Statistical Analysis of an Arrival Process

• **Part 1.** The Poisson Process (p.2-6)
  - Graphical Tests
• **Part 2.** How good is the fit? (p.7-12)
  - Q-Q Plots
  - Empirical CDF
  - Goodness-of-fit Tests
  - Ex. Fitting the Exponential Distribution
• **Part 3.** A Test for NHPP (p.13-14)
• **Appendix.** Main sample statistic (p.15-16)
Part 1. The Poisson Process (1)

• The Poisson process is a counting process with the properties:
  o The process has independent increments
  o Number of events in any time interval of length t has a Poisson distribution with mean $\lambda t$ (implies stationarity).

• Is the arrival process a Poisson Process?
Part 1. The Poisson Process (2)

- Graphical tests
  - Stationarity: Cumulative

Plot the cumulative arrival curve. Draw a straight line connecting the points $A(0)$ and $A(T)$. 

$max \left| \frac{A(t)}{A(T)} - \frac{t}{T} \right| = 0.061$
Part 1. The Poisson Process (3)

- Graphical tests
  - Stationarity/Independence: Interarrival Times
    Plot the interarrival times in serial order on an (x,y) graph.
Part 1. The Poisson Process (4)

- Graphical tests
  - Independence

Plot the points \(\{(X_n, X_{n-1}), \ n = 2, 3, \ldots\}\) on a graph where \(X_n\) represents the n-th arrival time.
Part 1. The Poisson Process (5)

- Graphical tests
  - Exponential Interarrival Distribution

Plot the empirical probability distribution for the interarrival times. Plot the exponential distribution with $\lambda = A(T)/T$. 

![Graph showing empirical distribution vs. exponential distribution]
Part 2. How good is the fit? (1)

- **Q-Q Plots**: Widely used to compare sample versus theoretical distribution or two sample distributions

- **How to plot?**
  - Order the sample data from smallest to largest. Denote the resulting order statistics by \( \{X_{(k)}, 1 \leq k \leq N\} \) \((X_{(1)}\) is smallest).
  - Divide \([0,1]\) into \(N+1\) intervals “in a uniform way”. Denote the division points by \(\{z_k, 1 \leq k \leq N\}\). The simplest way is to assume \(z_k = \frac{k}{N+1}, 1 \leq k \leq N\). (In practice, there can be minor changes in the definition of \(z_k\).)
  - Calculate \(\{Y_k, 1 \leq k \leq N\}\), where \(Y_k\) is \(z_k\)-quantile of the theoretical distribution. namely if \(X\) is a random variable with the underlying theoretical distribution:
    \[
P\{X < Y_k\} = z_k
    \]
  - Plot \((X_{(k)}, Y_k), 1 \leq k \leq N\).

If the points concentrate along a straight line \(y = x\), the theoretical distribution provides a good fit for the sample data.
Part 2. How good is the fit? (2)

- **Empirical cumulative distribution function (cdf)**
  - The empirical cdf of a data sample \( \{X_i\}_{i=1}^N \) (\( N \) is the number of observations) is
    \[
    \hat{F}_N(x) = \frac{1}{N} \sum_{i=1}^{N} I_{\{X_i \leq x\}}
    \]
  - If \( \{X_i\} \) are IID (independently and identically distributed with cdf \( F \)), \( \hat{F}_N \) is an unbiased estimator of the cdf \( F \):
    \[
    \mathbb{E}\hat{F}_N(x) = F(x), \text{ for all } x,
    \]
  - and a consistent estimator:
    \[
    \sup_x |\hat{F}_N(x) - F(x)| \to 0, \text{ as } N \to \infty.
    \]
Part 2. How good is the fit? (3)

- EX: the *empirical cdf* of a data sample with N=10

Data: (1, 2, 2, 4, 6, 5, 8, 8, 8, 9.5)
Part 2. How good is the fit? (4)

- **Goodness-of-fit tests**
  - Devised for checking the hypothesis
    - $H_0$: sample was taken from distribution (family of distributions) $F$
    - $H_1$: sample was *not* taken from $F$
  - Most popular tests are Chi-square and Kolmogorov-Smirnov tests
    - **Chi-square test**
      1. Samples with discrete values
      2. Samples with continuous values where parameter(s) of the distribution is(are) estimated from the sample
    - **Kolmogorov-Smirnov test**
      - Continuous distributions with a priori specified parameters
    - Refer to *chi_square_goodness_of_fit.pdf* and *tests_of_poisson_process.pdf*
Part 2. How good is the fit? (5)

EX. Fitting the Exponential Distribution

- If $\{X_i\}_{i=1}^N$ is assumed to have some distribution $F$, then its empirical cdf, $\hat{F}_N(x)$, must be close to the theoretical $F(x) = P(X \leq x)$.

- For the exponential distribution,

$$\hat{F}_N(x) \simeq F(x) = 1 - e^{-\lambda x}, \quad 1 - \hat{F}_N(x) \simeq e^{-\lambda x}.$$  

$$-\ln \left(1 - \hat{F}_N(x)\right) \simeq \lambda x.$$

- One can visually observe how close $-\ln(1 - \hat{F}_N)$ is to a straight line, whose slope can be used to estimate $\lambda$. 

Part 2. How good is the fit? (6)

**EX.** Fitting the Exponential Distribution- continued

- EX) Good picture (exp variables) vs. "Bad" picture (uniform r.v.)

