Computer Simulation

Chapter 8:
Input Modeling

Purpose & Overview

- Input models provide the driving force for a simulation model.
- The quality of the output is no better than the quality of inputs.
- In this chapter, we will discuss the 4 steps of input model development:
  - Collect data from the real system
  - Identify a probability distribution to represent the input process
  - Choose parameters for the distribution
  - Evaluate the chosen distribution and parameters for goodness of fit.

Data Collection (1)

- Be aware of:
  - Stale data
  - Unexpected Data
  - Time Varying Data
  - Dependent Data

Data Collection (2)

- One of the biggest tasks in solving a real problem. GIGO – garbage-in-garbage-out
- Suggestions that may enhance and facilitate data collection:
  - Plan ahead: begin by a practice or pre-observing session, watch for unusual circumstances
  - Analyze the data as it is being collected: check adequacy
  - Combine homogeneous data sets, e.g. successive time periods, during the same time period on successive days
  - Be aware of data censoring: the quantity is not observed in its entirety, danger of leaving out long process times
  - Check for relationship between variables, e.g. build scatter diagram
  - Check for autocorrelation
  - Collect input data, not performance data

Identifying the Distribution

- Histograms
- Selecting families of distribution
- Parameter estimation
- Goodness-of-fit tests
- Fitting a non-stationary process

Histograms (1)

- A frequency distribution or histogram is useful in determining the shape of a distribution
- 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:
  - Corresponds to the probability density function of a theoretical distribution
- For discrete data:
  - Corresponds to the probability mass function
- If few data points are available: combine adjacent cells to eliminate the ragged appearance of the histogram

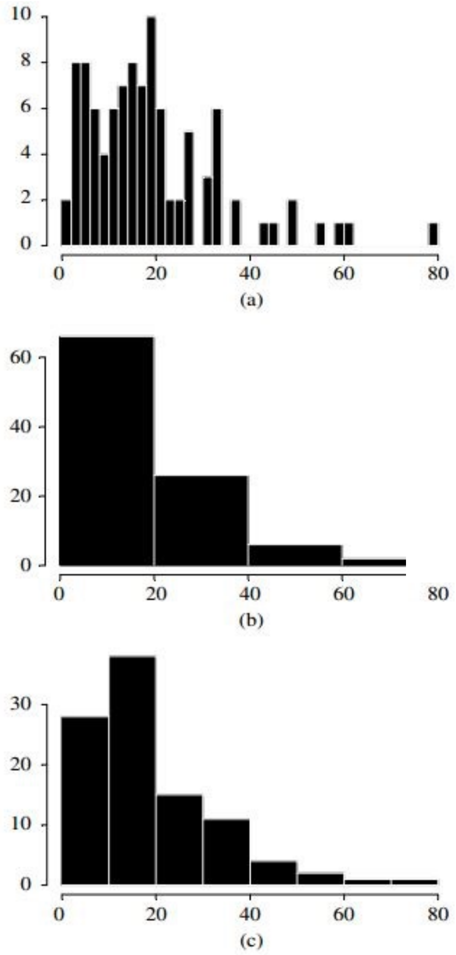
Histograms (2)

[Identifying the distribution]

The original data

Ragged Histogram

✓ **A perfect smooth plot**

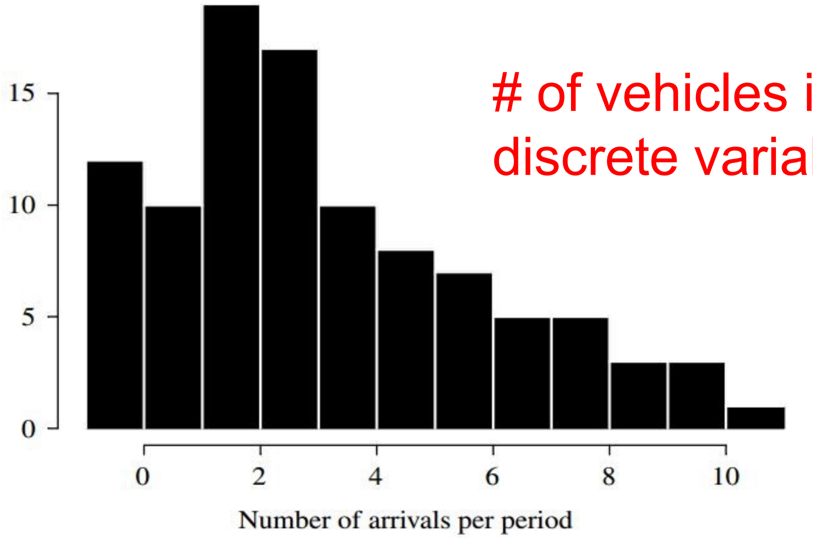


Same data with different interval sizes

Histograms (3)

- Vehicle Arrival Example: # of vehicles arriving at an intersection between 7 am and 7:05 am was monitored for *100* random workdays.

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



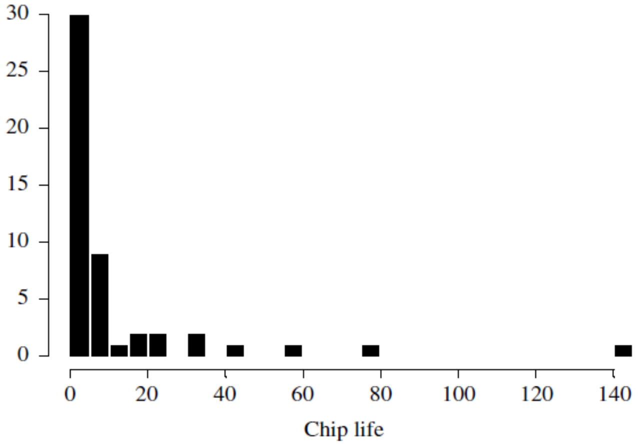
# of vehicles is a discrete variable

- There are ample data, so the histogram may have a cell for each possible value in the data range

Histogram (4)

- Chips Lifetime Example: Lifetime is recorded (in days), for 50 random chips

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



- 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

Selecting the Family of Distributions (1)

- A family of distributions is selected based on:
  - The context of the input variable
  - Shape of the histogram
- Frequently encountered distributions:
  - Easier to analyze: exponential, normal and Poisson
  - Harder to analyze: beta, gamma and Weibull

Selecting the Family of Distributions (2)

- **Binomial:** Models the number of successes in n trials, when the trials are independent with common success probability p
  - □ Example, the number of defective computer chips found in a lot of n chips.
- **Negative Binomial (includes the geometric distribution):** Models the number of trials required to achieve k successes
  - □ Example, the number of computer chips that we must inspect to find 4 defective chips.
- **Poisson:** Models the number of independent events that occur in a fixed amount of time or space
  - □ Example, the number of customers that arrive to a store during 1 hour, or the number of defects found in 30 square meters of sheet metal.
- **Normal:** Models the distribution of a process that can be thought of as the sum of a number of component processes
  - □ Example, a time to assemble a product that is the sum of the times required for each assembly operation. Notice that the normal distribution admits **negative values,** which could be impossible for process times.

