0405324: Stochastic System Simulation

Lecture 5: Inputs analysis

Last revision June 21, 2009

Content

- **Input analysis**
  - Specifying input distributions, parameters
  - Deterministic vs. random input
  - Collecting and using data
  - Fitting input distributions
  - Fitting input distributions via Arena Input Analyzer
  - No data?

Deterministic vs. Random Inputs

- In addition to the structural (conceptual) model, we need to model the ***mathematical aspects*** of the model inputs (interarrival time distribution,…)
  - Either observe the system if it exists
  - Or use specifications of the system, collect data, analyze them, and come up with reasonable models of how they will be specified in the simulation

Deterministic vs. Random Inputs

- ***Deterministic*: nonrandom, fixed values**
  - Number of units of a resource
  - Entity transfer time (?)
  - Interarrival, processing times (?)
- ***Random* (*stochastic*): model as a distribution, "draw" or "generate" values from to drive simulation**
  - Transfer, Interarrival, Processing times
  - What distribution? What distributional parameters?
  - Causes simulation output to be random, too
- **Don't just assume randomness away – validity**

Collecting Data

- **Generally hard, expensive, frustrating, boring**
  - System might not exist
  - Data available on wrong things (processing time including changeover time) – might have to change model according to what's available
  - Incomplete, "dirty" data (processing time that includes the failure time)
  - Usually either too little data (new system) or Too much data (!)
- **Sensitivity of outputs to uncertainty in inputs→ use sensitivity analysis to check the effect of wrong inputs on outputs**
- **Match model detail to quality of data: do not detail a part of the model without having the required detailed data (reliable)**
- **Cost of collecting the data – should be budgeted in the project**
- **Capture variability in data – model validity**
- **Garbage In, Garbage Out (GIGO)--It means that if invalid data is entered in a computer program, the resulting output will also be invalid.**

Using Data: Alternatives and Issues

- **Use data "directly" in simulation**
  - Read actual observed values to drive model inputs (interarrivals, service times, part types, …)
    - Arena ReadWrite module ... see Model 10-2 Input.txt (Chapter 10 in the textbook)
  - All values will be "legal" and realistic
  - But can never go outside your observed data
  - May not have enough data for long or many runs
  - The data may not be valid anymore (e.g. change in the product features)
  - Computationally slow (reading disk files)
- **Or, fit probability distribution to data**
  - "Draw" or "generate" synthetic observations (methods for generating various types of **observations** from simulation data) from this distribution to drive model inputs
  - We've done it this way so far
  - Can go beyond observed data (good and bad)
  - May not get a good "fit" to data – validity?

Using Data: Alternatives and Issues

- **Be careful when using data**
  - 1000 observations of the time to pass a security gate (airport) are collected
  - Mean = standard deviation = almost 30 seconds → maybe exponentially distributed (mean = std deviation, and interarrival time)
  - However, usually it takes few seconds to pass through a security gate
  - More analysis (Histogram: Figure below) → two populations: one without triggering the security alarm and one with the security alarm triggered
  - Need three models of data:
    - One for the time to pass through without setting the detector (alarm)
    - One for the time to pass through while triggering the detector (alarm)
    - One for the chance for a person to trigger the alarm

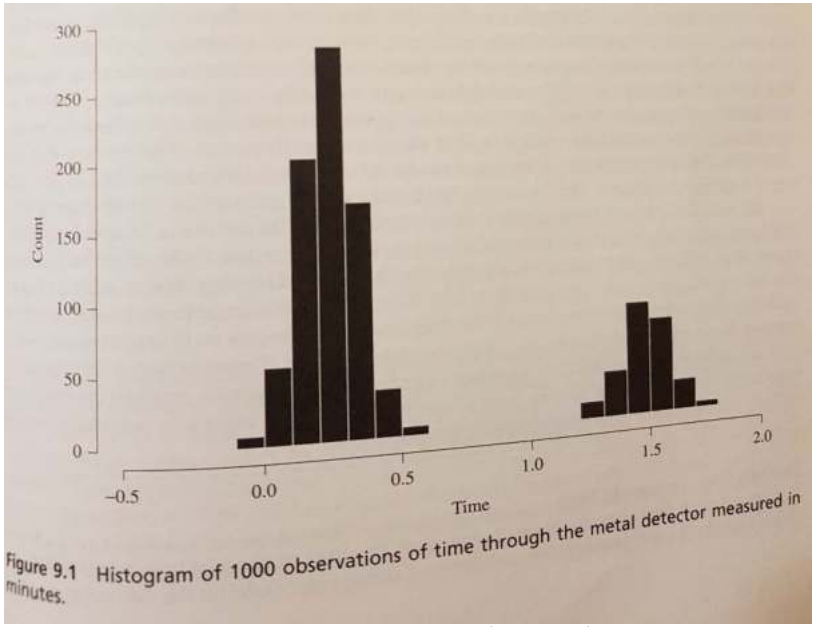


Figure 9.1 Histogram of 1000 observations of time through the metal detector measured in minutes.

Source: Banks et al., 2010

Fit probability distribution of data

- **Goodness-of-fit tests**
  - to determine whether data fits a specified distribution (observed and expected frequencies are significantly different or not)
  - to check whether data is independent
  - **Chi square test (large sample, e.g. N ≥ 50),**
  - **Kolmogorov-Smirnov (K-S) test (small sample size, e.g. N < 50)**
  - Most important part: ***p-value***, always between 0 and 1(**"Small *p*" (< 0.05 or so): poor fit (try again or give up))**

    - **Probability plot (good for small sample)**
    - **ARENA Input analyzer can help you to fit the probability distributions to data observed**

    Tools > Input Analyzer → load data from an existing data file or create a new one

- **Minitab/ SPSS software can also be used**

Chi-Square Test Statistic

H_0: The r.v. conforms to the distributional assumption

H_1: The r.v. does not conforms to the distributional assumption

- **The test statistic is**

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \qquad (\text{where } \mathrm{df} = k - p - 1)$$

Where:

k = number of categories or cells (classes in the histogram)

p = # of parameters of the hypothesized distribution estimated by sample statistics

O_i = observed frequency for category or cell (class) *i*

E_i = expected frequency for category or cell *i* *(should be at least five)*

The Rejection Region

H_0: The r.v. conforms to the distributional assumption

H_1: The r.v. does not conforms to the distributional assumption

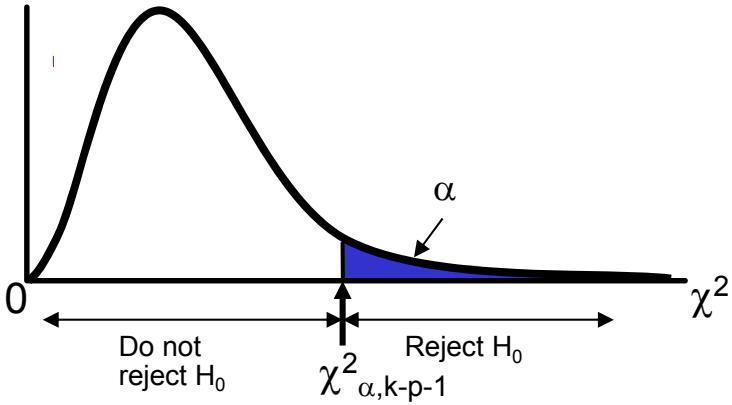
Test statistic: $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$

P-Value = Probability under the chi-square distribution (with (k-p-1) degrees of freedom), that is above the computed value of the test statistic χ^2

P-Value = Pr($\chi^2_{k-p-1} > \chi^2$)

For a given level of significance α, find the critical level: $\chi^2_{\alpha,k-p-1}$ form Chi-square table

Reject H_0, if $\chi^2 > \chi^2_{\alpha,k-p-1}$



Distribution often used in simulation

- **Chi-square test for discrete probability distribution**
  - Uniform distribution
  - Poisson distribution
- **Chi-square test for continuous probability distribution**
  - Normal distribution
  - Exponential distribution
  - Continuous Uniform distribution
  - Gamma distribution
  - Weibull distribution

Review of discrete distribution: Uniform

Meaning: models incomplete uncertainty → all outcomes are equally likely.

