What do we anticipate?

- We anticipate a **Nonhomogeneous Poisson Process (NHPP)**.
- For staffing, we may want **piecewise-constant arrival rate function**.
- Problem with multiple days: **day-to-day variation** (over-dispersion).

To confirm, we perform **statistical tests** *(could be whole course)*.

- Today: **On Kolmogorov-Smirnov Tests of NHPP**.
- Based on **2005 paper by Brown et al.** and two **2014 papers by Song-Hee Kim and WW** plus recent **2014 paper**

Given NHPP, we only need to **estimate the arrival rate function**.

- Using historical data: **forecasting** *(Friday)*. *(could be whole course)*.
Important Concepts Covered Today, I

1. statistical hypothesis testing
   - null hypothesis \((H_0)\)
   - significance level
   - p-value
   - alternative hypothesis \((H_1)\)
   - power of a statistical test \((P(\text{reject } H_0|H_1))\)

2. Comparing Cumulative Distribution Functions (CDF’s)
   - Q-Q plot
   - empirical cdf
   - Kolmogorov-Smirnov (KS) statistical test
Important Concepts Covered Today, II

1. statistical test of a Poisson process
   - the standard KS test (use iid exponential interarrival times)
   - conditional-uniform property of the Poisson process
   - CU KS test

2. statistical tests of an NHPP
   - CU KS test
   - the logarithm test from Brown (2005)
   - the Lewis (1965) test exploiting Durbin (1961) used in KW (2014a)
   - over-dispersion (relative to Poisson process), KW (2014b)
Identifier the Predictable and Unpredictable Variability

Details of Fitting Process

Plot for Aug 9 (Fri.)

- Divide day (7am to midnight) into time intervals of 150 seconds (≈2½ minutes)
- Count number arrivals in each interval, and make scatterplot
- Fit using a nonparametric regression smoothing
Look at Multiple Days: IID NHPP’s?

We address over-dispersion directly through daily totals.

- We ask if daily totals are consistent with Poisson.
- Is the variance equal to the mean?
- We estimate the distribution of the daily totals.

We avoid the issue in the test of an NHPP.

- We do so by using the conditional-uniform property of a PP.
  - (to be explained in following slides)
- An NHPP can pass the test even if there is over-dispersion.
- We thus test the NHPP property conditional on the daily total.
Statistical Tests of a NHPP

1. Reduce to statistical tests of a Poisson Process (PP).
   - Assume piecewise-constant arrival rate function.
   - Then independent PP’s over subintervals.

2. Interarrival times iid exponential on each interval, but we would need to estimate mean of each, so we do not use that approach.

3. Exploit Conditional Uniform (CU) Property of PP. (first key idea)
   - $n$ arrival times $A_k$ in $[0,T]$: $A_k/T$ are $n$ ordered iid uniforms on $[0,1]$.
   - No nuisance parameter: independent of arrival rate.
   - We can combine data from different intervals and days.
   - Use Kolmogorov-Smirnov test. (To be discussed in following slides.)
Theorem. Given \( n \) arrival times \( A_k \) of a PP in \([0,T]\), \( A_k/T \) are distributed as order statistics of \( n \) iid uniform variables on \([0, 1]\).

Proof. For \( 0 < t_1 < t_2 < \cdots < t_n < T \),

\[
f_{A_1, \ldots, A_n|A(T)}(t_1, \ldots, t_n|n) \\
\approx P(N(t_i + \delta) - N(t_i) = 1, 1 \leq i \leq n, \text{no other points}|N(T) = n) \\
\approx \frac{e^{-\lambda t_1}(\lambda \delta e^{-\lambda \delta})e^{-\lambda(t_2-t_1)}(\lambda \delta e^{-\lambda \delta})\ldots e^{-\lambda(T-t_n)}}{\delta^n e^{-\lambda T}(\lambda T)^n/n!} \\
\rightarrow \frac{n!}{T^n} \quad \text{as} \quad \delta \downarrow 0.
\]

(Limiting form of \( n \)-dimensional pdf. See §5.3.5 of Ross (2010).)
Compare Empirical CDF (ECDF) to Theoretical CDF

- Given \( n \) random variables \( X_1, X_2, \ldots, X_n \), (the data)
  - each with CDF (Cumulative Distribution Function) \( F(x) = P(X_k \leq x) \),
  - the empirical CDF (ECDF) is

\[
\hat{F}_n(x) = \frac{1}{n} \sum_{k=1}^{n} 1\{X_k \leq x\}
\]

(\( \hat{F}_n(x) \) is the proportion of the \( n \) variables less than or equal to \( x \).)

