The process of examining and interpreting data that is fed into a system, model, or analysis to gain insights, make predictions, or optimize decisions.

Input Data Analysis

IE 630 Simulation Modeling and Analysis

Jayendran V

IEOR @ IIT Bombay

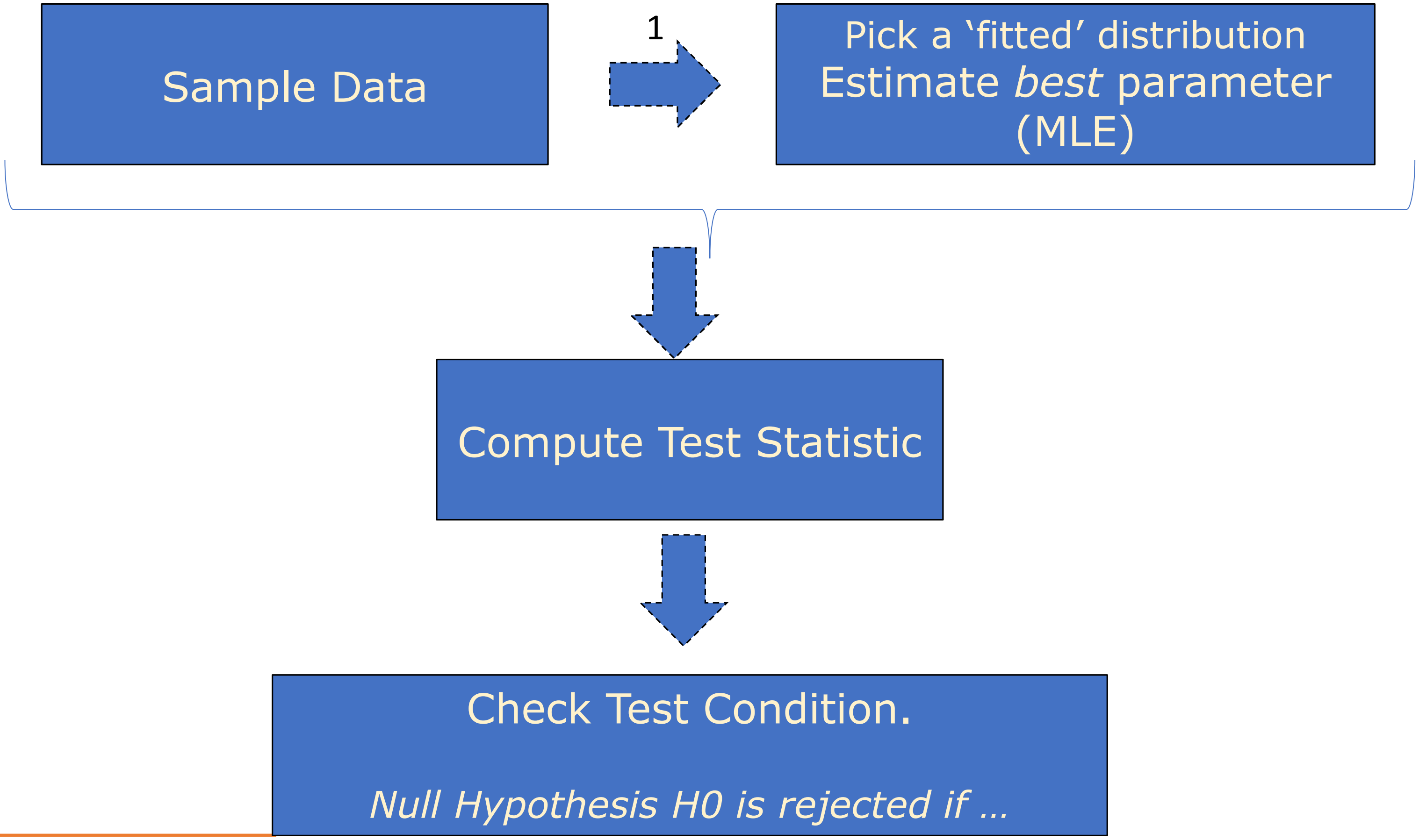
Goodness-of-Fit Tests

- Need to examine how well the 'fitted distribution' represent the true distribution of our data
  - **H_0**: X_is are IID with fitted distribution function & parameters given by estimators
  - **H_1**: Xis do not conform

- Chi-square Test
  - Chi-square test with equal width
  - Chi-square test with equal probabilities
- Kolmogorov Smirnov Test

- *p*-values and "Best Fit"

Goodness of Fit Test (contd.)



