Queuing theory to the rescue!

Ambulance fleet staffing to meet time-varying demand.

IEOR 267 Final Project Presentation

Riley Murray

Problem : Ambulance Fleet Staffing

Ambulance Fleet Staffing : Setting

Number of ambulances in service at any given time:

- ~ 15 to 25 SFFD operated ambulances
- ~ 3 to 10 privately operated ambulances

Mandates:

- "Lights and sirens" → 90th% ambulance response time < 10 minutes.
- *No* lights and sirens → 90th% ambulance response time < 20 minutes.
- Private ambulances handle < 20% of all calls.

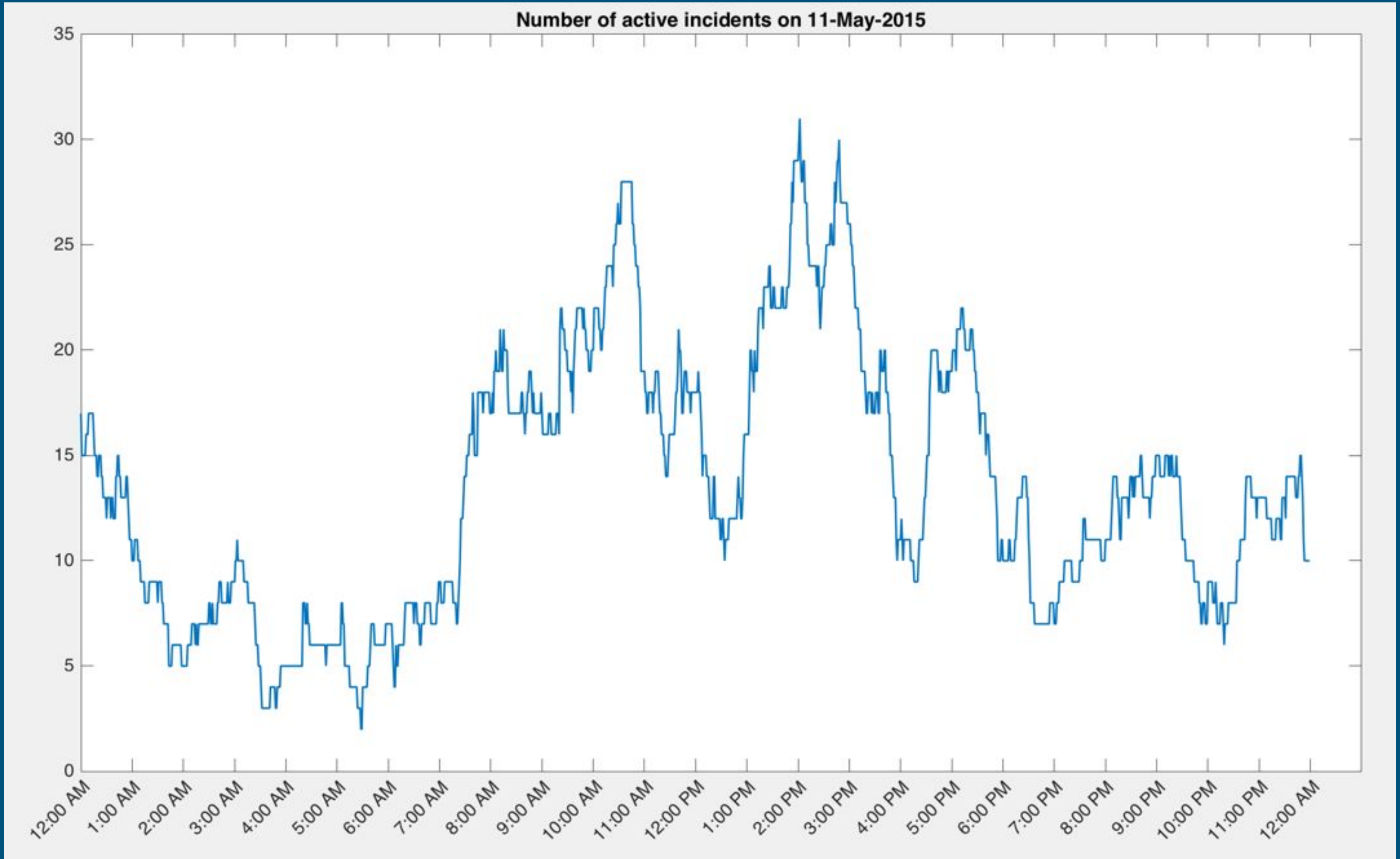
Ambulance Fleet Staffing : Project

Project components :

