SHARIF
UNIVERSITY OF
TECHNOLOGY

Computer Simulation

Dr. Bardia Safaei

Chapter Eight: Input Modeling

Purpose and Overview

SHARIF UNIVERSITY OF TECHNOLOGY

- Input modeling indicates the **preparation of a sequence of inputs** to be fed into our simulation model
- The quality of the output is no better than the quality of inputs
  - So, make sure to select good distributions for your inputs
  - Inappropriate and erroneous inputs results in misleading outputs
- In this chapter, we will discuss the 4 steps of input model development:
  - Collect data from the real system
  - Identify a probability distribution which represents the collected data
  - Choose the correct parameters for the distribution
  - Evaluate the chosen distribution and parameters for goodness of fit

Data Collection (1)

SHARIF UNIVERSITY OF TECHNOLOGY

- Data collection is the most **pivotal** task among 4 steps
- What should we care through data collection?
  - Stale data
    - Do not use old data, e.g., do not use data collected in summer to model number of patients in ICUs in winter
    - If the intended system is operating the same as it did 100 years ago, you can still use old collected data
    - Never forget: **New data is always preferrable**
  - Coping with unexpected data
    - Due to various reasons, we may face out of bound data while collecting data, e.g., error in measurements, tools, etc.
    - Do not use these data along with others, because:
      - It may complicate the process of finding the right distribution and model
      - Results in fake outputs

Data Collection (2)

SHARIF UNIVERSITY OF TECHNOLOGY

- What should we care through data collection? (Cont.)
  - Time-variation of data
    - During the collection process, behavior of data may change as time passes
      - For instance in Non-Stationary Poisson Processes (NSPP), where the rate of arrival is not constant
    - In such cases, the simulation shall not be done via a stationary model, and the model should be dynamic
  - Dependency of data
    - The dependency between chunks of data **MUST** be determined beforehand
    - If we use dependent input data, and do the analysis based on an independency assumption, our evaluation and conclusions will be incorrect
      - We must determine the dependency, and consider it in the modeling, unless we should use independent data

Data Collection (3)

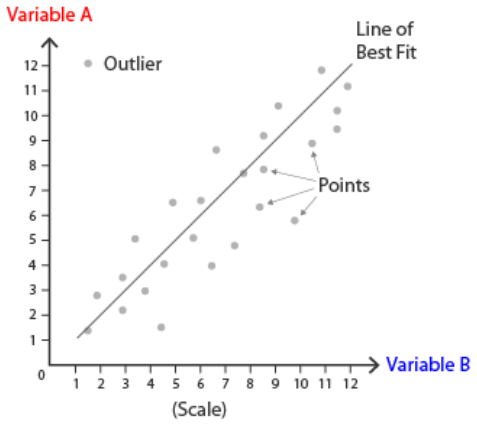
SHARIF UNIVERSITY OF TECHNOLOGY

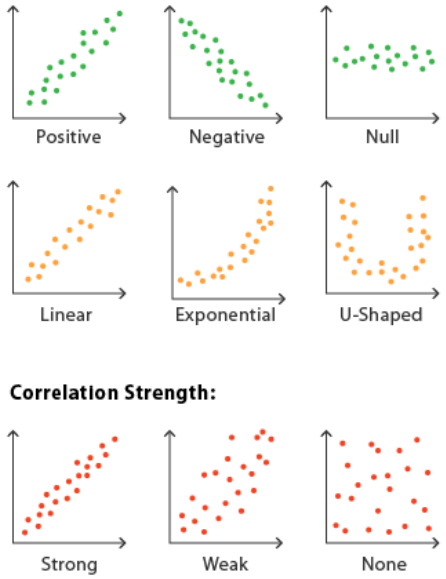
- So, be aware of GIGO (Garbage-In-Garbage-Out)
- Suggestions that may enhance and facilitate a correct data collection process include:
  - Plan ahead, and begin by a short **practice** or **pre-observing session**, watch for any unusual behaviors
  - Analyze the data as it is being collected not after completion
  - Combine **homogeneous data sets**
    - Example: Successive time periods
      - Check phone calls between 14:00 to 15:00, and 15:00 to 16:00 in a day
    - Or during the same time period on successive points of time
      - Number of parked vehicles in an intersection on Tuesday in 2 successive weeks
    - This is completely dependent on the target system
      - Example: number of workers enter a hypermarket at the beginning, end or middle of the month

Data Collection (4)

SHARIF UNIVERSITY OF TECHNOLOGY

- Suggestions that may enhance and facilitate a correct data collection process include (Cont.):
  - Be aware of data censoring
    - The simulation period must be in accordance with the type of data we want to collect, and also the parameter we want to evaluate
  - Check for **relationship** between variables
    - Use **scatter diagrams** to identify the relationship
      - Checkout Fig. 2 in page 335
  - Check for autocorrelation
  - Collect input data
    - Not performance data





