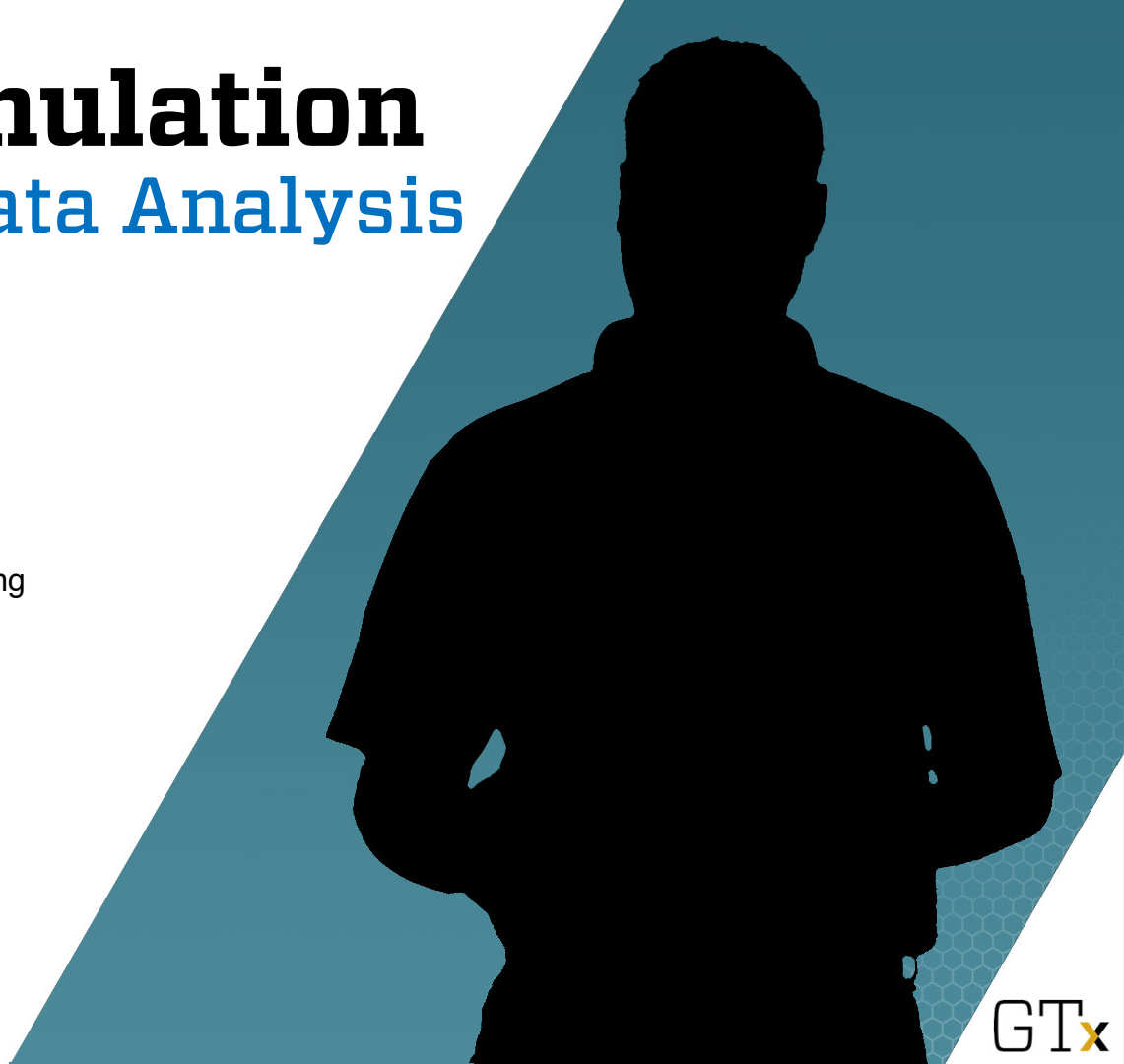
Computer Simulation

Module 8: Input Data Analysis

Dave Goldsman, Ph.D.
Professor
Stewart School of Industrial and Systems Engineering



Problem Children

GTx

Lesson Overview

Last Time: More g-o-f tests, including K-S, A-D, C-vM, S-W, etc.

This Time: Whatever shall we do with our problem children?

