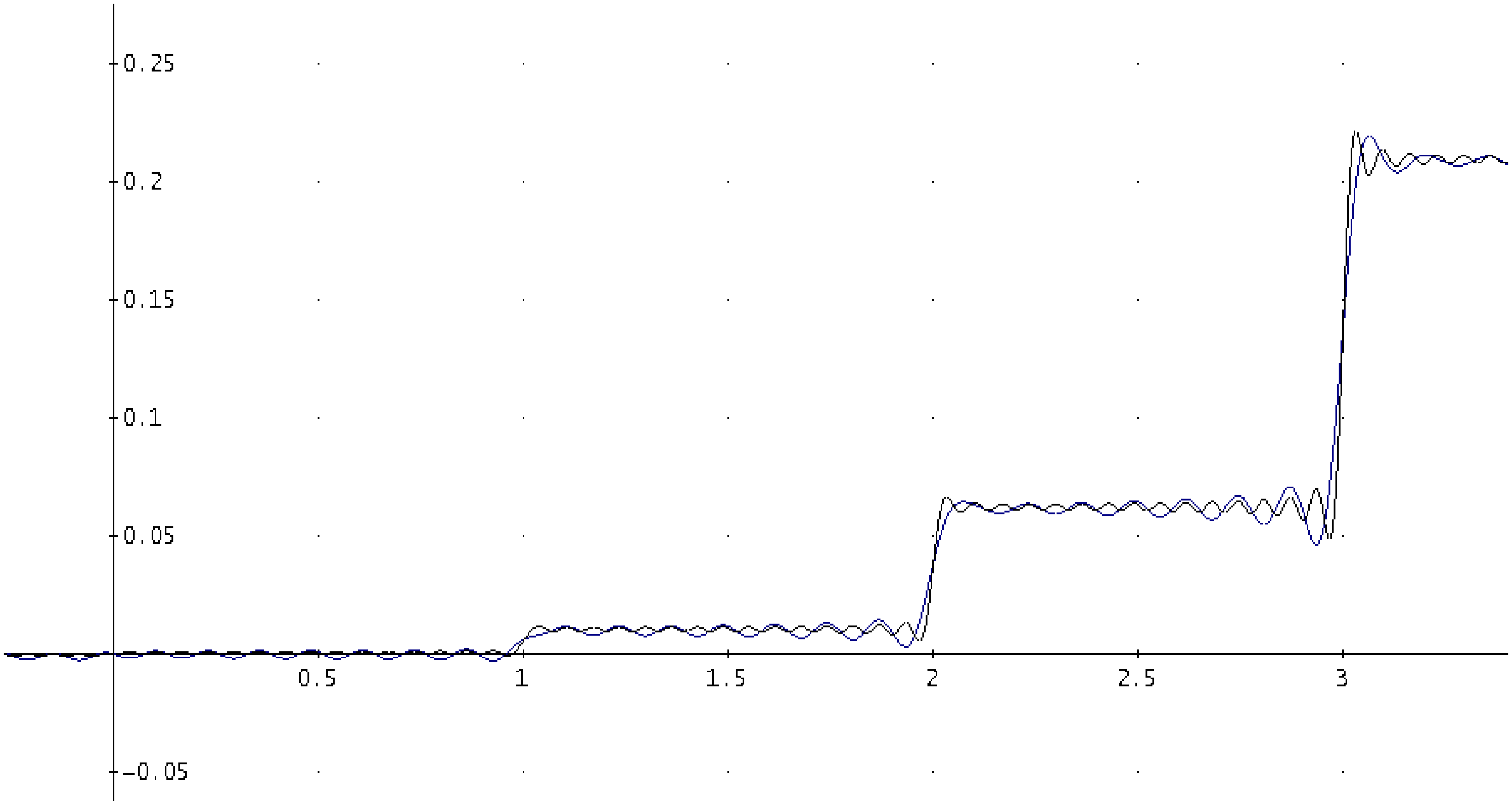
$$\frac{\mathrm{RE}\left(\displaystyle\int_{-n}^{n} \frac{e^{-i\cdot(-0.9)\cdot s} - e^{-i\cdot b\cdot s}}{i\cdot s}\cdot\left(0.5\cdot e^{i\cdot s} + 1 - 0.5\right)^{5}\cdot\left(0.7\cdot e^{i\cdot s} + 1 - 0.7\right)^{3} ds\right)}{2\cdot\pi}$$

$$\frac{\mathrm{RE}\left(\displaystyle\int_{-200}^{200} \frac{e^{-i\cdot(-0.9)\cdot s} - e^{-i\cdot b\cdot s}}{i\cdot s}\cdot\left(0.5\cdot e^{i\cdot s} + 1 - 0.5\right)^{5}\cdot\left(0.7\cdot e^{i\cdot s} + 1 - 0.7\right)^{3} ds\right)}{2\cdot\pi}$$