Selecting the Family of Distributions (3)

- **Lognormal:** Models the distribution of a process that can be thought of as the product of (meaning to multiply together) a number of component processes
  - Example, the rate of return on an investment, when interest is compounded, is the product of the returns for a number of periods.
- **Exponential:** Models the time between independent events, or a process time that is memoryless
  - Example, the times between the arrivals from a large population of potential customers who act independently of one another. The exponential is a highly variable distribution; **it is sometimes overused, because it often leads to mathematically tractable models.**
- **Gamma:** An extremely flexible distribution used to model nonnegative random variables. The gamma can be shifted away from 0 by adding a constant
- **Beta:** An extremely flexible distribution used to model bounded (fixed upper and lower limits) random variables. The beta can be shifted away from 0 by adding a constant and can be given a range larger than [0, 1] by multiplying by a constant.

Selecting the Family of Distributions (4)

- **Erlang:** Models processes that can be viewed as the sum of several exponentially distributed processes
  - Example, a computer network fails when a computer and two backup computers fail, and each has a time to failure that is exponentially distributed. The Erlang is a special case of the gamma.
- **Weibull:** Models the time to failure for components
  - Example, the time to failure for a disk drive. The exponential is a special case of the Weibull.
- **Discrete or Continuous Uniform:** Models complete uncertainty: All outcomes are equally likely. This distribution often is used inappropriately when there are no data.
- **Triangular:** Models a process for which only the minimum, most likely, and maximum values of the distribution are known
  - Example, the minimum, most likely, and maximum time required to test a product. **This model is often a marked improvement over a uniform distribution.**
- **Empirical:** Resamples from the actual data collected; often used when no theoretical distribution seems appropriate.

Selecting the Family of Distributions (5)

- Remember the physical characteristics of the process
  - Is the process naturally discrete or continuous valued?
  - Is it bounded?
- No “true” distribution for any stochastic input process
- Goal: obtain a good approximation

Quantile-Quantile Plots (1) [Identifying the distribution]

- *Q-Q* plot is a useful tool for evaluating distribution fit
- If *X* is a random variable with cdf *F*, then the *q*-quantile of *X* is the γ such that

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

  - □ When *F* has an inverse, $\gamma = F^{-1}(q)$

- Let $\{x_i,\ i = 1,2,\ \ldots.,\ n\}$ be a sample of data from *X* and $\{y_j,\ j = 1,2,\ \ldots,\ n\}$ be the observations in ascending order:

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

where *j* is the ranking or order number

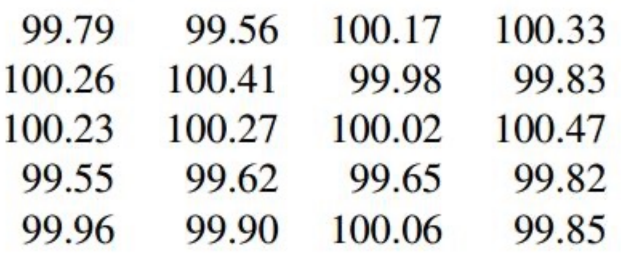
and y_j is an estimate of the *(j - 1/2)/n* quantile of *X.*

Quantile-Quantile Plots (2) [Identifying the distribution]

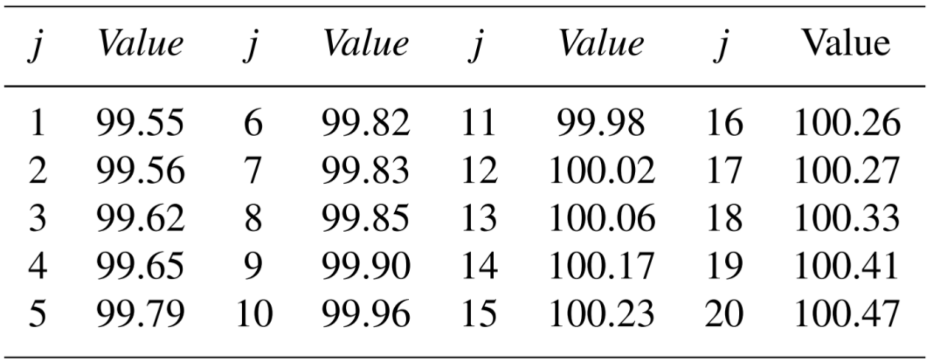
- What we should do now?
  - □ Plotting y_j versus $F^{-1}((j-1/2)/n)$ as our Q-Q plot
- If F is a member of appropriate distribution functions
  - □ Q-Q plot would be approximately a straight line
- If F is a member of appropriate distribution function and it has appropriate parameter values
  - □ Q-Q plot will be approximately a straight line with slope=1
- If F is an inappropriate distribution function for our collected data
  - □ The Q-Q plot will deviate from a straight line

Quantile-Quantile Plots (2) [Identifying the distribution]

- Example: Check whether the door installation times follows a normal distribution.
  - □ The observations are now ordered from smallest to largest:



99.79	99.56	100.17	100.33
100.26	100.41	99.98	99.83
100.23	100.27	100.02	100.47
99.55	99.62	99.65	99.82
99.96	99.90	100.06	99.85



