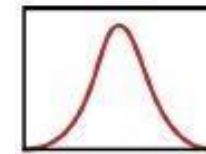


# Data Modeling and Probability Distributions

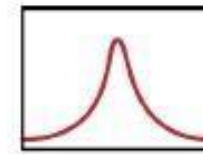
Dr. Ahmad S. Tarawneh



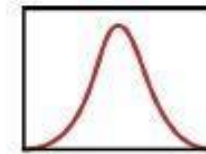
Normal Distribution



Uniform Distribution



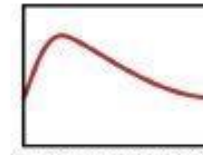
Cauchy Distribution



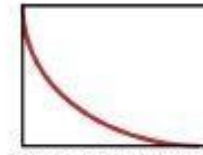
t Distribution



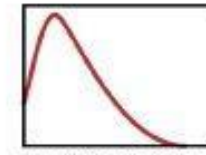
F Distribution



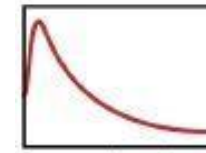
Chi-Square Distribution



Exponential Distribution



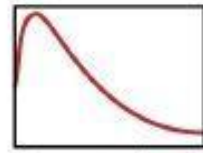
Weibull Distribution



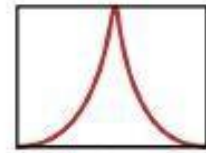
Lognormal Distribution



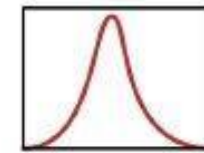
Birnbbaum-Saunders  
(Fatigue Life) Distribution



Gamma Distribution



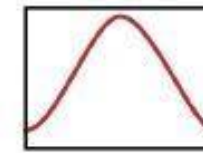
Double Exponential  
Distribution



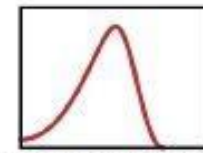
Power Normal Distribution



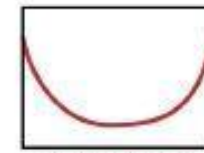
Power Lognormal  
Distribution



Tukey-Lambda Distribution



Extreme Value Distribution



Beta Distribution



# Outline

- Introduction
  - Data Collection
  - Probability Distributions
  - Fit a Distribution on Data
  - ExpertFit (Data modeling software)
-

# Introduction

- What we have discussed so far is a set of problems in which the distribution is given
- But from where did we get this table
- If we know the model that fits our data we do not need to use such tables, practically.

<i>Time between Arrivals (Minutes)</i>	<i>Probability</i>
1	0.25
2	0.40
3	0.20
4	0.15

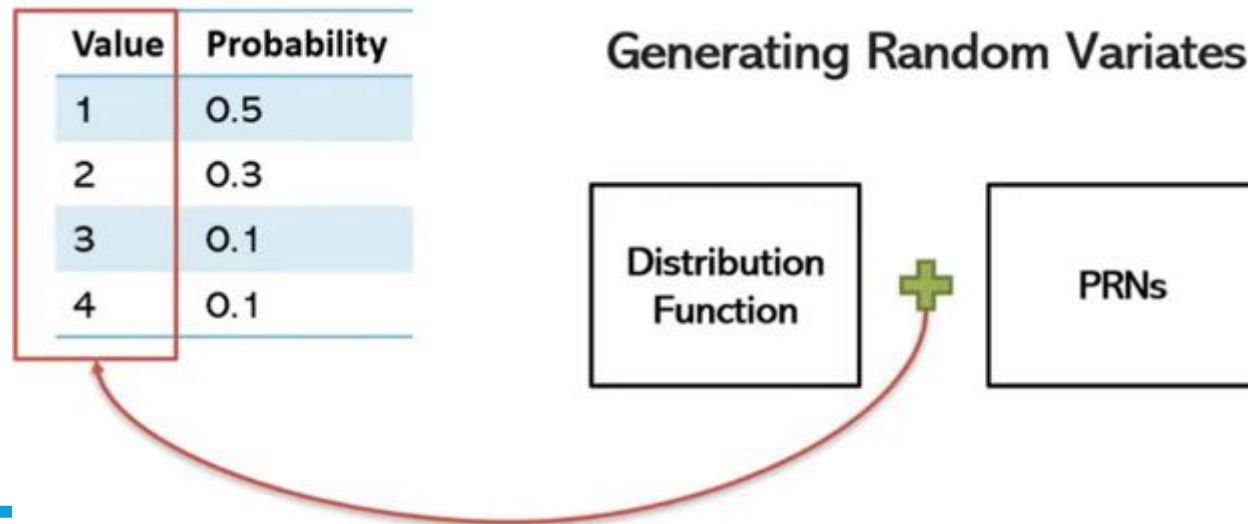
Queueing Sim.

<i>Demand</i>	<i>Probability</i>
0	0.10
1	0.25
2	0.35
3	0.21
4	0.09

Inventory Sim.

# Cont.

- Knowing the distribution will help us generate the values directly depending on the distribution of the data



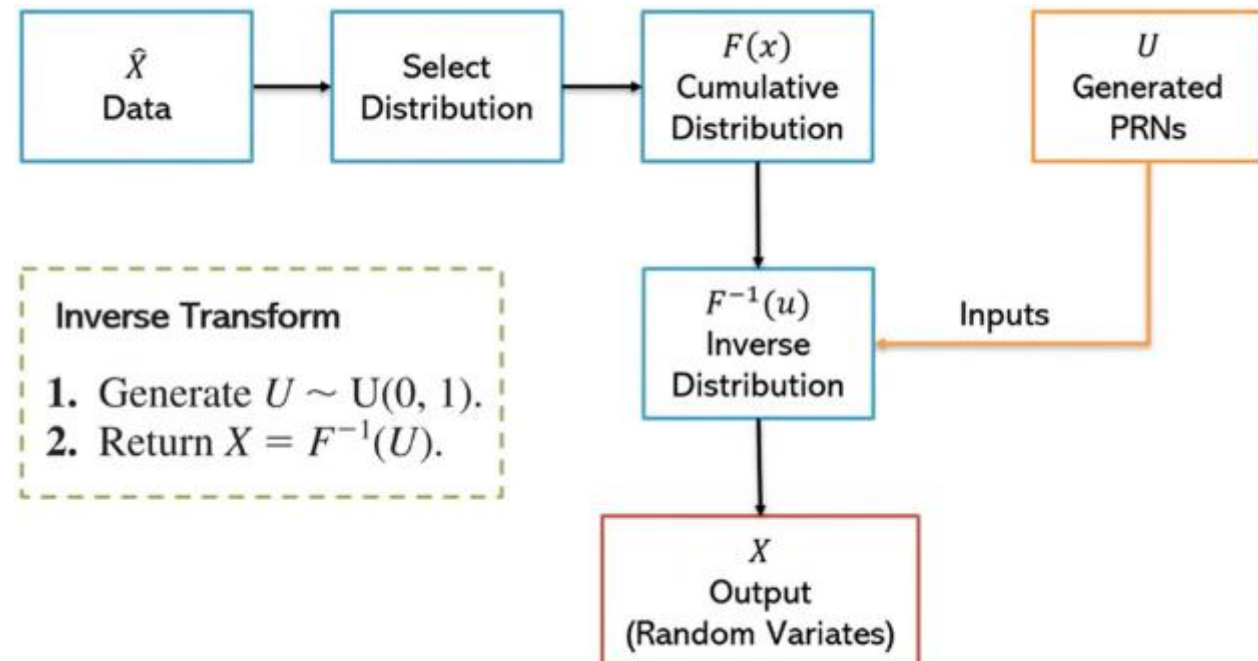


# Cont.

- There are several approaches to generate random variables (not numbers) given the distribution:
    1. Inverse Transform
    2. Composition
    3. Convolution
    4. Acceptance-Rejection
    5. Ratio of Uniforms
    6. Many more
-

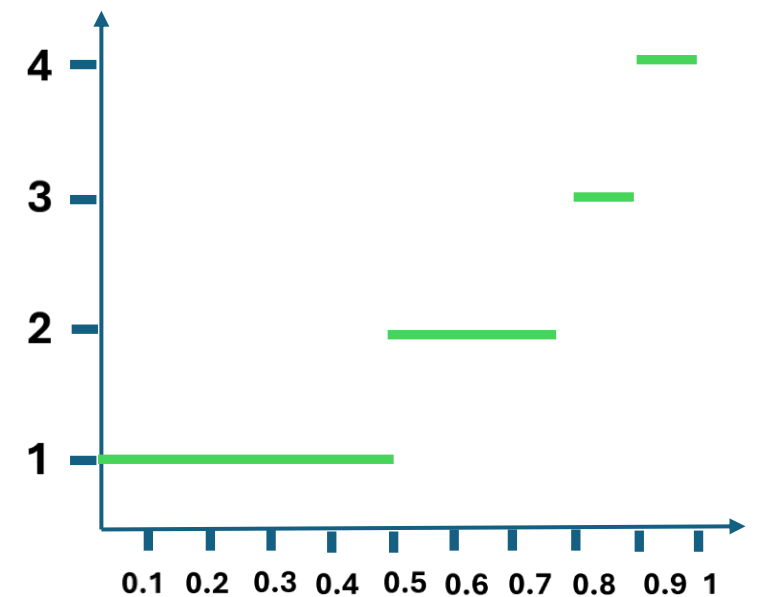
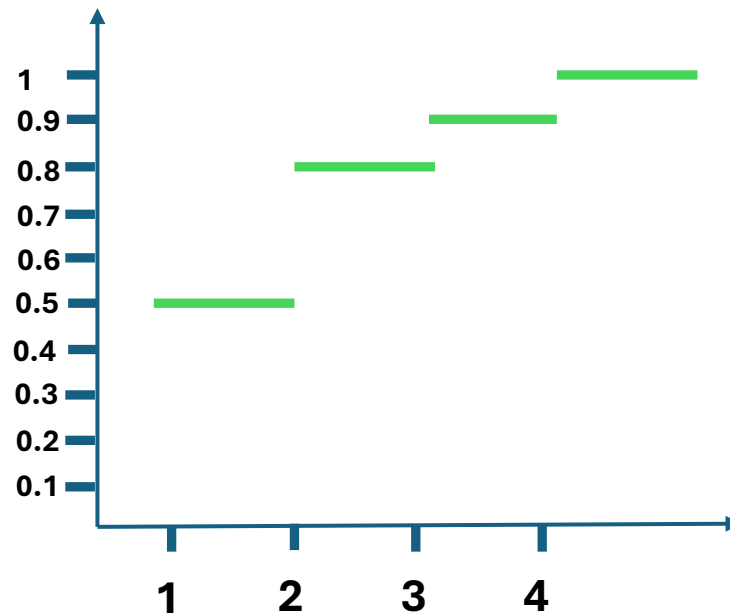
# Cont.