- The ECDF is an estimator of the CDF \( F \).
  - **unbiased**: For each \( x \), \( E[\hat{F}_n(x)] = F(x) \).
  - If \( \{X_k\} \) are iid, then consistent: absolute difference

\[
D_n \equiv \sup_x |\hat{F}_n(x) - F(x)| \to 0 \quad \text{as} \quad n \to \infty. \quad \text{(Glivenko-Cantelli Thm.)}
\]
Example of an Empirical cdf (ECDF)

Data: (1, 2, 2, 4, 6, 6.5, 8, 8, 8, 9.5)
Compare Two CDF’s

CDF

\[ F(x) \]

Quantiles \( F^{-1}(p) \) and \( G^{-1}(p) \)

\[ p = 0.30 \]

uniform CDF

\[ G(x) = x/b \]
Q-Q Plots: Comparing Two CDF’s Via Quantiles

- Given two CDF’s $F$ and $G$,

  1. Consider associated quantile functions (inverses) $F^{-1}(p)$ and $G^{-1}(p)$ for $0 \leq p \leq 1$.
  2. Construct function $h : [0, 1] \rightarrow \mathbb{R}^2$ mapping $p$ into $(F^{-1}(p), G^{-1}(p))$.
  3. Plot the image of this function: $\{(F^{-1}(p), G^{-1}(p)) : 0 \leq p \leq 1\}$.

- Common convention for empirical CDF’s:

  1. Let $\hat{F}_n^{-1}(k/(n+1)) = X_{(k)}$, $k^{th}$ smallest (order statistic)
  2. **Q-Q plot** is $\{(\hat{F}_n^{-1}(k/(n+1))), F^{-1}(k/(n+1)) : 1 \leq k \leq n\}$ or $\{(X_{(k)}, F^{-1}(k/(n+1)) : 1 \leq k \leq n\}$. 

Example. QQ-Plots comparing exponential data (good fit) and uniform data (bad fit) to the exponential distribution.
The Kolmogorov-Smirnov (KS) Statistical Test

- **Null Hypothesis, H0:** We have a sample of size $n$ from a sequence

  \( \{X_k : k \geq 1\} \) of i.i.d. random variables with continuous CDF $F$.

- **Alternative Hypothesis, H1:** We have a sample of size $n$ from a another (specified) sequence of random variables.

  - the CDF of $X_k$ might be **not** $F$.
  - There might be **dependence** among the random variables.

- KS test based on **absolute difference** $D_n \equiv \sup_x |\hat{F}_n(x) - F(x)|$.

  - Observe $D_n = \hat{D}_n$ for the data. Reject if $\hat{D}_n > x_\alpha$, where
  - $P(D_n > x_\alpha|H0) = \alpha = 0.05$ (**significance level**)
  - Compute **p-value:** $P(D_n > \hat{D}_n|H0)$ (level for rejection)
The KS Test Needs to be Modified for NHPP

- The KS Test can be applied to test the NHPP.
  1. Assume that the NHPP has a piecewise-constant arrival rate function.
  2. **Exploit the Conditional Uniform (CU) Property** to obtain sequence of i.i.d. random variables uniform on $[0, 1]$ (under $H_0$).
  3. Use code for computing $P(D_n > x_\alpha | H_0)$ (e.g. `ksstat` in matlab)

- **Problem**: KS test of NHPP using CU property has very low power.
  - **Power**: $P(\text{Reject } H_0 | H_1)$ (1 -type II error).
  - Low power means that alternatives pass too easily!

- **Solve** by applying KS test after data transformation. (Apply KS test after producing new sequence of i.i.d. variables under $H_0$.)
Why does the CU KS Test have low power?

- **The CU transformation focuses on the arrival times instead of the interarrival times.**
  - It is the arrival times that are uniformly distributed on $[0, T]$.
  - Asymptotically, the CU KS test can be shown to have no power. (See §7 and §8 of KW14.)

- **Solve** by applying KS test after **data transformation**. (Apply KS test after producing new sequence of i.i.d. variables under H0.)
The Log KS Test from §3 of Brown et al. (2005)

- Given \(n\) ordered arrival times \(0 < A_1 < \cdots < A_n < t\) in \([0, t]\), let

  \[
  X_{j}^{\text{Log}} \equiv -(n + 1 - j) \log_e \left( \frac{t - A_j}{t - A_{j-1}} \right), \quad 1 \leq j \leq n.
  \]

- **Under H0:** If these random variables are obtained from a PP over \([0, t]\) using the CU property, then \(\{X_{j}^{\text{Log}} : 1 \leq j \leq n\}\) are \(n\) i.i.d. mean-1 exponential random variables. (Proof in §2.2 of KW14a Appendix.)

- The \(- \log_e (1 - X_{j}^{\text{Log}})\) are \(n\) i.i.d. uniforms on \([0, 1]\).

- The KS test can also be applied in this new setting.