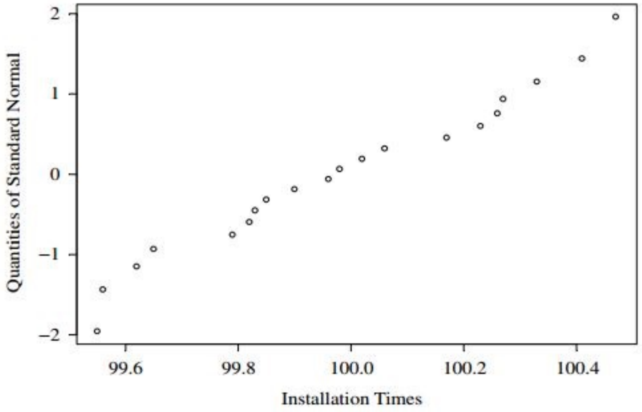
j	*Value*	*j*	*Value*	*j*	*Value*	*j*	Value
1	99.55	6	99.82	11	99.98	16	100.26
2	99.56	7	99.83	12	100.02	17	100.27
3	99.62	8	99.85	13	100.06	18	100.33
4	99.65	9	99.90	14	100.17	19	100.41
5	99.79	10	99.96	15	100.23	20	100.47

  - □ y_j are plotted versus $F^{-1}((j-0.5)/n)$ where *F* has a normal distribution with the sample mean *(99.99 sec)* and sample variance (*0.2832^2 sec^2*)

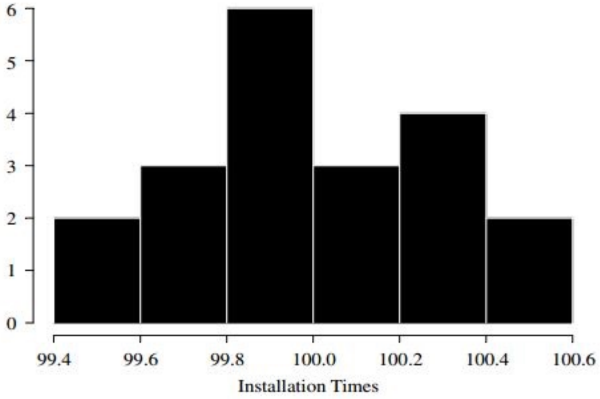
Quantile-Quantile Plots (3) [Identifying the distribution]

- Example (continued): Check whether the door installation times follow a normal distribution.

Straight line, supporting the hypothesis of a normal distribution



Superimposed density function of the normal distribution



Quantile-Quantile Plots (4) [Identifying the distribution]

- Consider the following while evaluating the linearity of a *q-q* plot:
  - The observed values never fall exactly on a straight line
  - The ordered values are ranked and hence not independent, unlikely for the points to be scattered 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* plot can also be used to check homogeneity
  - Check whether a single distribution can represent both sample sets
  - Plotting the order values of the two data samples against each other

Parameter Estimation (1) [Identifying the distribution]

- Next step after selecting a family of distributions
- If observations in a sample of size *n* are $X_1, X_2, \ldots, X_n$ (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 a frequency distribution:

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

where f_j is the observed frequency of value X_j

Parameter Estimation (2) [Identifying the distribution]

- Number of vehicles in the intersection (previous example): $n = 100, f_1 = 12, X_1 = 0, f_2 = 10, X_2 = 1, \ldots,$

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

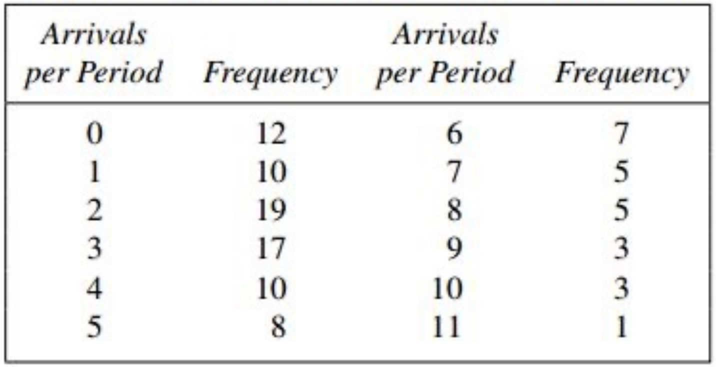
  - □ The sample mean and variance are

$$\bar{X} = \frac{364}{100} = 3.64$$

$$S^2 = \frac{2080 - 100 * (3.64)^2}{99}$$

$$= 7.63$$

Table 1 Number of Arrivals in a 5-Minute Period



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

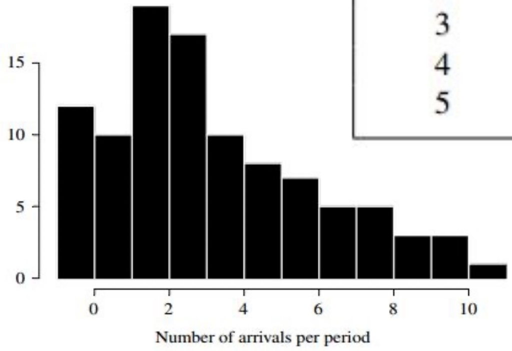


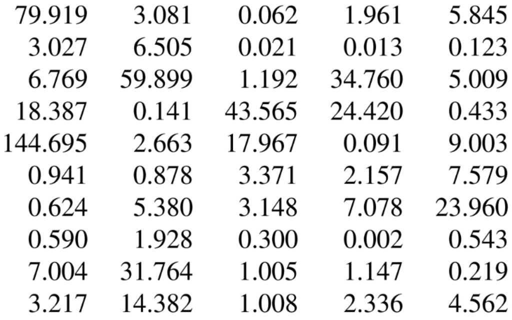
Figure 4 Histogram of number of arrivals per period.

