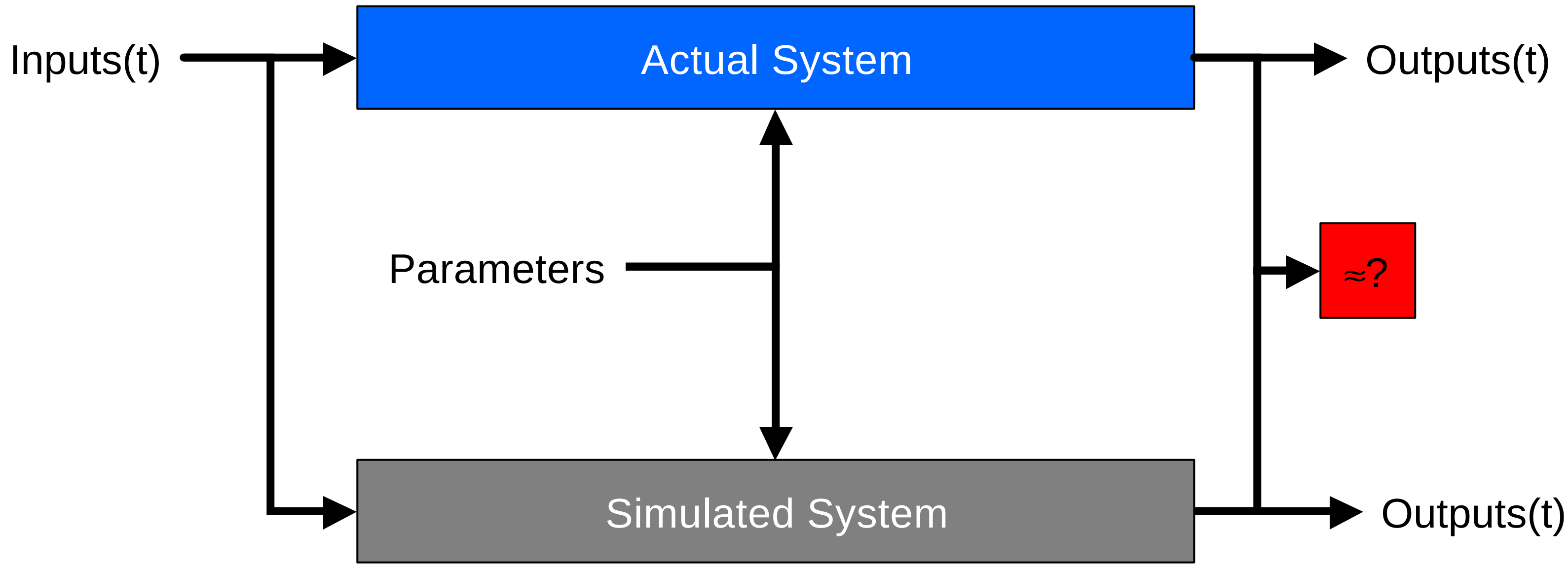
EE/CpE 345

Modeling and Simulation

Fall 2002

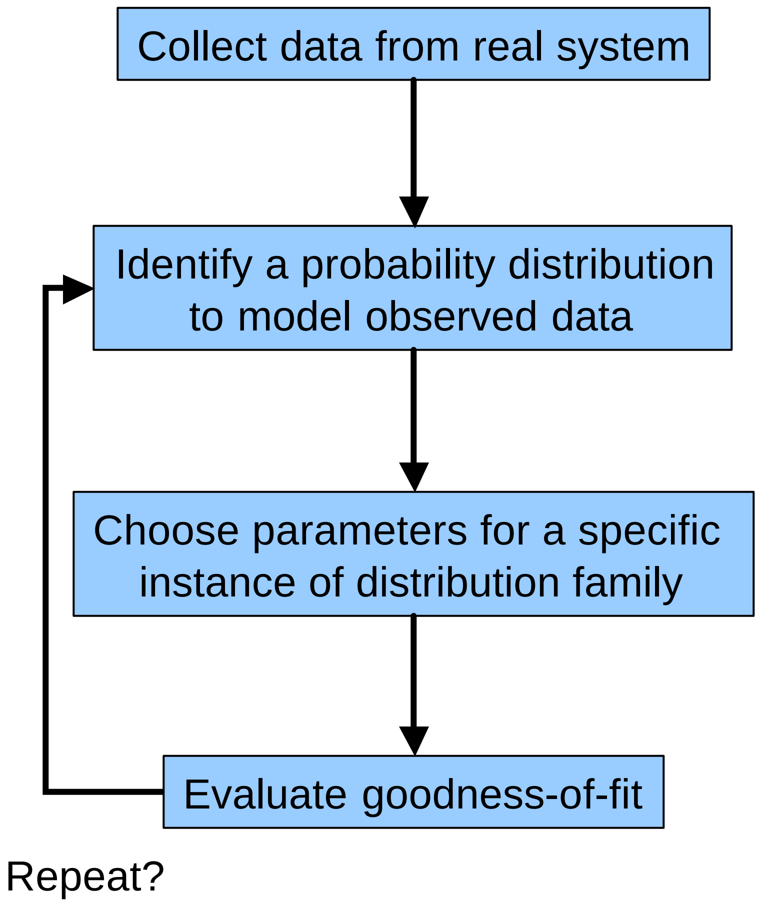
Class 10
November 18, 2002

Input Modeling



- The input data is the driving force for the simulation - the behavior of the simulation and all the results/conclusions that can be reached depend on appropriate inputs

Developing a Model of Input Data



- Is there time to collect enough data?
- Are there other sources available to obtain relevant information?

- Start with a histogram of data to enable visualization.
- Is anything known about process?

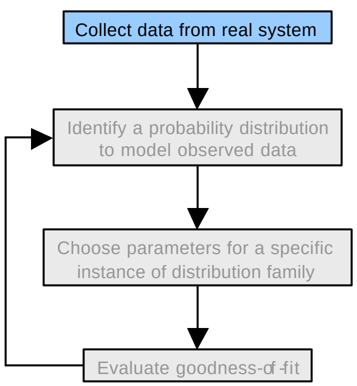
- Valid data is especially important for this step

- χ^2 and Kolmogorov-Smirnov tests

Data Collection

- Data collection can be the hardest tasks in solving a real problem
- Data collection is one of the most important and hardest tasks in simulation