Lack of data
Goofy data
Nonstationary data
Multivariate / correlated data

GTx

Problem Children

Nobody likes to talk about them, but every family has them. You'd think that after all of the theory we've done, we could always find good distributions to fit our data. Not exactly. Here are some cases that you have to be careful about.

1. No / little data
2. Data that doesn't look like one of the usual distributions
3. Nonstationary data (from distributions that change over time)
4. Multivariate / correlated data

1. No / Little Data

This issue turns up more often than you would expect. There could literally be no data available, or the data that you have is awful (goofy values, not cleaned properly, etc.). What to do? No great options — but here are some suggestions.

- Interview so-called “experts”.
  - Try to at least get min, max, and “most likely” distribution values out of them — then you can guess uniform or triangular distributions.
  - Getting quantiles from the expert even better.
  - At least discuss the nature of the observations.

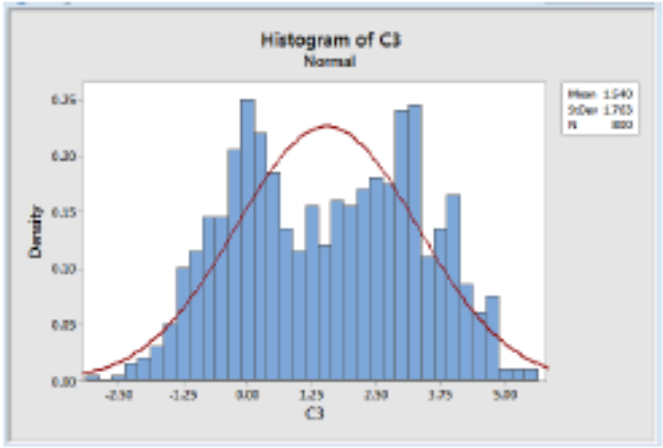
1. No / Little Data

If you have some idea about the nature of the RVs, maybe you can make a good guess as to the distribution.

- Discrete or continuous?
- Are observations successes / failures? Think Bernoulli, binomial,…
- Do observations adhere to Poisson assumptions? Then Poisson (if you're counting arrivals) or exponential (interarrival times).
- Are observations averages or sums? Then maybe normal.
- Are observations bounded? Then think beta.
- Reliability or job times? Maybe gamma, Weibull, lognormal, etc.
- Anything else from the physical characteristics underlying the RV?

2. Goofy Distributions

Here's a forced marriage of two normals — most packages can't pick this up or fit it properly. Example: Poorly designed 6644 exam has two modes!



Can attempt to model as a *mixture* of reasonable distributions.

Easier: Can sample from the empirical distribution or a smoothed version of the empirical. This is a form of *bootstrapping*.

3. Nonstationary Data

Arrival rates change over time — think restaurants, traffic on the highway, call center activity, seasonal demands for a product. You must take this variability into account, else GIGO!

- Suggestion: Nonhomogeneous Poisson process. (Recall from RV Generation module.)
- Need to model rate function properly.
- Arena uses piecewise-constant rate function; so specify a constant arrival rate for each separate period.

4. Multivariate / Correlated Data

Data don't have to be i.i.d.! What if data is multivariate and / or serially correlated in time? Examples:

- Multivariate RV — A person's height and weight are correlated.
- Serially correlated examples:
  - Monthly unemployment rates.
  - Arrivals to a social media site may be correlated if an interesting item appears there and the public gets wind of it.
  - A badly damaged part may require more service than usual at a series of stations.
  - If a server gets tired, his service times may be longer than usual.

4. Multivariate / Correlated Data

So what do you need to do?

- Identify multivariate / serial correlation situations.
- Propose appropriate models. Examples:
  - Multivariate normal for heights and weights.
  - Time series models for serially correlated observations, e.g., autoregressive-moving average ARMA(p, q), EAR(1), ARTOP, which we discussed back in Module 7.