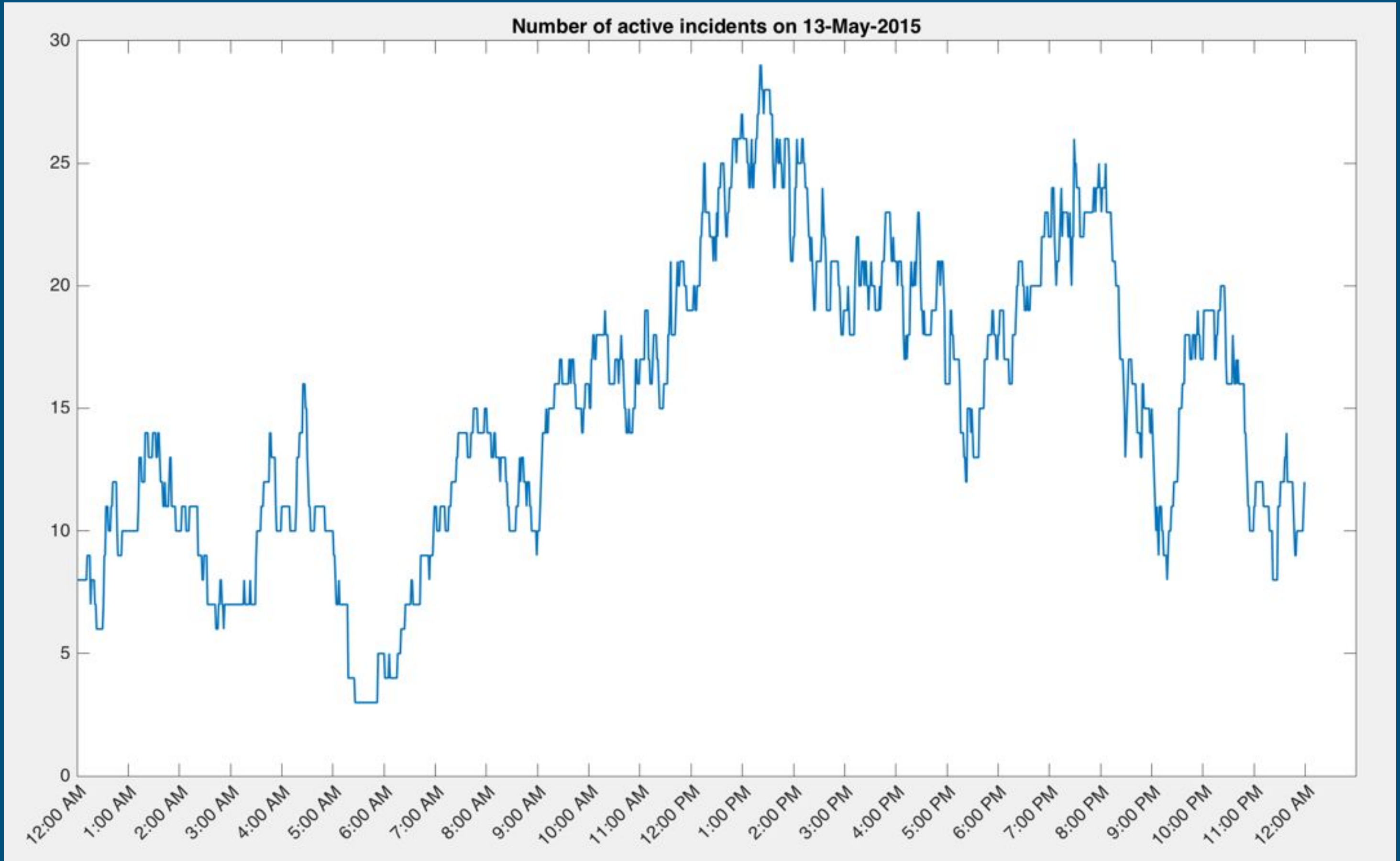
this presentation

- Model uncertainty in call volume.
- Find optimal ambulance fleet schedule given the uncertainty model.
- Develop a computer system to do this automatically, whenever SFFD wants.

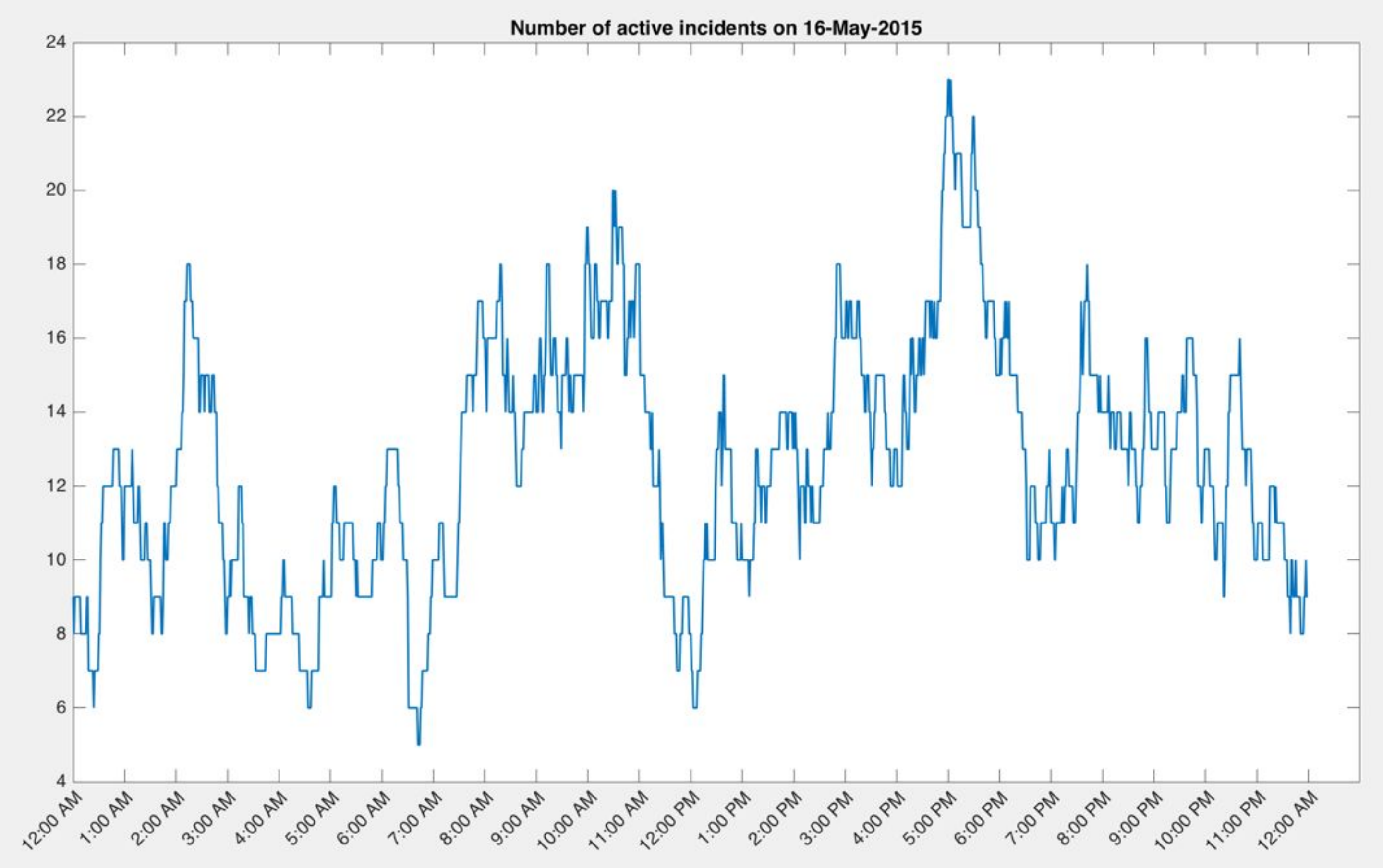
my senior project



calls in system : Mon., May 11, 2015



calls in system : Wed., May 13, 2015



calls in system : Sat., May 16, 2015

… how do you plan for that?

Key : ambulances need to be staffed at *high* levels of service.