  - Data is often either scarce or overly abundant

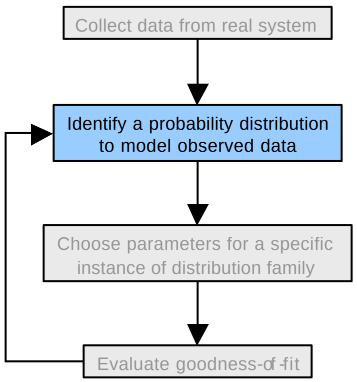
- "GIGO" (Garbage-In = Garbage-Out often applies
  - The simulation often abstracts real data, hiding its inadequacies



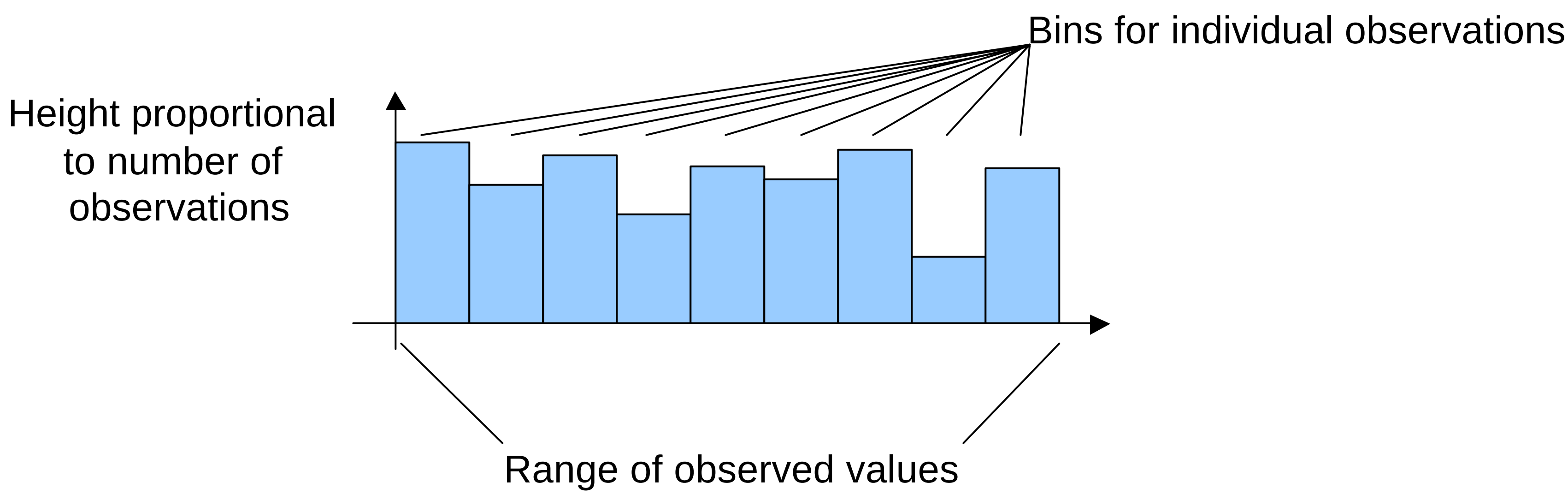
- Suggestions to improve data collection:
  - PLAN! Do some trial runs to see if there are any special circumstances that will have to be captured
  - Analyze/summarize data during collection - this might highlight a problem with data being collected
  - Look for homogeneity with plan to combine similar data sets
  - Watch for data censoring - is an observation of a process complete? Or are the long procedures truncated artificially?
  - Use a scatter plot to see relationships between variables - other senses help, as well
  - Look for correlation in data
  - Distinguish input data (independent variables) from performance data (dependent variables)