Identifying the Distribution

SHARIF UNIVERSITY OF TECHNOLOGY

- Post to data collection
  - We need to identify which families of distributions suits best
  - Parameter estimation
  - Goodness-of-fit tests
    - We may fit a non-stationary process
    - We will discuss this later in this chapter
- From now on, we assume that a set of i.i.d data has been collected
- Using the **frequency distribution** is an appealing approach to guess and identify the shape of the distribution
  - Histograms

Histogram (1)

SHARIF UNIVERSITY OF TECHNOLOGY

- A frequency distribution or histogram is useful in determining the shape of a distribution
- To create a histogram, 5 following steps must be applied:
  - Divide the range of data into intervals
    - **Usually** of equal length
    - We may also use variable lengths if we need to arrange the heights of frequencies
  - Label the horizontal axis to conform to the intervals selected
  - Find the frequency of occurrences within each interval
  - Label the vertical axis so that the total occurrences can be plotted for each interval
  - Plot the frequencies on the vertical axis

Histogram (2)

SHARIF UNIVERSITY OF TECHNOLOGY

- The number of class intervals depends on:
  - The number of observations
  - The dispersion of the data
  - Suggested: the **square root of the sample size**
- For continuous data:
  - The obtained histogram corresponds to the **probability density function** of a theoretical distribution
- For discrete data:
  - It corresponds to the **probability mass function**
- If few data points are available: combine adjacent cells to eliminate the ragged appearance of the histogram

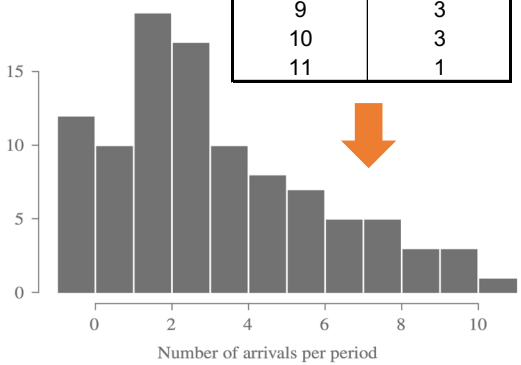
Vehicle Arrival Example

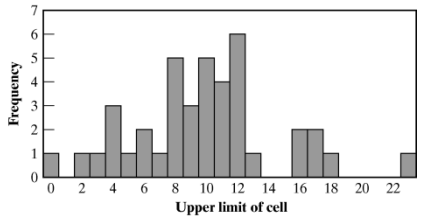
SHARIF UNIVERSITY OF TECHNOLOGY

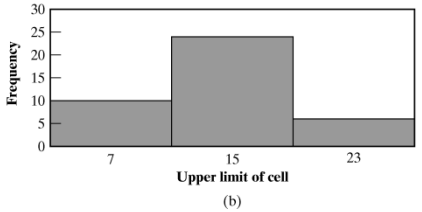
- Number of vehicles arriving at an intersection between 7:00 AM and 7:05 AM was monitored for 100 random workdays
- Collected data is few
  - The histogram may get ragged (~1st figure)
  - We need to use intervals (~2nd and 3rd figures)
    - 2nd figure is not eye-catching
    - Try to resolve the issue by increasing the intervals to some extent
- Ultimate figure has more similarity with known distributions

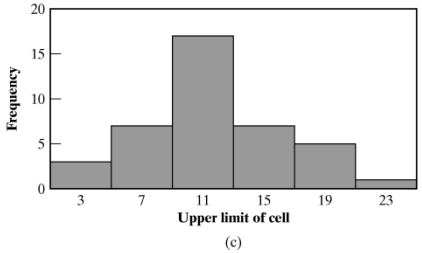
Same data with different interval sizes

Arrivals per Period	Frequency
0	12
1	10
2	19
3	17
4	10
5	8
6	7
7	5
8	5
9	3
10	3
11	1









Selecting the Family of Distributions (1)

SHARIF UNIVERSITY OF TECHNOLOGY

- A family of distributions is selected based on:
  - The context of the **input variable** and the **type of the application**
  - Shape of the histogram
- Frequently encountered distributions:
  - Easier to analyze: exponential, normal and Poisson

Continuous

Discrete

  - Harder to analyze: Beta, Gamma and Weibull
    - Keep in mind, they may have better flexibility
    - But, due to their complex equations, they are harder to utilize
      - Less interest (unless we require the **precision** provided by them)

Selecting the Family of Distributions (2)

SHARIF UNIVERSITY OF TECHNOLOGY

- Use the physical basis of the distribution as a guide, for example:
  - **Binomial:** Number of successes in n trials
  - **Poisson:** Number of independent events that occur in a fixed amount of time or space
  - **Normal:** Distribution of a process that is the sum of a number of component processes
  - **Exponential:** Time between independent events, or a process time that is memoryless
  - **Weibull:** Time to failure for components
  - **Discrete or continuous uniform:** Models complete uncertainty
  - **Triangular:** A process for which only the minimum, most likely, and maximum values are known
  - **Empirical:** Resamples from the actual data collected

Selecting the Family of Distributions (3)