Parameter Estimation (3) [Identifying the distribution]

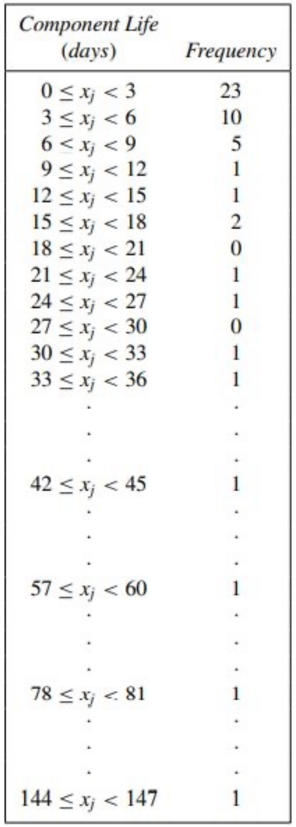
- When raw data are unavailable (data are grouped into class intervals), the approximate sample mean and variance are:

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

- ✓ The raw data in the Life test example is disordered
- ✓ But the table is still **available**
- ✓ Approximating the mean and variance as follows (n=50):



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
$6 \le x_j < 9$	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

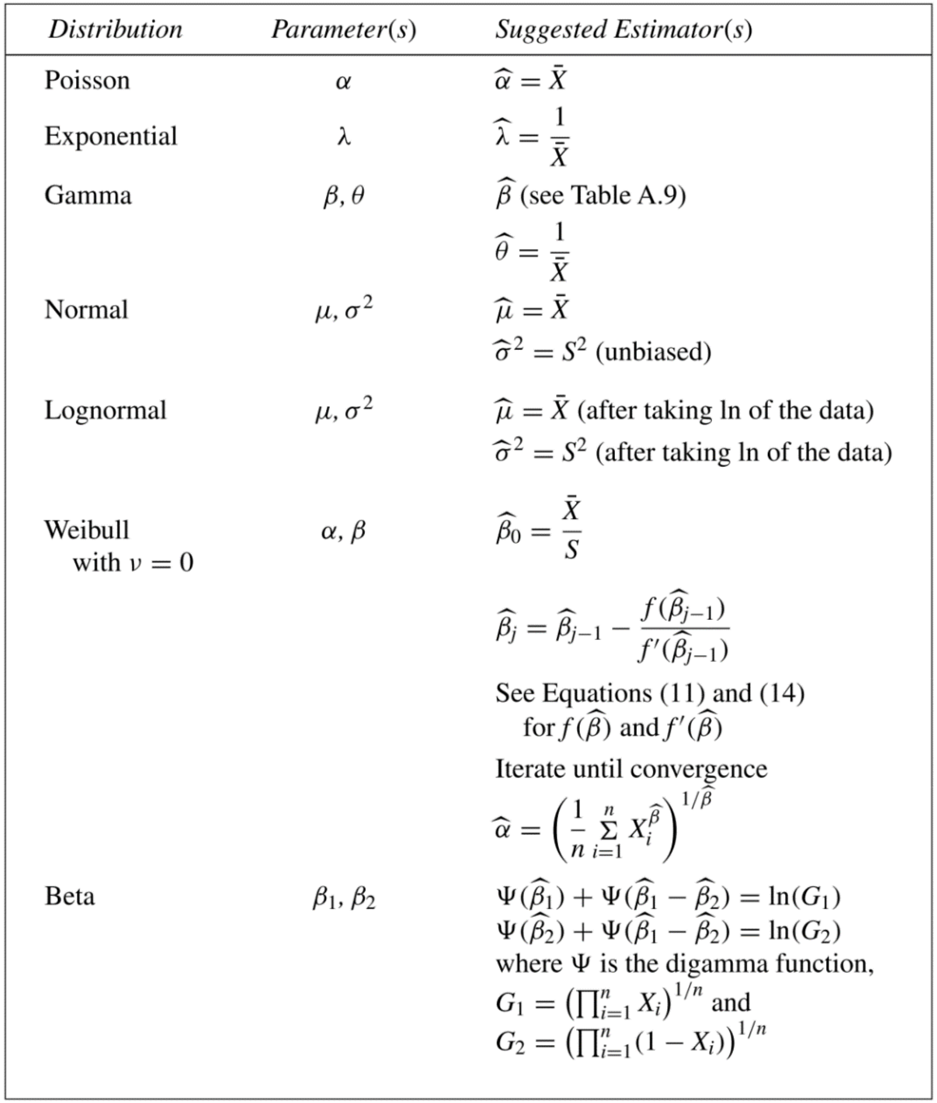
$$f_1 = 23, m_1 = 1.5, f_2 = 10, m_2 = 4.5, \ldots$$

$$\sum_{j=1}^{49} f_j m_j = 614 \quad \text{and} \quad \sum_{j=1}^{49} f_j m_j^2 = 37226.5 \qquad \bar{X} = \frac{614}{50} = 12.28$$

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

Parameter Estimation (4)

- Suggested estimators for distributions
  - Check the whiteboard



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 $G_2 = \left(\prod_{i=1}^{n} (1 - X_i)\right)^{1/n}$

Goodness-of-Fit Tests

[Identifying the distribution]

- Conduct hypothesis testing on input data distribution using:
  - Kolmogorov-Smirnov test
  - Chi-square test
- No single correct distribution in a real application exists.
  - If very little data are available, it is unlikely to reject any candidate distributions
  - If a lot of data are available, it is likely to reject all candidate distributions

Chi-Square test (1) [Goodness-of-Fit Tests]

- 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 statistics is:

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

Observed Frequency

Expected Frequency
$E_i = n^*p_i$
where p_i is the theoretical prob. of the *i*th interval.
Suggested Minimum = 5

which **approximately** follows the chi-square distribution with *k-s-1* degrees of freedom, where s = # of parameters of the hypothesized distribution estimated by the sample statistics.

Chi-Square test (2) [Goodness-of-Fit Tests]

- The hypothesis of a chi-square test is:

  H_0: The random variable, X, conforms to the distributional assumption with the parameter(s) given by the estimate(s).

  H_1: The random variable X does not conform.

- If the distribution tested is discrete and combining adjacent cell is not required (so that $E_i >$ minimum requirement):
  - □ Each value of the random variable would be a class interval, unless combining is necessary, and

$$p_i = p(x_i) = P(X = x_i)$$

If combining is needed, p_i is found by summing the probabilities of appropriate adjacent cells.

Chi-Square test (3) [Goodness-of-Fit Tests]

- If the distribution tested 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 i^{th} class interval and *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

- □ Caution: Different grouping of data (i.e., *k*) can affect the hypothesis testing result.

Chi-Square test (4) [Goodness-of-Fit Tests]

- Vehicle Arrival Example (continued):

  H_0: the random variable is Poisson distributed.

  H_1: the random variable 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

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

Combined because of min E_i **(=5)**

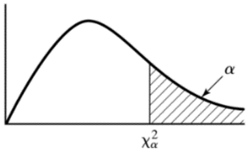
- □ Degree of freedom is *k-s-1 = 7-1-1 = 5*, hence, the hypothesis is rejected at the *0.05* level of significance.