Identifying the Distribution

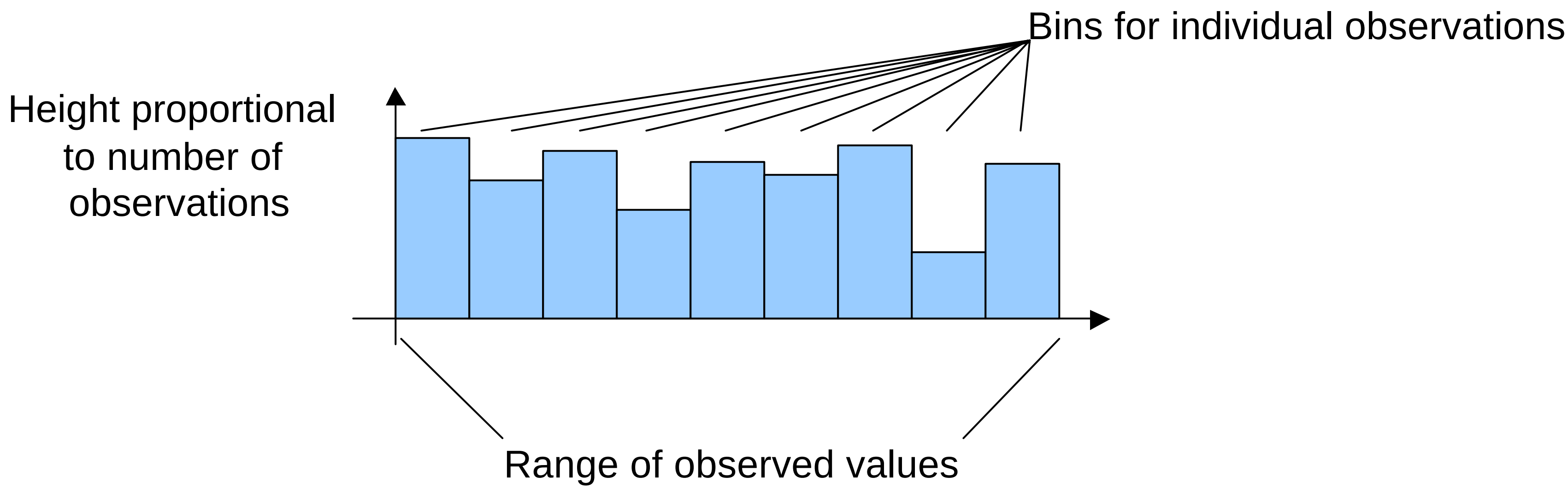
- Methods to identify distributions:
  - Histograms
  - Select the Family of Distributions
  - Quantile-quantile plots



Histograms


Bins for individual observations
Height proportional to number of observations
Range of observed values