SHARIF UNIVERSITY OF TECHNOLOGY

- To have a good guess about the distribution
  - Notice the physical characteristics of the process
  - Is the process naturally discrete or continuous valued?
  - Is it bounded?
    - Example: Normal or truncated normal?
  - Can it get positive or negative values?
- There is no **"true"** distribution for any stochastic input process
  - There is nothing regarded as 100% match
  - Goal: To obtain a good approximation

Quantile-Quantile Plots (1)

SHARIF UNIVERSITY OF TECHNOLOGY

- Q-Q plot is a useful tool for **evaluating the fit of the chosen distributions**
  - Histogram helps us to have a **first guess(s)** generally
  - How the guess is fit to collected data must be evaluated with Q-Q
- If X is a random variable with CDF F(X), then the q-quantile of X is γ such that:

$$F(\gamma) = P(X \leq \gamma) = q, \qquad for\ 0 < q < 1$$

  - When F(X) has an inverse, then $\gamma = F^{-1}(q)$
- Now, let's $\{x_i, i = 1,2, \ldots, n\}$ be a sample of collected data from X
  - Let $\{y_j, j = 1,2, \ldots, n\}$ be these observations in **ascending order**

Quantile-Quantile Plots (2)

SHARIF UNIVERSITY OF TECHNOLOGY

- So, we have $y_1 \leq y_2 \leq \cdots \leq y_j \leq \cdots \leq y_n$
  - y_1 is the smallest sample, while y_n represents the highest value
  - j indicates the **rank** or **order**
- Based on the principles of Q-Q plotting, y_j is an **estimated value** for the [j-0.5/n]-quantile of X:

$$y_j \approx F^{-1}\left(\frac{j-0.5}{n}\right)$$

  - Or, in other words:

$$F(y_j) = P(X \leq y_j) \approx \frac{j-0.5}{n}$$

- Now what? How could we use this to evaluate the fit?

Quantile-Quantile Plots (3)

SHARIF UNIVERSITY OF TECHNOLOGY

- Let's assume that we have selected a distribution with CDF F(X) for our collected data
- First, we plot y_j versus $F^{-1}[\frac{j-0.5}{n}]$
  - If F(X) is a member of an appropriate family of distributions
    - The resulting plot is approximately a **straight line**
  - If F(X) is a member of an appropriate family of distributions **with appropriate parameter** values
    - The line has slope 1
  - If the distribution is not selected wisely
    - The Q-Q plot will get deviated from the straight line

Example (1)

SHARIF UNIVERSITY OF TECHNOLOGY

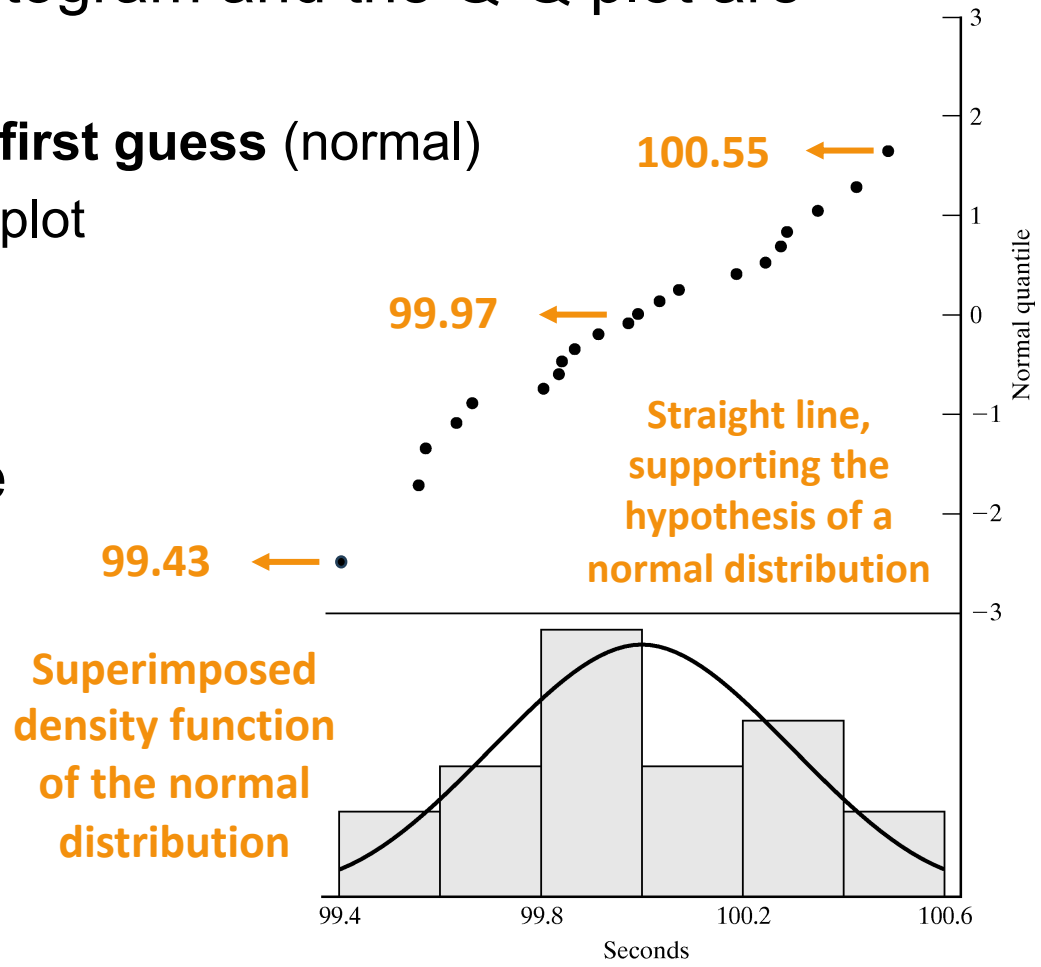
- Analyzing the production line in an automotive industry
  - Check whether the **door installation times** follows a normal distribution or not?
    - A sensor on the robot arm has measured 20 samples
  - The observations are **ordered** from the smallest to the largest:

j	Value	j	Value	j	Value	j	Value
1	99.55	6	99.82	11	99.98	16	100.26
2	99.56	7	99.83	12	100.02	17	100.27
3	99.62	8	99.85	13	100.06	18	100.33
4	99.65	9	99.9	14	100.17	19	100.41
5	99.79	10	99.96	15	100.23	20	100.47

  - Then, y_js are plotted versus $F^{-1}\left[\frac{j-0.5}{n}\right]$, where F(X) has a normal distribution
    - The sample mean is 99.99s and its variance is 0.2832^2 s^2

Example (2)

- In the following, the histogram and the Q-Q plot are both illustrated
  - Histogram provides the **first guess** (normal)
  - We test the fit with Q-Q plot
- If the collected data was 100% match with normal, we would have a precise straight line
  - While, the points are scattered around the line, their deviation is not much
    - Accept the guess



What else About Q-Q Plots?

SHARIF UNIVERSITY OF TECHNOLOGY

- Consider the following issues while evaluating the linearity of a Q-Q plots:
  - The observed values **never** fall exactly on a straight line
  - The ordered values are **ranked** and hence **not independent**
    - So, when a point resides above the line, the succeeding points will also reside above the line
      - It is unlikely for the points to be scattered evenly about the line
  - Variance of the extremes is higher than the middle
    - Linearity of the points in the **middle of the plot is more important**
- Q-Q plots can also be used to check **homogeneity**
  - Check whether a single distribution can be used for two sample sets (or two variables X, and Y)
    - By plotting the ordered values of the two data samples against each other

Parameter Estimation

SHARIF UNIVERSITY OF TECHNOLOGY

- After knowing that a family of distributions is good fit for our data, the next step is to **estimate its parameters**
- If observations in a sample of size n are $X_1, X_2, \dots, X_n$ (discrete or continuous), the **sample mean** and **variance** are:

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

- If the data are discrete and have been grouped in k different frequency distributions:

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

  - Where f_j is the observed frequency of value X_j

Example for Grouped Data

SHARIF UNIVERSITY OF TECHNOLOGY

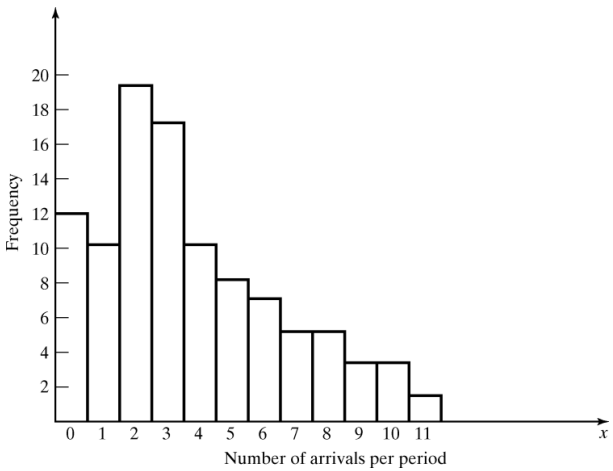
- Vehicle arrival example (continued)
  - Table in the histogram example on slide 11 can be analyzed to obtain:

$$n = 100, f_1 = 12, X_1 = 0, f_2 = 10, X_2 = 1, \ldots,$$

$$\text{and } \sum_{j=1}^{k} f_j X_j = 364, \text{ and } \sum_{j=1}^{k} f_j X_j^2 = 2080$$

- The sample mean and variance are:

$$\bar{X} = \frac{364}{100} = 3.64 \quad S^2 = \frac{2080 - 100 \times (3.64)^2}{99} = 7.63$$



- The histogram suggests X to have a Poisson distribution
  - However, note that sample mean is not equal to sample deviation!
  - Reason: each estimator is a random variable, and is **not perfect**