Often inappropriately used when there is no data

A random variable X has a **discrete uniform distribution** if each of the n values in its range, say, $x_1, x_2, \ldots, x_n$, has equal probability. Then,

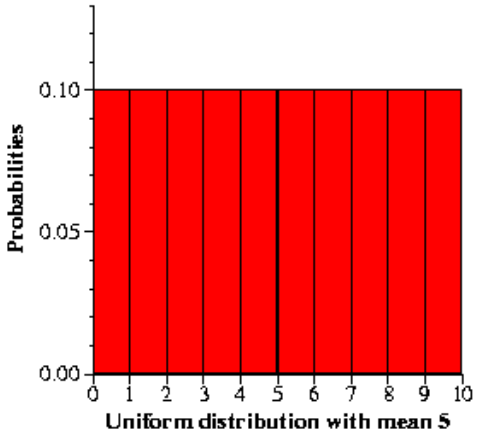
$$f(x_i) = 1/n \qquad (3\text{-}5)$$

Suppose X is a discrete uniform random variable on the consecutive integers $a, a+1, a+2, \ldots, b$, for $a \leq b$. The mean of X is

$$\mu = E(X) = \frac{b+a}{2}$$

The variance of X is

$$\sigma^2 = \frac{(b-a+1)^2 - 1}{12} \qquad (3\text{-}6)$$



Review of discrete distribution: Poisson

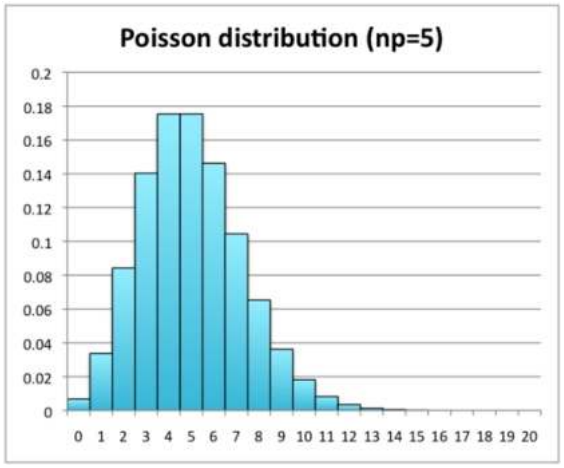
Meaning: models the number of independent events that happens in a fixed amount of time: number of customers arriving to a store in one hour

The random variable X that equals the number of events in the interval is a **Poisson random variable** with parameter $0 < \lambda$, and the probability mass function of X is

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad x = 0, 1, 2, \ldots \qquad (3\text{-}16)$$

If X is a Poisson random variable with parameter λ, then

$$\mu = E(X) = \lambda \quad \text{and} \quad \sigma^2 = V(X) = \lambda \qquad (3\text{-}17)$$



Review of continuous distribution: Uniform

Meaning: models incomplete uncertainty → all outcomes are equally likely.

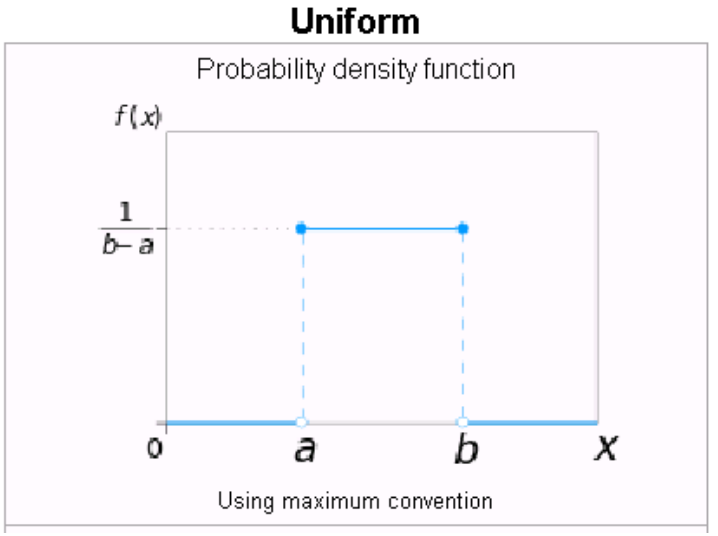
Often inappropriately used when there is no data

- The probability density function, $f(x)$ of U(a, b):

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

$$\mu = \frac{1}{2}(a+b)$$

$$\sigma^2 = \frac{1}{12}(b-a)^2$$



- The cumulative distribution function, $F(x)$ of U(a, b):

$$F(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a, b) \\ 1 & \text{for } x \geq b \end{cases}$$

Review of continuous distribution: Normal

Meaning: Models the distribution of a process that can be thought of as a sum of a number of component processes; for example, the time to assemble a product that is the sum of the times of each assembly operation.
- Admits negative values → impossible for time
- Widely used to model population's characteristics (weight, height,…)

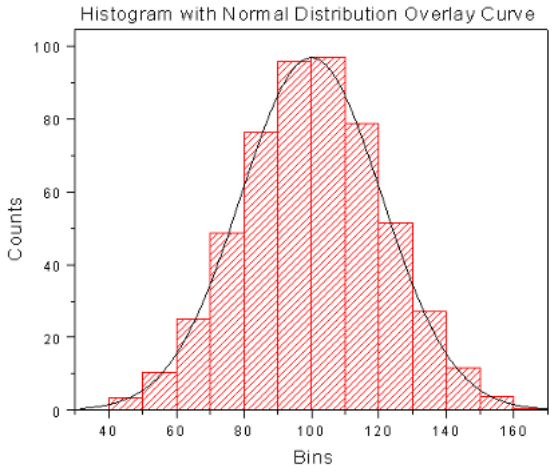
A random variable X with probability density function

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \qquad -\infty < x < \infty \qquad (4\text{-}8)$$

is a **normal random variable** with parameters μ, where $-\infty < \mu < \infty$, and $\sigma > 0$. Also,

$$E(X) = \mu \quad \text{and} \quad V(X) = \sigma^2 \qquad (4\text{-}9)$$

and the notation $N(\mu, \sigma^2)$ is used to denote the distribution. The mean and variance of X are shown to equal μ and σ^2, respectively, at the end of this Section 5-6.



Review of continuous distribution: Exponential

Meaning: Models the time between independent events. It is memoryless

If the number of arrivals in one hour (or time unit) follows an Poisson distribution with rate λ, then the time between two arrivals is exponential with rate λ, i.e. with mean of β=1/λ.