- Inverse Transform is the simplest



# Example

Value	Probability	Com
1	0.5	0.5
2	0.3	0.8
3	0.1	0.9
4	0.1	0.1





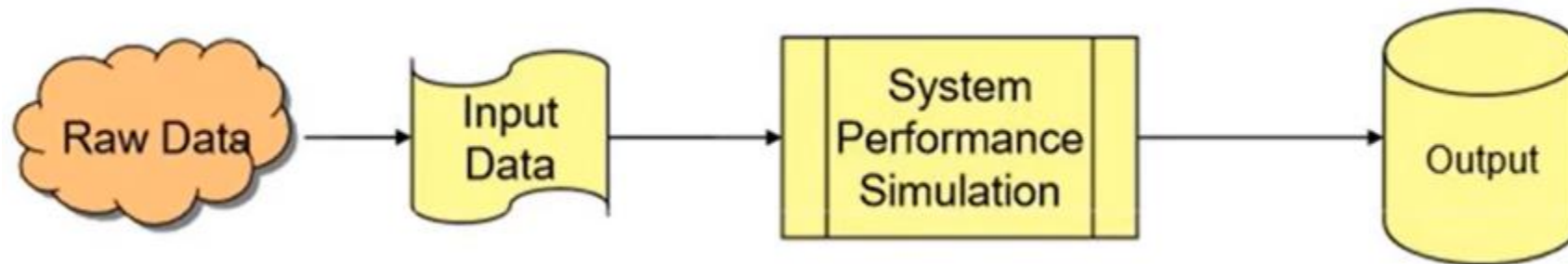
# Random variables

- Inputs are independent variables in the system:
    1. Interarrival time, which is typically a continuous random variable
    2. Amount of demand is a discrete random variable
-

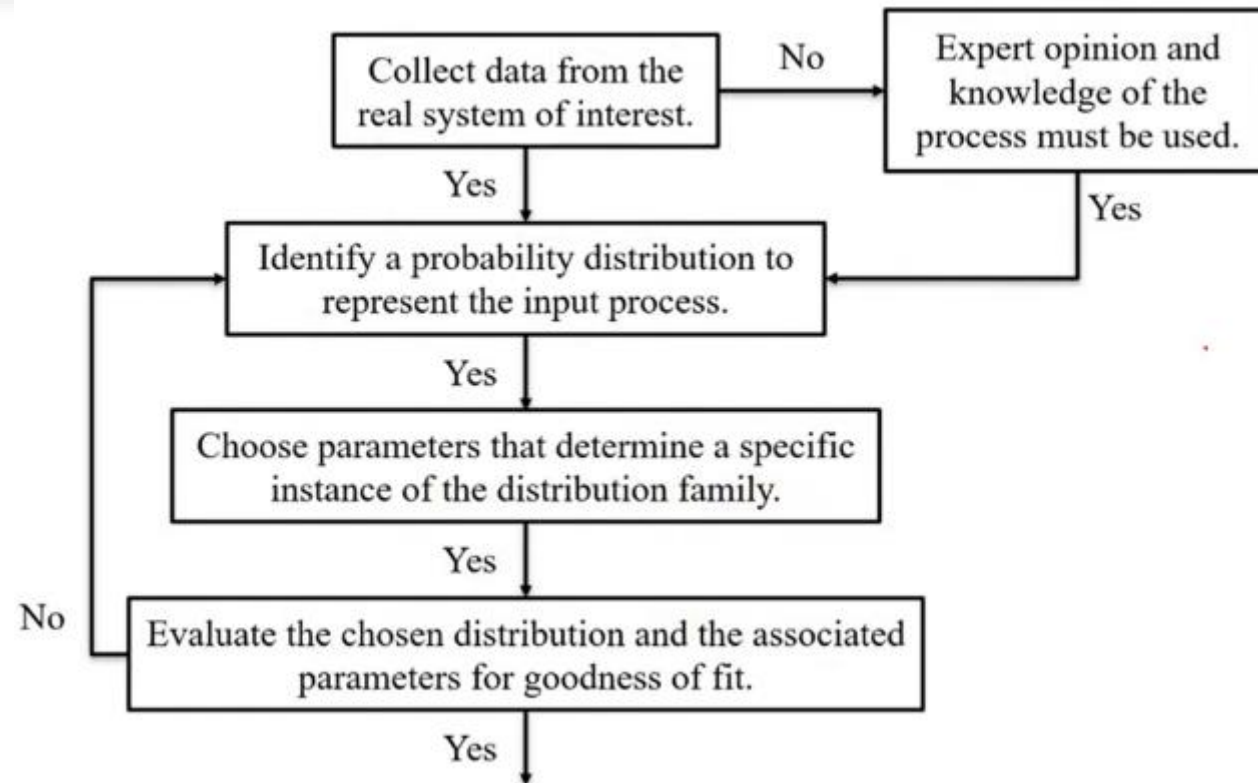


# Data modeling

- Data model (prob distribution of the data) selection is very important for the simulation model, as all the simulation depends upon these models
  - Good data modeling means good simulation performance



# Steps of data modeling





# Data Collection

- The tool used for simulation should have several features to support building a good performing simulations
    1. Must have a good random number generator
    2. Should be able to create a probability distribution for each source of randomness of the system
    3. Gives the ability to perform several, independent, runs of the simulation, using different random numbers, to validate the stability of the performance
-



# Cont.

- In simulation systems, typically, we try to find the **best standard probability distribution** that fits the source of randomness
    - If it can be found, then it should be used
  - If a standard distribution cannot be found, then we can use an **empirical distribution, aka custom distribution or used defined distribution.**
    - This custom distribution is created based on the data
  - **Using standard distribution is always preferable over empirical distributions**
-



# Cont.

- The data collected should be:
    1. Accurately collected
    2. Representative of the environment
    3. Analyzed correctly
  - Otherwise, the simulation will give a misleading results or wrong conclusions
-



# Cont.

- When the system exists, the data is better to be collected from this system, real system
  - **If the system does not exist or no time to collect the data from existed system, an informed guess can be made, by the help of experts.**
-

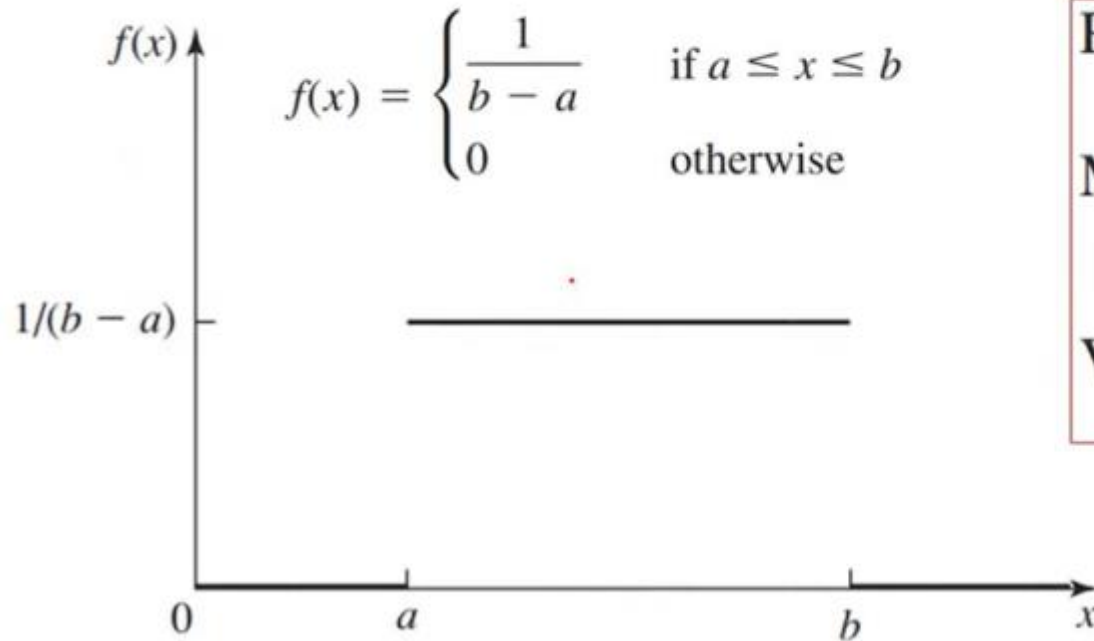


# Standard Distributions

- Many standard theoretical distributions are there.
  - Some of these distributions work for continuous data and some work for discrete
  - **Continuous Distributions include but are not limited to**
    1. Uniform distribution
    2. Exponential distribution
    3. Gamma distribution
    4. Gaussian (Normal) distribution
-

# Uniform Distribution

- $a$  and  $b$  real numbers with  $a < b$ .



Range	$[a, b]$
Mean	$\frac{a+b}{2}$
Variance	$\frac{(b-a)^2}{12}$



# Cont.

## Solution steps

Density

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

Distribution

$$F(x) = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } a \leq x \leq b \\ 1 & \text{if } b < x \end{cases}$$

$$\frac{(x-a)}{b-a} = u$$

Multiply both sides by  $b-a$

$$x-a = u(b-a)$$

Move  $a$  to the right side

PRN

$$F^{-1}(u) = a + (b-a)u$$

Generating Random Variates  $X'$

$$x = u(b-a) + a; \quad b-a \neq 0$$

# Exponential Distribution

- Scale parameter  $\beta > 0$ .

Density

$$f(x) = \begin{cases} \frac{1}{\beta} e^{-x/\beta} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Distribution

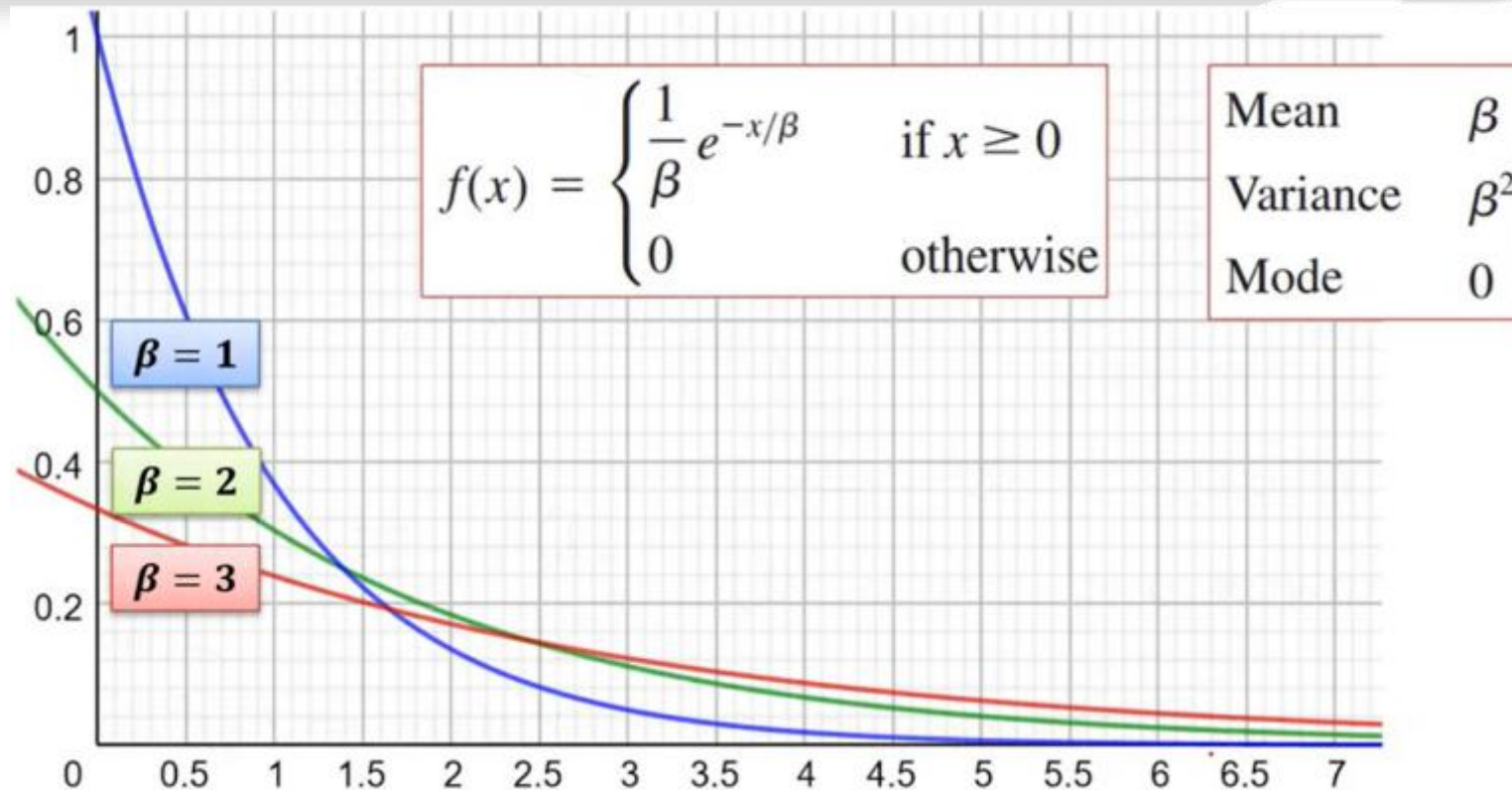
$$F(x) = \begin{cases} 1 - e^{-x/\beta} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