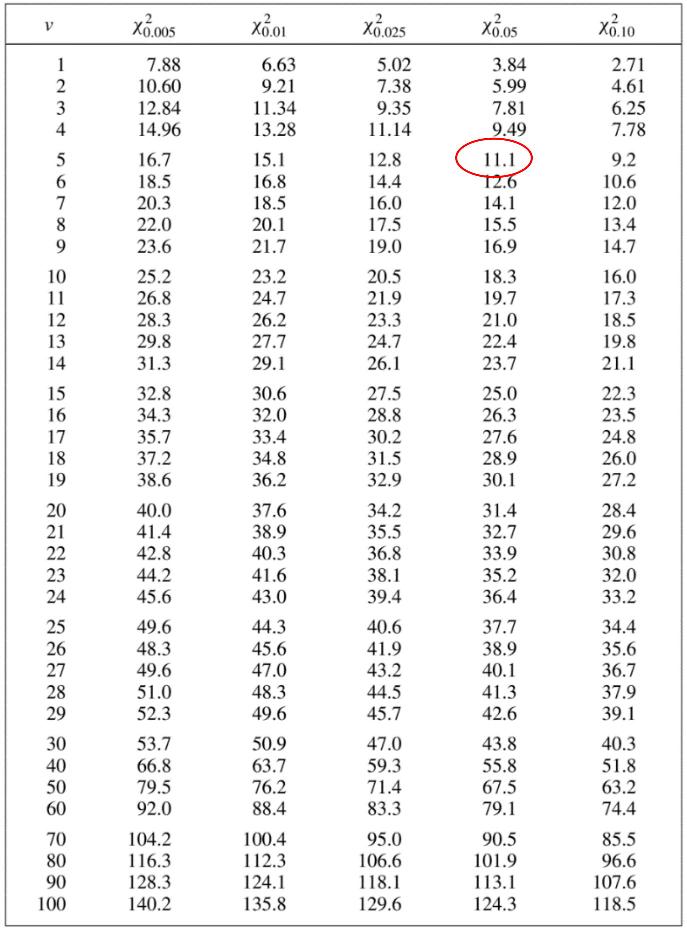
s is the number of estimated parameters

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

Chi-Square test (5)

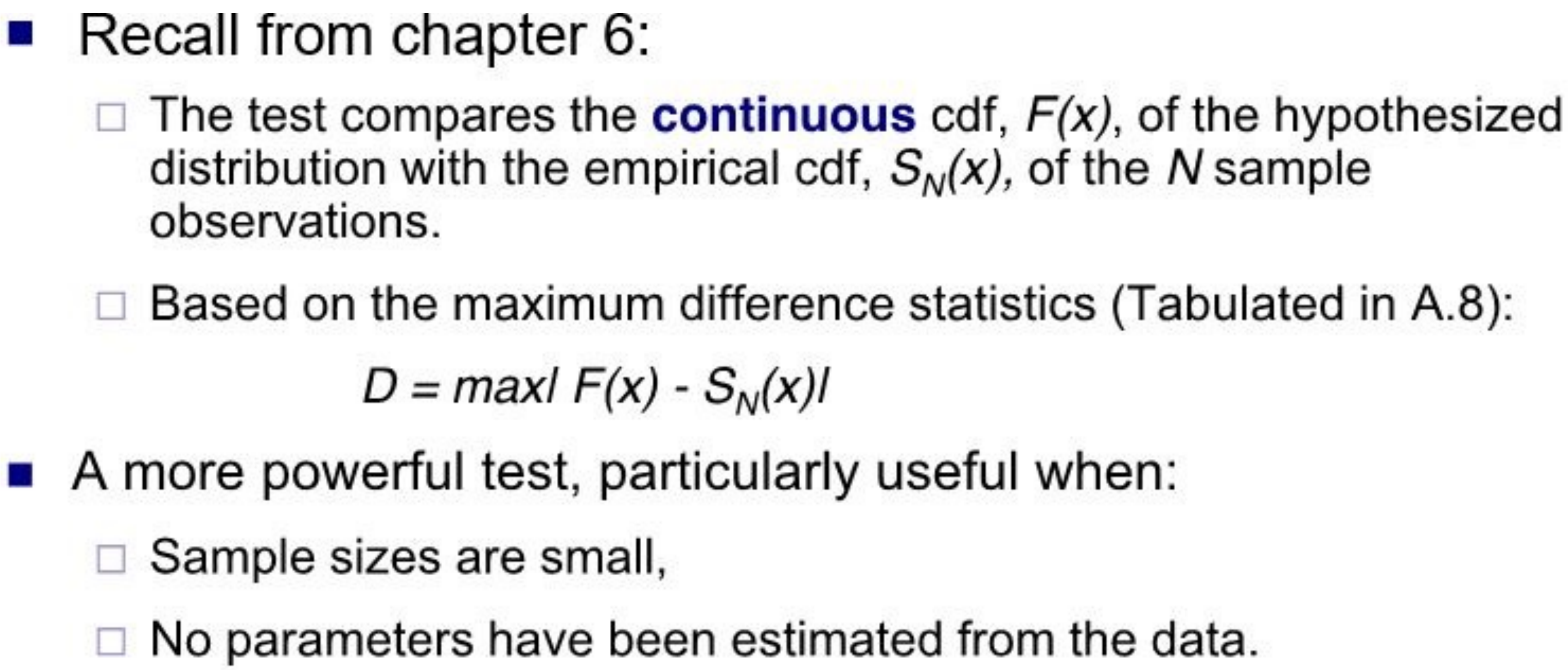
- Percentage points of the Chi-Square distribution with υ degree of freedom





v	$\chi^2_{0.005}$	$\chi^2_{0.01}$	$\chi^2_{0.025}$	$\chi^2_{0.05}$	$\chi^2_{0.10}$
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 (1)