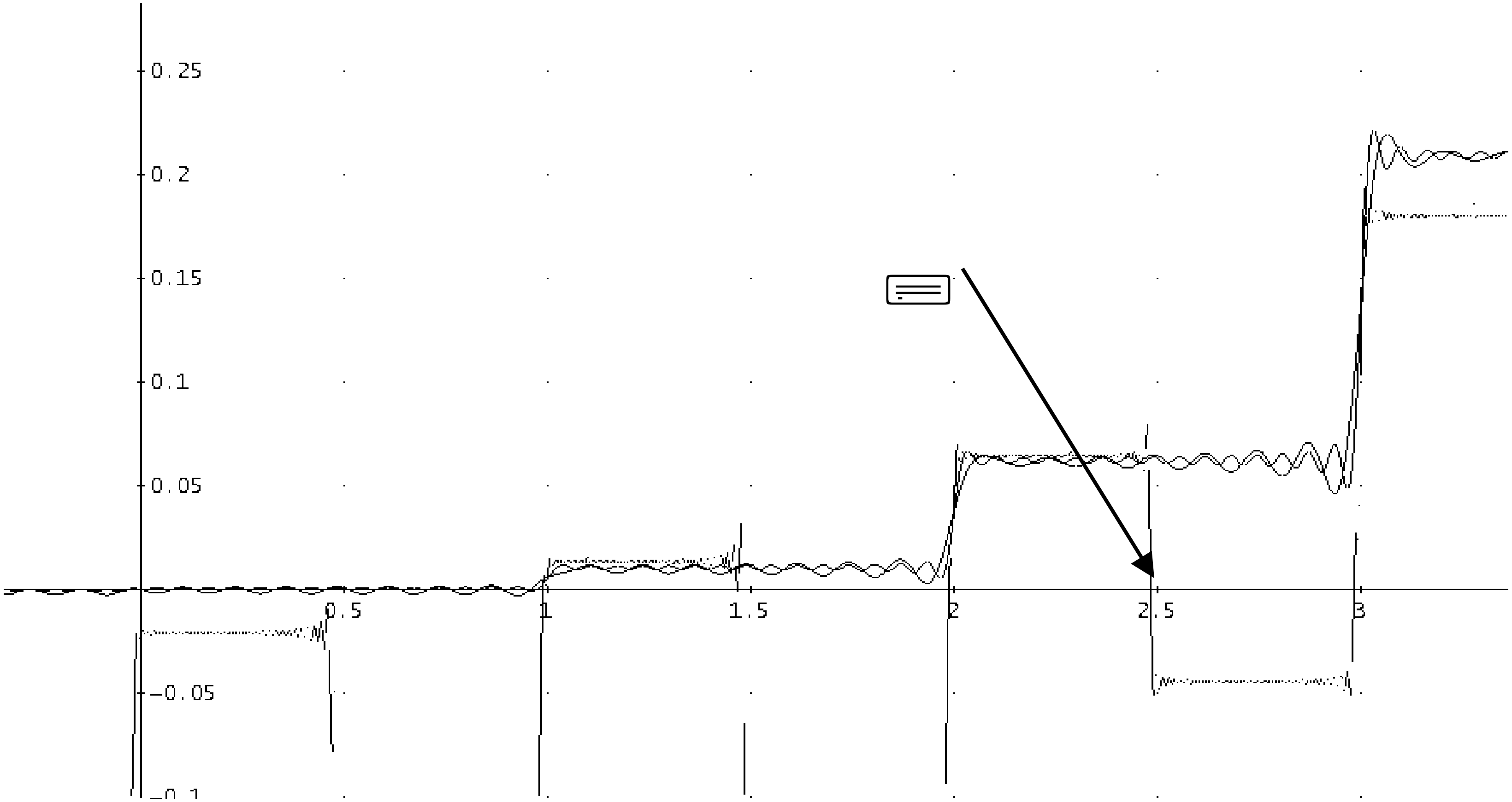
c = 50



c = 50 and c = 100



c = 50, c= 100, c=400



Mutation 2: Correlated Trials

- Linear algebra
- Binomial probabilities
- Simplest discrete random vectors
- Covariance and correlation
- Basic ideas of time series

Conclusions

- Mathematical statistics is alive and well
  - Inferred in part from this conference
- Mathematical computing packages are surprisingly useful pedagogical tools for undergraduates
- Mathematical computing packages are of some use in data analysis