PRN  $\rightarrow$

$$F^{-1}(u) = -\beta \ln(1 - u)$$

Generating Random Variates  $X$ 's

# Cont.

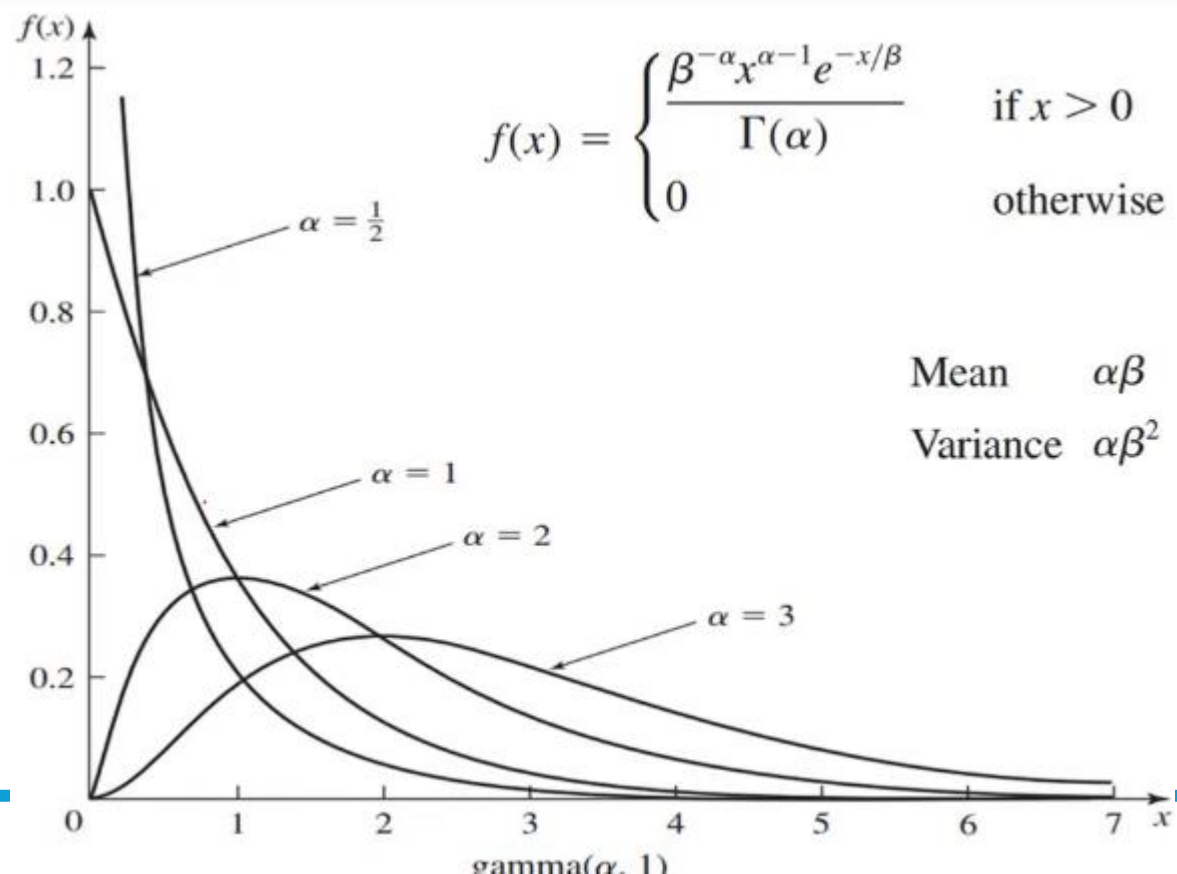


# Gamma Distribution

- Shape parameter  $\alpha > 0$ , Scale parameter  $\beta > 0$ .

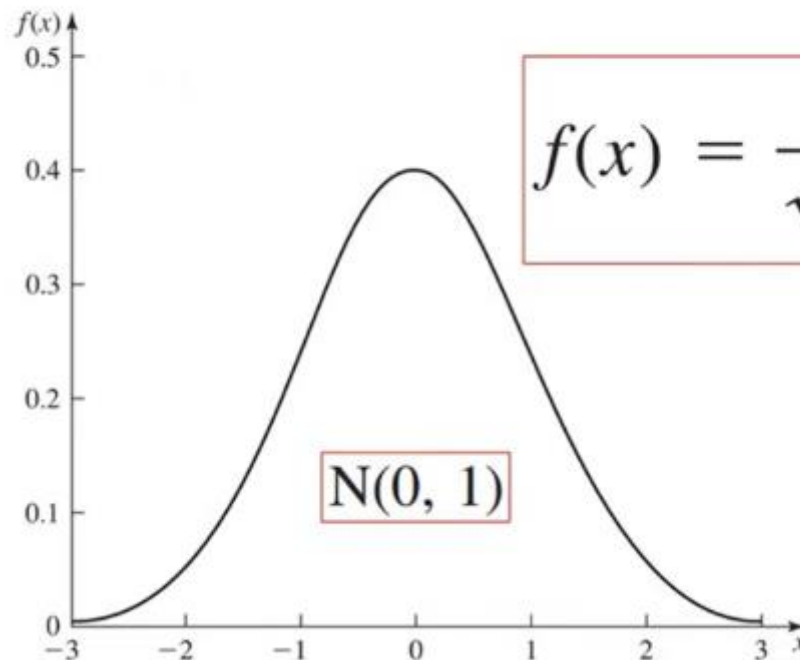
$$f(x) = \begin{cases} \frac{\beta^{-\alpha} x^{\alpha-1} e^{-x/\beta}}{\Gamma(\alpha)} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

# Graph Gamma



# Normal Distribution

- Location parameter  $\mu \in \mathbb{R}$ , Scale parameter  $\sigma > 0$ .



$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$$

Mean	$\mu$
Variance	$\sigma^2$
Mode	$\mu$

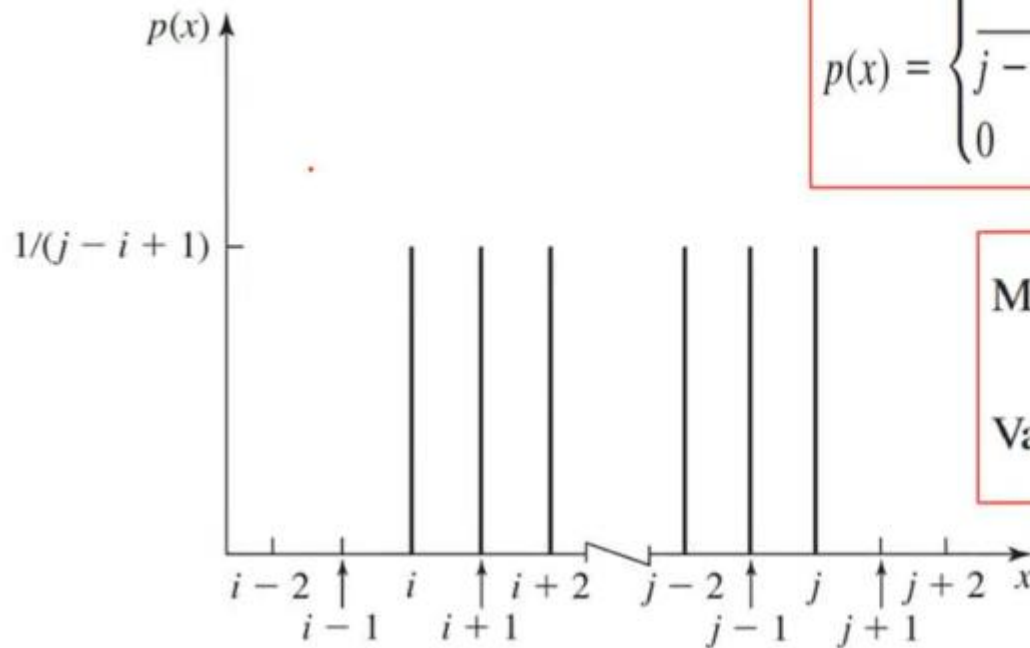


# Discrete Probability Distributions

- Discrete Uniform Distribution
  - Geometric Distribution
  - Poisson Distribution
-

# Discrete Uniform Distribution

- $i$  and  $j$  integers with  $i \leq j$ .



$$p(x) = \begin{cases} \frac{1}{j-i+1} & \text{if } x \in \{i, i+1, \dots, j\} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Mean} \quad \frac{i+j}{2}$$

$$\text{Variance} \quad \frac{(j-i+1)^2 - 1}{12}$$



# Cont.

- $i$  and  $j$  integers with  $i \leq j$ .

Mass

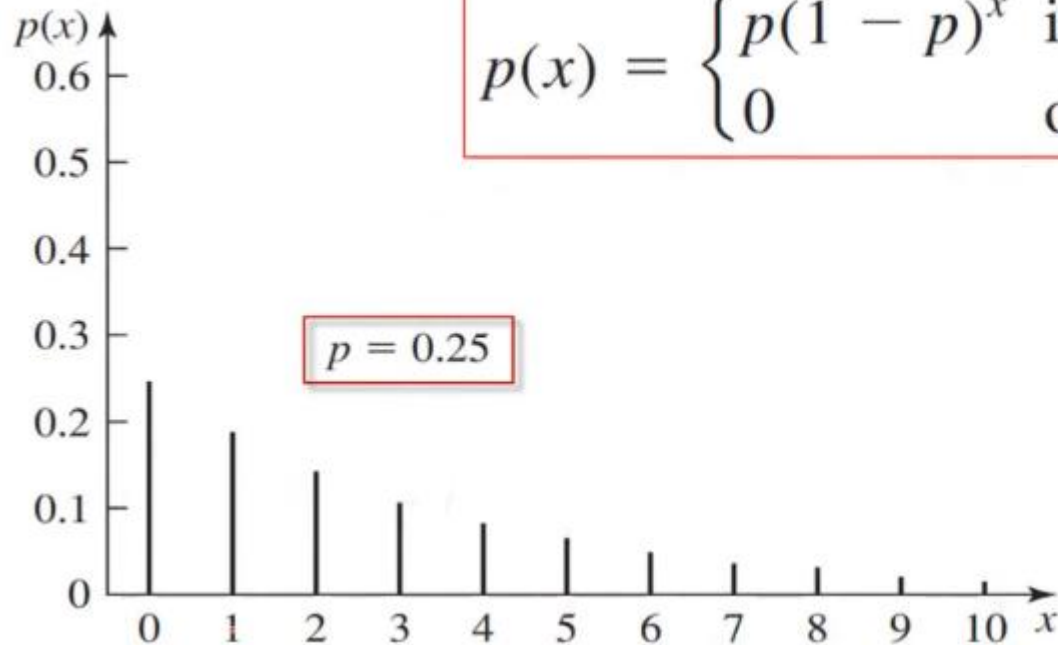
$$p(x) = \begin{cases} \frac{1}{j - i + 1} & \text{if } x \in \{i, i + 1, \dots, j\} \\ 0 & \text{otherwise} \end{cases}$$

Distribution

$$F(x) = \begin{cases} 0 & \text{if } x < i \\ \frac{\lfloor x \rfloor - i + 1}{j - i + 1} & \text{if } i \leq x \leq j \\ 1 & \text{if } j < x \end{cases}$$

# Geometric Distribution

- $p \in (0, 1)$ .