4. Multivariate / Correlated Data

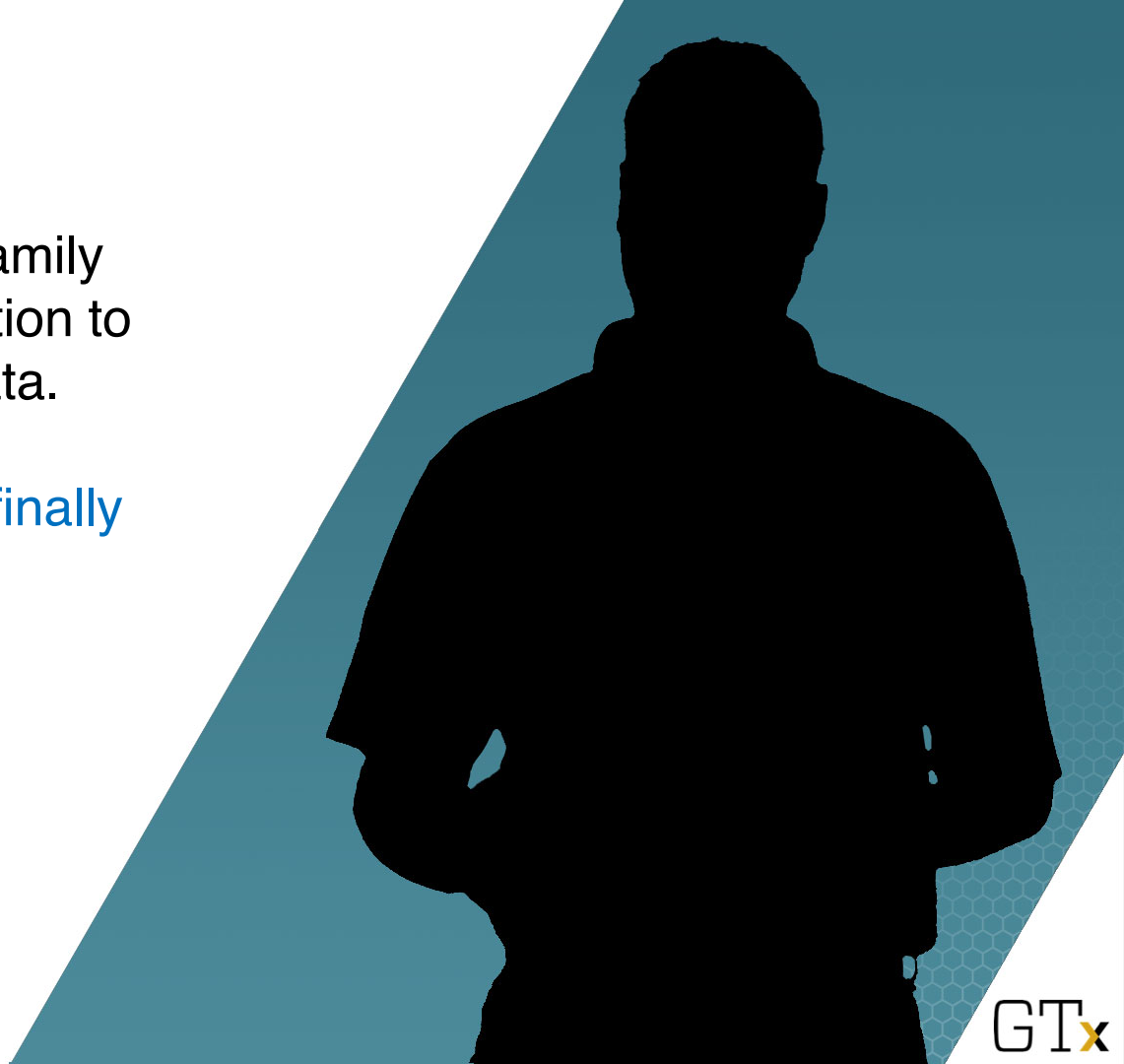
So what do you need to do (cont'd)?

- Estimate relevant parameters (easier said than done). Examples:
  - Multivariate normal: Marginal means and variances (no big deal) plus covariances (maybe not so easy).
  - Time series: For simple models like the AR(1), estimating the coefficient ϕ is easy (just like covariance). But coefficients for more-complicated models need to be estimated using available software, e.g., Box-Jenkins technology.
- Validate to see if your estimated model is actually any good.
- Alternative: Can bootstrap samples from an empirical distribution (if you have enough data).

Summary

This Time: Looked into our family secrets – things to pay attention to when thinking about input data.

Next Time: Demo time! We finally get a chance to have a fit!