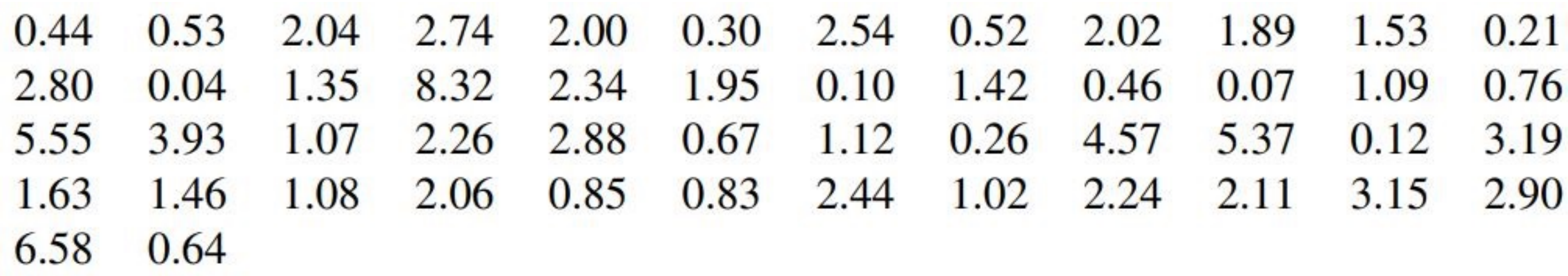
- Recall from chapter 6:
  - The test compares the **continuous** cdf, *F(x)*, of the hypothesized distribution with the empirical cdf, $S_N(x)$, of the *N* sample observations.
  - Based on the maximum difference statistics (Tabulated in A.8):

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

- A more powerful test, particularly useful when:
  - Sample sizes are small,
  - No parameters have been estimated from the data.

Kolmogorov-Smirnov Test (2)

- Suppose that 50 interarrival times (in minutes) are collected over a **100-minute** interval (arranged in order of occurrence):

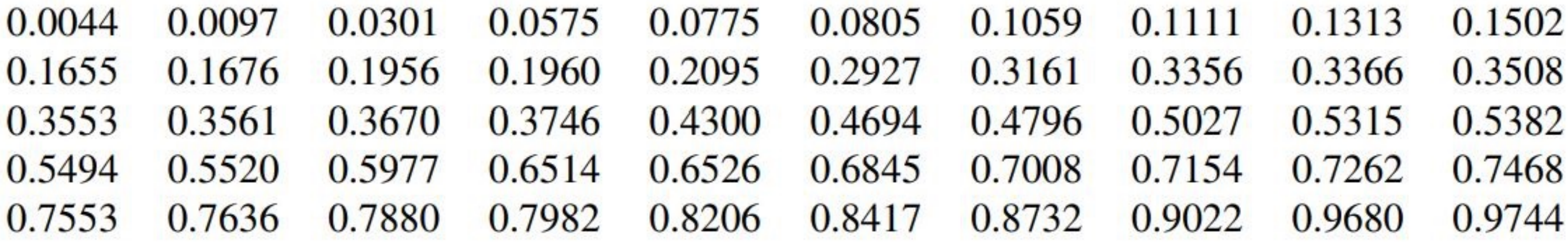


0.44	0.53	2.04	2.74	2.00	0.30	2.54	0.52	2.02	1.89	1.53	0.21
2.80	0.04	1.35	8.32	2.34	1.95	0.10	1.42	0.46	0.07	1.09	0.76
5.55	3.93	1.07	2.26	2.88	0.67	1.12	0.26	4.57	5.37	0.12	3.19
1.63	1.46	1.08	2.06	0.85	0.83	2.44	1.02	2.24	2.11	3.15	2.90
6.58	0.64										

- The null hypothesis and its alternate are formed as follows:
  - □ H0: The interarrival times are exponentially distributed.
  - □ H1: The interarrival times are not exponentially distributed.
- **If** interarrivals are **exponential** → the arrivals would be **uniform** on the [0,T] interval ∈ $\{T_1, T_1 + T_2, T_1 + T_2 + T_3, \dots, T_1 + T_2 + T_3 + \cdots + T_n\}$

Kolmogorov-Smirnov Test (3)

- The arrivals are then normalized on [0,1]
  - □ **Why?** To be able to use the Kolmogorov-Smirnov test
  - □ $\{T_1/T, (T_1+T_2)/T, (T_1+T_2+T_3)/T, \dots, (T_1+T_2+T_3+\cdots+T_n)/T\}$
- The uniform arrivals are arranged as follows:



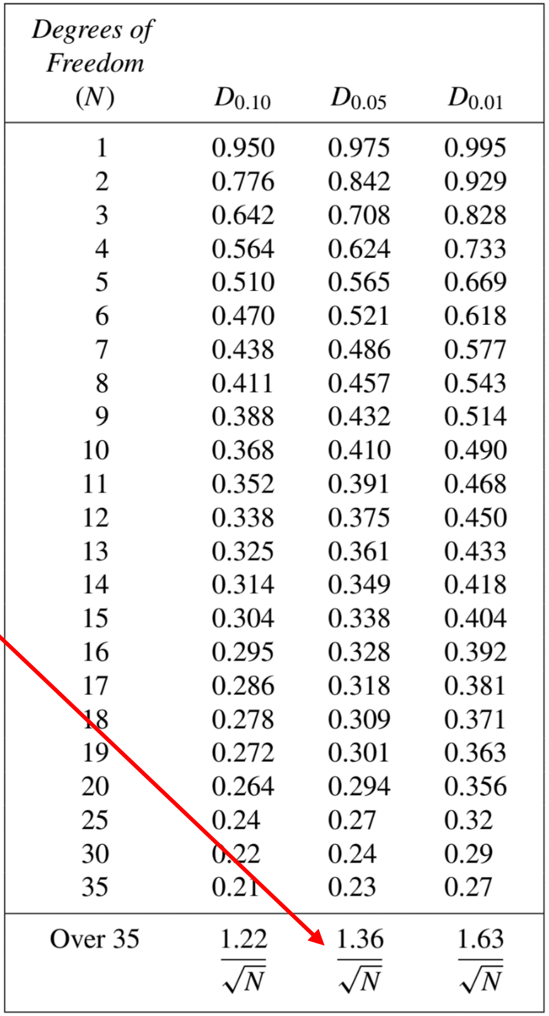
0.0044	0.0097	0.0301	0.0575	0.0775	0.0805	0.1059	0.1111	0.1313	0.1502
0.1655	0.1676	0.1956	0.1960	0.2095	0.2927	0.3161	0.3356	0.3366	0.3508
0.3553	0.3561	0.3670	0.3746	0.4300	0.4694	0.4796	0.5027	0.5315	0.5382
0.5494	0.5520	0.5977	0.6514	0.6526	0.6845	0.7008	0.7154	0.7262	0.7468
0.7553	0.7636	0.7880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