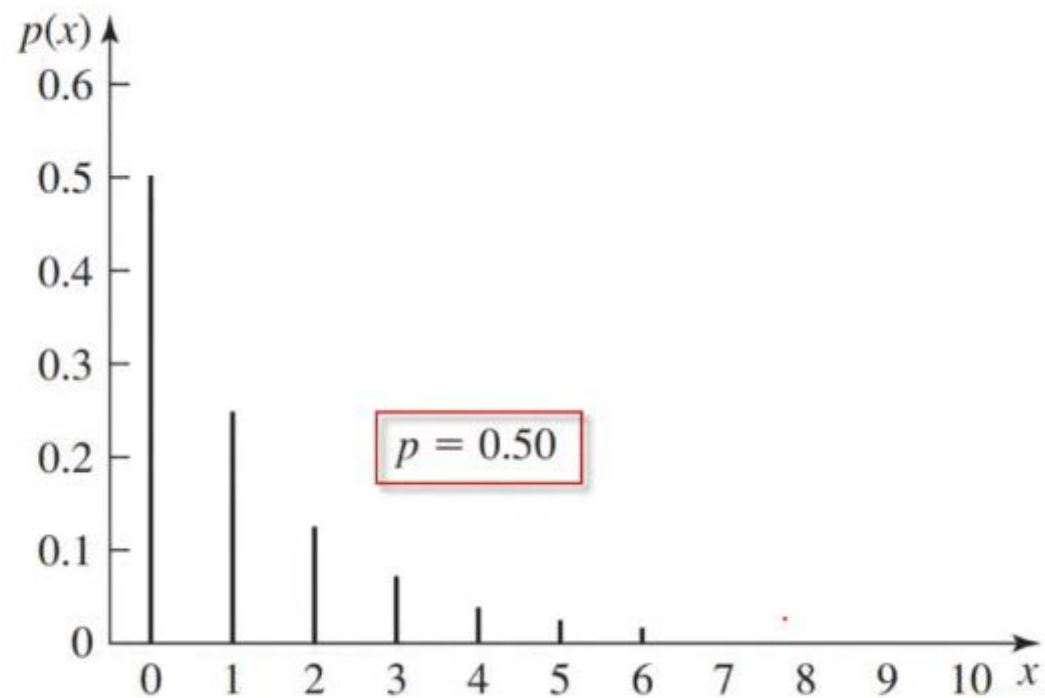


$$p(x) = \begin{cases} p(1 - p)^x & \text{if } x \in \{0, 1, \dots\} \\ 0 & \text{otherwise} \end{cases}$$

Mean	$\frac{1 - p}{p}$
Variance	$\frac{1 - p}{p^2}$
Mode	0

# Geometric Graph

- $p \in (0, 1)$ .



# Cont.

- $p \in (0, 1)$ .

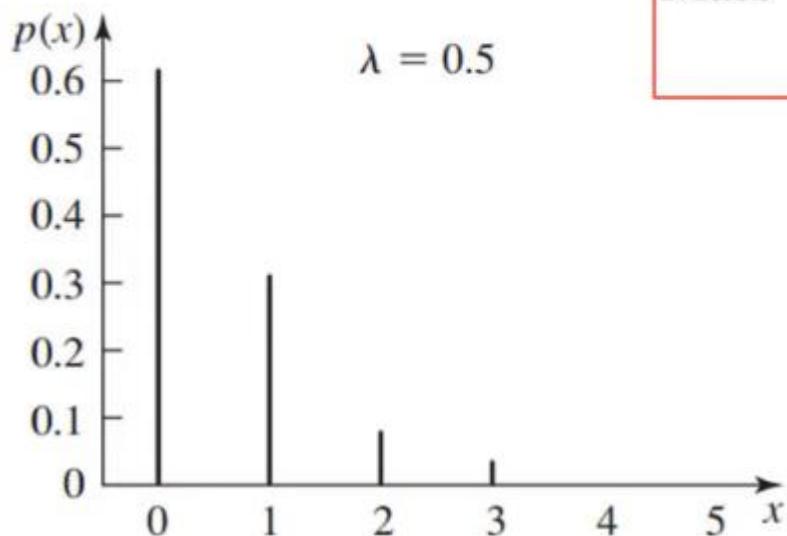
Mass  $p(x) = \begin{cases} p(1 - p)^x & \text{if } x \in \{0, 1, \dots\} \\ 0 & \text{otherwise} \end{cases}$

Distribution  $F(x) = \begin{cases} 1 - (1 - p)^{\lfloor x \rfloor + 1} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$

PRN  $F^{-1}(u) = \lfloor \ln u / \ln (1 - p) \rfloor$ . **Generating Random Variates  $X$ 's**

# Poisson Distribution

- $\lambda > 0$ .



$$\text{Mass } p(x) = \begin{cases} \frac{e^{-\lambda} \lambda^x}{x!} & \text{if } x \in \{0, 1, \dots\} \\ 0 & \text{otherwise} \end{cases}$$

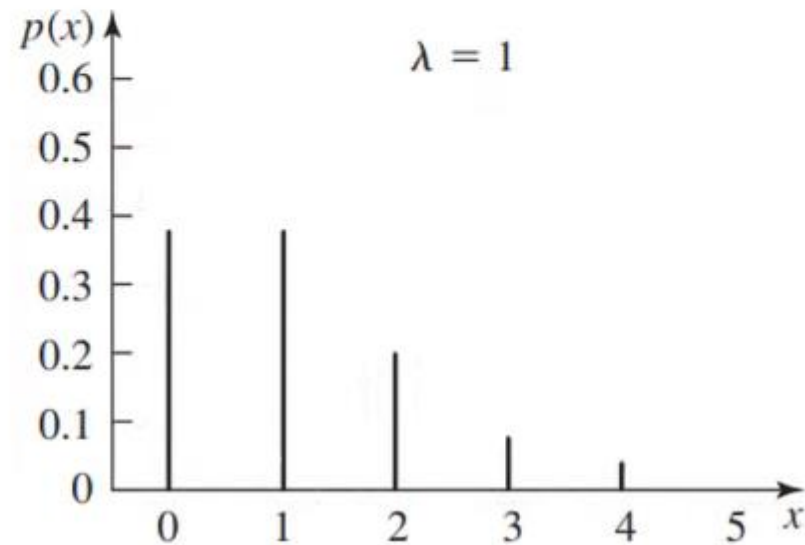
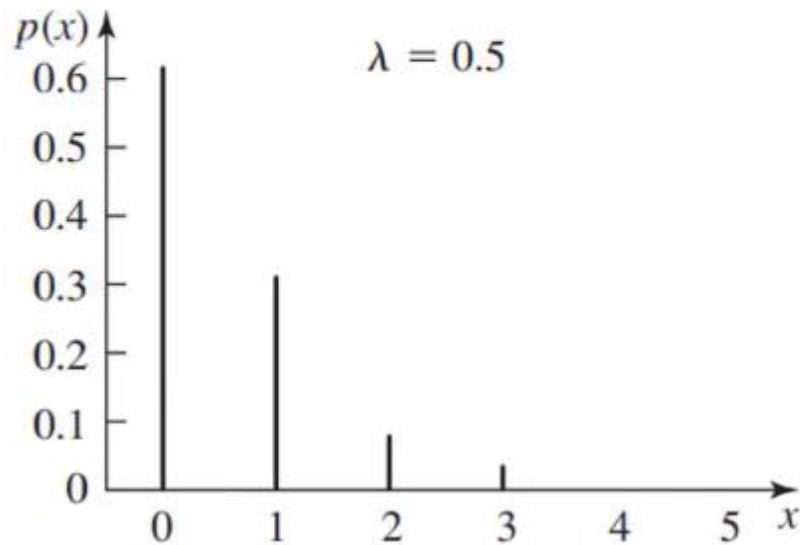
Mean  $\lambda$

Variance  $\lambda$

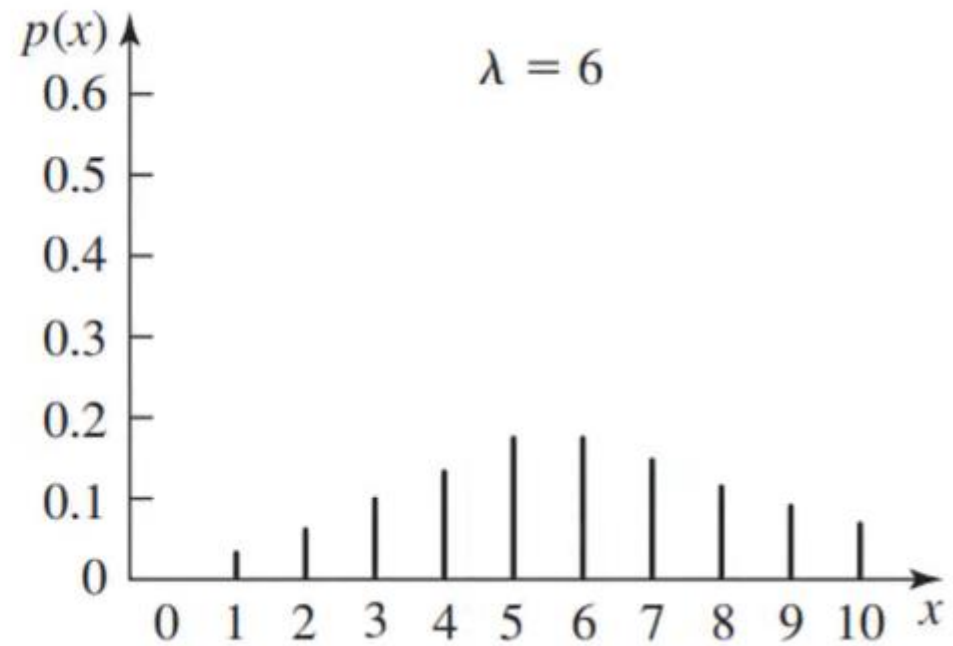
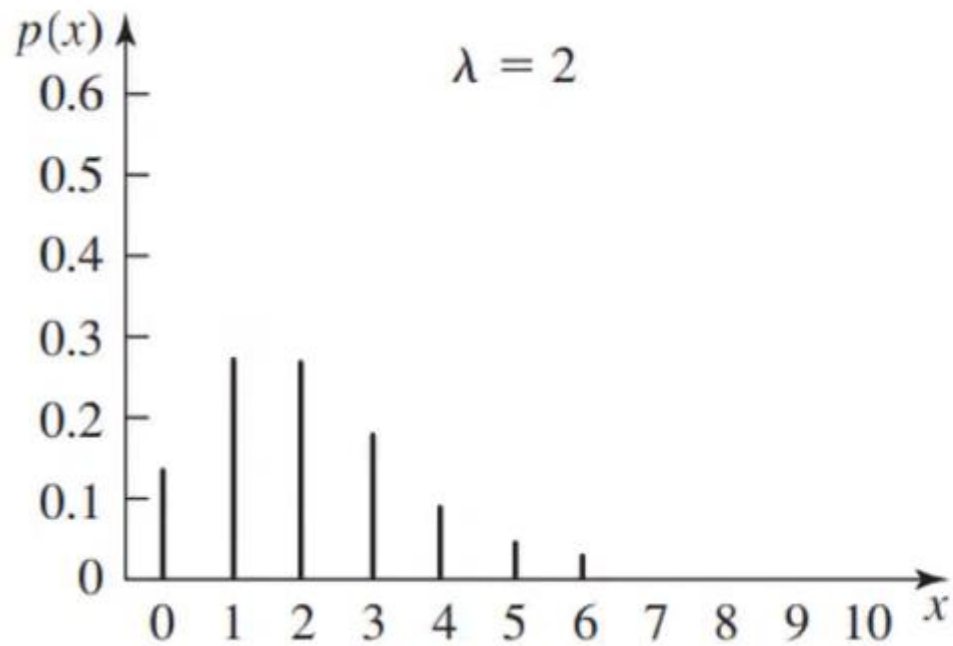
Mode  $\begin{cases} \lambda - 1 \text{ and } \lambda & \text{if } \lambda \text{ is an integer} \\ \lfloor \lambda \rfloor & \text{otherwise} \end{cases}$

# Poisson Graph 1

- $\lambda > 0$ .