Estimators for Malformed/Classified Data

SHARIF UNIVERSITY OF TECHNOLOGY

- The indicated equations in slide 21 are expressed with the assumption that all the raw data is available
  - Sometime, the entire raw data is N/A or **malformed**
  - But, we have classified the data prior to this
  - In this case, the approximate sample mean and variance are:

$$\bar{X} = \frac{\sum_{j=1}^{c} f_j m_j}{n} \longrightarrow \text{middle point of j-th class} \qquad S^2 = \frac{\sum_{j=1}^{c} f_j m_j^2 - n\bar{X}^2}{n-1}$$

- Example: Chips' lifetime is recorded (in days), for 50 random chips
  - Lifetime is usually considered as a continuous variable
    - It is recorded with three decimal-place accuracy
  - The histogram is prepared by placing the data in class intervals

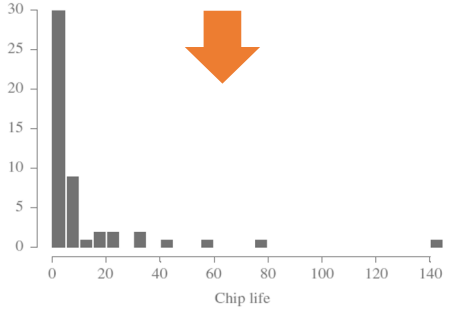
Chips' Lifetime Example

SHARIF UNIVERSITY OF TECHNOLOGY

- Later, raw data in the life test is assumed to be malformed
  - But the old table is still available
  - Approximating the mean and variance as follows (n=50):

This is the original data, which is no longer available

79.919	3.081	0.062	1.961	5.845
3.027	6.505	0.021	0.013	0.123
6.769	59.899	1.192	34.760	5.009
18.387	0.141	43.565	24.420	0.433
144.695	2.663	17.967	0.091	9.003
0.941	0.878	3.371	2.157	7.579
0.624	5.380	3.148	7.078	23.960
0.590	1.928	0.300	0.002	0.543
7.004	31.764	1.005	1.147	0.219
3.217	14.382	1.008	2.336	4.562



Component Life (days)	*Frequency*
$0 \le x_j < 3$	23
$3 \le x_j < 6$	10
$6 < x_j < 9$	5
$9 \le x_j < 12$	1
$12 \le x_j < 15$	1
$15 \le x_j < 18$	2
$18 \le x_j < 21$	0
$21 \le x_j < 24$	1
$24 \le x_j < 27$	1
$27 \le x_j < 30$	0
$30 \le x_j < 33$	1
$33 \le x_j < 36$	1
⋮	⋮
$42 \le x_j < 45$	1
⋮	⋮
$57 \le x_j < 60$	1
⋮	⋮
$78 \le x_j < 81$	1
⋮	⋮
$144 \le x_j < 147$	1

$$f_1 = 23, m_1 = 1.5, f_2 = 10, m_2 = 4.5, \ldots$$

$$\sum_{j=1}^{49} f_j m_j = 614 \qquad \sum_{j=1}^{49} f_j m_j^2 = 37226.5$$

$$\bar{X} = \frac{614}{50} = 12.28 \qquad S^2 = \frac{37226.5 - 50(12.28)^2}{49} = 605.849$$

Suggested Estimators (1)

SHARIF UNIVERSITY OF TECHNOLOGY

- The following table represents the estimates of the distribution's parameters which are needed to:
  - Reduce the family of probable distributions to a specific distribution
  - Test the resulting hypothesis in the next step
- Checkout the estimates for Poisson and exponential

Distribution	*Parameter(s)*	*Suggested Estimator(s)*
Poisson	α	$\widehat{\alpha} = \bar{X}$
Exponential	λ	$\widehat{\lambda} = \frac{1}{\bar{X}}$
Gamma	β, θ	$\widehat{\beta}$ (see Table A.9) $\widehat{\theta} = \frac{1}{\bar{X}}$
Normal	μ, σ^2	$\widehat{\mu} = \bar{X}$ $\widehat{\sigma}^2 = S^2$ (unbiased)
Lognormal	μ, σ^2	$\widehat{\mu} = \bar{X}$ (after taking ln of the data) $\widehat{\sigma}^2 = S^2$ (after taking ln of the data)
Weibull with $\nu = 0$	α, β	$\widehat{\beta}_0 = \frac{\bar{X}}{S}$ $\widehat{\beta}_j = \widehat{\beta}_{j-1} - \frac{f(\widehat{\beta}_{j-1})}{f'(\widehat{\beta}_{j-1})}$ See Equations (11) and (14) for $f(\widehat{\beta})$ and $f'(\widehat{\beta})$ Iterate until convergence $\widehat{\alpha} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^{\widehat{\beta}}\right)^{1/\widehat{\beta}}$
Beta	β_1, β_2	$\Psi(\widehat{\beta}_1) + \Psi(\widehat{\beta}_1 - \widehat{\beta}_2) = \ln(G_1)$ $\Psi(\widehat{\beta}_2) + \Psi(\widehat{\beta}_1 - \widehat{\beta}_2) = \ln(G_2)$ where Ψ is the digamma function, $G_1 = \left(\prod_{i=1}^{n} X_i\right)^{1/n}$ and $G_2 = \left(\prod_{i=1}^{n} (1 - X_i)\right)^{1/n}$