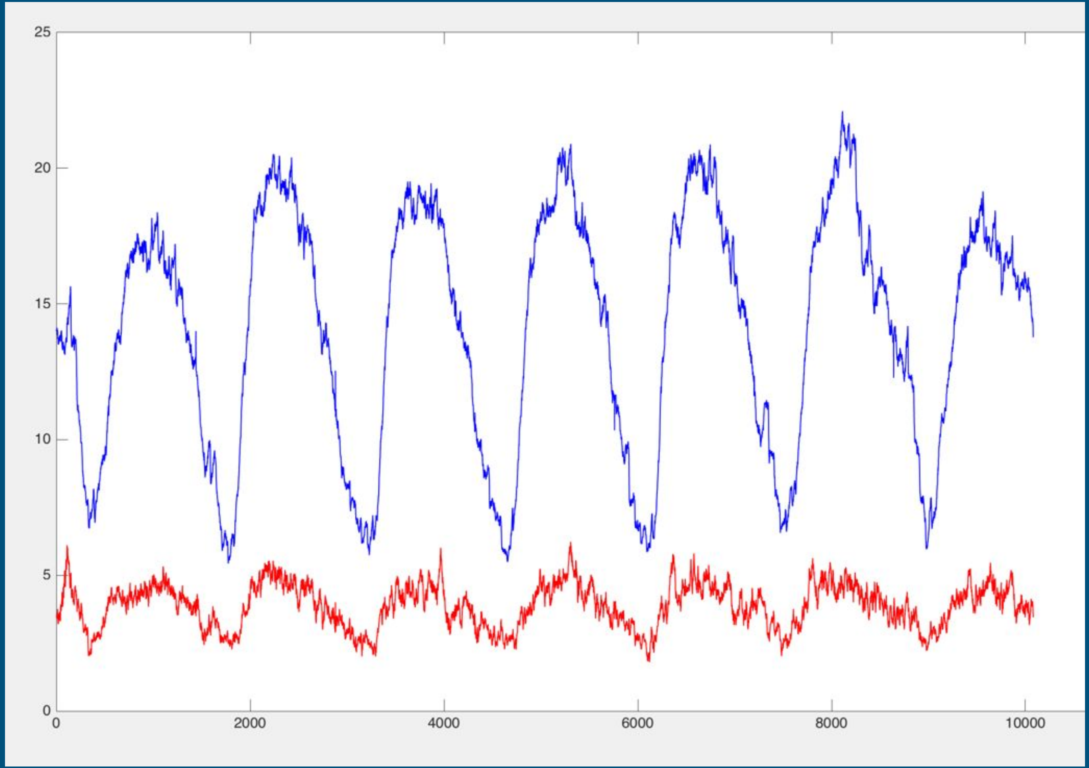
What distributional information is available to ensure high service level?

Treat each *minute of the week* as having its own (unknown) distribution.

Distribution of #calls in system - 2015

~50 samples for each minute of the week (10,080 minutes / week)

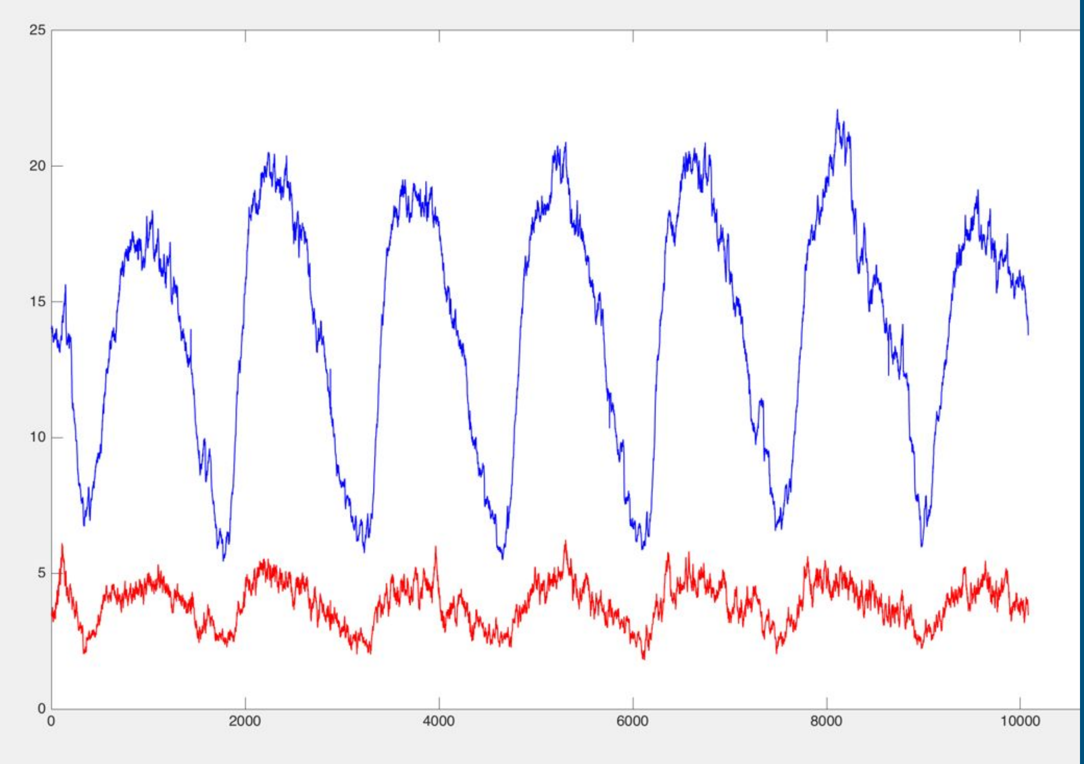
- Blue = mean for each minute
- Red = std dev for each minute