# Poisson Graph 2





# Checking the distribution

- After collecting the data and selecting the standard probability distribution, you need to check if the selected distribution is appropriate and gives good fit of your data.
  - This can be done either:
    1. Graphically using graph comparison
    2. Using statistical tests such as K-S test, Chi-square test, etc.
-





# Selecting The Distribution

- But how to select the most appropriate theoretical, standard distribution for your data?
  - There are several methods to use to select a standard distribution for your data:
    1. Summary Statistics
    2. Histograms
    3. Quantile- Quantile Plot
    4. Boxplot
    5. And other methods
-

# Summary Statistics

Function	Sample estimate (summary statistic)	Continuous (C) or discrete (D)
Minimum, maximum	$X_{(1)}, X_{(n)}$	C, D
Mean $\mu$	$\bar{X}(n)$	C, D
Median $x_{0.5}$	$\hat{x}_{0.5}(n) = \begin{cases} X_{((n+1)/2)} & \text{if } n \text{ is odd} \\ [X_{(n/2)} + X_{((n/2)+1)}]/2 & \text{if } n \text{ is even} \end{cases}$	C, D
Variance $\sigma^2$	$S^2(n) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$	C, D

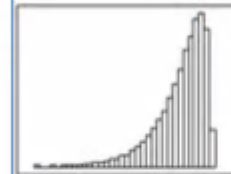
# Cont.

Function	Sample estimate	Continuous (C) or discrete (D)
Coefficient of variation, $cv = \frac{\sqrt{\sigma^2}}{\mu}$	$\widehat{cv}(n) = \frac{\sqrt{S^2(n)}}{\bar{X}(n)}$	C
Lexis ratio, $\tau = \frac{\sigma^2}{\mu}$	$\hat{\tau}(n) = \frac{S^2(n)}{\bar{X}(n)}$	D
Skewness, $\nu = \frac{E[(X - \mu)^3]}{(\sigma^2)^{3/2}}$	$\hat{\nu}(n) = \frac{n^2}{(n-1)(n-2)} \frac{\sum_{i=1}^n [X_i - \bar{X}(n)]^3 / n}{[S^2(n)]^{3/2}}$	C, D

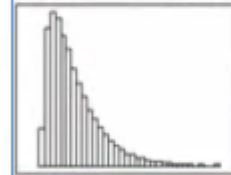
# Cont.



Symmetric  
Bell shaped



Skewed to  
the Left



Skewed to  
the Right

$$\text{Skewness, } \nu = \frac{E[(X - \mu)^3]}{(\sigma^2)^{3/2}} \quad \hat{\nu}(n) = \frac{n^2}{(n-1)(n-2)} \frac{\sum_{i=1}^n [X_i - \bar{X}(n)]^3 / n}{[S^2(n)]^{3/2}} \quad \text{C, D}$$



# Example

- If the median is equal or near equal to the mean, this indicates symmetric, (e.g., normal) distribution.
  - If the coefficient of variation (cv) is close to 1 this indicates exponential distribution because its cv is 1.
-



# Example

- A simulation model was developed for a drive-up banking facility, and data were collected on the arrival pattern for cars. Over a fixed 90-minute interval, 220 cars arrived, and we noted the (continuous) interarrival time  $X_i$  (in minutes) between cars  $i$  and  $i + 1$ , for  $i = 1, 2, \dots, 219$ .
-

**$n = 219$  interarrival times (minutes) sorted into increasing order**

0.01	0.06	0.12	0.23	0.38	0.53	0.88
0.01	0.07	0.12	0.23	0.38	0.53	0.88
0.01	0.07	0.12	0.24	0.38	0.54	0.90
0.01	0.07	0.13	0.25	0.39	0.54	0.93
0.01	0.07	0.13	0.25	0.40	0.55	0.93
0.01	0.07	0.14	0.25	0.40	0.55	0.95
0.01	0.07	0.14	0.25	0.41	0.56	0.97
0.01	0.07	0.14	0.25	0.41	0.57	1.03
0.02	0.07	0.14	0.26	0.43	0.57	1.05
0.02	0.07	0.15	0.26	0.43	0.60	1.05
0.03	0.07	0.15	0.26	0.43	0.61	1.06
0.03	0.08	0.15	0.26	0.44	0.61	1.09
0.03	0.08	0.15	0.26	0.45	0.63	1.10
0.04	0.08	0.15	0.27	0.45	0.63	1.11
0.04	0.08	0.15	0.28	0.46	0.64	1.12
0.04	0.09	0.17	0.28	0.47	0.65	1.17
0.04	0.09	0.18	0.29	0.47	0.65	1.18
0.04	0.10	0.19	0.29	0.47	0.65	1.24
0.04	0.10	0.19	0.30	0.48	0.69	1.24
0.05	0.10	0.19	0.31	0.49	0.69	1.28
0.05	0.10	0.20	0.31	0.49	0.70	1.33
0.05	0.10	0.21	0.32	0.49	0.72	1.38
0.05	0.10	0.21	0.35	0.49	0.72	1.44
0.05	0.10	0.21	0.35	0.50	0.72	1.51
0.05	0.10	0.21	0.35	0.50	0.74	1.72
0.05	0.10	0.21	0.36	0.50	0.75	1.83
0.05	0.11	0.22	0.36	0.51	0.76	1.96
0.05	0.11	0.22	0.36	0.51	0.77	
0.05	0.11	0.22	0.37	0.51	0.79	
0.06	0.11	0.23	0.37	0.52	0.84	
0.06	0.11	0.23	0.38	0.52	0.86	
0.06	0.12	0.23	0.38	0.53	0.87	

# Cont.

## Summary statistics for the interarrival-time data

Summary statistic	Value
Minimum	0.010
Maximum	1.960
Mean	0.399
Median	0.270
Variance	0.144
Coefficient of variation	0.953
Skewness	1.478



# Cont.

Since:

$\bar{X}(219) = 0.399 > 0.270 = \hat{x}_{0.5}(219)$   
and  $\hat{v}(219) = 1.478$ , this suggests that the underlying distribution is skewed to the right, rather than symmetric.

Furthermore,  $\widehat{cv}(219) = 0.953$ , which is close to the theoretical value of 1 for the *exponential* distribution.

Summary statistics for the interarrival-time data

Summary statistic	Value
Minimum	0.010
Maximum	1.960
Mean	0.399
Median	0.270
Variance	0.144
Coefficient of variation	0.953
Skewness	1.478

# Histograms

- To make a histogram, we break up the range of values covered by the data into  $k$  disjoint adjacent intervals  $[b_0, b_1), [b_1, b_2), \dots, [b_{k-1}, b_k)$ .
- All the intervals should be the same width  $\Delta b$ , which might necessitate throwing out a few extremely large or small  $X_i$ 's to avoid getting an unwieldy-looking histogram plot.

$$\Delta b = \frac{\max(x) - \min(x)}{k}$$

$$k = \frac{\max(x) - \min(x)}{\Delta b}$$

- Selecting the best  $k$  or  $\Delta b$  is trial and error process, although there are some rules that helps to approximate the value of  $k$

$$k = \lfloor 1 + \log_2 n \rfloor = \lfloor 1 + 3.322 \log_{10} n \rfloor$$

**$n = 219$  interarrival times (minutes) sorted into increasing order**

0.01	0.06	0.12	0.23	0.38	0.53	0.88
0.01	0.07	0.12	0.23	0.38	0.53	0.88
0.01	0.07	0.12	0.24	0.38	0.54	0.90
0.01	0.07	0.13	0.25	0.39	0.54	0.93
0.01	0.07	0.13	0.25	0.40	0.55	0.93
0.01	0.07	0.14	0.25	0.40	0.55	0.95
0.01	0.07	0.14	0.25	0.41	0.56	0.97
0.01	0.07	0.14	0.25	0.41	0.57	1.03
0.02	0.07	0.14	0.26	0.43	0.57	1.05
0.02	0.07	0.15	0.26	0.43	0.60	1.05
0.03	0.07	0.15	0.26	0.43	0.61	1.06
0.03	0.08	0.15	0.26	0.44	0.61	1.09
0.03	0.08	0.15	0.26	0.45	0.63	1.10
0.04	0.08	0.15	0.27	0.45	0.63	1.11
0.04	0.08	0.15	0.28	0.46	0.64	1.12
0.04	0.09	0.17	0.28	0.47	0.65	1.17
0.04	0.09	0.18	0.29	0.47	0.65	1.18
0.04	0.10	0.19	0.29	0.47	0.65	1.24
0.04	0.10	0.19	0.30	0.48	0.69	1.24
0.05	0.10	0.19	0.31	0.49	0.69	1.28
0.05	0.10	0.20	0.31	0.49	0.70	1.33
0.05	0.10	0.21	0.32	0.49	0.72	1.38
0.05	0.10	0.21	0.35	0.49	0.72	1.44
0.05	0.10	0.21	0.35	0.50	0.72	1.51
0.05	0.10	0.21	0.35	0.50	0.74	1.72
0.05	0.10	0.21	0.36	0.50	0.75	1.83
0.05	0.11	0.22	0.36	0.51	0.76	1.96
0.05	0.11	0.22	0.36	0.51	0.77	
0.05	0.11	0.22	0.37	0.51	0.79	
0.06	0.11	0.23	0.37	0.52	0.84	
0.06	0.11	0.23	0.38	0.52	0.86	
0.06	0.12	0.23	0.38	0.53	0.87	