- The general formula for the probability density function of the exponential distribution is

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

where β is the scale parameter (e.g. mean time between successive events). The scale parameter is often referred to as λ = 1/β.

Thus, $f(x) = \lambda e^{-\lambda x}$ **for x > 0, λ > 0**

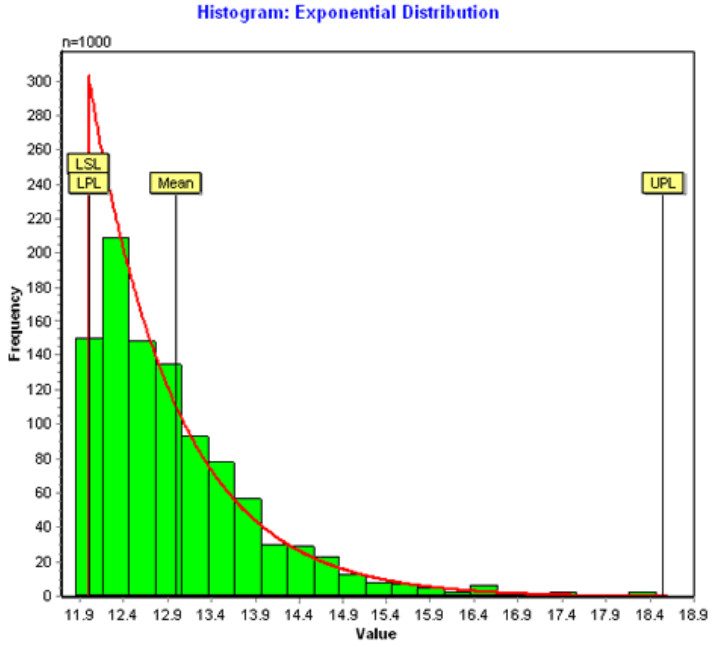
where λ is the mean number of events per unit time.

Review of continuous distribution: Exponential

If x is a exponential random variable with parameter λ (or β)

$$\beta = \frac{1}{\lambda}$$

$$\sigma^2 = \beta^2 = \frac{1}{\lambda^2}$$



The cumulative distribution function is given by,

$$F(x) = P(X \le x) = 1 - e^{-\lambda x}$$

Parameter estimation from sample data: ungrouped data

After the distribution has been selected, its parameters should be estimated.
Preliminary statistics: sample mean and variance as estimators of the hypothesized distribution parameters.

For ungrouped (into a frequency histogram) discrete or continuous data:

The sample mean, $\bar{x}$, is defined by $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

And the sample variance, s^2, is defined by

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2}{n-1}$$

Parameter estimation from sample data: discrete grouped data

The sample mean, $\bar{x}$, is defined by

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{n}, \quad n = \sum_{i=1}^{k} f_i$$

And the sample variance, s^2, is defined by

$$s^2 = \frac{\sum_{i=1}^{k} f_i (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{k} f_i x_i^2 - \frac{\left(\sum_{i=1}^{k} f_i x_i\right)^2}{n}}{n-1} = \frac{\sum_{i=1}^{k} f_i x_i^2 - n \cdot \bar{x}^2}{n-1}$$

k is the number of different values of the variable x (classes) and f_i is the observed frequency.

Suggested estimators for distribution often used in Simulation

Distribution	Parameters	Estimators
Uniform on (0, b)	b	$\hat{b} = [(n+1)/n] \cdot [\max(x_i)]$
Poisson	λ	$\hat{\lambda} = \bar{x}$
Normal	μ, σ^2	$\hat{\mu} = \bar{x},\ \hat{\sigma}^2 = s^2$
Exponential	β	$\hat{\beta} = \bar{x}$ **(note :** $\hat{\lambda} = 1/\hat{\beta} = 1/\bar{x}$**)**

Note: x is a random variable that depends on the sample values

Chi-Square Goodness-of-Fit Test: Uniform

- **Are technical support calls equal across all days of the week? (i.e., do calls follow a uniform distribution? $\alpha = 0.05$)**
  - **Sample data for 7 days per of week:**

	Sum of calls for this day:
Monday	290
Tuesday	250
Wednesday	238
Thursday	257
Friday	265
Saturday	230
Sunday	192
	$\Sigma = 1722$ (= n)

Chi-Square Goodness-of-Fit Test: Uniform

Solution:

- **If calls** are **uniformly distributed, the 1722 calls would be expected to be equally divided across the 7 days (E = np(x) = 1722 x 1/7 = 246):**

$$\frac{1722}{7} = 246 \quad \text{expected calls per day if uniform}$$

- **Chi-Square Goodness-of-Fit Test: test to see if the sample results are consistent with the expected results**

Chi-Square Goodness-of-Fit Test: Uniform

Observed vs. Expected Frequencies

	Observed O_i	Expected E_i
Monday	290	246
Tuesday	250	246
Wednesday	238	246
Thursday	257	246
Friday	265	246
Saturday	230	246
Sunday	192	246
TOTAL	1722	1722

Chi-Square Goodness-of-Fit Test: Uniform

H_0: The distribution of calls is uniform over days of the week

H_1: The distribution of calls is not uniform

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

$$\chi^2 = \frac{(290-246)^2}{246} + \frac{(250-246)^2}{246} + \ldots + \frac{(192-246)^2}{246} = 23.05$$

k-p-1 = 7-0-1 = 6 degrees of freedom:

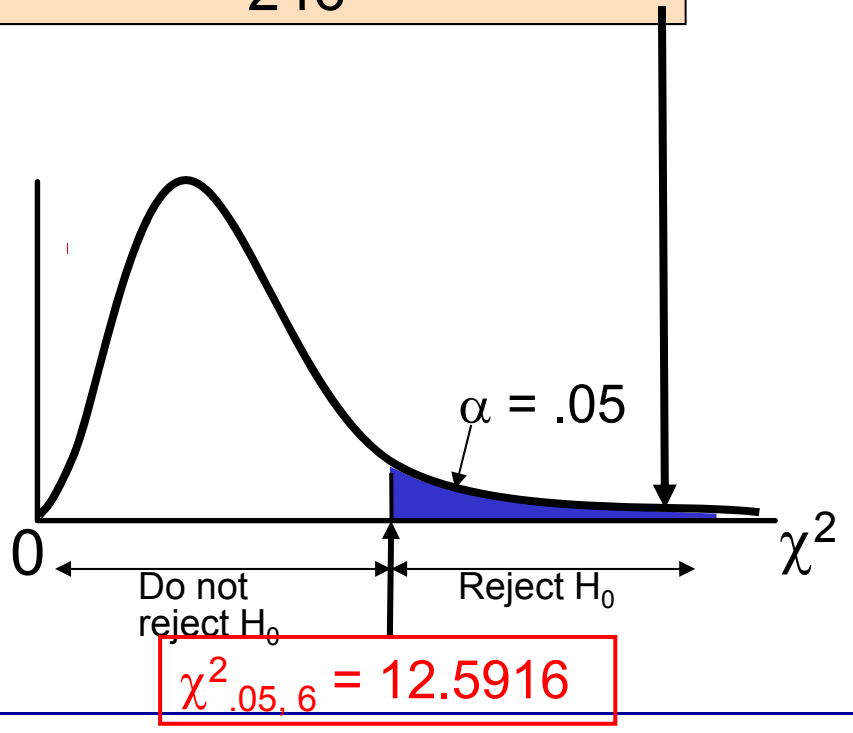
$$\chi^2_{\alpha,k-p-1} = \chi^2_{.05,\,6} = 12.5916$$

