CS626 Data Analysis and Simulation

Instructor: Peter Kemper
R 104A, phone 221-3462, email:kemper@cs.wm.edu
Office hours: Monday, Wednesday 2-4 pm

Today:
Stochastic Input Modeling

NIST/SEMATECH e-Handbook of Statistical Methods,
http://www.itl.nist.gov/div898/handbook/
What is input modeling?

**Input modeling**

- Deriving a representation of the uncertainty or randomness in a stochastic simulation.

**Common representations**

- Measurement data
- Distributions derived from measurement data — focus of “Input modeling”
  - usually requires that samples are i.i.d and corresponding random variables in the simulation model are i.i.d
- i.i.d. = independent and identically distributed
- theoretical distributions
- empirical distribution
- Time-dependent stochastic process
- Other stochastic processes

**Examples include**

- time to failure for a machining process;
- demand per unit time for inventory of a product;
- number of defective items in a shipment of goods;
- times between arrivals of calls to a call center.
Overview of fitting with data

- Check if key assumptions hold (i.i.d)
- Select one or more candidate distributions
  - based on physical characteristics of the process and
  - graphical examination of the data.
- Fit the distribution to the data
  - determine values for its unknown parameters.
- Check the fit to the data
  - via statistical tests and
  - via graphical analysis.
- If the distribution does not fit,
  - select another candidate and repeat the process, or
  - use an empirical distribution.

from WSC 2010 Tutorial by Biller and Gunes, CMU, slides used with permission
Check the fit to the data

**Graphical analysis**

- Plot fitted distribution and data in a way that differences can be recognized
  - beyond obvious cases, there is a grey area of subjective acceptance/rejection

**Challenges**

- How much difference is significant enough to trash a fitted distribution?
- Which graphical representation is easy to judge?

**Options:**

- Histogram-based plots
- Probability plots: P-P plot, Q-Q plot

**Statistical tests**

- define a measure $X$ for the difference between fitted distribution & data
- $X$ is an RV, so if we find an argument what distribution $X$ has, we get a statistical test to see if in a concrete case a value of $X$ is significant

**Goodness-of-fit tests:**

- Chi-square test($\chi^2$), Kolmogorov-Smirnov test(K-S), Anderson Darling test(AD)
Check the fit to the data:

**Statistical tests**
- define a measure $X$ for the difference between fitted distribution & data
- Test statistic $X$ is an RV
  - say small $X$ means small difference, high $X$ means huge difference
- if we find an argument what distribution $X$ has, we get a statistical test to see if in a concrete case a value of $X$ is significant or not
  - Say $P(X \leq x) = (1-\alpha)$, and e.g. this holds for $x=10$ and $\alpha=.05$, then we know that if data is sampled from a given distribution and this is done $n$ times ($n\to\infty$), this measure $X$ will be below 10 in 95% of those cases.
  - If in our case, the sample data yields $x=10.7$, we can argue that it is too unlikely that the sample data is from the fitted distribution.

**Concepts, Terminology**
- Hypothesis $H_0$, Alternative $H_1$
- Power of a test: $(1-\beta)$, probability to correctly reject $H_0$
- Alpha / Type I error: reject a true hypothesis
- Beta / Type II error: not rejecting a false hypothesis
- P-value: probability of observing result at least as extreme as test statistic assuming $H_0$ is true
Sample test characteristic for Chi-Square test (all parameters known)

One-sided
Right side:
- critical region
- region of rejection
Left side:
- region of acceptance
where we fail to reject hypothesis
P-value of $x$: $1 - F(x)$

Chi-square density with $k - 1$ df

Shaded area = $\alpha$

Do not reject
Reject
Tests and p-values

- In the typical test...
  - $H_0$: the chosen distribution fits
  - $H_1$: the chosen distribution does not fit

- P-value of a test is:
  - the probability of observing a result at least as extreme as test statistic assuming $H_0$ is true (hence $1-F(x)$ on previous slide)
  - is the Type I error level (significance) at which we would just reject $H_0$ for the given data.