Distribution of #calls in system - 2015

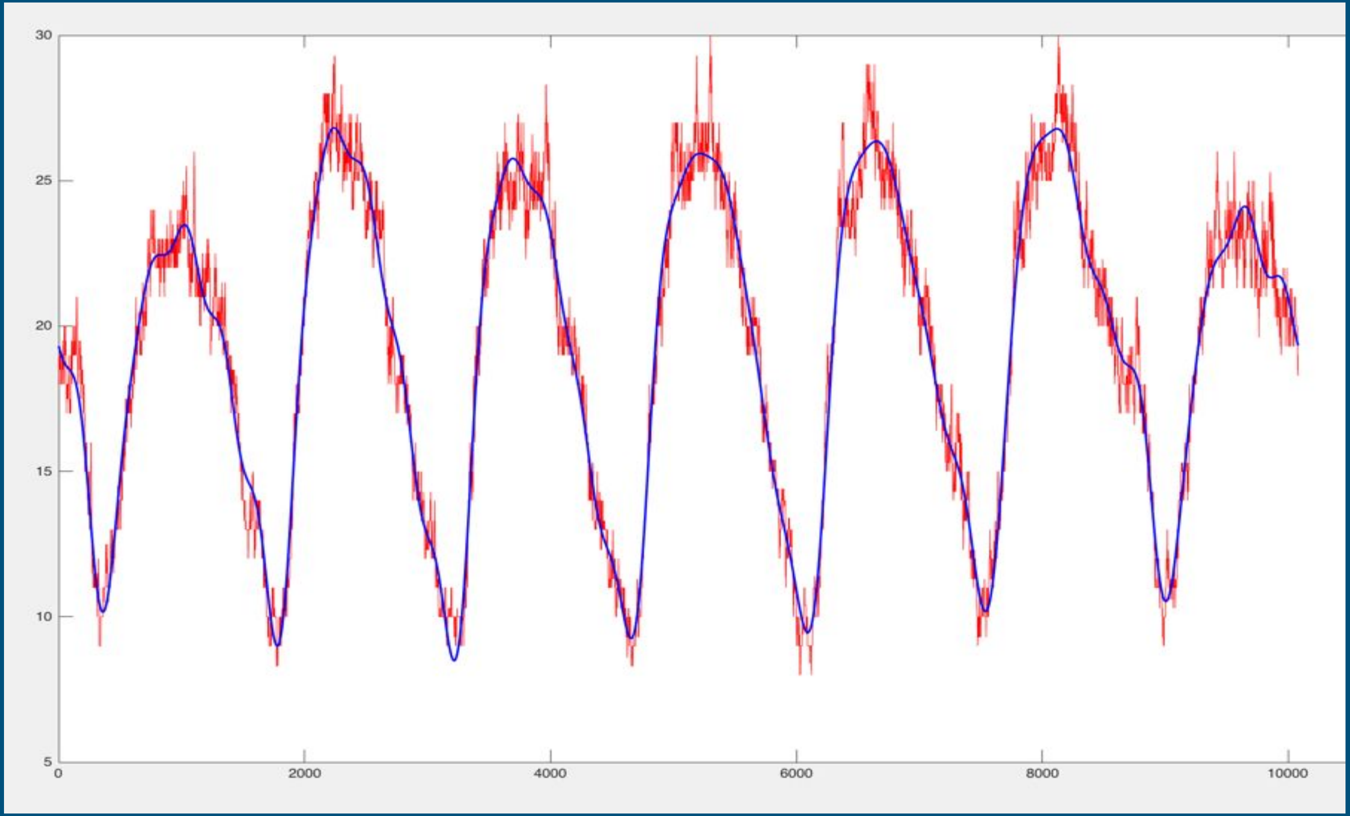
PROBLEM : curve is too jagged. Taking this as gospel will have us “staffing to fit noise.”

SOLUTION : smooth the curve … somehow.



How does one fit periodic functions?

FOURIER SERIES !



Mean # of Calls in System (with 16th order Fourier approximation).

How will we set robust target staffing level?

- Want target to be "90th percentile of demand" + [transit time buffer]
  - Ensure > 90% of all calls have ambulance on scene in required timeframe.
- How estimate 90th percentile of demand?
  - Lots of data → empirical distribution
  - Less data → **queueing theoretic model(s)**

Est. 90th%ile of demand - empirical dist.

One row == one week
(~ 12 or ***~50*** rows)

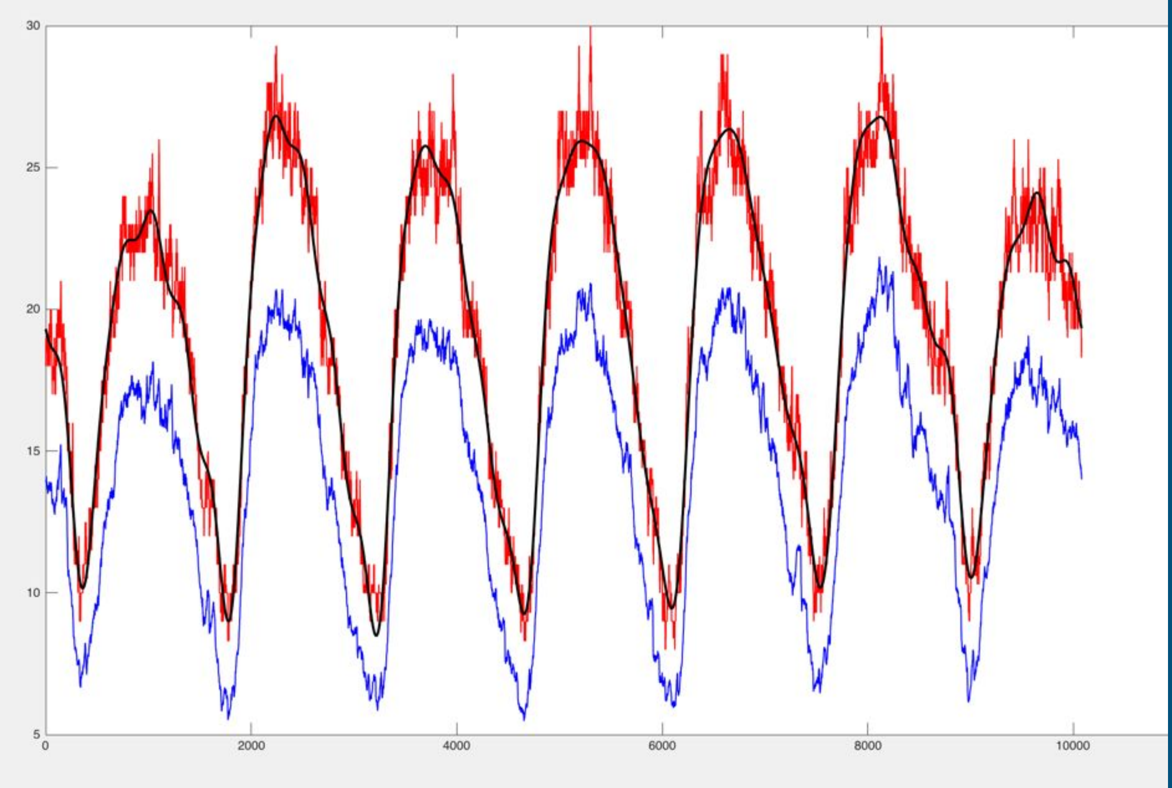
"big_matrix" containing "# active calls" data

One column == one minute (10,080 columns)

Definition : 90% of data in column i of "big_matrix" is <= "**P90_raw(i)**"

Run fourier approximation on "P90_raw" to get "**P90_nominal**".

P90_nominal(i) is nominal 90th%ile of demand at minute "i" of a non-holiday week.