Suggested Estimators (2)

SHARIF UNIVERSITY OF TECHNOLOGY

- In the **Gamma distribution**, we need to estimate two parameters, β and θ
  - To estimate β, an other parameter called M is required:

$$M = Ln\bar{X} - \frac{1}{n}\sum_{i=1}^{n} LnX_i$$

  - After obtaining M, lookup the **Table A.9** of the textbook to find the corresponding $\hat{\beta}$
  - To estimate θ, we use $\hat{\theta} = 1/\bar{X}$
- In the **normal distribution**, we also need to estimate two parameters μ, and σ^2
  - From the past, we use $\hat{\mu} = \bar{X}$, and $\hat{\sigma}^2 = S^2$
    - Don't forget to how to use S^2

Suggested Estimators (3)

SHARIF UNIVERSITY OF TECHNOLOGY

- The act is almost the same for **lognormal distribution**
  - The only difference with the normal distribution is that we need to first get Ln from all of the collected data
  - Then apply the equations
- For the Weibull distribution, two parameters including α, and β must be estimated
  - Assume $v = 0$
  - $\hat{\beta}_0 = \frac{\bar{X}}{S}$, and $\hat{\beta}_j = \hat{\beta}_{j-1} - \frac{f(\hat{\beta}_{j-1})}{\acute{f}(\hat{\beta}_j - 1)}$ **What are $f()$ and $\acute{f}()$?**

$$f(\beta) = \frac{n}{\beta} + \sum_{i=1}^{n} \ln X_i - \frac{n \sum_{i=1}^{n} X_i^{\beta} \ln X_i}{\sum_{i=1}^{n} X_i^{\beta}}$$

Suggested Estimators (4)

SHARIF UNIVERSITY OF TECHNOLOGY

$$f'(\beta) = -\frac{n}{\beta^2} - \frac{n\sum_{i=1}^{n} X_i^{\beta}(\ln X_i)^2}{\sum_{i=1}^{n} X_i^{\beta}} + \frac{n\left(\sum_{i=1}^{n} X_i^{\beta}\ln X_i\right)^2}{\left(\sum_{i=1}^{n} X_i^{\beta}\right)^2}$$

- In order to estimate β, you must iterate the process of increasing j, until $f(\hat{\beta}_j) = 0$
  - Then we can put $\hat{\beta} = \hat{\beta}_j$
  - We can usually stop the iteration when $|f(\hat{\beta}_j)| \leq 0.001$
- By obtaining $\hat{\beta}$, it is turn for estimating $\hat{\alpha}$ with the following:

$$\hat{\alpha} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^{\hat{\beta}}\right)^{1/\hat{\beta}}$$

Suggested Estimators (5)

SHARIF UNIVERSITY OF TECHNOLOGY

- In the **Beta distribution**, we need to estimate β_1, and β_2
  - These two parameters are estimated by using the following

$$\psi(\hat{\beta}_1) + \psi(\hat{\beta}_1 - \hat{\beta}_2) = Ln(G_1)$$

$$\psi(\hat{\beta}_2) + \psi(\hat{\beta}_1 - \hat{\beta}_2) = Ln(G_2)$$

  - Where:

$$\psi(x) = \frac{\Gamma(\grave{x})}{\Gamma(x)} \quad , \quad G_1 = \left(\prod_{i=1}^{n} X_i\right)^{1/n} \quad , \quad G_2 = \left(\prod_{i=1}^{n} [1 - X_i]\right)^{1/n}$$

  - Recall: The gamma function is defined as $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx \, , x > 0$

Goodness-of-Fit Tests

SHARIF UNIVERSITY OF TECHNOLOGY

- Post to the **distribution selection** and its **parameter(s) estimation**
  - We test the **goodness-of-fit** to the collected data
- Conduct hypothesis testing on input data distribution using:
  - Kolmogorov-Smirnov test
  - Chi-square test
- Note: There may exist **no single correct distribution** for a real application if:
  - Very little data are available → it is unlikely to reject any candidate distributions
  - A lot of data are available → it is likely to reject all candidate distributions

Chi-Square Test (1)

SHARIF UNIVERSITY OF TECHNOLOGY

- Intuition: Comparing the histogram of the data to the shape of the **candidate** density or mass function
- Valid for **large sample sizes** when parameters are estimated by maximum likelihood
- By arranging the n observations into a set of k class intervals or cells, the test criterion is:

Observed Frequency

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Expected Frequency $E_i = n \times p_i$ where p_i is the theoretical probability of the ith interval

  - X_0^2 approximately follows the chi-square distribution with k-s-1 degrees of freedom
    - s is number of parameters of the hypothesized distribution **estimated by the sample statistics**

Chi-Square Test (2)

SHARIF UNIVERSITY OF TECHNOLOGY

- The hypothesis of the chi-square test is:
  - H0: The random variable, X, conforms to the distributional assumption with the parameter(s) given by the estimate(s)
  - H1: The random variable X does not conform
- Use the info provided in table A.6 to obtain the test critical value $X^2_{\alpha,k-s-1}$
  - If $X_0^2 > X^2_{\alpha,k-s-1}$ → H0 is rejected
- Suggested **Minimum E_i** in this test is 5
  - If the distribution under test is **discrete** and $E_i \geq 5$
    - Combining adjacent cells is not required
    - And $p_i = p(x_i) = P(X = x_i)$
  - If $E_i < 5$
    - Combining is necessary

Number of classes is determined after combining (if needed)!