- Implications
  - If the $\alpha$ level (common values: 0.01, 0.05, 0.1) < p-value, then we do not reject $H_0$ otherwise, we reject $H_0$.
  - If the p-value is large (> 0.10)
    - then more extreme values than our current one are still reasonably likely
    - so we fail to reject $H_0$
    - in this sense it supports $H_0$ that the distribution fits (but not more than that!)
Chi-Square Test

Histogram-based test

\[ \chi^2_0 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \]

- **Observed Frequency**
- **Expected Frequency**
  \[ E_i = n \times p_i \]
  where \( p_i \) is the theoretical prob. of the \( i \)th interval.

Sums the squared differences
Chi-Square Test

- Arrange n observations into k cells, test statistics:
  \[ \chi^2_0 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \]

  - 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.

- Valid only for large sample size
- Each cell has at least 5 observations for both \( O_i \) and \( E_i \)
- Result of the test depends on grouping of the data
- Example: #vehicles arriving at an intersection between 7-7.05 am for 100 random workdays

<table>
<thead>
<tr>
<th>Arrivals per Period</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>19</td>
</tr>
<tr>
<td>3</td>
<td>17</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
</tr>
</tbody>
</table>
Chi-Square Test

**Example continued: Sample mean 3.64**

**\( H_0 \):** Data are Poisson distributed with mean 3.64  
**\( H_1 \):** Data are not Poisson distributed with mean 3.64

<table>
<thead>
<tr>
<th>( x_i )</th>
<th>Observed Frequency, ( O_i )</th>
<th>Expected Frequency, ( E_i )</th>
<th>( \frac{(O_i - E_i)^2}{E_i} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>12</td>
<td>2.6</td>
<td>7.87</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>9.6</td>
<td>0.15</td>
</tr>
<tr>
<td>2</td>
<td>19</td>
<td>17.4</td>
<td>0.8</td>
</tr>
<tr>
<td>3</td>
<td>17</td>
<td>21.1</td>
<td>4.41</td>
</tr>
<tr>
<td>4</td>
<td>19</td>
<td>19.2</td>
<td>2.57</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>14.0</td>
<td>0.26</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>8.5</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>4.4</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>2.0</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>&gt; 11</td>
<td>1</td>
<td>0.1</td>
<td>11.62</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>100.0</td>
<td>27.68</td>
</tr>
</tbody>
</table>

Degree of freedom is \( k-s-1 = 7-1-1 = 5 \) and so the \( p \)-value is 0.00004.
Chi-Square Test

- What if m parameters estimated by MLEs?
- Chi-Square distributions lose m degrees of freedom (df)
Goodness-of-fit tests

Chi-square test
Features:
• A formal comparison of a histogram or line graph with the fitted density or mass function
• Sensitive to how we group the data.

K-S and A-D tests
Features:
• Comparison of an empirical distribution function with the distribution function of the hypothesized distribution.
• Does not depend on the grouping of data.
• A-D detects discrepancies in the tails and has higher power than K-S test

from WSC 2010 Tutorial by Biller and Gunes, CMU, slides used with permission
Kolmogorov-Smirnov Test

K-S test is useful when sample size is small

Test statistic

\[ D = \max | F(x) - S_n(x) | \]

KS-Test detects the max difference

- If we have \( n \) observations \( x_1, x_2, \ldots, x_n \), then

\[ S_n(x) = \frac{\text{(number of } x_1, x_2, \ldots, x_n \text{ that are } \leq x)}{n} \]
Sometimes a bit tricky: geometric meaning of test statistic

\[ D_n^+ = \max_{1 \leq i \leq n} \left\{ \frac{i}{n} - \hat{F}(X_{(i)}) \right\}, \]

\[ D_n^- = \max_{1 \leq i \leq n} \left\{ \hat{F}(X_{(i)}) - \frac{i - 1}{n} \right\} \]

\[ D_n = \max \{ D_n^+, D_n^- \} \]

but not

\[ D'_n = \max_{1 \leq i \leq n} \left\{ \frac{i}{n} - \hat{F}(X_{(i)}) \right\} \]

for details, see Law/Kelton, Chap. 6
Anderson-Darling test (AD test)

Test statistic is a weighted average of the squared differences

\[ A_n^2 = n \int_{-\infty}^{\infty} [F_n(x) - \hat{F}(x)]^2 \psi(x)f(x) \, dx \]

with weights

\[ \psi(x) = 1/\{\hat{F}(x)[1 - \hat{F}(x)]\} \]

such that weights are largest for \( F(x) \) close to 0 and 1.