# Cont.

- Let us use the previous formula to find  $k$

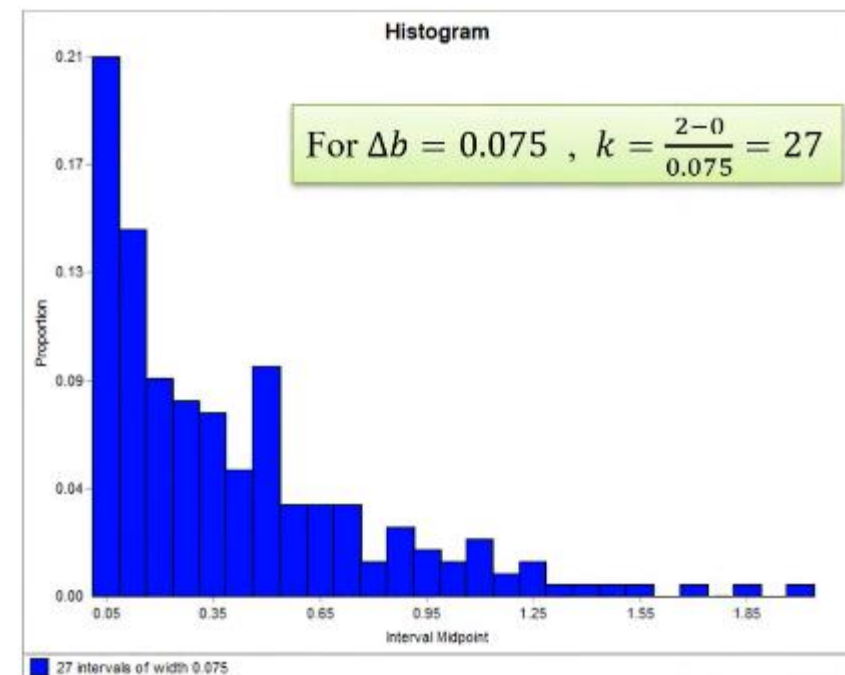
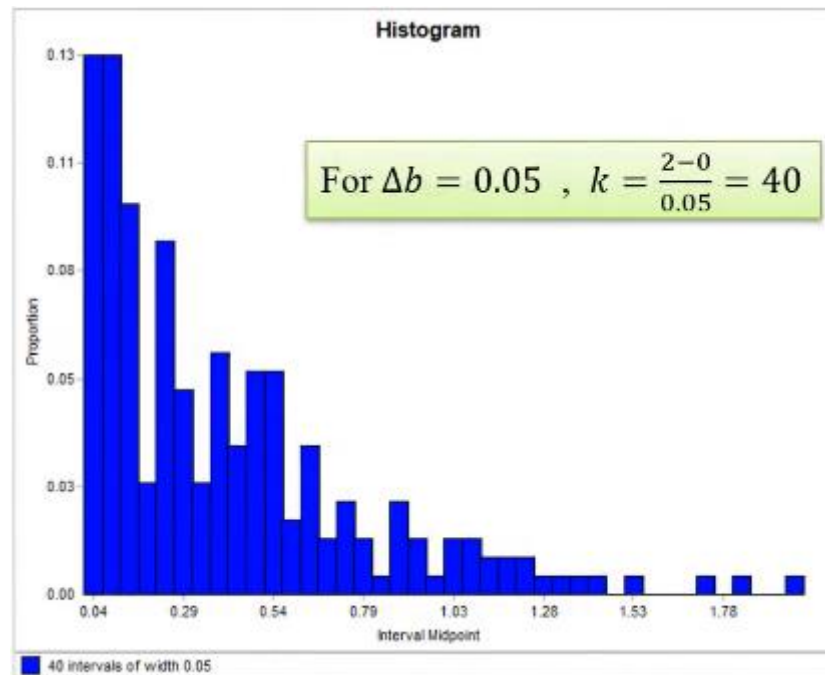
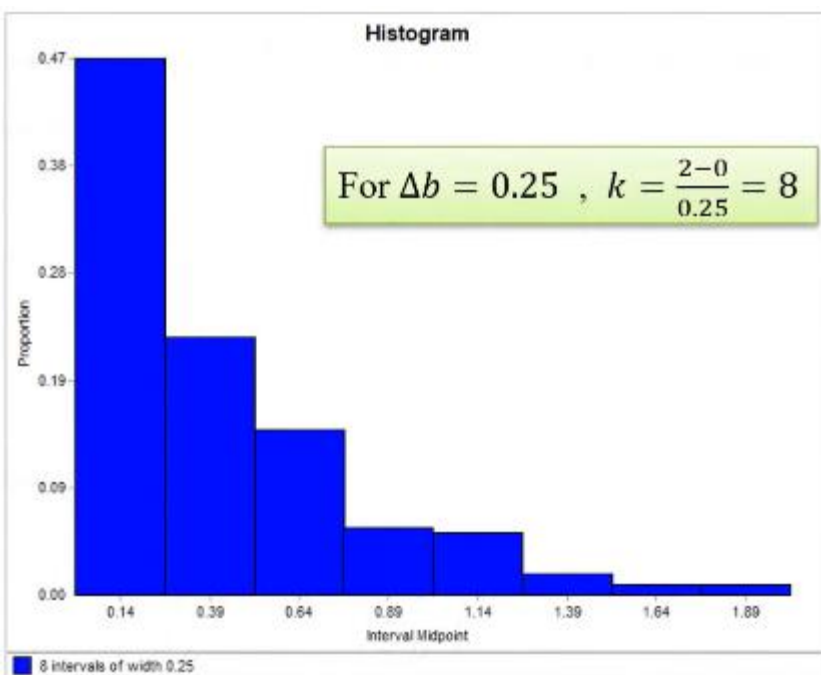
$$k = \lfloor 1 + \log_2 n \rfloor = \lfloor 1 + 3.322 \log_{10} n \rfloor$$

$$k = \lfloor 1 + \log_2 219 \rfloor = 8$$

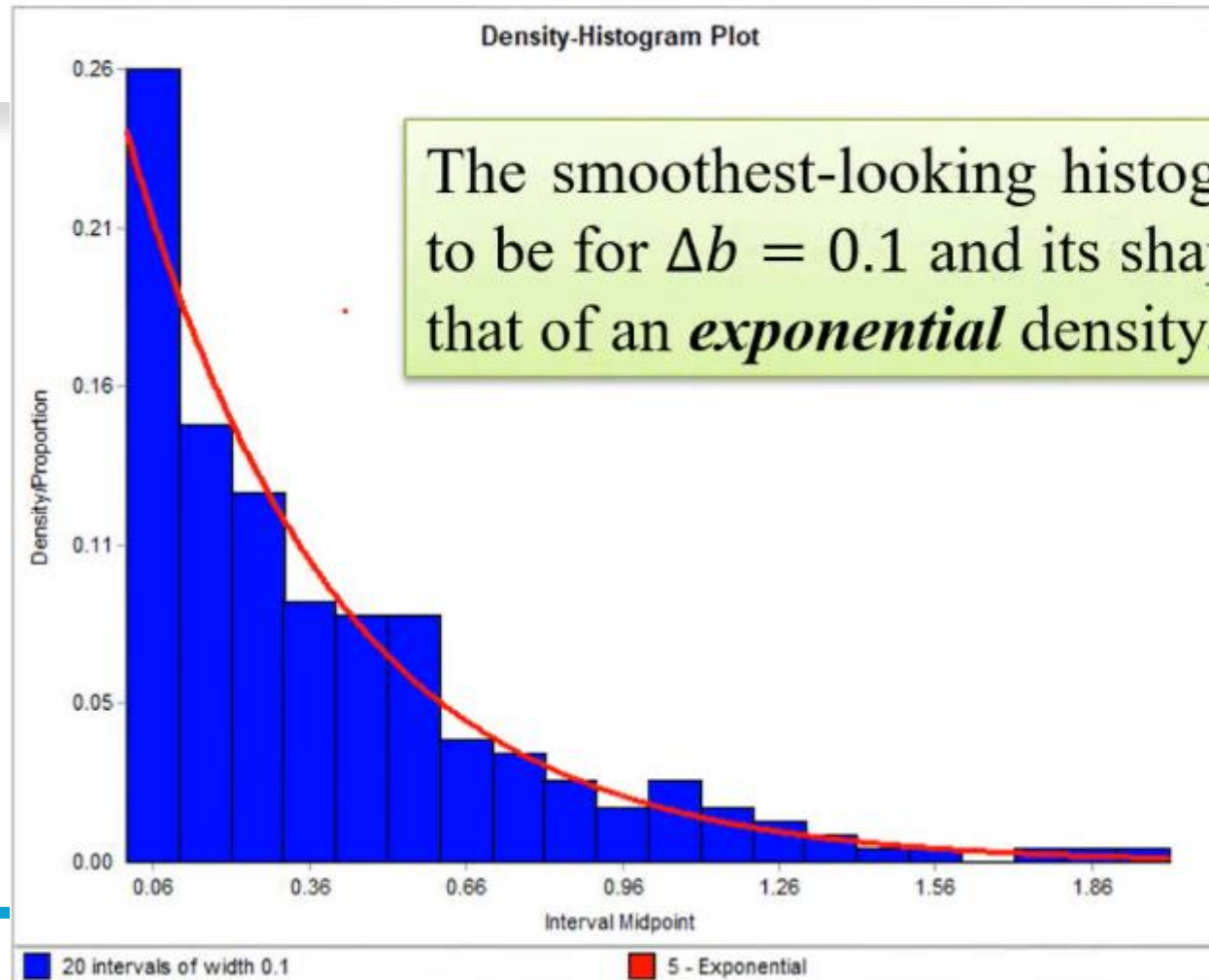
$$\min X = 0.01 \approx 0, \max X = 1.96 \approx 2, \quad \Delta b = \frac{2 - 0}{8} = 0.25$$

$$[0, 0.25), [0.25, 0.5), [0.5, 0.75), [0.75, 1), \\ [1, 1.25), [1.25, 1.5), [1.5, 1.75), [1.75, 2),$$

# Cont.



# Example



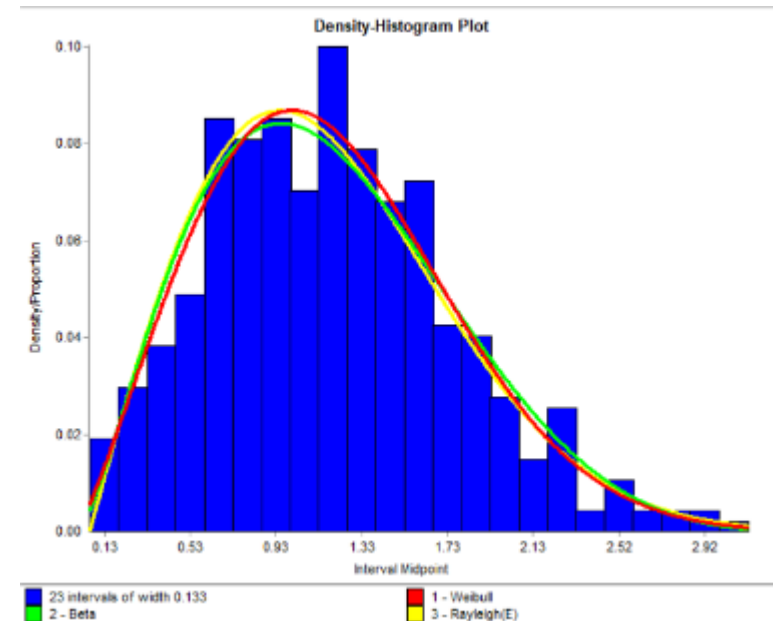
The smoothest-looking histogram appears to be for  $\Delta b = 0.1$  and its shape resembles that of an *exponential* density.



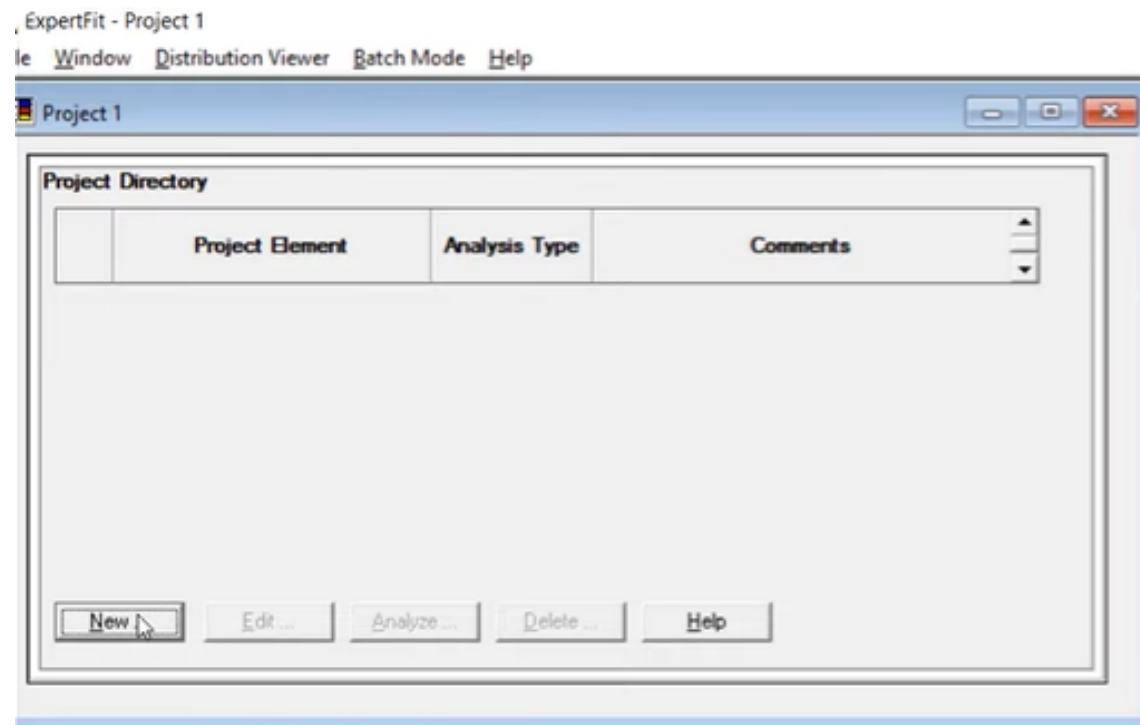
# ExpertFit

- There are several software that help in modeling a given data, making the selection of the appropriate distribution an easy task.
- ExpertFit is one of these software