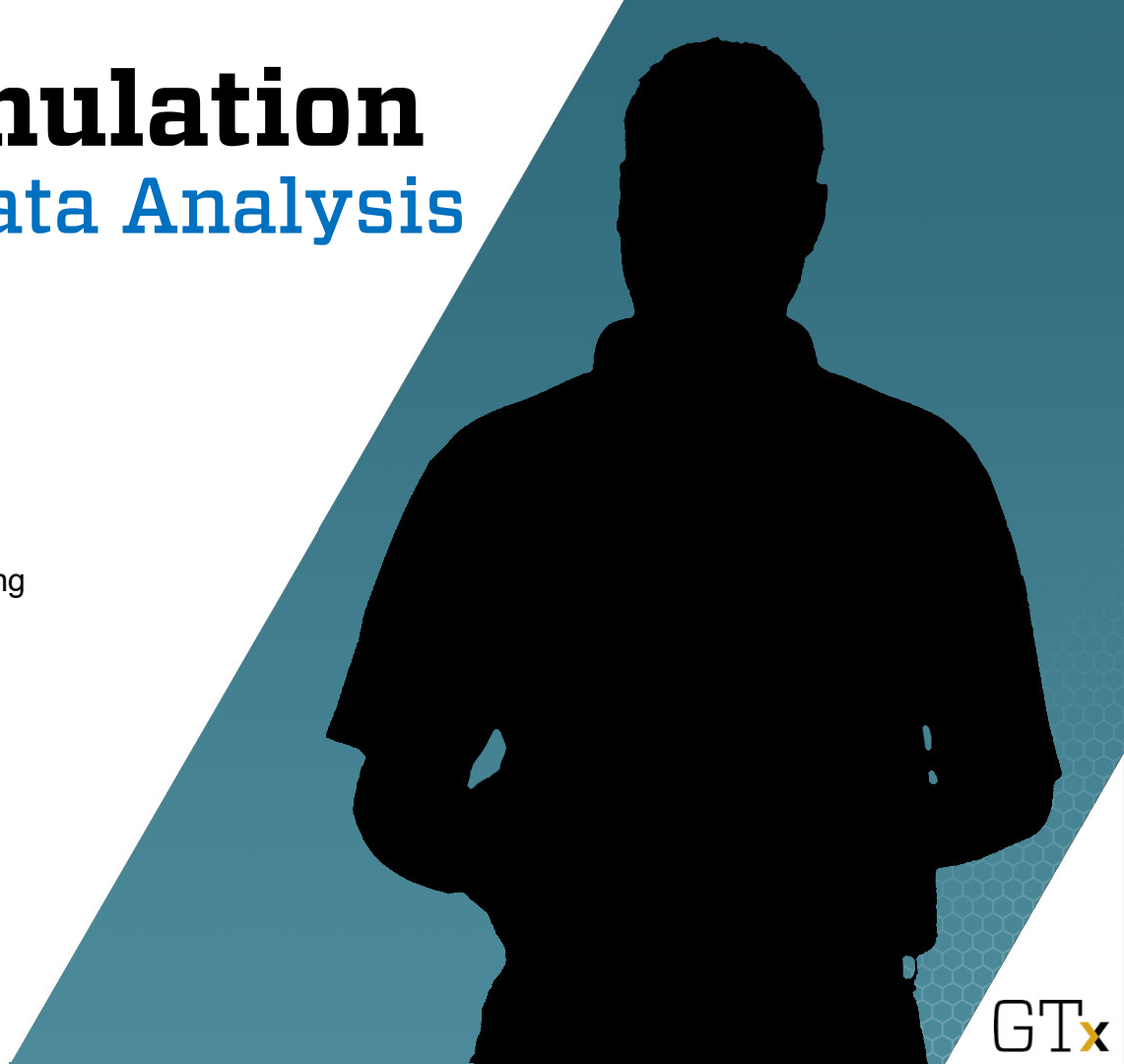
GTx

Computer Simulation

Module 8: Input Data Analysis

Dave Goldsman, Ph.D.
Professor
Stewart School of Industrial and Systems Engineering

Demo Time!



GTx

Lesson Overview

Last Time: Looked at some “problem children” issues related to input analysis.

This Time: Let’s finally do a demo to see how one can carry out an elementary input analysis!

Demo Time!

Arena has very nice functionality that automatically fits simple distributions to your data. Just go to Tools > Input Analyzer.

The Input Analyzer gives you the “best” distribution from its library, along with relevant sample and goodness-of-fit statistics.