Modified critical values for adjusted A-D test statistics, reject \( H_0 \) if \( A_n^2 \) exceeds critical value.

<table>
<thead>
<tr>
<th>Case</th>
<th>Adjusted test statistic</th>
<th>0.900</th>
<th>0.950</th>
<th>0.975</th>
</tr>
</thead>
<tbody>
<tr>
<td>All parameters known</td>
<td>( A_n^2 ) for ( n \geq 5 )</td>
<td>1.933</td>
<td>2.492</td>
<td>3.070</td>
</tr>
<tr>
<td>N(( \bar{X}(n) ), ( S^2(n) ))</td>
<td>( 1 + \frac{4}{n} - \frac{25}{n^2} ) ( A_n^2 )</td>
<td>0.632</td>
<td>0.751</td>
<td>0.870</td>
</tr>
<tr>
<td>Expo(( \bar{X}(n) ))</td>
<td>( 1 + \frac{0.6}{n} ) ( A_n^2 )</td>
<td>1.070</td>
<td>1.326</td>
<td>1.587</td>
</tr>
<tr>
<td>Weibull(( \hat{\alpha} ), ( \hat{\beta} ))</td>
<td>( 1 + \frac{0.2}{\sqrt{n}} ) ( A_n^2 )</td>
<td>0.637</td>
<td>0.757</td>
<td>0.877</td>
</tr>
<tr>
<td>Log-logistic(( \hat{\alpha} ), ( \hat{\beta} ))</td>
<td>( 1 + \frac{0.25}{\sqrt{n}} ) ( A_n^2 )</td>
<td>0.563</td>
<td>0.660</td>
<td>0.769</td>
</tr>
</tbody>
</table>
Goodness-of-fit tests

- Beware of goodness-of-fit tests because they are unlikely to reject any distribution when you have little data, and are likely to reject every distribution when you have lots of data.

---

**Chi-square test**

- A formal comparison of a histogram or line graph with the fitted density or mass function
- Sensitive to how we group the data.

**K-S and A-D tests**

- Comparison of an empirical distribution function with the distribution function of the hypothesized distribution.
- Does not depend on the grouping of data.
- A-D detects discrepancies in the tails and has higher power than K-S test

---

from WSC 2010 Tutorial by Biller and Gunes, CMU, slides used with permission
Graphic Analysis vs Goodness-of-fit tests

- Graphic analysis includes:
  - Histogram with fitted distribution
  - Probability plots: P-P plot, Q-Q plot.

- Goodness-of-fit tests
  - represent lack of fit by a summary statistic, while plots show where the lack of fit occurs and whether it is important.
  - may accept the fit, but the plots may suggest the opposite, especially when the number of observations is small.

A data set of 30 observations is believed to be from a normal distribution. The following are the p-values from chi-square test and K-S test:

Chi-square test: 0.166
K-S test: >0.15

What is your conclusion?
Density Histogram

- compares sample histogram (mind the bin sizes) with fitted distribution
Frequency Histogram

compares histogram from data with histogram according to fitted distribution
Differences in distributions are easier to see along a straight line:
Graphical comparisons

<table>
<thead>
<tr>
<th>Frequency Comparisons</th>
<th>Probability Plots</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Features:</strong></td>
<td><strong>Features:</strong></td>
</tr>
<tr>
<td>• Graphical comparison of a histogram of the data with the density function of the fitted distribution.</td>
<td>• Graphical comparison of an estimate of the true distribution function of the data with the distribution function of the fit.</td>
</tr>
<tr>
<td>• Sensitive to how we group the data.</td>
<td>• $Q-Q \ (P-P)$ plot amplifies differences between the tails (middle) of the model and sample distribution functions.</td>
</tr>
</tbody>
</table>

- Use every graphical tool in the software to examine the fit.
- If histogram-based tool, then play with the widths of the cells.
- $Q-Q$ plot is very highly recommended!

from WSC 2010 Tutorial by Biller and Gunes, CMU, slides used with permission