<https://www.averill-law.com/distribution-fitting/>



# Cont.






# Cont.

Project-Element Editing

Project-Element Name:  
P1

Analysis Type

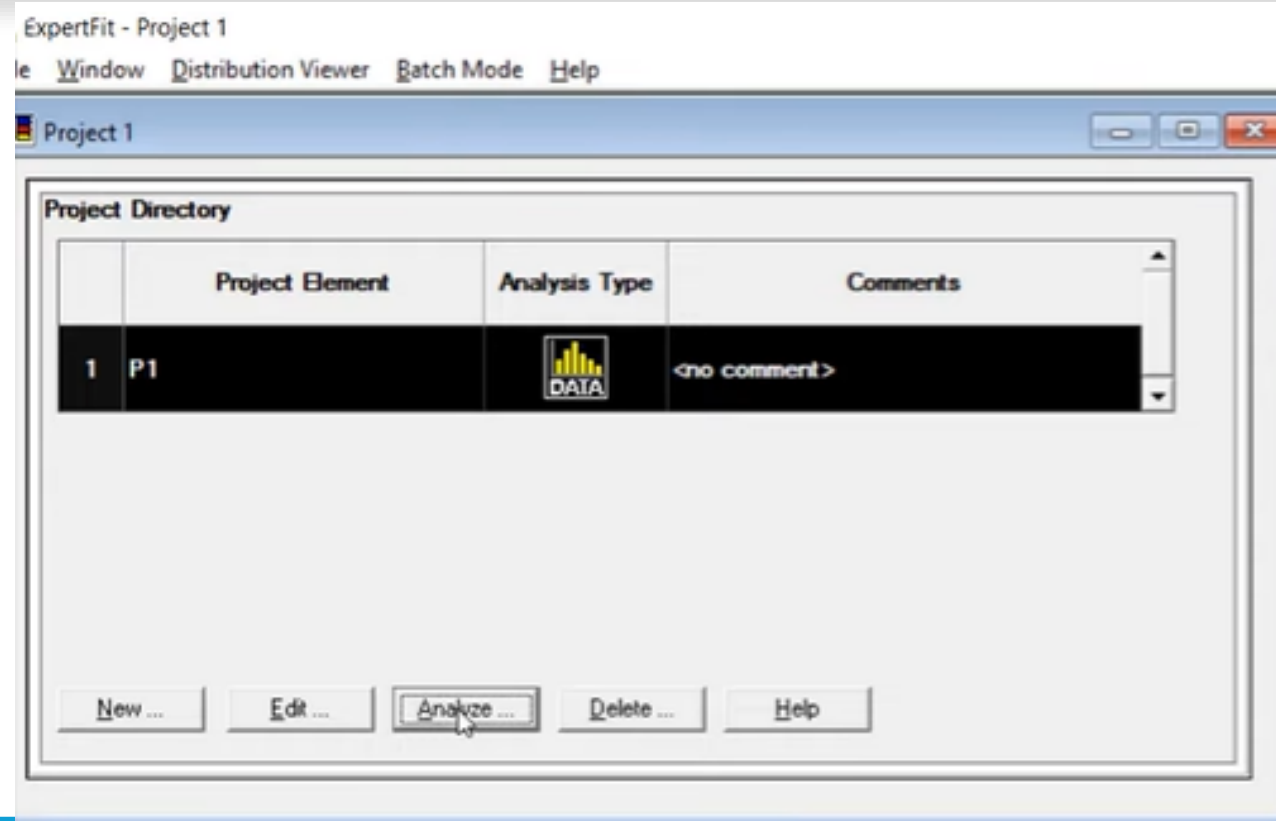
 ☒ Fit distributions to data

 ☐ Construct distributions in the absence of data

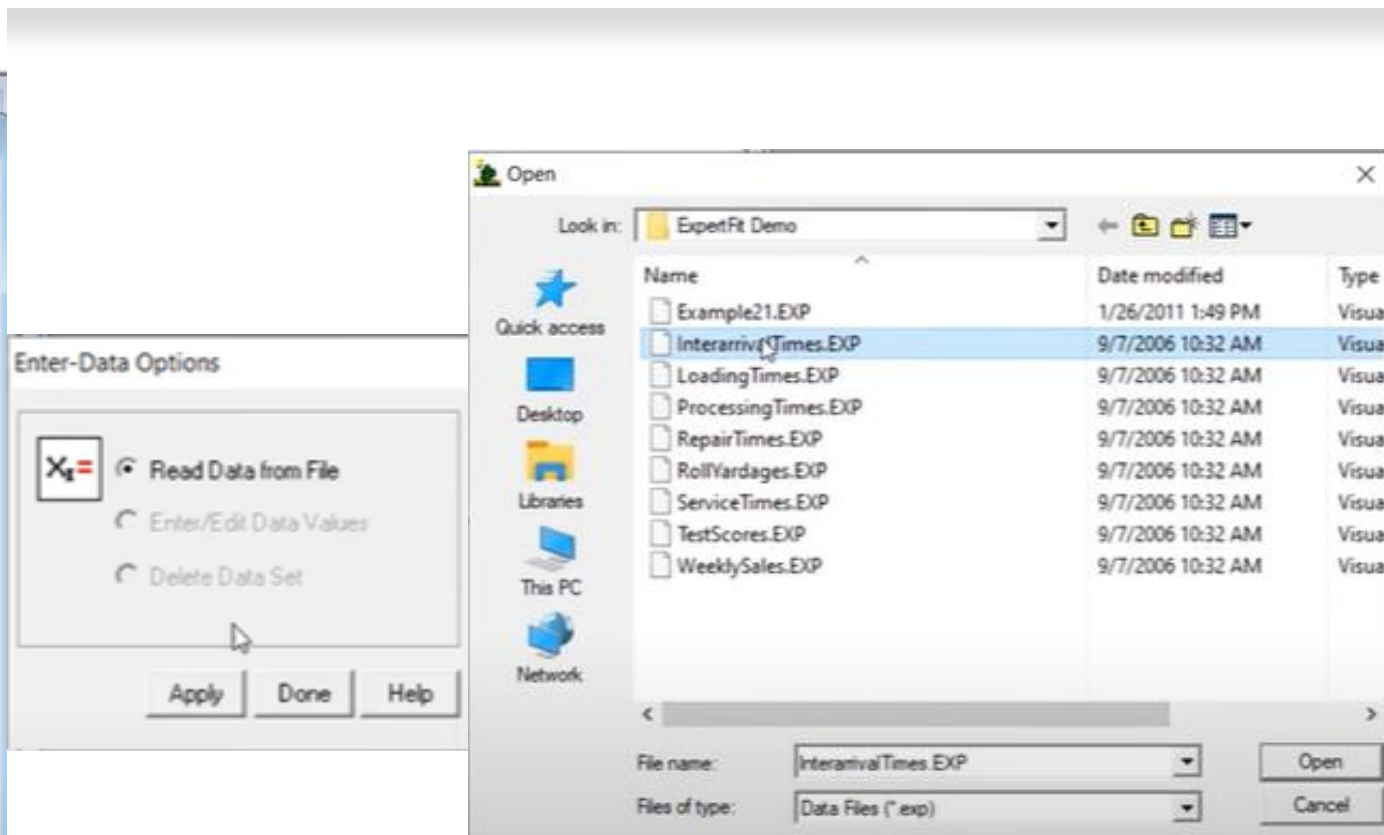
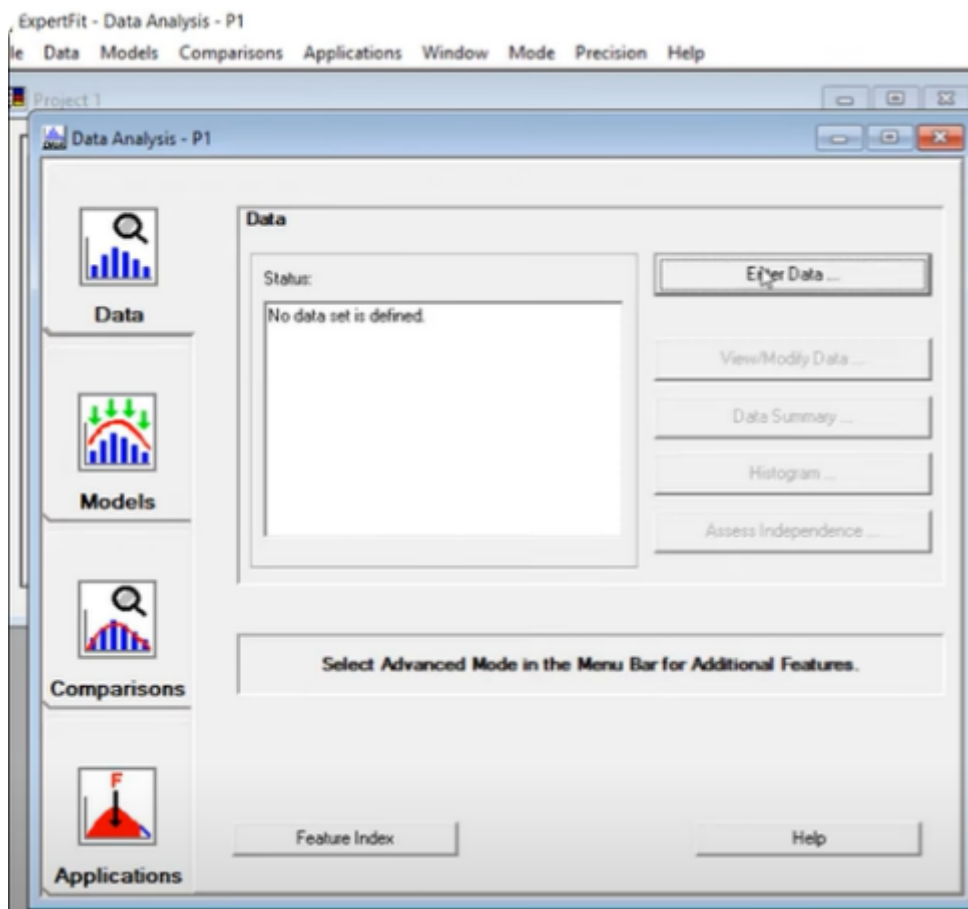
Comments:  
<no comment>

Help  
Cancel  
OK

# Cont.



# Cont.



# Cont.

Data-Summary Table

Data Characteristic	Value
Source file	InterarrivalTimes
Observation type	Real valued
Number of observations	219
Minimum observation	0.01000
Maximum observation	1.96000
Mean	0.39881
Median	0.27000
Variance	0.14439
Coefficient of variation	0.95281
Skewness	1.47845

Copy  
Print  
Help  
Done

Project 1

Data Analysis - P1

Data

Status:

219 real values  
Minimum value 0.01  
Maximum value 1.96  
Mean value 0.39881

Source file name InterarrivalTimes

Enter Data ...

View/Modify Data ...

Data Summary ...

Histogram ...

Assess Independence ...

Select Advanced Mode in the Menu Bar for Additional Features.