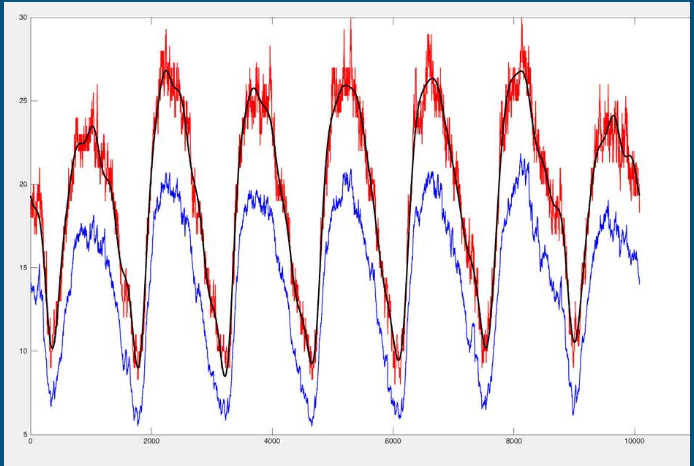
blue → mean
red → P90_raw
blk → P90_nominal

Calls in System

Est. 90th%ile of demand - empirical dist.

Looks good, but has a drawback.

- Estimating 90th%ile takes a lot of data.
- Each *week* gets us a single datapoint.
- We would like to capture seasonality in data (if it exists).



Est. 90th%ile of demand

* WANT *

to estimate P90 with significantly fewer data points (~15 to 20)

* HAVE *

Queueing Theory!

Est. 90th%ile of demand - queueing theory

My procedure at a high level :

- Estimate (via data + smoothing) a mean # in system function $m(t)$
- Set target using square root staffing level *, plus transit factor.

theory

Other things one can do :

- Estimate (via data + smoothing) an arrival *rate function* $\lambda(t)$ & service rate function.
- Use above to compute mean function $m(t)$ by solving a diff.eq.
- Set target staffing level appropriately (potentially estimate variance function $v(t)$).

Queueing **theory** : Square Root Staffing

- Approximate M/G/k as M/G/∞
- Steady state # in system for M/G/∞ ~ Poisson(m) with $m \triangleq \lambda / \mu >> 1$
- Approximate Poisson(m) with $\mathcal{N}(m, m)$ (with heavy load; use continuity correction)
- → Approximate steady state # in system for M/G/k as $\mathcal{N}(m, m)$
- Staff at m + c · m ^ (1/2)
  - Pick c to solve : c Φ(c)/ϕ(c) = (1- δ) / δ

Queueing **theory** : Non-Stationary Staffing

- $M/G/\infty$ approximation used stationary distribution arguments.
- Use normal approx for $[M/G/\infty](t)$ too! (even though it doesn't have stationary distribution…)
  - i.e. approx $[M/G/\infty](t)$ with $\mathcal{N}(m(t), m(t))$ (we'll only use this)
  - More generally, approx $[G/G/\infty](t)$ with $\mathcal{N}(m(t), v(t))$ (for suitable $v(t)$)
- Set "c" in the square-root staffing rule a little differently.

Queueing **theory** : Non-Stationary Staffing

Set $m(t)$ as the smoothed fourier approximation of historical mean # in system.

$$s(t) = \lceil m(t) + 0.5 + z_\alpha \sqrt{\nu(t)} \rceil,$$

(We don't actually take the ceiling.)

$$\nu(t) = m(t).$$

Since 911 calls follow NHPP

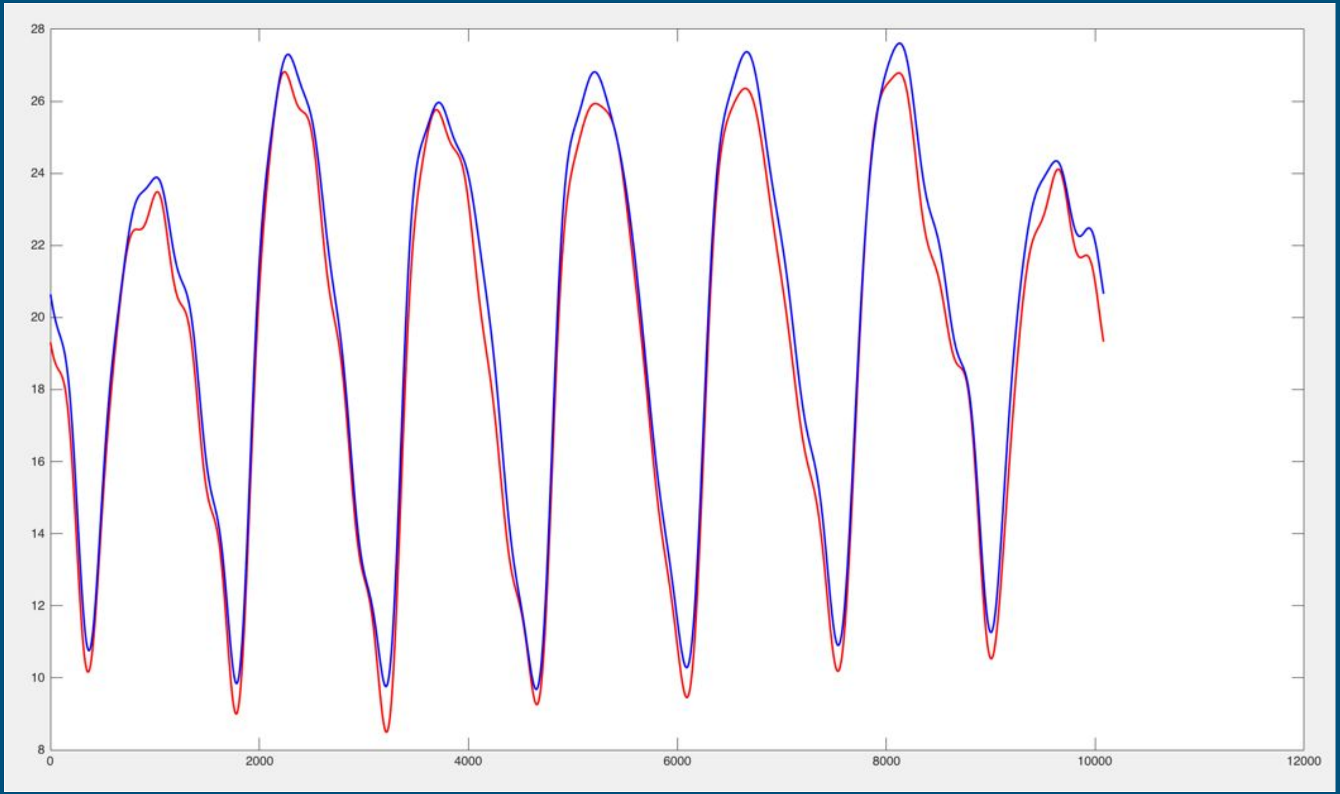
$$p_D(\alpha) \equiv [1 + \sqrt{2\pi} z_\alpha (1 - \alpha) \exp(z_\alpha^2/2)]^{-1}$$