Chi-Square Test (3)

SHARIF UNIVERSITY OF TECHNOLOGY

- If the distribution under test is **continuous**:

$$p_i = \int_{a_{i-1}}^{a_i} f(x)\,dx = F(a_i) - F(a_{i-1})$$

  - Where:
    - a_{i-1} and a_i are the endpoints of the ith class interval $[a_{i-1}, a_i)$
    - f(x) is the assumed PDF
    - F(x) is the assumed CDF
- Recommended number of class intervals (k):

Sample Size, n	Number of Class Intervals, k
20	Do not use the chi-square test
50	5 to 10
100	10 to 20
> 100	$n^{1/2}$ to n/5

Note: The Number of classes (k) affects the result of our test!

Example (1)

- Let's get back to the vehicle arrival example (Discrete)
  - H0: The collected data is Poisson distributed
  - H1: The collected data is not Poisson distributed

x_i	Observed Frequency, O_i	Expected Frequency, E_i	$(O_i - E_i)^2/E_i$
0	12	2.6	7.87
1	10	9.6	
2	19	17.4	0.15
3	17	21.1	0.8
4	19	19.2	4.41
5	6	14.0	2.57
6	7	8.5	0.26
7	5	4.4	11.62
8	5	2.0	
9	3	0.8	
10	3	0.3	
> 11	1	0.1	
	100	100.0	27.68

α Must be estimated according to slide 24 which will be 3.64

$$E_i = np(x_i) = n\frac{e^{-\alpha}\alpha^{x_i}}{x_i!}$$

Example 1: $P(0) = 0.026$

Example 2: $P(1) = 0.096$

...

Combined because of min E_i

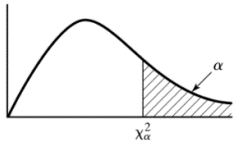
- Degree of freedom is $k - s - 1 = 7 - 1 - 1 = 5$
  - Hence, the hypothesis is **rejected** at the 0.05 level of significance

$$\chi_0^2 = 27.68 > \chi_{0.05,5}^2 = 11.1$$

Example (2)

- Percentage points of the Chi-Square distribution with υ degree of freedom

$$X^2_{0.05,5} = 11.1$$



ν	$\chi^2_{0.005}$	$\chi^2_{0.01}$	$\chi^2_{0.025}$	$\chi^2_{0.05}$	$\chi^2_{0.10}$
1	7.88	6.63	5.02	3.84	2.71
2	10.60	9.21	7.38	5.99	4.61
3	12.84	11.34	9.35	7.81	6.25
4	14.96	13.28	11.14	9.49	7.78
5	16.7	15.1	12.8	11.1	9.2
6	18.5	16.8	14.4	12.6	10.6
7	20.3	18.5	16.0	14.1	12.0
8	22.0	20.1	17.5	15.5	13.4
9	23.6	21.7	19.0	16.9	14.7
10	25.2	23.2	20.5	18.3	16.0
11	26.8	24.7	21.9	19.7	17.3
12	28.3	26.2	23.3	21.0	18.5
13	29.8	27.7	24.7	22.4	19.8
14	31.3	29.1	26.1	23.7	21.1
15	32.8	30.6	27.5	25.0	22.3
16	34.3	32.0	28.8	26.3	23.5
17	35.7	33.4	30.2	27.6	24.8
18	37.2	34.8	31.5	28.9	26.0
19	38.6	36.2	32.9	30.1	27.2
20	40.0	37.6	34.2	31.4	28.4
21	41.4	38.9	35.5	32.7	29.6
22	42.8	40.3	36.8	33.9	30.8
23	44.2	41.6	38.1	35.2	32.0
24	45.6	43.0	39.4	36.4	33.2
25	49.6	44.3	40.6	37.7	34.4
26	48.3	45.6	41.9	38.9	35.6
27	49.6	47.0	43.2	40.1	36.7
28	51.0	48.3	44.5	41.3	37.9
29	52.3	49.6	45.7	42.6	39.1
30	53.7	50.9	47.0	43.8	40.3
40	66.8	63.7	59.3	55.8	51.8
50	79.5	76.2	71.4	67.5	63.2
60	92.0	88.4	83.3	79.1	74.4
70	104.2	100.4	95.0	90.5	85.5
80	116.3	112.3	106.6	101.9	96.6
90	128.3	124.1	118.1	113.1	107.6
100	140.2	135.8	129.6	124.3	118.5