- Following the procedure in chapter 6:
  - □ $D^+ = 0.1054\ and\ D^- = 0.0080$ → $D = Max(D^+, D^-) = 0.1054$
  - □ Based on the Kolmogorov crucial table (next slide), for α=0.05 and n=50, the value of $D_{0.05} = 1.36/\sqrt{n}$=0.1923

Since $D < D_\alpha$ → H0 could not be rejected

Kolmogorov-Smirnov Test (4)

To obtain D_α



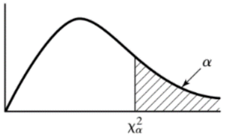
Degrees of Freedom (N)	$D_{0.10}$	$D_{0.05}$	$D_{0.01}$
1	0.950	0.975	0.995
2	0.776	0.842	0.929
3	0.642	0.708	0.828
4	0.564	0.624	0.733
5	0.510	0.565	0.669
6	0.470	0.521	0.618
7	0.438	0.486	0.577
8	0.411	0.457	0.543
9	0.388	0.432	0.514
10	0.368	0.410	0.490
11	0.352	0.391	0.468
12	0.338	0.375	0.450
13	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	$\frac{1.22}{\sqrt{N}}$	$\frac{1.36}{\sqrt{N}}$	$\frac{1.63}{\sqrt{N}}$

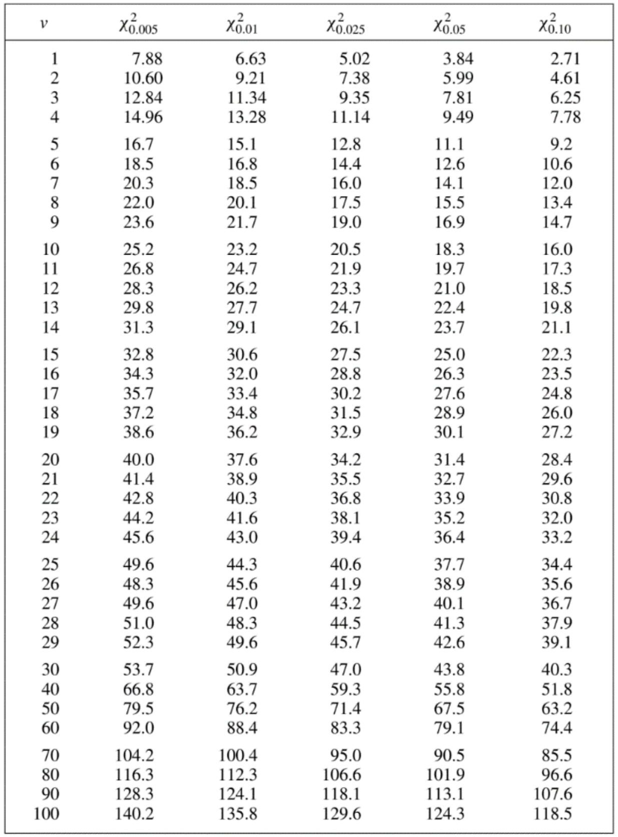
p-Values and “Best Fits” (1)

- *p-value* for the test statistics
  - □ The significance level at which one would just reject H_0 for the given test statistic value.
  - □ A measure of fit, the larger the better
  - □ Large *p-value*: good fit
  - □ Small *p-value*: poor fit

- Vehicle Arrival Example (cont.):
  - □ H_0: data is Possion
  - □ Test statistics: $\chi_0^2 = 27.68$, with *5* degrees of freedom
  - □ *p-value = 0.00004*, meaning we would reject H_0 with *0.00004* significance level, hence Poisson is a poor fit.

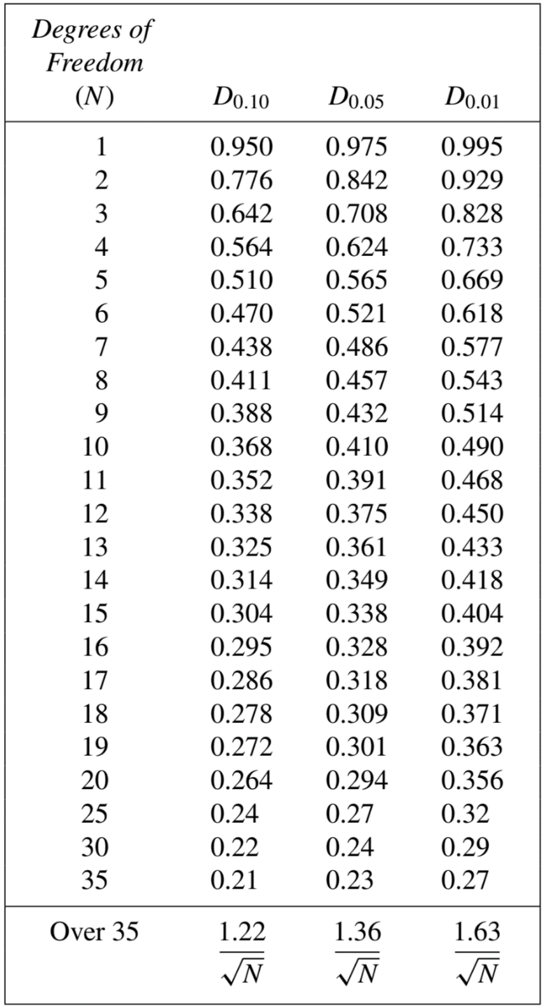
P-values and “Best Fits” (2)





v	$\chi^2_{0.005}$	$\chi^2_{0.01}$	$\chi^2_{0.025}$	$\chi^2_{0.05}$	$\chi^2_{0.10}$
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

Percentage points in Chi-Square



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	$\frac{1.22}{\sqrt{N}}$	$\frac{1.36}{\sqrt{N}}$	$\frac{1.63}{\sqrt{N}}$

Critical values in Kolmogorov-Smirnov

p-Values and “Best Fits” (3)

- Many software use *p-value* as the ranking measure to automatically determine the “best fit”. Things to be cautious about:
  - □ Software may not know about the physical basis of the data, distribution families it suggests may be inappropriate.
  - □ Close conformance to the data does not always lead to the most appropriate input model.
  - □ *p-value* does not say much about where the lack of fit occurs
- Recommended: always inspect the automatic selection using graphical methods.