Pick α so that $P_D(\alpha) = 0.1$

This is same as before : $c\,\Phi(c)/\phi(c) = (1- \delta) / \delta$

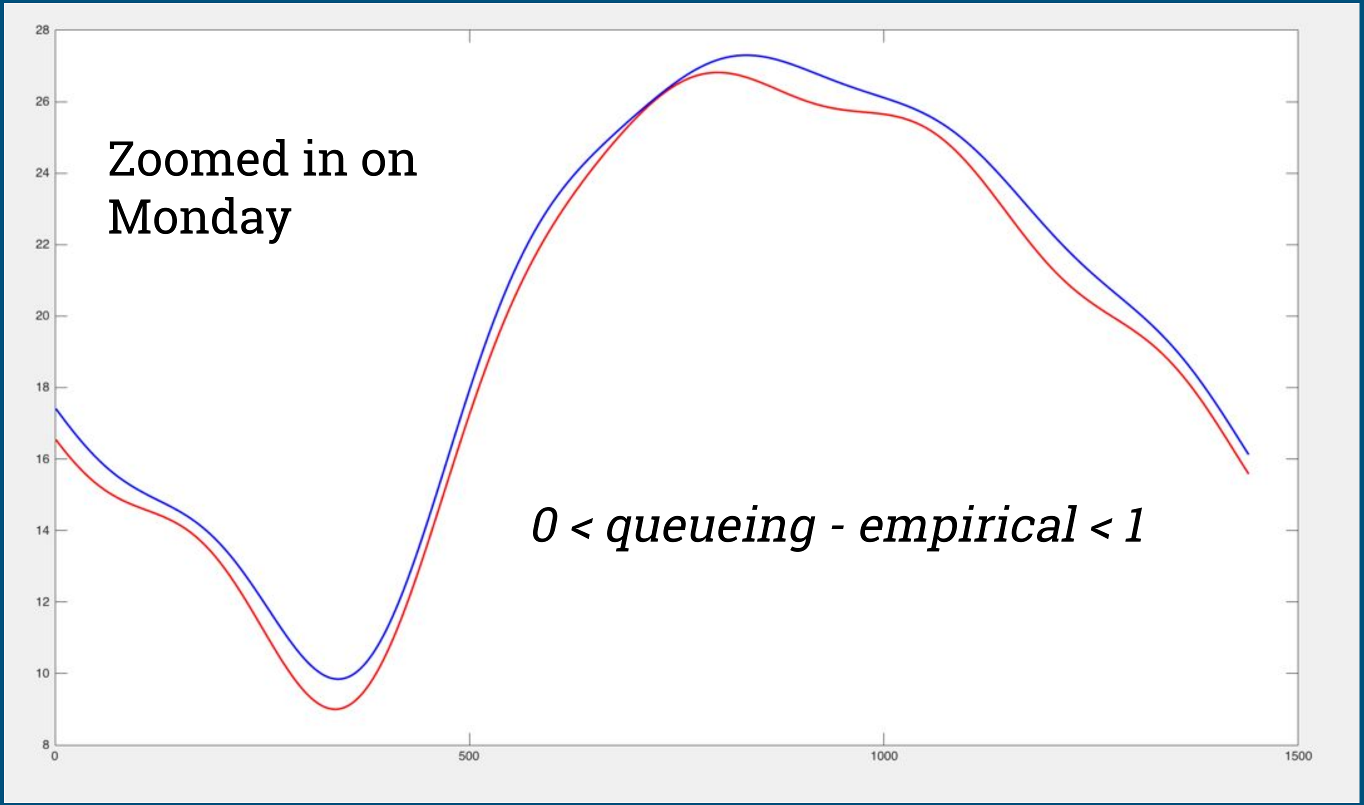
Queueing **theory** : Non-Stationary Staffing

… but does it work?



(*YES !*)

Staffing for 90%ile of demand w/ empirical dist. (red), queueing model (blue)



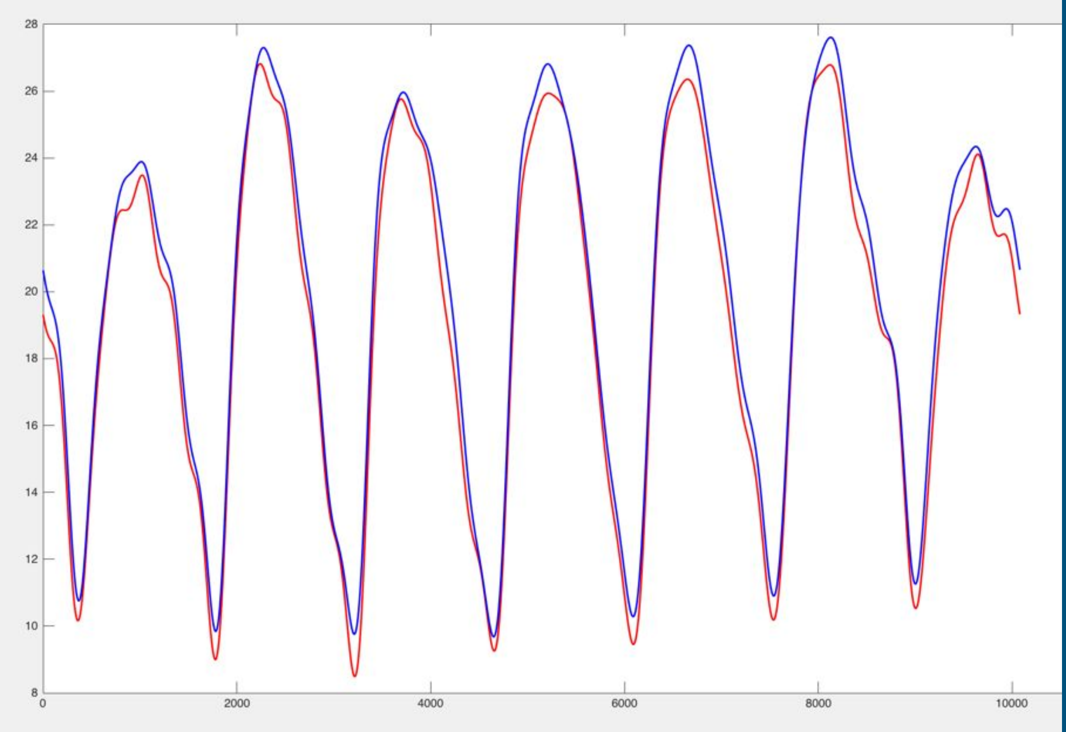
It *really* works.