Conclusion:

$\chi^2 = 23.05 > \chi^2_{\alpha,k\text{-}p\text{-}1} = 12.5916$ so reject H_0 and conclude that the distribution is not uniform

P-value (at 6 df and $\chi^2 = 23.05$) < 0.005

Decision: P-value < α (= 0.05)—reject H0



Chi-Square Goodness-of-Fit Test: Poisson

The number of vehicles arriving at the northwest corner of an intersection in a 5-minute period between 7:00 a.m. and 7:05 a.m. was monitored for five workdays over a 20-week period. The following table shows the resulting data. The first entry in the table indicates that there were 12 five-minute periods during which zero vehicles arrived, 10 five-minute periods during which one vehicle arrived, and so on.

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

Does the number of vehicle arrivals follow a Poisson distribution?
Consider the significance level $\alpha = 0.05$

Chi-Square Goodness-of-Fit Test: Poisson

Solution:

The random variable (vehicle arrival data) follow a Poisson distribution with parameter,

$$\lambda = \bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{n} = \frac{0\times 12 + 1\times 10 + \ldots + 1\times 11}{100} = 3.64$$

$$n = \sum_{i=1}^{k} f_i = 100$$

H0: the random variable (vehicle arrival data) is Poisson distributed
H1: the random variable (vehicle arrival data) is not Poisson distributed

Chi-Square Goodness-of-Fit Test: Poisson

$$P(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, P_1 = P(x=0) = \frac{e^{-3.64}3.64^{0}}{0!} = 0.026$$

$$E_1 = n \times P(x=0) = 100 \times 0.026 = 2.6$$ **(Rounded up to one decimal point)**

x_i	Observed Frequency, O_i	Expected Frequency, E_i
0	12	2.6
1	10	9.6
2	19	17.4
3	17	21.1
4	10	19.2
5	8	14.0
6	7	8.5
7	5	4.4
8	5	2.0
9	3	0.8
10	3	0.3
11	1	0.1
	100	100.0

Chi-Square Goodness-of-Fit Test: Poisson

Since $E_1 = 2.6 < 5$, E_1 and E_2 are combined and that O_1 and O_2 are also combined. The last five class intervals are also combined for the same reason. Therefore, k is also reduced by five.

x_i	Observed Frequency, O_i	Expected Frequency, E_i	$(O_i - E_i)^2/E_i$
0	12 } 22	2.6 } 12.2	7.87
1	10	9.6	
2	19	17.4	0.15
3	17	21.1	0.80
4	10	19.2	4.41
5	8	14.0	2.57
6	7	8.5	0.26
7	5	4.4	
8	5	2.0	
9	3 } 17	0.8 } 7.6	11.62
10	3	0.3	
11	1	0.1	
	100	100.0	27.68

Chi-Square Goodness-of-Fit Test: Poisson

Decision:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = 27.68$$

The calculated χ_0^2 is 27.68. The degrees of freedom for the tabulated value of χ^2 is k-p-1 = 7-1-1 = 5. Here, p = 1, since one parameter (λ) was estimated from the data. At $\alpha = 0.05$, the critical value $\chi^2_{0.05,5}$ is 11.1. Thus, H_0 would be rejected at level of significance 0.05, i.e., the distribution is not Poisson. The analyst must now search for a better-fitting model or use the empirical distribution of the data.

P-value (at 5 df and $\chi^2 = 27.68$) < 0.005

Decision: P-value < α (= 0.05)—reject H0

Chi-Square test for continuous distribution: formation of class intervals

A common practice in constructing the class intervals (or cells) for the frequency distribution is to choose the cell boundaries so that the expected frequencies $E_i = np_i$ are equal for all cells.

We assume that the Class intervals are equal in probability (instead of being equal in width), i.e.,

$$p_i = 1 / k$$

since $E_i = np_i = n/k \geq 5 \Rightarrow n / k \geq 5$

and solve for k yields

$$k \leq n / 5$$

Where, k is the number of cells and n is the sample size.

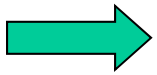
Chi-Square Goodness-of-Fit Test: Normal

- **Do measurements from a production process follow a normal distribution with μ = 50 and σ = 15?**
- **Steps:**
  - Get sample data
  - Group sample results into classes (cells)
  - (Expected cell frequency must be at least 5 for each cell)
  - Compare actual cell frequencies with expected cell frequencies

Chi-Square Goodness-of-Fit Test: Normal

- **Sample data (integer data) and values grouped into classes:**

150 Sample Measurements
80
65
36
66
50
38
57
77
59
…etc…



Class	Frequency
$x < 30$	10
$30 \le x < 40$	21
$40 \le x < 50$	33
$50 \le x < 60$	41
$60 \le x < 70$	26
$70 \le x < 80$	10
$80 \le x < 90$	7
$90 \le x$	2
TOTAL	n=150

Chi-Square Goodness-of-Fit Test: Normal

- What are the expected frequencies for these classes for a normal distribution with $\mu = 50$ and $\sigma = 15$?

Class	Frequency	Expected Frequency
$x < 30$	10	?
$30 \le x < 40$	21	
$40 \le x < 50$	33	
$50 \le x < 60$	41	
$60 \le x < 70$	26	
$70 \le x < 80$	10	
$80 \le x < 90$	7	
$90 \le x$	2	
TOTAL	n=150	

Chi-Square Goodness-of-Fit Test: Normal

Value	P(X < value)	Expected frequency
x < 30	0.09121	13.68
30 ≤ x < 40	0.16128	24.19
40 ≤ x < 50	0.24751	37.13
50 ≤ x < 60	0.24751	37.13
60 ≤ x < 70	0.16128	24.19
70 ≤ x < 80	0.06846	10.27
80 ≤ x < 90	0.01892	2.84
90 ≤ x	0.00383	0.57
TOTAL	1.00000	150.00

Expected frequencies in a sample of size n=150, from a normal distribution with μ=50, σ=15

Example:

$$P_1 = P(x < 30) = P\left(z < \frac{30-50}{15}\right) = P(z < -1.3333) = .0912$$

$$E_1 = P_1 \cdot n = (0.0912)(150) = 13.68$$

Chi-Square Goodness-of-Fit Test: Normal

Class	Frequency (observed, O_i)	Expected Frequency, E_i
x < 30	10	13.68
30 ≤ x < 40	21	24.19
40 ≤ x < 50	33	37.13
50 ≤ x < 60	41	37.13
60 ≤ x < 70	26	24.19
70 ≤ x < 80	10	10.27
80 ≤ x < 90	7	2.84
90 ≤ x	2	0.57
TOTAL	150	150.00