Histograms



Number of bins ~ sqrt(observations)

Histograms - Picking the Number of Bins

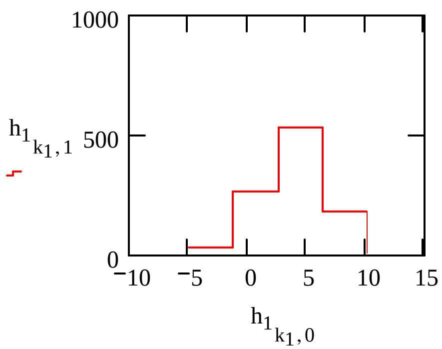
$\mu := 2 \qquad N := 1000$

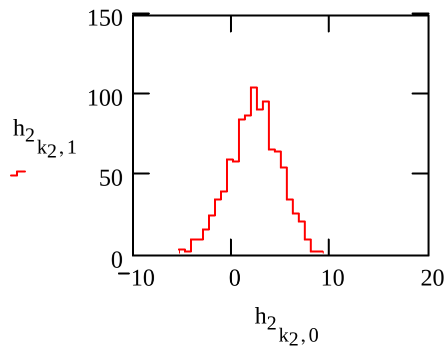
$\sigma := 2.5 \qquad R := \mathrm{rnorm}(N, \mu, \sigma)$

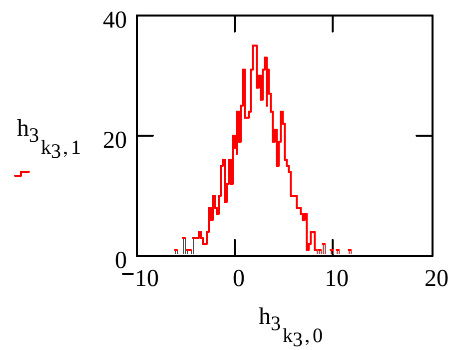
$M_1 := 5 \qquad M_2 := \mathrm{floor}(\sqrt{N}) \qquad M_3 := 100$

$h_1 := \mathrm{histogram}(M_1, R) \qquad h_2 := \mathrm{histogram}(M_2, R) \qquad h_3 := \mathrm{histogram}(M_3, R)$

$k_1 := 0 .. M_1 - 1 \qquad k_2 := 0 .. M_2 - 1 \qquad k_3 := 0 .. M_3 - 1$







• All Mathcad examples are in the file “Class10.mcd”

Meaning of the Histogram for Input Modeling

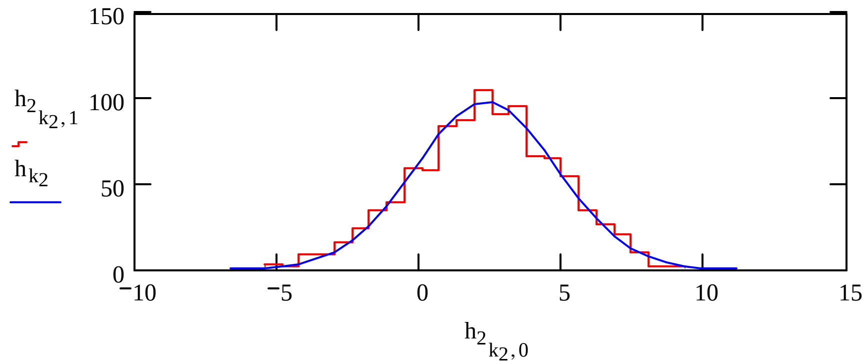
$H_{k_2} := \text{pnorm}\left(h2_{k_2,0}, \mu, \sigma\right)$ pnorm returns the cdf

$h_0 := H_0$

$k := 1 .. M_2 - 2$

$h_k := H_k - H_{k-1}$ convert to pdf

$h := h \cdot N$



- Scaled by the total number of points, the histogram approximates the p.d.f. of the input distribution

- Use the histogram to visualize the p.d.f. of the observed data to enable selection of a known distribution function

Selecting the Family of the Distribution

- There are a large number of probability distributions
  - generated from observations of the real world
  - proposed as theoretical models

- What is known about the physical characteristics of the input process?
  - Is it naturally discrete or continuous valued?
  - Are the observable values inherently bounded or is there no natural bound?
  - Can you infer a distribution from what you know about the process that generates input values?
    - E.g., Normal (Gaussian) process is derived from the sum of a large number of independent random variables
    - E.g., Erlang process is sum of several exponential processes
    - E.g., Lognormal process is derived from product of several component processes
    - E.g., Poisson process models the number of independent events that occur in a bounded period of time or area in space.