ExpertFit is a specialty product that does distribution fitting with a larger library of distributions.

Minitab and R have distribution fitting functionality, though not quite convenient as the above tools.

Summary

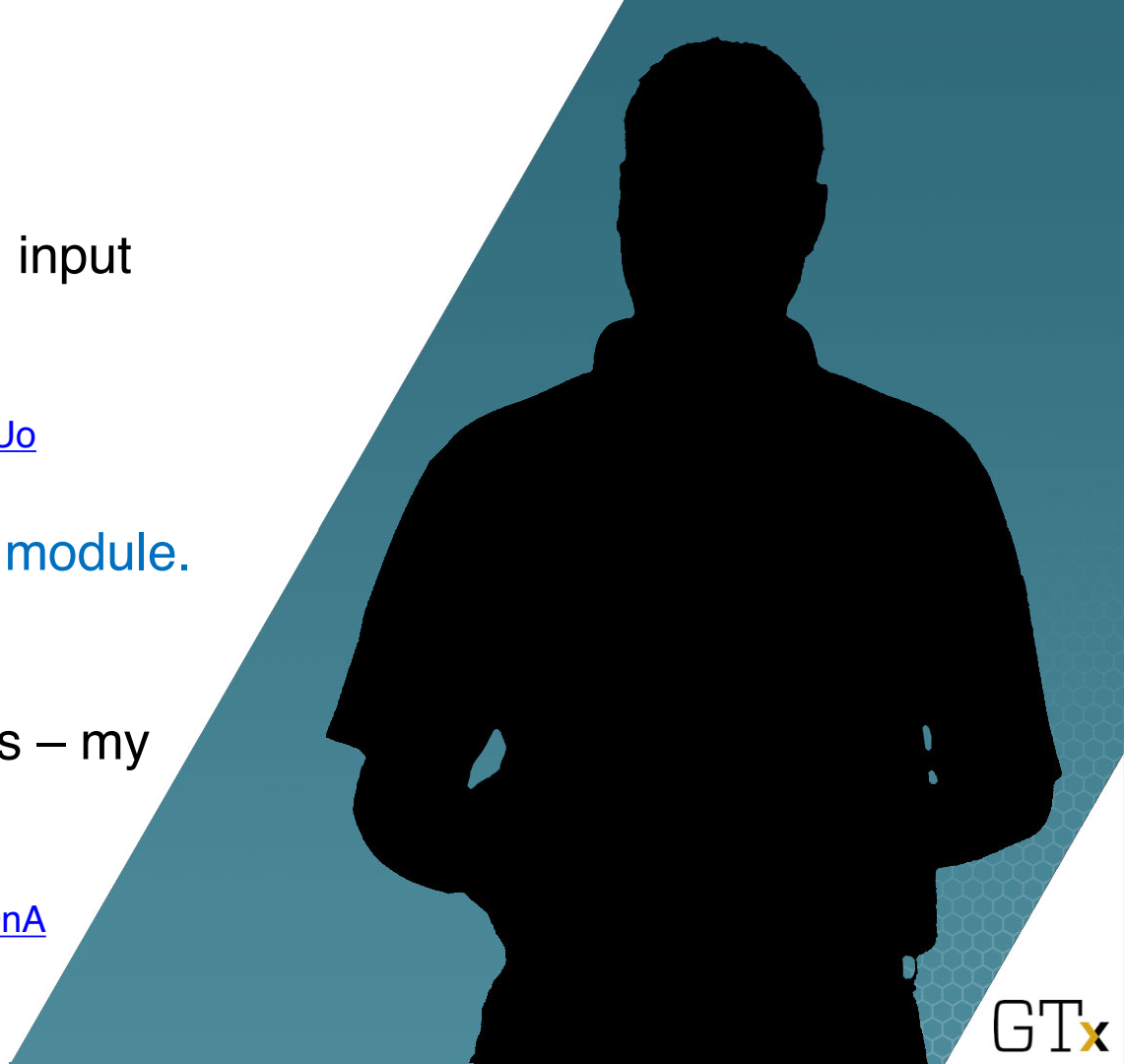
This Time: We carried out an input analysis Devo!

www.youtube.com/watch?v=IIEVqFB4WUo

This ends our Input Analysis module.

Next Module: Output Analysis – my fave topic!

www.youtube.com/watch?v=DGABqdbtQnA



GTx