- And the power is greater than direct KS + CU for most alternatives.
Given \( n \) ordered arrival times \( A_j, 0 < A_1 < \cdots < A_n \), from a Poisson process over \([0, t]\), apply the Conditional Uniform (CU) property to deduce that \( U(j) \equiv A_j/t \) are \( n \) ordered uniforms in \([0, 1]\).

Construct the successive intervals between these ordered uniforms, getting \( C_1 = U(1), C_j = U(j) - U(j) \) and \( C_{n+1} = 1 - U(n) \).

Let \( C(j) \) be the associated ordered intervals from \( \{C_j\} \), so that

\[
0 < C(1) < C(2) < \cdots < C(n+1) < 1.
\]

Finally, let \( Z_j = (n + 2 - j)(C(j) - C(j-1)) \) be the scaled intervals between), and let \( S_k = Z_1 + \cdots + Z_k \) be the associated partial sums.
Remarkably, Durbin (1961) showed that under H0, \((Z_1, \ldots, Z_n)\) is distributed the same as \((C_1, \ldots, C_n)\).

Hence, \((S_1, \ldots, S_n)\) is distributed the same as \((U_{(1)}, \ldots, U_{(n)})\).

Hence, \(\hat{F}_n(x) = n^{-1} \sum_{k=1}^{n} 1\{S_k \leq x\}\) is ECDF of uniform CDF, i.e., the ECDF of i.i.d. random variables uniformly distributed on \([0, 1]\).

Hence we can apply KS test under H0: i.i.d. uniforms on \([0, 1]\).

Why? The Lewis KS test has even more power! Under alternative hypotheses, the constructed ECDF tends to be more distant from the uniform CDF \(F(x) = x\).
Sanity Check

\[ S_k \equiv \sum_{j=1}^{k} Z_j = \sum_{j=1}^{k} (n + 2 - j)(C_j - C_{j-1}) \]
\[ = (n + 1)C(1) + n(C(2) - C(1)) + (n - 1)(C(3) - C(2)) \]
\[ + \cdots + (n + 2 - k)(C(k) - C(k-1)) = C(1) + C(2) + \cdots + C(k) \]
\[ = U(1) + (U(2) - U(1)) + (U(3) - U(2)) + \cdots + (U(k) - U(k-1)) \]
\[ = U(k) = C_1 + \cdots + C_k \leq 1, \quad 1 \leq k \leq n + 1. \]

Hence, \( Z_k \geq 0 \) and \( \sum_{n+1}^{j=1} Z_j = 1 \). But that does not explain the key property that \((Z_1, \ldots, Z_n)\) is distributed the same as \((C_1, \ldots, C_n)\) under the null hypothesis.
**Example: Different KS Tests Applied to an Alternative**

**Simulation Experiment:** Apply the KS test to the *alternative:* a non-Poisson renewal process with interarrival times having an $H_2$ (hyperexponential) CDF (mixture of two exponentials) with a squared coefficient of variation $c_X^2 = \frac{Var(X)}{(E[X])^2} = 2.0$.

**Table:** Performance of alternative KS tests of a rate-1 Poisson process for the time interval $[0, 200]$ with significance level $\alpha = 0.05$: the case of a renewal process with $H_2$ interarrival times having $c_X^2 = 2$, based on $10^4$ replications.

<table>
<thead>
<tr>
<th>KS test</th>
<th>Lewis</th>
<th>Standard</th>
<th>Log</th>
<th>CU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power</td>
<td>0.94</td>
<td>0.63</td>
<td>0.51</td>
<td>0.28</td>
</tr>
<tr>
<td>Average $p$ value</td>
<td>0.01</td>
<td>0.10</td>
<td>0.13</td>
<td>0.23</td>
</tr>
</tbody>
</table>
Insightful Plots: Average of ECDF over $10^4$ Replications

$H_2 \ (c^2 = 2)$: Standard KS, Conditional-Uniform, Lewis, Log Tests (clockwise)
The banking call center data passes all the KS tests for a NHPP, as described above, 

provided we adjust for rounding to nearest second.

We adjust by un-rounding, i.e., by adding small independent uniform random variables, to undo the rounding.

- Rounding causes rejection by Lewis KS test (but not CU KS test).
- Unrounding avoids problem.
- Unrounding does not change non-Poisson into Poisson.
Insightful Plots: Rounding a Poisson Process

Figure 1  Comparison of the average ecdf for a rate-1000 Poisson process. From top to bottom: CU, Lewis test. From left to right: Raw, Rounded, Un-rounded.
Insightful Plots: Rounding an $H_2$ Renewal Process

Figure 2  Comparison of the average ecdf of a rate-1000 arrival process with $H_2$ interarrival times. From top to bottom: CU, Lewis test. From left to right: Raw, Rounded, Un-rounded.
References


More Work

  - The impact of data rounding and correcting for it.
  - How to choose subintervals to make arrival rate piecewise constant.
  - Testing for over-dispersion in the daily totals. (in the call center data)