Quantile-Quantile Plots

$N := 30$

$R := \text{rnorm}(N, \mu, \sigma)$

$S := \text{sort}(R)$

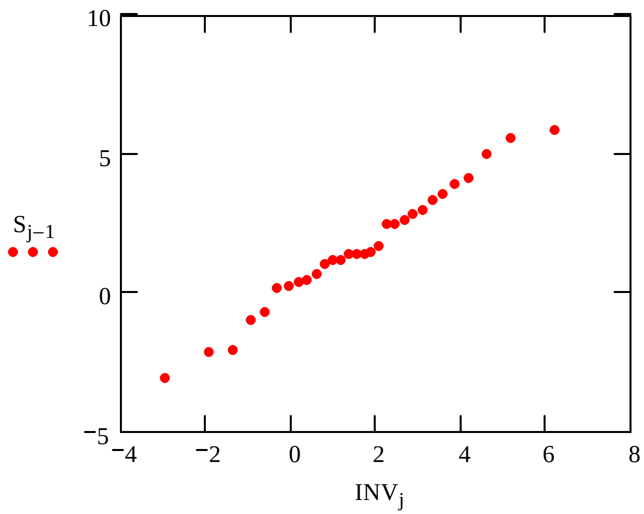
$\text{mean}(R) = 1.604$

$\text{var}(R) = 5.458$

$j := 1..\, N$

$$\gamma_j := \frac{j - \frac{1}{2}}{N}$$

$$INV_j := \text{qnorm}\left(\gamma_j, \text{mean}(R), \sqrt{\text{var}(R)}\right)$$



- Given a set of input data, *R*
- Calculate m and s^2
- Sort *R* (*S* in this Mathcad file)
- Generate a set of g_j, evenly distributed between 0 and 1
- For a given assumed distribution (using calculated m and s^2), calculate the inverse of the c.d.f. for each g
- Plot the sorted data vs. the calculated values

- If the assumed distribution matches, the plot should be a straight line with slope=1, intercept=0

Q-Q Plots with Incorrect Assumptions

- Generating Exponentially distributed random numbers
- Matching a Normal distribution

- The resulting plot is not linear, as expected.

- If you have Mathcad or the Mathcad viewer*, try experimenting with other distributions, parameters

Using the wrong distribution:

$N := 1000$

$R := rexp(N, \mu)$

$S := sort(R)$

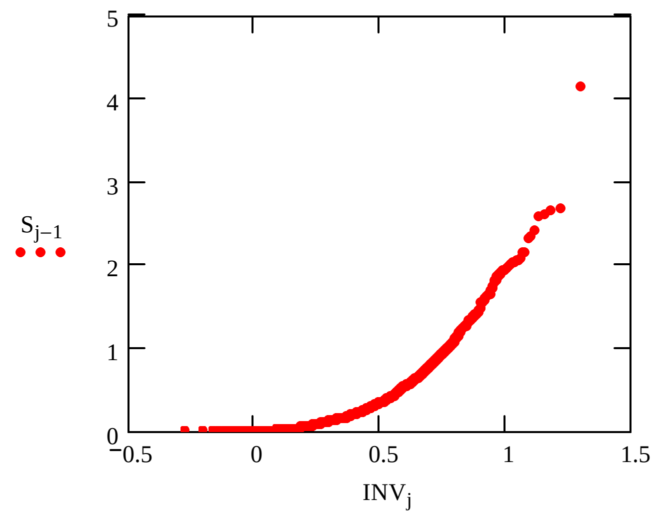
$mean(R) = 0.504$

$var(R) = 0.24$

$j := 1 .. N$

$$\gamma_j := \frac{j - \frac{1}{2}}{N}$$

$INV_j := qnorm(\gamma_j, mean(R), var(R))$



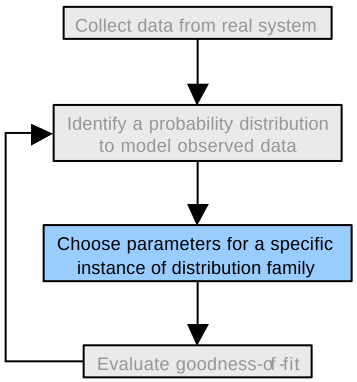
Parameter Estimation

- Mean and variance:
  - n samples:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$S^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}{n-1}$$