Staffing for 90%ile of demand w/ empirical dist. (red), queueing model (blue)

Queueing theory : Why does it work so well?

An observation: the queueing theoretic model matches the empirical distribution best when demand is rising and falling (not at peak or anti-peak hours).

Let’s investigate that.



Queueing theory : Why does it work so well?

Blue = mean, Red = variance

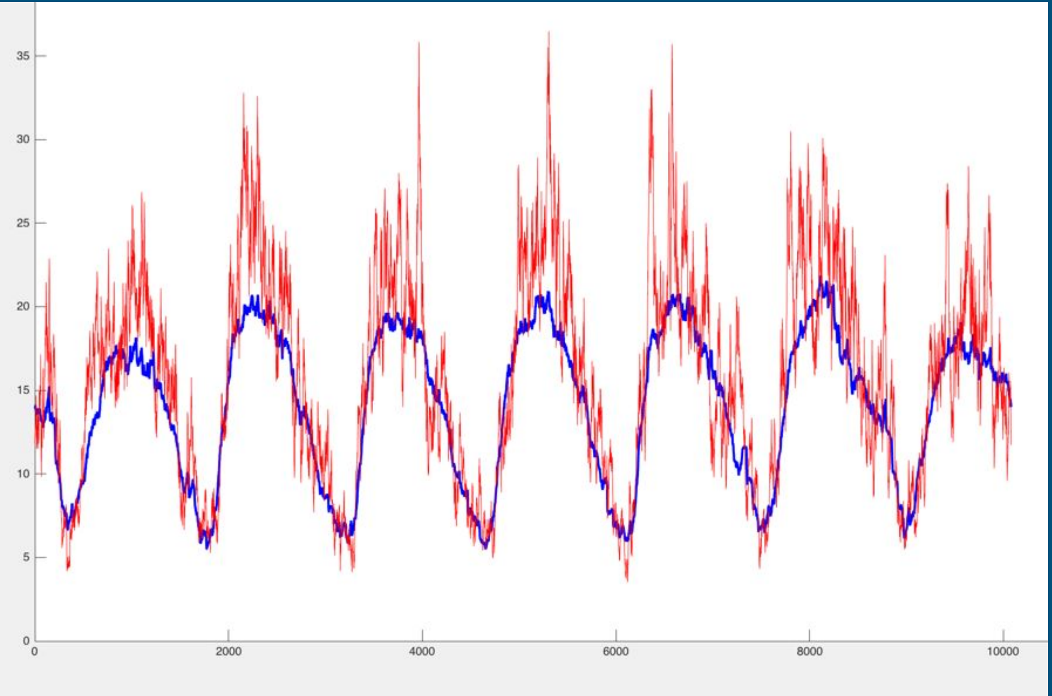
… other than peak hours, it looks like the mean tracks the variance!

Remind you of anything?

Steady state # in system for

M/G/∞ ~ ***Poisson(m)***

Poisson → mean == variance

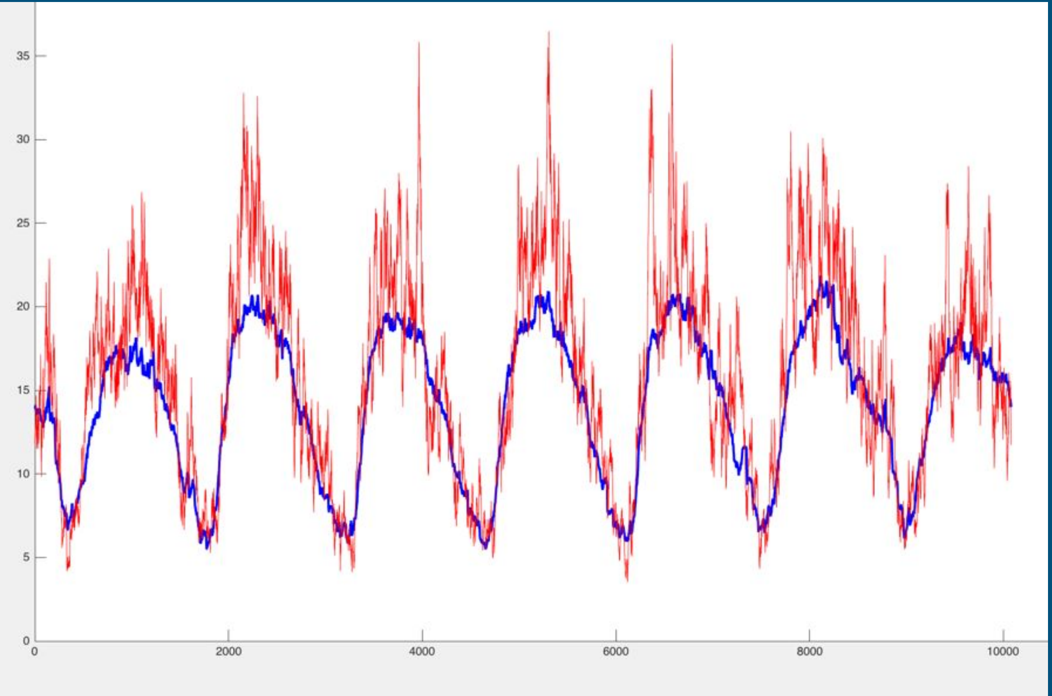


Queueing theory : Why does it work so well?

Blue = mean, Red = variance

You can run hypothesis tests and find # in system is...

- *Normally distributed* during peak hours.
- *Poisson distributed* for off-peak hours.