Sample Data

1

Pick a ‘fitted’ distribution
Estimate *best* parameter (MLE)

Compute Test Statistic

Check Test Condition.

Null Hypothesis H0 is rejected if …

Goodness-of-Fit Tests

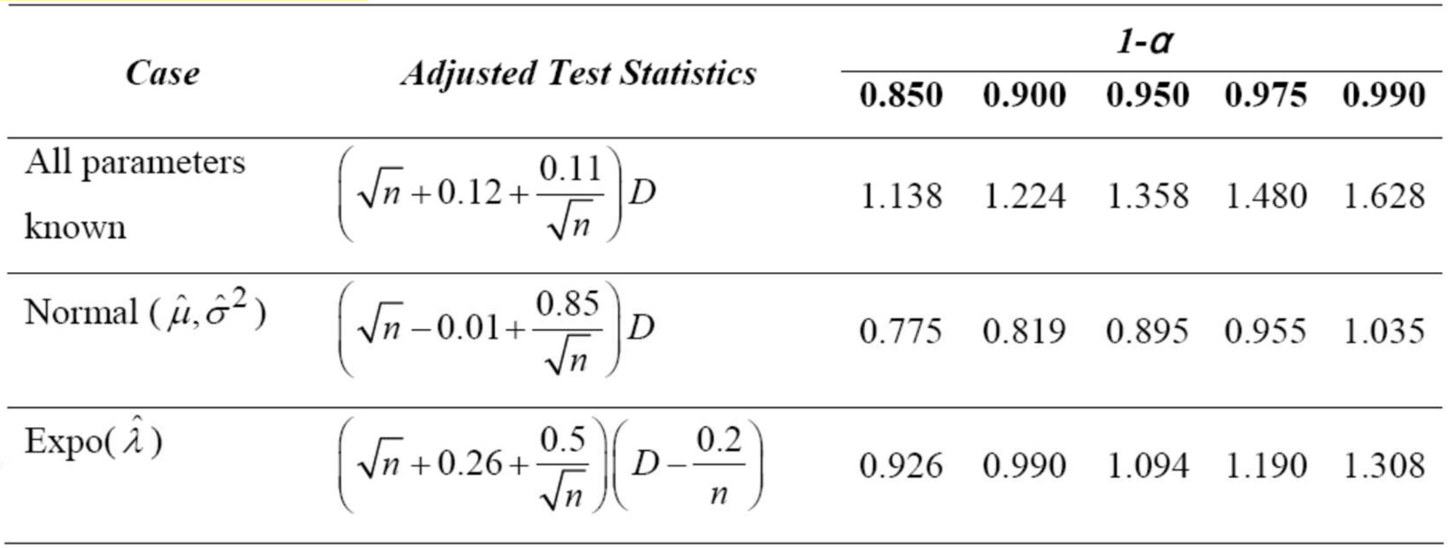
- Need to examine how well the ‘fitted distribution’ represent the true distribution of our data
  - **H_0**: X_is are IID with fitted distribution function & parameters given by estimators
  - **H_1**: Xis do not conform

- Chi-square Test
  - Chi-square test with equal width
  - Chi-square test with equal probabilities
- **Kolmogorov Smirnov Test**

- *p*-values and “Best Fit”

Kolmogorov-Smirnov (K-S) Test

- Test fitted cdf with empirical cdf
  - Step 1: Rank the data smallest to largest
  - Step 2: Compute D^+ and D^- → Construct a Table
  - Step 3: $D = \text{Max}\{D^+, D^-\}$
  - Step 4: Test Statistics
- Modified critical values of $c_{1-\alpha}$



Case	*Adjusted Test Statistics*	***1-α***				
		0.850	**0.900**	**0.950**	**0.975**	**0.990**
All parameters known	$\left(\sqrt{n}+0.12+\frac{0.11}{\sqrt{n}}\right)D$	1.138	1.224	1.358	1.480	1.628
Normal $(\hat{\mu},\hat{\sigma}^2)$	$\left(\sqrt{n}-0.01+\frac{0.85}{\sqrt{n}}\right)D$	0.775	0.819	0.895	0.955	1.035
Expo$(\hat{\lambda})$	$\left(\sqrt{n}+0.26+\frac{0.5}{\sqrt{n}}\right)\left(D-\frac{0.2}{n}\right)$	0.926	0.990	1.094	1.190	1.308