The test statistic is

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Reject H_0, if

$$\chi^2 > \chi^2_{\alpha,k-p-1}$$

Chi-Square Goodness-of-Fit Test: Normal

H_0: The distribution of the process is normal with μ = 50 and σ = 15

H_1: The distribution of process does not have normal distribution

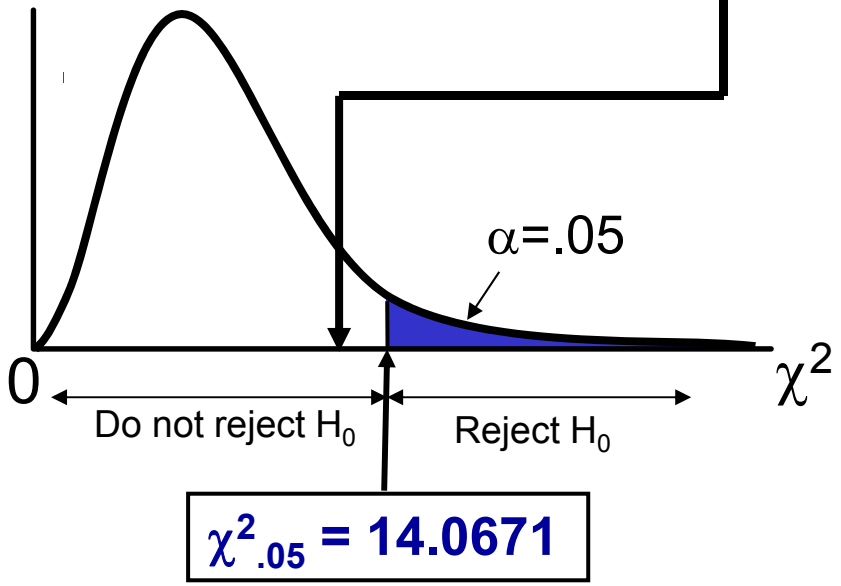
$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i} = \frac{(10-13.68)^2}{13.68} + \ldots + \frac{(2-0.57)^2}{0.57} = 12.097$$

df = 8 – 0 – 1 = 7 (p = 0, μ and σ given)

$\chi^2_{.05,7} = 14.0671$

Conclusion:

$\chi^2 = 12.097 < \chi^2_\alpha = 14.0671$ so do not reject H_0 and that the distribution is normal



Chi-Square Goodness-of-Fit Test: Normal (class intervals are not given)

- **A manufacturing engineer is testing a power supply used in a notebook computer and wishes to determine whether output voltage is adequately described by a normal distribution. To test the hypothesis, a sample of 100 units are tested and the output voltage are recorded as follows (all data are not shown here):**

  3.919, 4.081, 0.062, 1.961, 5.845, 3.027, 6.505, 0.021, 1.192……..

  Mean and standard deviation of the output voltage are estimated based on sample data and are 5.04 Volt and 0.08 Volt, respectively.

  (i) Construct the class intervals so that the expected frequencies are equal for all cells.

  (ii) Conduct the chi-square goodness of fit test to check whether the engineer's assumption is correct (α = 0.05).

Chi-Square Goodness-of-Fit Test: Normal (class intervals are not given)

Solution:

(i) Class intervals

In order to perform the chi-square test with *intervals of equal probability (uniform distribution)*, the endpoints of the class intervals must be determined.

The number of intervals k is given by (n = 100), k ≤ n / 5 ➔ k ≤ 100 / 5 = 20.

Suppose k = 8, then each interval will have probability, p = 1/k = 1/8 = 0.125. The endpoints for each interval a_i are computed from the cdf for the normal distribution, as follows:

$$F(a_i) = P(x \le a_i) = \Phi\left(\frac{a_i - \mu}{\sigma}\right), \quad -\infty \le a_i \le +\infty$$

$$\Rightarrow i \cdot p = \Phi\left(\frac{a_i - \mu}{\sigma}\right), F(a_i) \text{ is the cumulative area from 0 to } a_i, \text{ thus, } F(a_i) = i \cdot p$$

$$\Rightarrow \Phi^{-1}(i \cdot p) = \frac{a_i - \mu}{\sigma}$$

$$\Rightarrow a_i = \Phi^{-1}(i \cdot p) \cdot \sigma + \mu; \quad \text{note, } p = 1/k = 1/8 = 0.125$$

$$\Rightarrow a_i = \Phi^{-1}(i \cdot 0.125) \cdot \sigma + \mu, \; i = 1,2,\ldots,k-1; \; a_0 = -\infty, a_k = +\infty$$

Replace i by 1, 2, …, 7 and form the class intervals (see next slide)

Chi-Square Goodness-of-Fit Test: Normal (class intervals are not given)

$$a_i = \Phi^{-1}(i \cdot 0.125) \cdot \sigma + \mu,$$

$$a_1 = \Phi^{-1}(1 \times 0.125) \cdot \sigma + \mu = \Phi^{-1}(0.125) \cdot 0.08 + 5.04$$
$$= -1.15 \times 0.08 + 5.04 = 4.948$$

Class Interval
$x < 4.948$
$4.948 \le x < 4.986$
$4.986 \le x < 5.014$
$5.014 \le x < 5.040$
$5.040 \le x < 5.066$
$5.066 \le x < 5.094$
$5.094 \le x < 5.132$
$5.132 \le x$

Chi-Square Goodness-of-Fit Test: Normal (class intervals are not given)

(ii) Chi-square test

Class Interval	Observed Frequency o_i	Expected Frequency E_i
$x < 4.948$	12	12.5
$4.948 \leq x < 4.986$	14	12.5
$4.986 \leq x < 5.014$	12	12.5
$5.014 \leq x < 5.040$	13	12.5
$5.040 \leq x < 5.066$	12	12.5
$5.066 \leq x < 5.094$	11	12.5
$5.094 \leq x < 5.132$	12	12.5
$5.132 \leq x$	14	12.5
Totals	100	100

Chi-Square Goodness-of-Fit Test: Normal (class intervals are not given)

1. The variable of interest is the form of the distribution of power supply voltage.
2. H_0: The form of the distribution is normal.
3. H_1: The form of the distribution is nonnormal.
4. $\alpha = 0.05$
5. The test statistic is

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

6. Since two parameters in the normal distribution have been estimated, the chi-square statistic above will have $k - p - 1 = 8 - 2 - 1 = 5$ degrees of freedom. Therefore, we will reject H_0 if $\chi_0^2 > \chi_{0.05,5}^2 = 11.07$.
7. Computations:

$$\begin{aligned}\chi_0^2 &= \sum_{i=1}^{8} \frac{(o_i - E_i)^2}{E_i} \\ &= \frac{(12 - 12.5)^2}{12.5} + \frac{(14 - 12.5)^2}{12.5} + \cdots + \frac{(14 - 12.5)^2}{12.5} \\ &= 0.64\end{aligned}$$

Chi-Square Goodness-of-Fit Test: Normal (class intervals are not given)

8. Conclusions: Since $\chi_0^2 = 0.64 < \chi^2_{0.05,5} = 11.07$, we are unable to reject H_0, and there is no strong evidence to indicate that output voltage is not normally distributed.