  - Discrete data, grouped in *k* frequency distribution classes:



$$\bar{X} = \frac{\sum_{j=1}^{k} f_j X_j}{n}$$

$$S^2 = \frac{\sum_{j=1}^{k} f_j X_j^2 - n\bar{X}^2}{n-1}$$

Parameter Estimation

- Mean and variance:
  - Continuous data in *c* frequency classes when raw data is not available. m_j are the midpoints of the frequency classes:

$$\bar{X} \doteq \frac{\sum_{j=1}^{c} f_j m_j}{n}$$

$$S^2 \doteq \frac{\sum_{j=1}^{c} f_j m_j^{\;2} - n\bar{X}^2}{n-1}$$

Parameter Estimation

Distribution	Parameter(s)	Suggested Estimator(s)
Poisson	a	$\hat{a} = \bar{X}$
Exponential	l	$\hat{l} = \frac{1}{\bar{X}}$
Normal	m, s^2	$\hat{m} = \bar{X}$ $\hat{s}^2 = S^2$

- The estimated parameters using the suggested estimators are maximum-likihood estimators, based on raw data.
- The true parameters, assuming that the distribution was known, are not expected to be the same as the experimentally measured parameters
  - small sample size
  - noise, randomness in measurements

Parameter Estimation

$\mu := 3.5$

$\sigma := 2$

$N_{norm} := 50$

$R_{norm} := \text{rnorm}(N_{norm}, \mu, \sigma)$

$\text{mean}(R_{norm}) = 3.883$

$\sqrt{\text{var}(R_{norm})} = 1.653$

$\lambda := .1$

$N_{exp} := 50$

$R_{exp} := \text{rexp}\left(N_{exp}, \frac{1}{\lambda}\right)$

$\text{mean}(R_{exp}) = 0.093$

$\sqrt{\text{var}(R_{exp})} = 0.138$

These should be equal

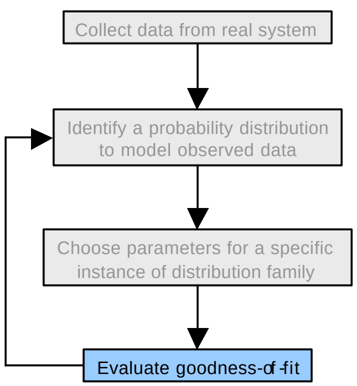
- Distinguish the parameter of the distribution: a
- From the estimator or statistic: $\hat{a}$

Goodness-of-Fit

- Hypothesis testing was introduced 2 weeks ago to test random number distributions
  - Kolmogorov-Smirnov
  - c^2

- Goodness-of-fit tests should *guide* the choice of a distribution, not *establish* it: there is often no perfect answer with real-world data.

- Sample size can have a significant effect on results:
  - With a small number of data points, few or no candidate distributions will be rejected
  - With a large number of data points, almost all candidate distributions will be rejected.



Goodness-of-Fit Tests

- c^2 test:
  - Compares the histogram of candidate density function
  - Valid for large sample sizes
  - Assumes parameters are estimated by maximum likelihood function

- c^2 test with equal probabilities:
  - If distribution is assumed to be continuous, class intervals should be equal probability, rather than equal width. Example 9.14 (next slide)

- Kolmogorov-Smirnov:
  - With c^2 test, grouping of data may influence accept/reject decision
  - K-S test is based on examining a q-q plot
  - Especially useful when no parameters have been estimated from data

Example 9.14 - c^2 Test for Exponential Distribution

- The data, *X* is generated from an exponential distribution, and the c^2 test is applied.
- Here, c_o^2 =15.92, which exceeds the tabulated value for a significance of 0.01, but not at a significant of 0.05 - we might reject this distribution if the level of significance were tight enough

