Statistical Analysis of an Arrival Process

- Part 1. Main Sample statistics (p.2-3)
- Part 2. Graphical Representation of Data (p.4-6)
  - Histograms
  - Empirical CDF
- Part 3. Fitting the Exponential Distribution (p.7-9)
- Part 4. A Test for NHPP (p.10-11)

Part 1. Main Sample Statistics

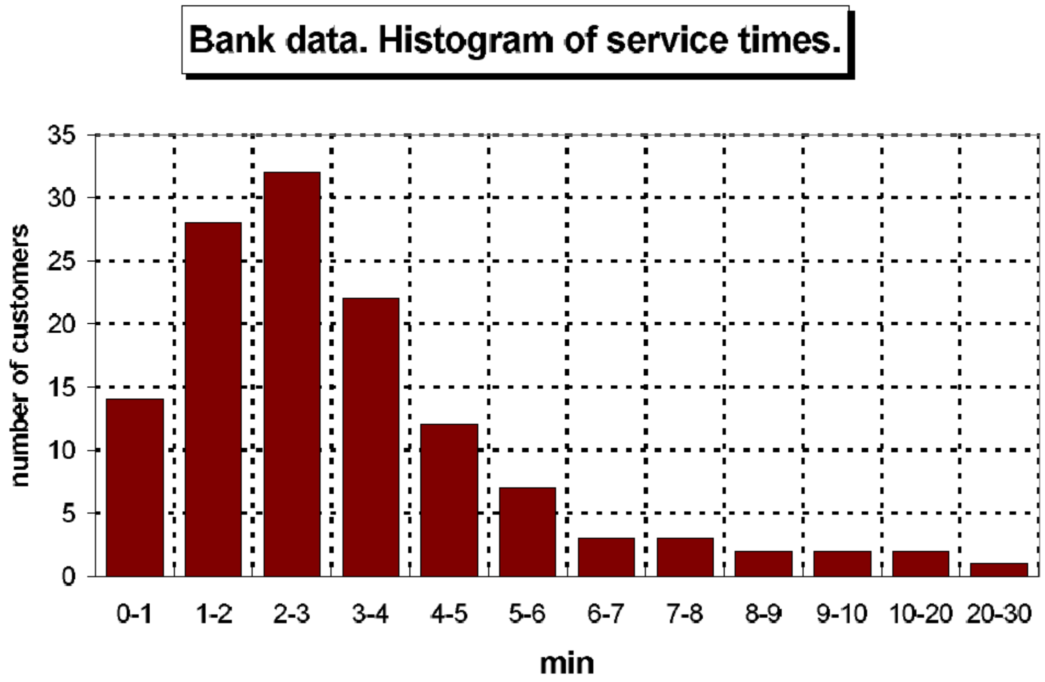
- Statistical Analysis of a data sample $\{X_i\}_{i=1}^{N}$
  - For example, interarrival times of customers
  - Average: $\widehat{m} = \frac{\sum_{i=1}^{N} X_i}{N}$ (Function AVERAGE in EXCEL)
  - Standard deviation: $\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \widehat{m})^2}{N-1}}$ (Function STDEV in EXCEL)
  - Coefficient of Variation (CV): $\hat{c} = \frac{\widehat{\sigma}}{\widehat{m}}$
    - 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

Part 1. 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)
    - $$\left[\hat{m} + \frac{Z_{\alpha/2}}{\sqrt{N}} \cdot \hat{\sigma}, \hat{m} - \frac{Z_{\alpha/2}}{\sqrt{N}} \cdot \hat{\sigma}\right] = \left[\hat{m} - \frac{Z_{1-\alpha/2}}{\sqrt{N}} \cdot \hat{\sigma}, \hat{m} + \frac{Z_{1-\alpha/2}}{\sqrt{N}} \cdot \hat{\sigma}\right]$$
    - $Z_{\frac{\alpha}{2}}$ is the $\frac{\alpha}{2}$ 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 χ^2 values
    - EX: [0.783, 1.109], very close to the normal approximation

Part 2. Graphical Representation of Data

- Histograms
  - A standard method for data representation
  - Divide the data into categories in a *reasonable* way
  - EX: Histogram of 128 service times
    - the bin width is not constant (1 min in [0,10] and 10 min in [10, 30]
    - Usually the bin width is taken to be a constant.