P-value (at 5 df and $\chi^2 = 0.64$) = between 0.975 and 0.99

Decision: P-value > α (= 0.05)—do not reject H_0

Chi-Square Goodness-of-Fit Test: Exponential

- **Life tests were performed on a random sample of 50 PDP-11 electronic chips at 1.5 times the normal voltage, and their lifetime (or time to failure) in days was recorded:**

79.919	3.081	0.062	1.961	5.845	3.027	6.505	0.021	0.012	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

Check whether the life of chips follow an exponential distribution (α = 0.05)

Chi-Square Goodness-of-Fit Test: Exponential

Solution:

The parameter of the exponential distribution is estimated by sample data,

$$\beta = \bar{x} = \frac{79.919 + 3.081 + ... + 4.562}{50} = 11.90 \text{ days/chip}$$

$$\lambda = \frac{1}{\beta} = 0.084 \text{ chips/day}$$

H_0: the chips life is exponentially distributed

H_1: the chips life is not exponentially distributed

Chi-Square Goodness-of-Fit Test: Exponential

In order to perform the chi-square test with intervals of equal probability, the endpoints of the class intervals must be determined. The number of intervals should be less than or equal to n/5. Here, n=50, so that k ≤ 10. Let k = 8, then each interval will have probability p =1/k =1/8= 0.125. The endpoints a_i for each interval are computed from the cdf for the exponential distribution, as follows:

$$F(a_i) = P(x \le a_i) = 1 - e^{-\lambda a_i}, a_i \ge 0$$

$$\Rightarrow i \cdot p = 1 - e^{-\lambda a_i}, F(a_i) \text{ is the cumulative area from } 0 \text{ to } a_i, \text{ thus, } F(a_i) = i \cdot p$$

$$\Rightarrow e^{-\lambda a_i} = 1 - i \cdot p$$

$$\Rightarrow -\lambda \cdot a_i = \ln(1 - i \cdot 0.125), \quad \text{note, } p = 1/k = 1/8 = 0.125$$

$$\Rightarrow a_i = -\frac{1}{\lambda}\ln(1 - i \cdot 0.125), \; i = 1,2,...,k-1, \; a_0 = 0, a_k = \infty$$

Replace *i* by 1, 2, …, 7 (λ = 0.084): we get a_1,…a_7 as 1.590, 3.425, 5.595, 8.252, 11.677, 16.503, and 24.755. Now form the class intervals (see next slide)

Chi-Square Goodness-of-Fit Test: Exponential

Class Intervlas	Observed Frequency, O_i	Expected Frequency, E_i	$(O_i - E_i)^2 / E_i$
$x < 1.590$	19	6.25	26.01
$1.590 \leq x < 3.42$	10	6.25	2.25
$3.425 \leq x < 5.59$	3	6.25	0.81
$5.595 \leq x < 8.25$	6	6.25	0.01
$8.252 \leq x < 11.677$	1	6.25	4.41
$11.677 \leq x < 16.503$	1	6.25	4.41
$16.503 \leq x < 24.755$	4	6.25	0.81
$x > 24.75$	6	6.25	0.81
	50	50	39.6

Chi-Square Goodness-of-Fit Test: Exponential

The calculated value of χ_0^2 is 39.6. The degrees of freedom are given by k - p - 1 = 8 - 1 - 1 = 6.

At $\alpha = 0.05$, the tabulated value of $\chi^2_{0.05,6}$ is 12.6.

Since $\chi^2_{0.05,6} \leq \chi_0^2$, the null hypothesis is rejected, i.e. the distribution of chip's life is not exponential.

P-value (at 6 df and χ^2 = 39.6) < 0.005

Decision: P-value < α (= 0.05)—reject H0

Goodness-of-fit Test: Kolmogorov-Smirnov (K-S)

Assumption:

The assumptions are

1. Sample is a random sample

2. Hypothesized distribution $F_T(x)$ is continuous.

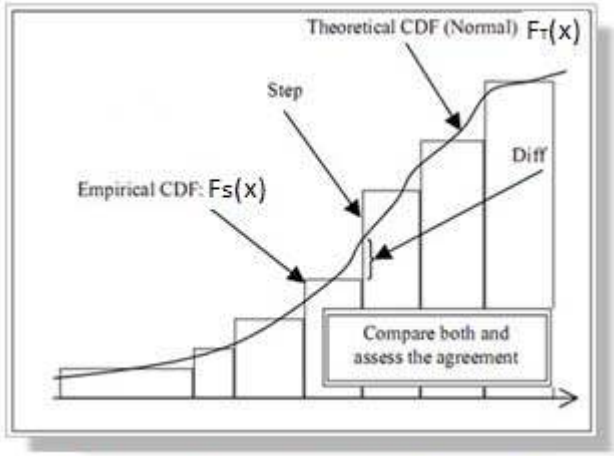


Figure 1. Distance Goodness of Fit Test Conceptual Approach

Test Statistic:

- **Kolmogorov-Smirnov (K-S) goodness-of-fit test a comparison between some theoretical cumulative distribution function, $F_T(x)$, and a sample cumulative distribution function $F_S(x)$.**
- **The difference between $F_T(x)$ and $F_S(x)$ is measured by the statistic *D* which is the greatest vertical distance between $F_S(x)$ and $F_T(x)$ (see the figure)**

Goodness-of-fit Test: Kolmogorov-Smirnov (K-S)

- When a two sided test is appropriate, that is when the hypothesis are

  H_0: $F_S(x) = F_T(x)$ for all x from $-\infty$ to $+\infty$

  H_1: $F_S(x) \neq F_T(x)$ for at least one x

then the statistic is :

$$D = \sup_x |F_s(x) - F_t(x)|$$

which is read, " D equals the supremum (greatest, or maximum) over all x, of the absolute value of the difference $F_s(x) - F_T(x)$."

- The H_0 is rejected at the α level of significance if the computed value of D exceeds the critical value (from table at given α and n).

- When values of D are based on a discrete theoretical distribution, the test is conservative.

- When the test is used with discrete data, then, the true probability of committing a type-I error is at most equal to α, the stated level of significance.

Goodness-of-fit Test: Kolmogorov-Smirnov (K-S)—Uniform dist.

Example:

The following data points come from an experiment:

1.41 0.26 1.97 0.33 0.55 0.77 1.46 1.18

Is there any evidence to say that the data are randomly sampled from a continuous Uniform(0, 2) distribution? **(α = 0.05).**

Goodness-of-fit Test: Kolmogorov-Smirnov (K-S)—Uniform dist.

Solution:

We assume that the sample available is a random sample from a continuous population distribution.

Hypothesis:

H_0: $F_S(x) = F_T(x)$ for all x from a to b

H_1: $F_S(x) \neq F_T(x)$ for at least one x.

Goodness-of-fit Test: Kolmogorov-Smirnov (K-S)—Uniform dist.

Solution:

We assume that the sample available is a random sample from a continuous population distribution.

Hypothesis:

H_0: $F_S(x) = F_T(x)$ for all x from a to b

H_1: $F_S(x) \neq F_T(x)$ for at least one x.

We get from cont. uniform dist (a = 0, b = 2):

$$f(x) = 1/(b-a) = 1/2$$

$$F_T(x) = (x-a)/(b-a)=x/2$$

Goodness-of-fit Test: Kolmogorov-Smirnov (K-S)—Uniform dist.