Kolmogorov-Smirnov Test

SHARIF UNIVERSITY OF TECHNOLOGY

- Intuition: Formalize the idea behind examining a Q-Q plot
  - Recall from Chapter 6
    - It compares the continuous CDF, F(x), of the hypothesized distribution with the empirical CDF, SN(x), of the N sample observations
  - How?
    - Based on the **maximum difference statistics** (Tabulated in A.8)

$$D = \max |F(x) - SN(x)|$$

- A more powerful test, particularly useful when:
  - Sample sizes are **small**
  - No parameters have been estimated from the data
- When parameter estimates have been made:
  - Critical values in Table A.8 are biased, too large.
- More conservative, i.e., smaller Type I error than specified

Example (1)

SHARIF UNIVERSITY OF TECHNOLOGY

- Suppose that 50 interarrival times (in minutes) are collected over a 100 minute interval
  - Below, they are indicated in order of their occurrences

0.44	0.53	2.04	2.74	2.00	0.30	2.54	0.52	2.02	1.89	1.53	0.21
2.80	0.04	1.35	8.32	2.34	1.95	0.10	1.42	0.46	0.07	1.09	0.76
5.55	3.93	1.07	2.26	2.88	0.67	1.12	0.26	4.57	5.37	0.12	3.19
1.63	1.46	1.08	2.06	0.85	0.83	2.44	1.02	2.24	2.11	3.15	2.90
6.58	0.64										

  - We know that the **exponential** distribution is a **good guess**
  - We test the goodness of fit with Kolmogorov-Smirnov
    - H0, and H1 hypotheses
- If our guess on the exponentially distributed interarrivals is true → Arrivals would be uniform on the [0,T] interval

Example (2)

SHARIF UNIVERSITY OF TECHNOLOGY

- Therefore, we test the uniformity of the arrival times according to the following:

$$Arrivals \in \{t_1, t_1 + t_2, t_1 + t_2 + t_3, \dots, t_1 + t_2 + t_3 + \cdots t_n\}$$

- Similar to lecture 6, in order to use Kolmogorov test, these data must be $\in [0,1]$
  - We should **normalize** these numbers to T (simulation period)

$$normalaized\ Arrivals \in$$
$$\{t_1/\mathbf{T}, (t_1+t_2)/\mathbf{T}, (t_1 + t_2 + t_3)/\mathbf{T}, \dots, (t_1+t_2 + t_3 + \cdots t_n)/\mathbf{T}\}$$

  - **Sort** the obtained values

0.0044	0.0097	0.0301	0.0575	0.0775	0.0805	0.1059	0.1111	0.1313	0.1502
0.1655	0.1676	0.1956	0.1960	0.2095	0.2927	0.3161	0.3356	0.3366	0.3508
0.3553	0.3561	0.3670	0.3746	0.4300	0.4694	0.4796	0.5027	0.5315	0.5382
0.5494	0.5520	0.5977	0.6514	0.6526	0.6845	0.7008	0.7154	0.7262	0.7468
0.7553	0.7636	0.7880	0.7982	0.8206	0.8417	0.8732	0.9022	0.9680	0.9744

Example (3)

SHARIF UNIVERSITY OF TECHNOLOGY

- Now, follow the instructions for this test in lecture 6
  - $D^+ = 0.1054$, and $D^- = 0.0080$
    - $D = D_{max} = Max\{D^-, D^+\} = 0.1054$
  - According to the Kolmogorov-Smirnov test crucial table
    - For $\alpha = 0.05$, n=50, the value of $D_{0.05}$ will be obtained as $\frac{1.36}{\sqrt{n}} = 0.1923$

Since $D < D_\alpha$ H0 cannot be rejected

We accept the uniformity of arrivals

We accept the exponentiality of interarrivals

Table A.8 Kolmogorov--Smirnov Critical Values

Degrees of Freedom (N)	$D_{0.10}$	$D_{0.05}$	$D_{0.01}$
1	0.950	0.975	0.995
2	0.776	0.842	0.929
3	0.642	0.708	0.828
4	0.564	0.624	0.733
5	0.510	0.565	0.669
6	0.470	0.521	0.618
7	0.438	0.486	0.577
8	0.411	0.457	0.543
9	0.388	0.432	0.514
10	0.368	0.410	0.490
11	0.352	0.391	0.468
12	0.338	0.375	0.450
13	0.325	0.361	0.433
14	0.314	0.349	0.418
15	0.304	0.338	0.404
16	0.295	0.328	0.392
17	0.286	0.318	0.381
18	0.278	0.309	0.371
19	0.272	0.301	0.363
20	0.264	0.294	0.356
25	0.24	0.27	0.32
30	0.22	0.24	0.29
35	0.21	0.23	0.27
Over 35	$\frac{1.22}{\sqrt{N}}$	$\frac{1.36}{\sqrt{N}}$	$\frac{1.63}{\sqrt{N}}$