Part 2. Graphical Representation of Data (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:

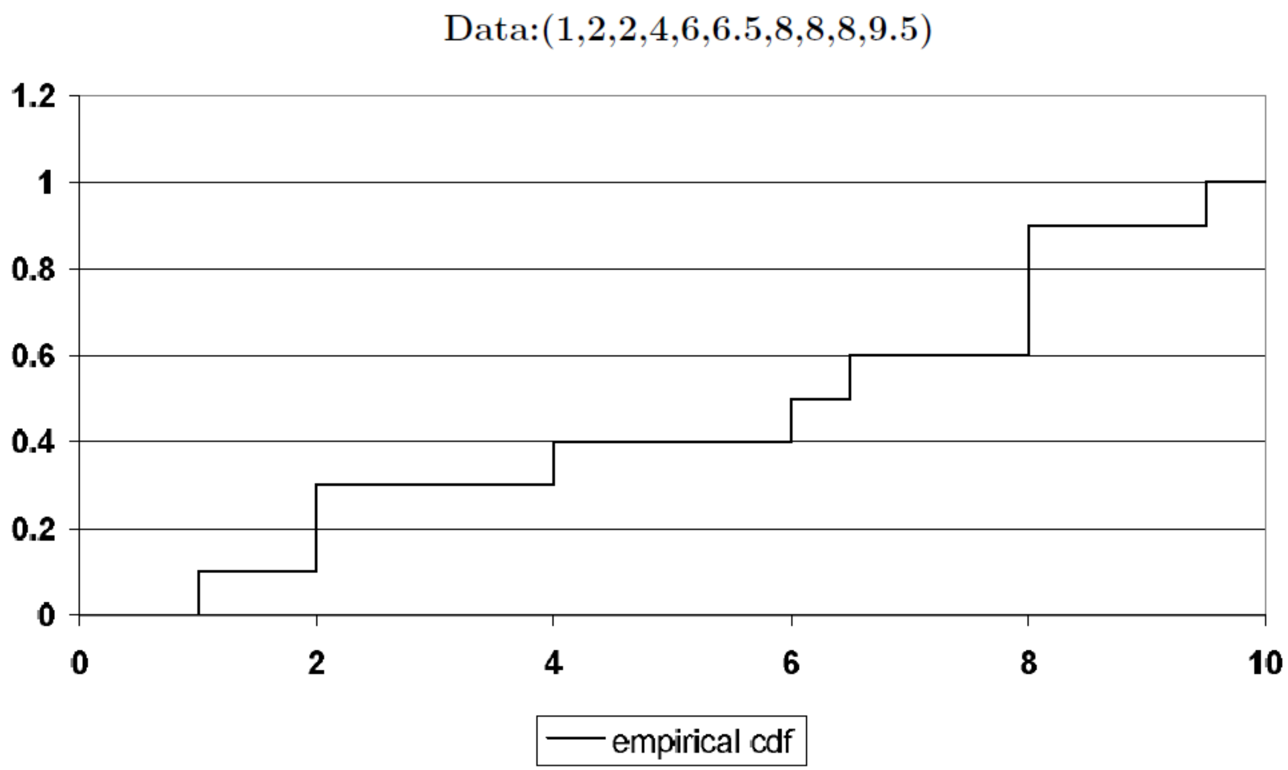
$$\mathrm{E}\hat{F}_N(x) = F(x), \text{ for all } x,$$

  - and a consistent estimator:

$$\sup_x |\hat{F}_N(x) - F(x)| \to 0 \,, \text{ as } N \to \infty.$$

Part 2. Graphical Representation of Data (2)

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



Part 3. 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 λ

Part 3. Fitting the Exponential Distribution (2)

- 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 $\{X_{(k)}, 1 \le k \le N\}$ ($X_{(1)}$ is smallest).
  - Divide [0,1] into $N+1$ intervals "in a uniform way". Denote the division points by $\{z_k, 1 \le k \le N\}$. The simplest way is to assume $z_k = \dfrac{k}{N+1}, 1 \le k \le N$. (In practice, there can be minor changes in the definition of z_k.)
  - Calculate $\{Y_k, 1 \le k \le N\}$, where Y_k is z_k-quantile of the theoretical distribution, namely if X is a random variable with the underlying theoretical distribution:

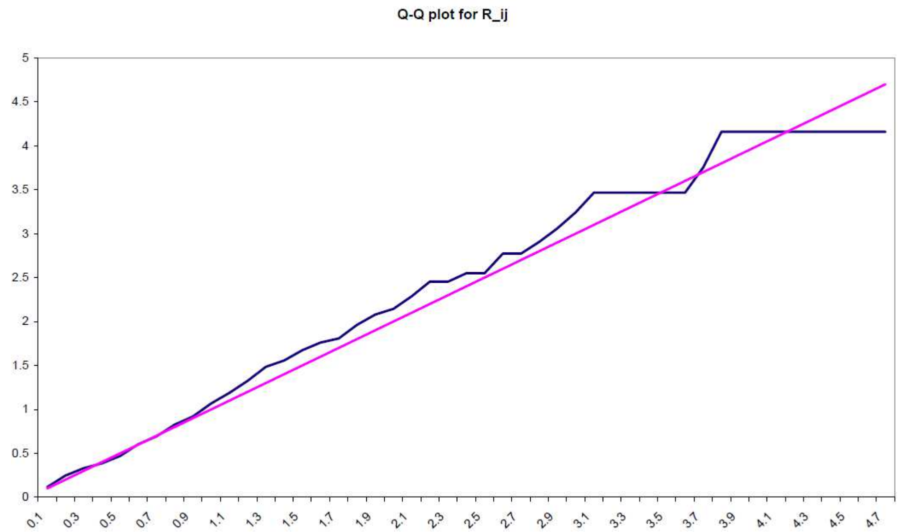
$$P\{X < Y_k\} = z_k$$

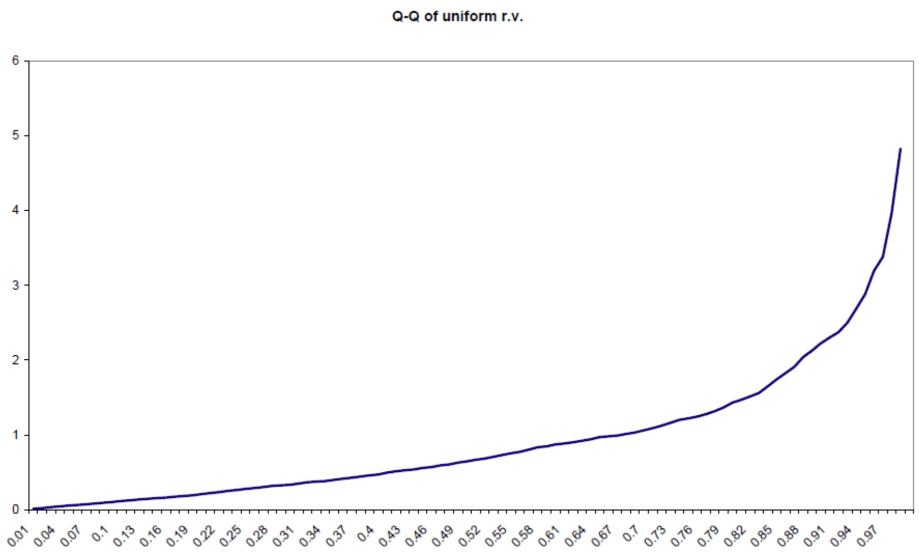
  - Plot $(X_{(k)}, Y_k)$, $1 \le k \le N$.

If the points concentrate along a straight line $y=x$, the theoretical distribution provides a good fit for the sample data.

Part 3. Fitting the Exponential Distribution (3)

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





- But still hard to decided

Part 4. A Test for 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 $T_{i\,0} = 0$. 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

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

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

	time	block i		L=0.1	0.1	
	10		j per block	J(i)	T ij	R ij
1	10.0172	1	1	6	0.0172	1.1325
2	10.0178	1	2	6	0.0178	0.0364
3	10.0208	1	3	6	0.0208	0.1487
4	10.0478	1	4	6	0.0478	1.2507
5	10.0789	1	5	6	0.0789	1.8116
6	10.095	1	6	6	0.095	1.4398
7	10.1064	2	1	4	0.0064	0.2646
8	10.1086	2	2	4	0.0086	0.0714
9	10.1611	2	3	4	0.0611	1.7085
10	10.1853	2	4	4	0.0853	0.9731
11	10.21	3	1	5	0.01	0.5268
12	10.2231	3	2	5	0.0231	0.6292

Use Q-Q plot to check if of R_ij is exponentially distributed