In ascending order

X	F	C.F	$F_S(x)$	X	$F_T(x)$	$\lvert F_S(x) - F_T(x) \rvert$
0.26	1	1	0.125	0.26	0.130	0.005
0.33	1	2	0.250	0.33	0.165	0.085
0.55	1	3	0.375	0.55	0.275	0.100
0.77	1	4	0.500	0.77	0.385	0.115
1.18	1	5	0.625	1.18	0.590	0.035
1.41	1	6	0.750	1.41	0.705	0.045
1.46	1	7	0.875	1.46	0.730	0.145
1.97	1	8	1.000	1.97	0.985	0.015

Decision:

Here, D (= 0.145) < critical value (= 0.46—from K-S table at n =8, α = 0.05). So we are not willing to reject H_0, that is the sample may have come from uniform distribution.

Try yourself:

Test whether the data points (slide # 50) come from an exponential distribution with parameter of 2?

Goodness-of-fit Test: Kolmogorov-Smirnov (K-S)—Normal dist.

Example:

Fasting, blood glucose determinations made on 36 non-obese, apparently healthy, adult males are given in the following table. We have to test if we may conclude that these data are not from a normally distributed population with a mean of 80 and a standard deviation of 6 (α = 0.05).

75	92	80	80	84	72
84	77	81	77	75	81
80	92	72	77	78	76
77	86	77	92	80	78
68	78	92	68	80	81
87	76	80	87	77	86

Goodness-of-fit Test: Kolmogorov-Smirnov (K-S)—Normal dist.

Solution:

We assume that the sample available is a random sample from a continuous population distribution.

Hypothesis:

H_0: $F_S(x) = F_T(x)$ for all x from $-\infty$ to $+\infty$

H_1: $F_S(x) \neq F_T(x)$ for at least one x.

Goodness-of-fit Test: Kolmogorov-Smirnov (K-S)—Normal dist.

Distribution of test statistic:

Critical value of the test statistic for selected values of *α* and *n* are given in K-S Table.

Decision rule:

Reject H_0 if the computed value of D exceeds 0.226 (the critical value of D for n = 36 and α = 0.05– from Table).

Calculation of test statistic:

First step is to compute values of $F_S(x)$. Each value of $F_S(x)$ is obtained by dividing the corresponding cumulative frequency by the sample size like first value of $F_S(x) = 2/36 = 0.0556$

Value of $F_T(x)$ can be obtained by converting each observed value of x to a value of the standard normal variable, z. Then find the area between $-\infty$ and z.

Goodness-of-fit Test: Kolmogorov-Smirnov (K-S)—Normal dist.

In ascending order

X	F	C.F	$F_S(x)$	X	Z = (x - 80)/6	$F_T(x)$	$\lvert F_S(x) - F_T(x) \rvert$
68	2	2	.0556	68	-2.00	.0228	.0328
72	2	4	.1111	72	-1.33	.0918	.0193
75	2	6	.1667	75	-.83	.2033	.0366
76	2	8	.2222	76	-.67	.2514	.0292
77	6	14	.3889	77	-.50	.3085	.0804
78	3	17	.4722	78	-.33	.3707	.1015
80	6	23	.6389	80	.00	.5000	.1389
81	3	26	.7222	81	.17	.5675	.1547
84	2	28	.7778	84	.67	.7486	.0292
86	2	30	.8333	86	1.00	.8413	.0080
87	2	32	.8889	87	1.17	.8790	.0099
92	4	36	1.000	92	2.00	.9772	.0228

Decision:

Here, D (= 0.1547) < critical value (= 0.226). So we are not willing to reject H_0, that is the sample may have come from normal distribution.

Fitting Distributions to Data with Arena Input Analyzer

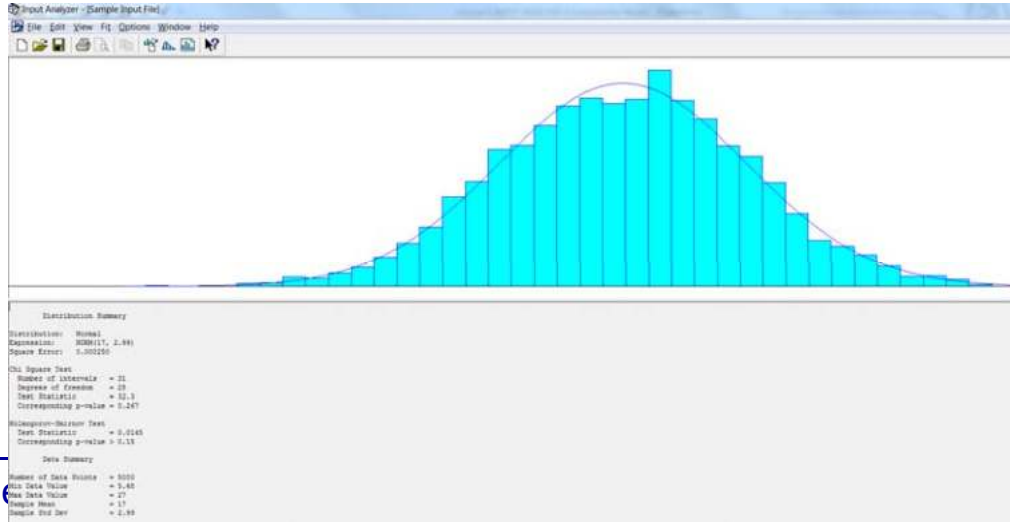
- **Assume:**
  - Have sample data: Independent and Identically Distributed (IID) list of observed values from actual physical system
  - Want to select or fit a probability distribution for use in generating inputs for simulation model
- **Arena Input Analyzer**
  - Separate application, also via Tools menu in Arena
  - Fits distributions, gives valid Arena expression for generation to paste directly into simulation model

Fitting Distributions to Data with Arena Input Analyzer (cont'd.)

- **Fitting = deciding on distribution form (exponential, normal, empirical, etc.) and estimating its parameters**
  - Several different methods (Maximum likelihood, moment matching, least squares, …)
  - Assess goodness of fit via hypothesis tests
    - H_0: fitted distribution adequately represents data
    - Get *p* value for test (small = poor fit)
- **Fitted "theoretical" vs. empirical distribution**
- **Continuous vs. discrete data, distribution**
- **"Best" fit from among several distributions**

Data Files for Arena Input Analyzer

- **Create data file**
  - Editor, word processor, spreadsheet, ...
  - Plain ASCII text – save as text or export)
  - Values separated by white space – blanks, tabs, linefeeds
  - Otherwise free format
- **Open data file from within Input Analyzer**
  - *File* > *New* or 
  - *File* > *Data File* > *Use Existing* or 
  - Get histogram, basic summary of data
  - To see data file: *Window* > *Input Data*
- **Generate "fake" data file to play around**
  - *File* > *Data File* > *Generate New*