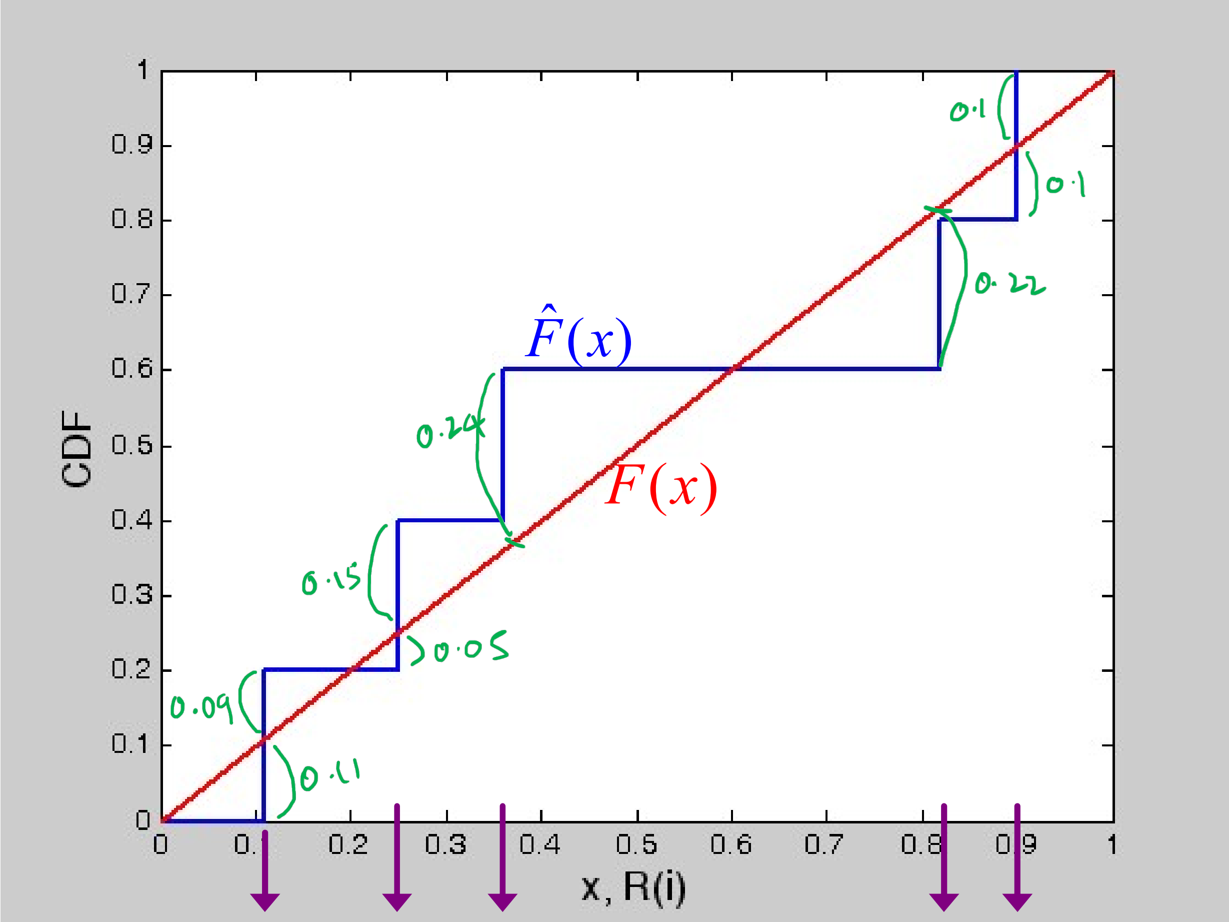
KS Test Example

- Using KSTest, check if the sample are UNIFORM[0,1] at $\alpha = 0.05$
  - 0.36, 0.90, 0.25, 0.11, 0.82

- Table

i	$R_{(i)}$	i/N	$i/N - R_{(i)}$	$R_{(i)} - (i-1)/N$

Visualizing KS Test (Example)



KS Test: Example 2

- Suppose you are observing arrivals at an ATM. You noted the time you started your observation as 0.00. The first arrival happened at time 3.93 min, second at 5.09 and so on, as shown in table below. Check if the interarrival times are exponentially distributed.