$n := 50 \qquad \lambda := .1$

$X := \text{rexp}(n, \lambda) \qquad X_s := \text{sort}(X)$

$\lambda_{hat} := \dfrac{1}{\text{mean}(X)} \qquad \lambda_{hat} = 0.102$

$k := 8 \qquad p := \dfrac{1}{k} \qquad p = 0.125$

$i := 0..\, k-1 \qquad a_i := \dfrac{-1}{\lambda_{hat}} \cdot \ln(1 - i \cdot p)$

$a_k := \infty \qquad E_i := p \cdot n$

$O := \text{hist}(a, X_s) \qquad \text{term}_i := \dfrac{(O_i - E_i)^2}{E_i}$

$$a = \begin{pmatrix} 0 \\ 1.306 \\ 2.814 \\ 4.597 \\ 6.78 \\ 9.593 \\ 13.559 \\ 20.339 \\ 1 \times 10^{307} \end{pmatrix} \quad O = \begin{pmatrix} 4 \\ 1 \\ 9 \\ 13 \\ 7 \\ 8 \\ 4 \\ 4 \end{pmatrix} \quad E = \begin{pmatrix} 6.25 \\ 6.25 \\ 6.25 \\ 6.25 \\ 6.25 \\ 6.25 \\ 6.25 \\ 6.25 \end{pmatrix} \quad \text{term} = \begin{pmatrix} 0.81 \\ 4.41 \\ 1.21 \\ 7.29 \\ 0.09 \\ 0.49 \\ 0.81 \\ 0.81 \end{pmatrix}$$

$$\chi_{sq} := \sum_{i=0}^{k-1} \frac{(O_i - E_i)^2}{E_i} \qquad \chi_{sq} = 15.92$$

p-Value and "Best Fits"

- In the prior example, with the particular data examined, the hypothesis that the data came from an exponential distribution would have been rejected at a significance level of 0.01, but not at a significance level of 0.05.
- How should you choose a significance level? 0.01, 0.05, & 0.10 are commonly used.
- The significance level is equivalent to the probability of falsely rejecting H_0
- Many software packages compute a *p*-value:
  - The *p*-value is the significance level which just rejects H_0
  - The *p*-value can be viewed as a measure of fit: larger *p*-value indicates a better fit
  - One possible approach to choosing a distribution:
    - Test every distribution available, choose the one that has largest *p*-value

Selecting Input Models without Data

- What about where there is no input data available to model, e.g., for a preliminary study when no time or funds are available to gather data?

- Creating data where none exists:
  – base it on published performance data
  – get expert opinion for the same or similar systems - this might help to bound or otherwise characterize inputs
  – use physical limitations to bound problem (e.g., maximum possible car arrivals at an intersection is related to minimum car spacing and maximum velocity)
  – The nature of the process: use the descriptions of various distributions to pick the one most closely related to the underlying input process

- Uniform, triangular, and beta distributions are used when nothing else is known

Multivariate and Time Series Input Models

- Previous discussion dealt with independent processes
- What if multiple inputs to the simulation are related to each other?
  - multivariate input models with a fixed, finite number of random variables
  - time-series input models of a sequence of related random variables
- There are two measures of dependence between random variables:
  - Covariance
  - Correlation

Multivariate and Time Series Input Models

- X_1 and X_2 are two random variables:
  - $m_i = E(X_i)$ and $s^2{}_i$ = Var(X_i)

- Covariance and correlation define how well the relationship between X_1 and X_2 is described by:

$$(X_1 - m_1) = b\,(X_2 - m_2) + e$$

- e is a zero mean R.V., independent of X_2
- Covariance:

$$\text{cov}(X_1, X_2) = E\left[(X_1 - m_1)(X_2 - m_2)\right] = E(X_1 X_2) - m_1 m_2$$

- Correlation:

$$r = \text{corr}(X_1, X_2) = \frac{\text{cov}(X_1, X_2)}{s_1 s_2}$$

- Observations about cov( ) and corr( )

Multivariate and Time Series Input Models

- If X_1, X_2,… are identically distributed (but possibly dependent), this is referred to as a time-series.

- cov(X_t, X_{t+h}) and corr(X_t, X_{t+h}) are the *lag-h* covariance and correlation
- If cov(X_t, X_{t+h}) depends on *h* and not *t*, the time series is covariance-stationary and can be represented as:

$$r_h = corr(X_t, X_{t+h})$$

Homework 10

- Ch. 9, p. 361 exercises 8, 9, 13