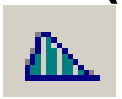
Fit Menu

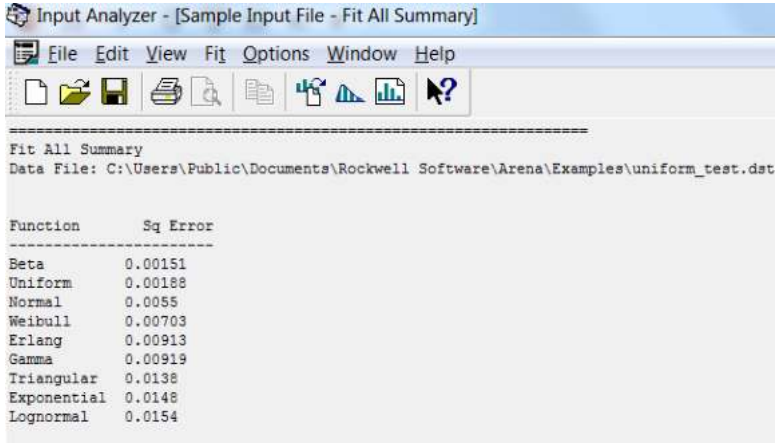
- **Fits distributions, does goodness-of-fit tests**
- **Fit a specific distribution form**
  - Plots density over histogram for visual "test"
  - Gives exact expression to Copy and Paste (Ctrl+C, Ctrl+V) over into simulation model
  - May include "offset" depending on distribution
  - Gives results of goodness-of-fit tests
    - Chi square, Kolmogorov-Smirnov tests
    - Most important part: *p-value*, always between 0 and 1:

      Probability of getting a data set that's more inconsistent with fitted distribution than data set you actually have, if fitted distribution is truly "the truth"

      "Small" *p* (< 0.05 or so): poor fit (try again or give up)

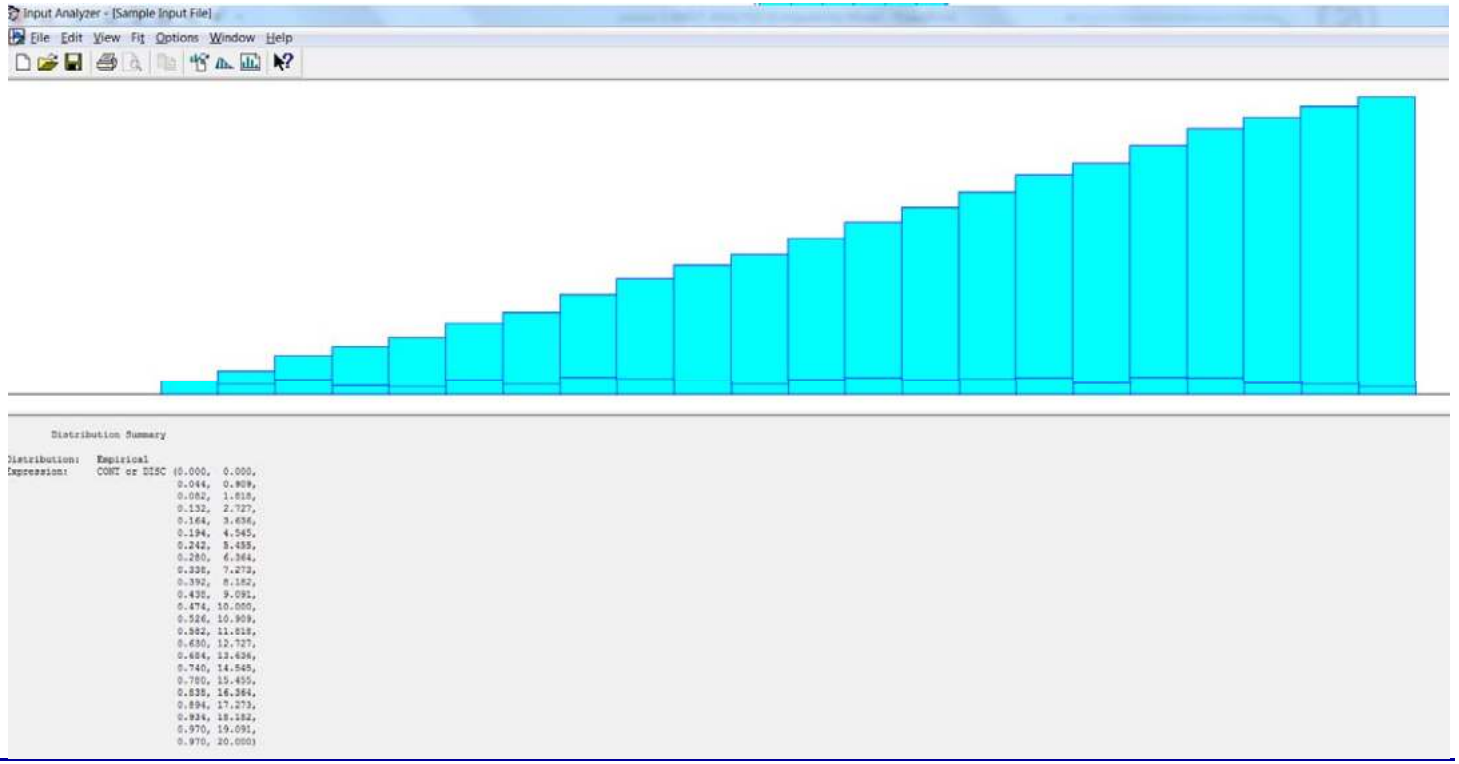
Fit Menu (cont'd.)

- **Fit all of Arena's (theoretical) distributions at once**
  - *Fit* > *Fit All* or 
  - Returns *minimum square-error* distribution
    - Square error = sum of squared discrepancies between histogram frequencies and fitted-distribution frequencies
    - Can depend on histogram intervals chosen: different intervals can lead to different "best" distribution
  - Could still be a poor fit, though (check *p* value)
  - To see all distributions, ranked: *Window* > *Fit All Summary* or 



Fit Menu (cont'd.)

- **"Fit" Empirical distribution (continuous or discrete):** ***Fit > Empirical***
  - Can interpret results as Discrete or Continuous distribution
    - Discrete: get pairs (*Cumulative* Probability, Value)
    - Continuous: Arena will linearly interpolate *within* data range according to these pairs (so you can never generate values outside range, which might be good or bad)
  - Empirical distribution can be used when "theoretical" distributions fit poorly, or intentionally



Issues in Fitting Input Distributions

- **Not an exact science – no "right" answer**
- **Consider theoretical vs. empirical**
- **Consider range of distribution**
  - Infinite both ways (e.g., normal)
  - Positive (e.g., exponential, gamma)
  - Bounded (e.g., beta, uniform)
- **Consider ease of parameter manipulation to affect means, variances**
- **Simulation model sensitivity analysis**
- **Outliers, multimodal data**
  - Maybe split data set (details in text)

No Data?

- **Happens more often than you'd like**
- **No good solution; some (bad) options:**
  - Interview "experts"
    - Min, Max: Uniform
    - Avg., % error or absolute error: Uniform
    - Min, Mode, Max: Triangular

      Mode can be different from Mean – allows asymmetry
  - Interarrivals – independent, stationary
    - Exponential – still need some value for mean
  - Number of "random" events in an interval: Poisson
  - Sum of independent "pieces": normal (heed left tail ...)
  - Product of independent "pieces": lognormal

Cautions on Using Normal Distributions

- **Probably most familiar distribution – normal "bell curve" used widely in statistical inference**
- **But it has infinite tails in both directions … in particular, has an infinite left tail so can always (theoretically) generate negative values**
  - Many simulation input quantities (e.g., time durations) must be positive to make sense – Arena truncates negatives to 0
- **If mean μ is big relative to standard deviation σ, then P(negative) value is small … one in a million**
  - But in simulation, *one in a million can happen*
- **Moral – avoid normal as input distribution**

See Tutorial-2 (Lecture-5)

Continued in Lecture 6