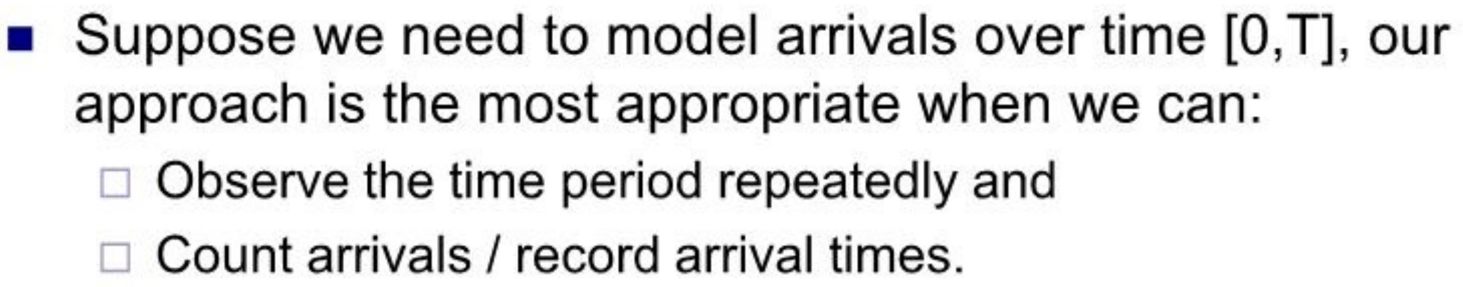
Fitting a Non-stationary Poisson Process (1)

- Fitting a NSPP to arrival data is difficult, possible approaches:
  - □ Fit a very flexible model with lots of parameters or
  - □ Approximate constant arrival rate over some basic interval of time, but vary it from time interval to time interval.

Our focus

Challenge: What is the length of the intervals, and how the λ(t) in each interval should be approximated?

Example: arrival of e-mail throughout the business day (8 A.M. to 6 P.M.)



- Suppose we need to model arrivals over time [0,T], our approach is the most appropriate when we can:
  - □ Observe the time period repeatedly and
  - □ Count arrivals / record arrival times.

Fitting a Non-stationary Poisson Process (2)

- The estimated arrival rate during the *i*th time period is:

$$\hat{\lambda}(t) = \frac{1}{n\Delta t}\sum_{j=1}^{n} C_{ij}$$

where n = # of observation periods, *Δt = time interval length*

C_{ij} = # of arrivals during the i^{th} time interval on the j^{th} observation period

- Example: Divide a *10*-hour business day [*8am,6pm*] into equal intervals *k = 20* whose length *Δt = ½*, and observe over n =3 days

Time Period	Number of Arrivals Day 1	 Day 2	 Day 3	Estimated Arrival Rate (arrivals/hr)
8:00 - 8:30	12	14	10	24
8:30 - 9:00	23	26	32	54
9:00 - 9:30	27	18	32	52
9:30 - 10:00	20	13	12	30

For instance, 1/3(0.5)*(23+26+32) = 54 arrivals/hour

Selecting Model without Data

- If data is not available, some possible sources to obtain information about the process are:
  - Engineering data: often product or process has performance ratings provided by the manufacturer or company rules specify time or production standards.
  - Expert opinion: people who are experienced with the process or similar processes, often, they can provide optimistic, pessimistic and most-likely times, and they may know the variability as well.
  - Physical or conventional limitations: physical limits on performance, limits or bounds that narrow the range of the input process.
  - The nature of the process.
- The uniform, triangular, and beta distributions are often used as input models.

Multivariate and Time-Series Input Models

- The input random variables were considered to be independent of any other variables within the context of the problem.
  - When this is not the case it is critical that input models be used that account for dependence; otherwise, a highly inaccurate simulation may result.

- Example 1
  - A supply-chain simulation includes the lead time and annual demand for industrial robots. An increase in demand results in an increase in lead time. The final assembly of the robots must be made according to the specifications of the purchaser. Therefore, rather than treat lead time and demand as independent random variables, a multivariate input model should be developed.
- Example 2
  - A simulation of the web-based trading site of a stock broker includes the time between arrivals of orders to buy and sell. Investors tend to react to what other investors are doing, so these buy and sell orders arrive in bursts. Therefore, rather than treat the time between arrivals as independent random variables, a time-series model should be developed.

Covariance and Correlation (1)

- Consider the model that describes relationship between X_1 and X_2:

$$(X_1 - \mu_1) = \beta(X_2 - \mu_2) + \varepsilon$$

ε is a random variable with mean *0* and is independent of X_2

  - □ $\beta = 0$, X_1 and X_2 are statistically independent
  - □ $\beta > 0$, X_1 and X_2 tend to be above or below their means together
  - □ $\beta < 0$, X_1 and X_2 tend to be on opposite sides of their means

- Covariance between X_1 and X_2 :

$$\operatorname{cov}(X_1, X_2) = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E(X_1 X_2) - \mu_1 \mu_2$$

  - □ where $cov(X_1, X_2) \begin{cases} = 0, \\ < 0, \\ > 0, \end{cases}$ then $\beta \begin{cases} = 0 \\ < 0 \\ > 0 \end{cases}$

Covariance and Correlation (2)

- Correlation between X_1 and X_2 (values between *-1* and *1*):

$$\rho = \text{corr}(X_1, X_2) = \frac{\text{cov}(X_1, X_2)}{\sigma_1 \sigma_2}$$

  - □ where $corr(X_1, X_2) \begin{cases} = 0, \\ < 0, \\ > 0, \end{cases}$ then $\beta \begin{cases} = 0 \\ < 0 \\ > 0 \end{cases}$
  - □ The closer ρ is to *-1* or *1*, the stronger the linear relationship is between X_1 and X_2.

Summary

- In this chapter, we described the 4 steps in developing input data models:
  - Collecting the raw data
  - Identifying the underlying statistical distribution
  - Estimating the parameters
  - Testing for goodness of fit