Feature Index

Help

View/Modify Options

☒ View Data

☒ Sort by Value

☒ Modify Data

☐ Create a Subset

☐ Perform a Transformation

Apply Done Help

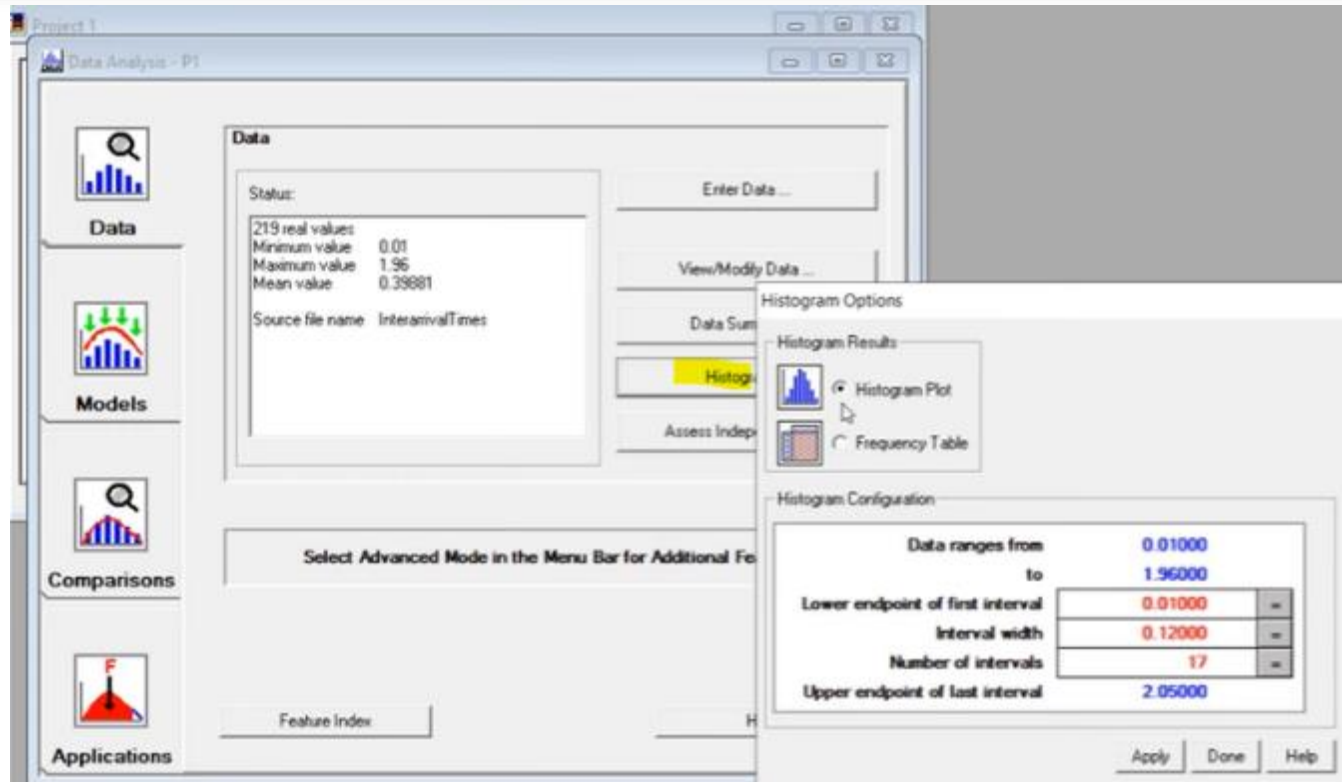
# Cont.

Listing of Data Values

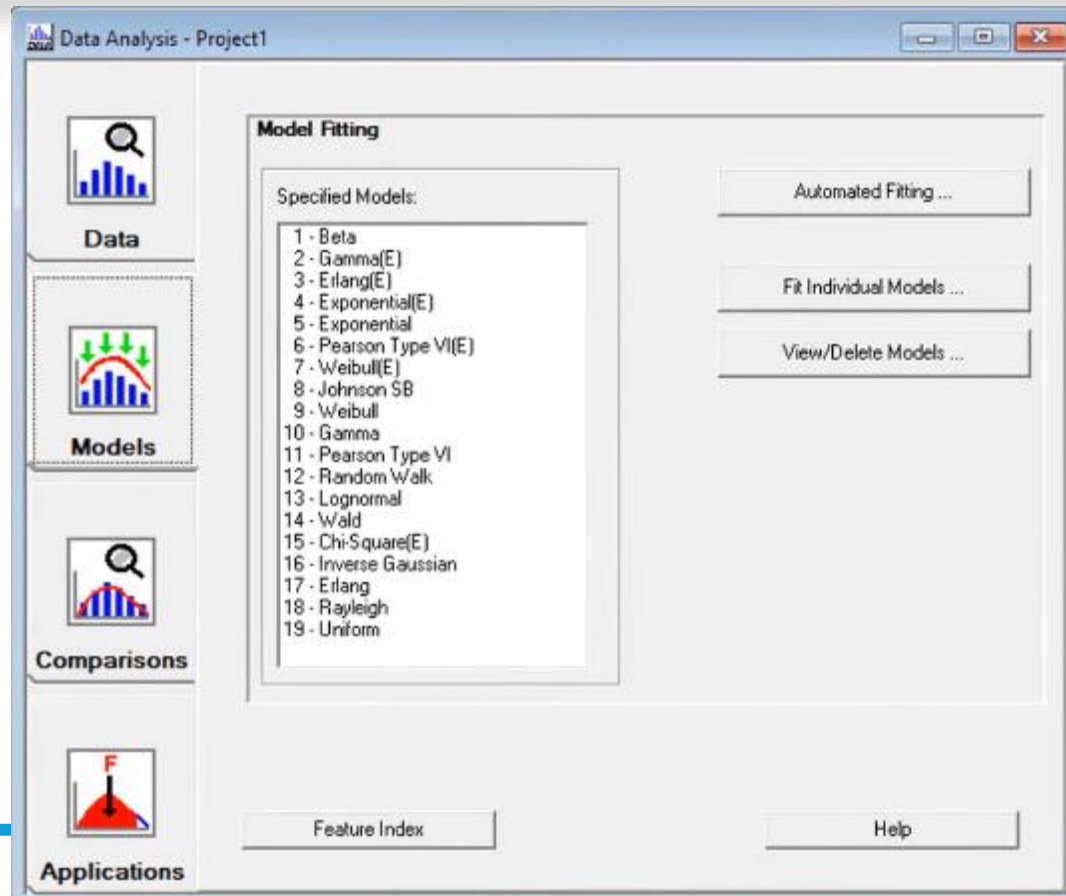
92	0.22000	202	1.05000
93	0.22000	203	1.06000
94	0.23000	204	1.09000
95	0.23000	205	1.10000
96	0.23000	206	1.11000
97	0.23000	207	1.12000
98	0.23000	208	1.17000
99	0.24000	209	1.18000
100	0.25000	210	1.24000
101	0.25000	211	1.24000
102	0.25000	212	1.28000
103	0.25000	213	1.33000
104	0.25000	214	1.38000
105	0.26000	215	1.44000
106	0.26000	216	1.51000
107	0.26000	217	1.72000
108	0.26000	218	1.83000
109	0.26000	219	1.96000
110	0.27000		

Copy  
Print  
Help  
Done

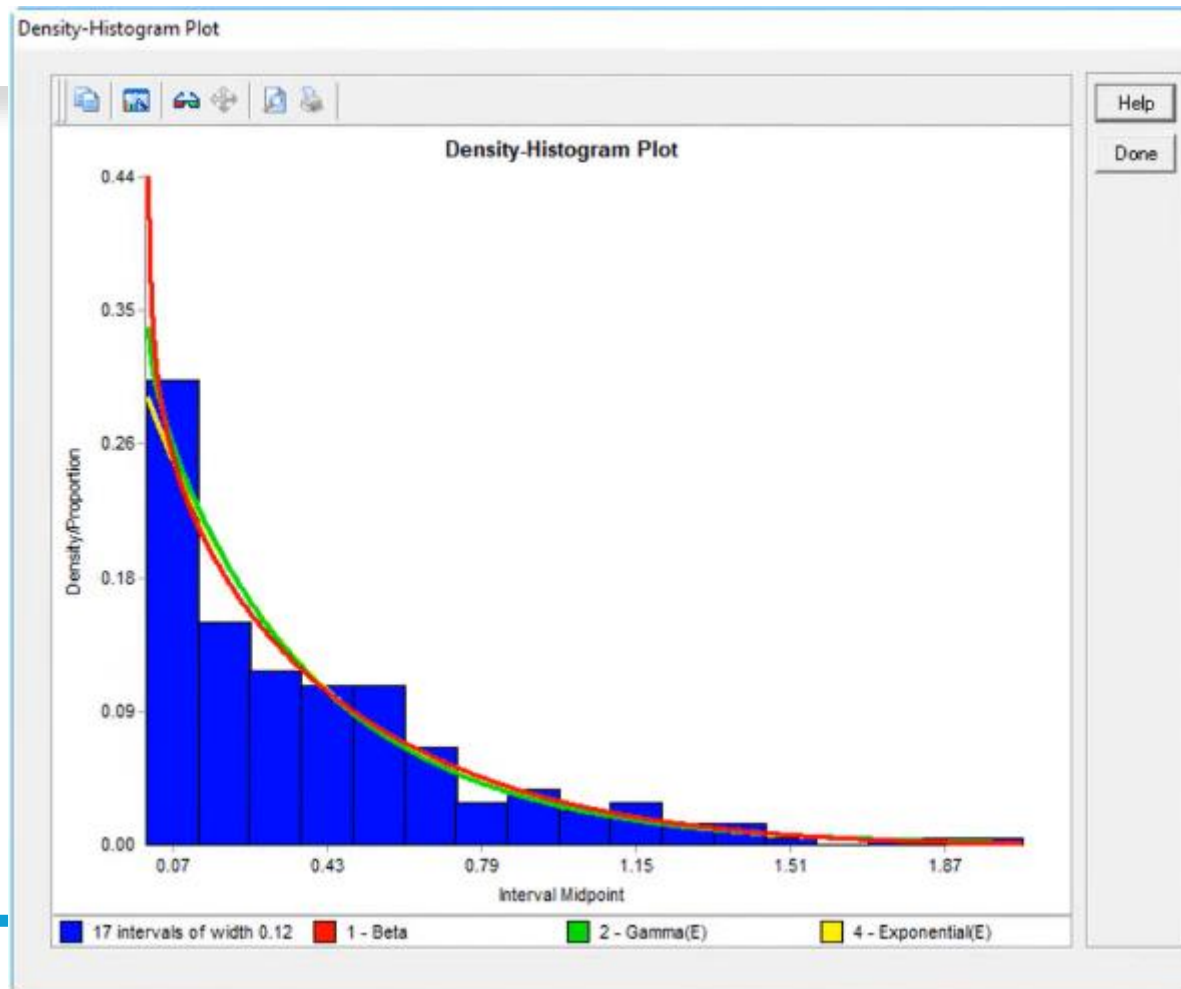
# Cont.



# Cont.



# Cont.





# Cont.

Options for Goodness-of-Fit Tests

Goodness-of-Fit Tests

☒ A-D Anderson-Darling Test

☒ K-S Kolmogorov-Smirnov Test

☒  $\chi^2$  Chi-Square Test

Model for Tests

1 - Beta

Chi-Square Test Configuration

Number of intervals 40 =

Expected model count 5.47500

Apply Done Help

Anderson-Darling Test

Anderson-Darling Test with Model 1 - Beta

Sample size 219

Test statistic 0.44196

Note: No critical values exist for this special case.  
The following critical values are for the case where all parameters are known, and are conservative.

	Critical Values for Level of Significance (alpha)					
Sample Size	0.250	0.100	0.050	0.025	0.010	0.005
219	1.248	1.933	2.492	3.070	3.857	4.500
Reject?	No					

Copy

Print

Help

Done

# Cont.

Simulation-Software Representation

☒ AnyLogic Representation of Model 1 - Beta

Use:

beta(0.855247, 7.048525, 0.007073, 3.627949)

Copy  
Print  
Help  
Done

Simulation-Software Representation

☒ Flexsim Representation of Model 1 - Beta

Use:

When using a picklist option:	
Distribution	Beta
Minimum	0.007073
Maximum	3.627949
Shape1	0.855247
Shape2	7.048525

When using code:  
beta( 0.007073, 3.627949, 0.855247, 7.048525, <stream>)



Cont.

**Thank you!**

**Hope you Enjoyed the course**

---