- But still hard to decided
Part 3. A Test for Non-Homogeneous Poisson Process

• Assume that time-homogeneity hypothesis for the arrival process is rejected. The test below is designed to check whether the arrival process is a non-homogeneous Poisson process.

1. Break up the given time period(s) into short blocks of time, say \( I \) (equal-length) blocks of length \( L \).

2. Let \( T_{i0} = 0 \). For \( i = 1, ..., I \) and \( j = 1, ..., J(i) \) define \( T_{ij} \) to be the \( j^{th} \) ordered relative arrival time in the \( i^{th} \) block. Define

\[
R_{ij} = (J(i) + 1 - j) \left( -\ln \left( \frac{L - T_{ij}}{L - T_{ij-1}} \right) \right)
\]

3. Under the null hypothesis that the arrival process is Poisson, with the arrival rates being constant within each given block of time, the \( \{R_{ij}\} \) will be i.i.d. \( \exp(1) \) random variables.

• NOTE: \( L \) must be chosen
  - large enough to include at least 5-7 observations
  - small enough to assume that the arrival rates are constant within each time block
### Part 3. A Test for Non-Homogeneous Poisson Process

**EX**

<table>
<thead>
<tr>
<th>time</th>
<th>block</th>
<th>i</th>
<th>j per block</th>
<th>J(i)</th>
<th>T ij</th>
<th>R ij</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.0172</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>0.0172</td>
<td>1.1325</td>
<td></td>
</tr>
<tr>
<td>10.0178</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>0.0178</td>
<td>0.0364</td>
<td></td>
</tr>
<tr>
<td>10.0208</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>0.0208</td>
<td>0.1487</td>
<td></td>
</tr>
<tr>
<td>10.0478</td>
<td>1</td>
<td>4</td>
<td>6</td>
<td>0.0478</td>
<td>1.2507</td>
<td></td>
</tr>
<tr>
<td>10.0789</td>
<td>1</td>
<td>5</td>
<td>6</td>
<td>0.0789</td>
<td>1.8116</td>
<td></td>
</tr>
<tr>
<td>10.095</td>
<td>1</td>
<td>6</td>
<td>6</td>
<td>0.095</td>
<td>1.4398</td>
<td></td>
</tr>
<tr>
<td>10.1064</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>0.0064</td>
<td>0.2646</td>
<td></td>
</tr>
<tr>
<td>10.1086</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>0.0086</td>
<td>0.0714</td>
<td></td>
</tr>
<tr>
<td>10.1611</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>0.0611</td>
<td>1.7085</td>
<td></td>
</tr>
<tr>
<td>10.1853</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>0.0853</td>
<td>0.9731</td>
<td></td>
</tr>
<tr>
<td>10.21</td>
<td>3</td>
<td>1</td>
<td>5</td>
<td>0.01</td>
<td>0.5268</td>
<td></td>
</tr>
<tr>
<td>10.2231</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>0.0231</td>
<td>0.6292</td>
<td></td>
</tr>
</tbody>
</table>

Use Q-Q plot to check if $R_{ij}$ is exponentially distributed.
Appendix. Main sample statistics

• How to estimate the parameters?

• Statistical Analysis of a data sample \( \{X_i\}_{i=1}^{N} \)
  - For example, interarrival times of customers

  - **Average:** \( \hat{m} = \frac{\sum_{i=1}^{N} X_i}{N} \) (Function AVERAGE in EXCEL)

  - **Standard deviation:** \( \hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \hat{m})^2}{N-1}} \) (Function STDEV in EXCEL)

  - **Coefficient of Variation (CV):** \( \hat{c} = \frac{\hat{\sigma}}{\hat{m}} \)

    - CV of the exponential distribution is equal to one.
    - CV should be close to one if the sample is taken from IID exponential random variables
Appendix. Main sample statistics (2)

- **Confidence intervals** (CI) for the average
  - Help to quantify the precision of the estimate
  - Choose a significance level
    - $\alpha = 0.05$ gives us 95% CI
  - **Normal Approximation**
    - CI for the Normal distribution (approximate in other cases)
      $$\left[ \hat{m} + \frac{Z_{\alpha/2}}{\sqrt{N}} \cdot \hat{\sigma}, \hat{m} - \frac{Z_{\alpha/2}}{\sqrt{N}} \cdot \hat{\sigma} \right] = \left[ \hat{m} - \frac{Z_{1-\alpha/2}}{\sqrt{N}} \cdot \hat{\sigma}, \hat{m} + \frac{Z_{1-\alpha/2}}{\sqrt{N}} \cdot \hat{\sigma} \right]$$
    - $Z_{\alpha/2}$ is the $\alpha/2$ quantile of the Standard Normal distribution
    - EX: $\left[ 0.939 - \frac{1.96 \cdot 0.859}{\sqrt{127}}, 0.939 + \frac{1.96 \cdot 0.859}{\sqrt{127}} \right] = [0.780, 1.088]$
  - **Exponential Distribution**
    - $$\left[ \hat{m} \cdot \frac{\chi^2_{1-\alpha/2}(2N)}{2N}, \hat{m} \cdot \frac{\chi^2_{\alpha/2}(2N)}{2N} \right]$$
    - Use EXCEL function CHIINV to compute $\chi^2$ values
    - EX: $[0.783, 1.109]$, very close to the normal approximation