Customer	Arrival time (in mins)
1	3.93
2	5.09
3	7.10
4	13.68
5	14.21

Anderson-Darling Tests

- Based on difference between empirical cdf and fitted cdf
- Drawback of K-S Test
  - KS test gives same weightage to difference $|F(x) - \hat{F}(x)|$

    for all *x*, whereas many distributions differ primarily in their tails

- A-D designed to detect discrepancy in tails & has higher power than K-S test

p-value

- To apply goodness-of-fit tests, we need to know the level of significance

- p-value is the level of significance at which one would just reject H_0 for the given value of the test statistic
  - Large *p-value* indicates a ___good___ fit
  - Small *p-value* indicates a ___poor___ fit

Caution using commercial software

- Many packages recommend a ‘best fit’ distribution for the data
  - Packages choose general purpose distributions
  - Erlang/ Weibull over Expo;
  - Beta over Uniform/ Normal
  - *Many package rank the fit based on MSE or SSE between fitted CDF and empirical CDF! → Check hypothesis test results!*
- **CHECK results of tests**
  - See which tests are done. Are they statistically valid?
  - Compare *p-value* for different tests across distributions
  - See where the lack of fit occurs
- Use your judgment to select right distribution
  - Consider the physical basis of the data

What if data does not fit any known distribution?

- Use data values themselves to define a ***empirical distribution***

Non-Stationary Arrival Process

Non-Stationary Arrival Process

- Event (often arrival) rates varies over time
  - Rush hour traffic
  - Arrivals at restaurants vary over the day, higher during meal times
  - Seasonal demand swings …
- Important to capture this nonstationarity
  - So that model is valid
  - We don’t miss peaks and valleys, important behavior
- Model: *Non-stationary Poisson Process*
  - Modeling a time varying general arrival process is tougher

Stationary Poisson Process

- PROPERTIES

1. Arrivals occur one at a time
2. Number of arrivals in the interval (t, t+s] is independent of number of arrivals in earlier time interval [0, t]; and also of the times at which these arrivals occur.
3. Distribution of number of arrivals is independent of time.

- (1) will not hold true if..
- (2) will not hold true if..
- (3) will not hold true if..

- For Poisson process with rate λ, the inter-arrival times are IID Exponential random variables with mean 1/λ.

Non-Stationary Poisson Process (NSPP)

- PROPERTIES

1. Arrivals occur one at a time
2. Number of arrivals in the interval (t, t+s] is independent of number of arrivals in earlier time interval [0, t]; and also of the times at which these arrivals occur.

- For NSPP arrival rate λ(t), is a function of time.

➔ How to estimate the λ(t) function from data collected?

NSPP: Piecewise-constant method

- Divide time frame of simulation into subintervals of time over which the rates are fairly flat
- Compute observed rates within each subinterval

How to approx. λ(t) from data?

- Data collected on customer arrivals, say from 8am to 10am for *n* days.
- Observation of arriving customers: NSPP may be applicable
- Piecewise constant method
  - Divide time into subintervals over which the rates are fairly flat
  - For each subinterval
    - Average number of arrivals over *n* days
    - Arrival rate (=average number/ subinterval length)
  - Arrival rate λ(t) varies over the time horizon, but constant in each time interval

How to approx. λ(t) from data? (contd.)

Time period	Number of arrivals			Estimated arrival rate (arrivals/ hour)
	Day1	Day2	Day3	
8:00-8:30	12	14	10	24
8:30-9:00	3	4	2	6
9:00-9:30	20	13	12	30
9:30-10:00	30	27	33	60
10:00-10:30	27	19	32	52
10:30-11:00	20	13	12	30

How to model NSPP in simulation correctly?

No data or Insufficient data

No Data?

- Possible if
  - … modeling a brand new system or planning a significant change in existing one
  - … need to develop a demo/ preliminary model
- To “get” data
  - Engineering data, standards, performance ratings
  - Expert opinion
  - Physical or conventional limitations
  - Nature of process
    - Deterministic data/ distributions?
    - Does model output change drastically to minor change in input?
  - TRY SENSITIVITY ANALYSIS

Possible “no data” distributions

- EXPONENTIAL ________________ :
  - High variance, Bounded on left, unbounded on right
  - Inter-arrival times, time to machine failure
- TRIANGULAR ________________ :
  - Symmetric or non-symmetric, bounded both sides
  - Activity times
- UNIFORM ________________ :
  - All values equally likely, bounded both sides
  - Little know about the process

Caution on using Normal Distributions

- Most familiar
- It has infinite tail in both directions
  - Has infinite LEFT tail… so can (theoretically) generate negative values!
  - But most input quantities MUST be positive !

- If μ is relatively big compared to σ, then P(negative value) is small.. One-in-million

- Avoid Normal Distribution as far as possible