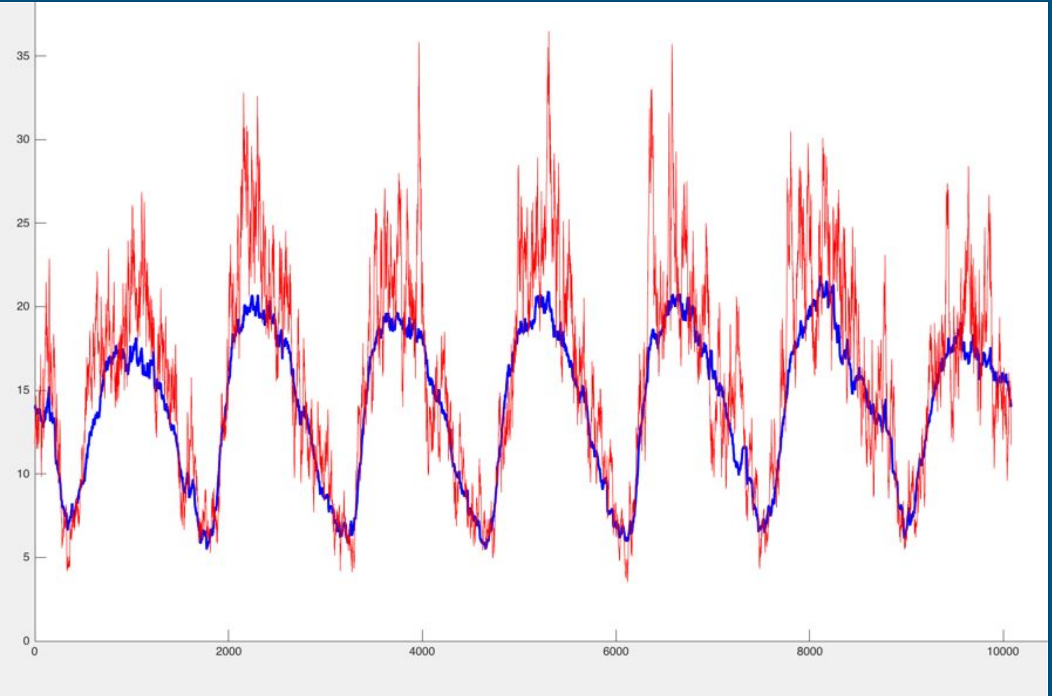
Queueing theory : Why does it work so well?

Blue = mean, Red = variance

Peak hours → $\mathcal{N}$

Off-peak hours → Poisson

Interpretation : off-peak hours see sufficient ambulances for system to appear as M/G/∞ .



Queueing theory : Is everyone so lucky?

I'm in a nice situation

- My arrivals are actually Poisson (what if they weren't?).
- We plan for 90%ile + [transit factor].
    - As a side effect, *transit factor enhances validity of M/G/∞ approximation*
    - What if we didn't have the transit factor? What if QoS was lower?
- I'm not making huge changes to the [queueing] system (what if I was?).

Non-Stationary Staffing for a [G/G/k](t)

What if my arrivals weren't Poisson?

Paralleling the treatment of the stationary model in Whitt (1992, §2), we suggest the approximation

$$\nu(t) \approx z(t)m(t), \quad \text{where} \tag{17}$$

$$z(t) = 1 + \frac{(c_a^2(t) - 1)}{E[S(t)]} \int_0^\infty [1 - G_t(x)]^2 dx, \tag{18}$$

$$c_a^2(t) \approx \frac{\mathrm{Var}[A(t) - A(t-\eta)]}{\int_{t-\eta}^{t} \lambda(u)du}, \tag{19}$$

There's something for that too!

Modifying an [M/G/k](t)

What if I was considering modifications to the [M/G/k](t) system?

I couldn't use historical data on number in system to estimate $m(t)$.

But I'd still have a shot!

$$m(t) = \int_{-\infty}^{t} G_u^c(t-u)\lambda(u)du,$$

$$G_t^c(x) = e^{-\int_t^{t+x} \mu(u)du}, \quad x \geq 0.$$

$G^c_u(t) \triangleq$ P{Service of an arrival at time "u" lasts longer than "t" time units }

… but this requires solving a differential equation.

Modifying an [M/M/k](t) : Overview

If services are exponential at rate μ, then we can recover $m(t)$ easily (with an ODE).

$$m'(t) = \lambda(t) - m(t)\mu,$$

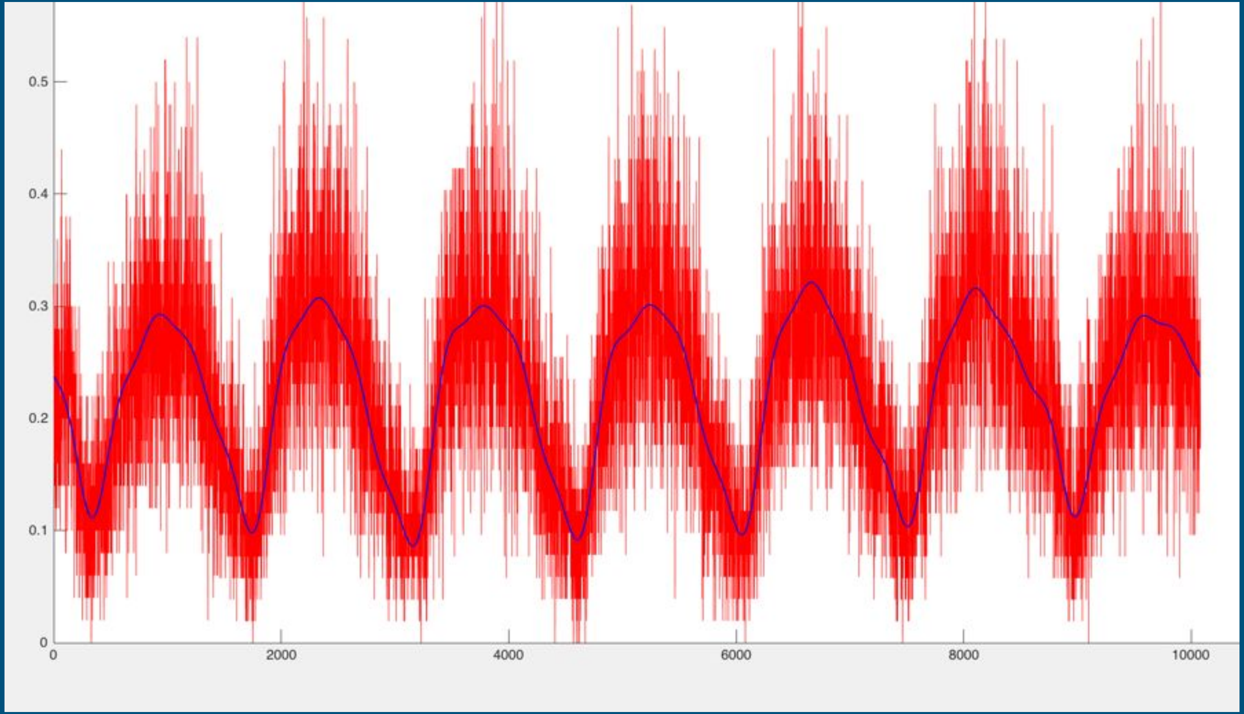
This could be useful when considering system modifications that could affect μ.

Although, now we have to estimate $\lambda(t)$

Modifying an [M/M/k](t) : Estimating $\lambda(t)$

MUCH more variation in measurements for $\lambda(t)$ than for $m(t)$.

Fourier approx. is decent, but there is a bigger concern of underestimating $\lambda(t)$.



Modifying an [M/M/k](t) : Solving the ODE

Google how to use MATLAB's "ODE45" function.

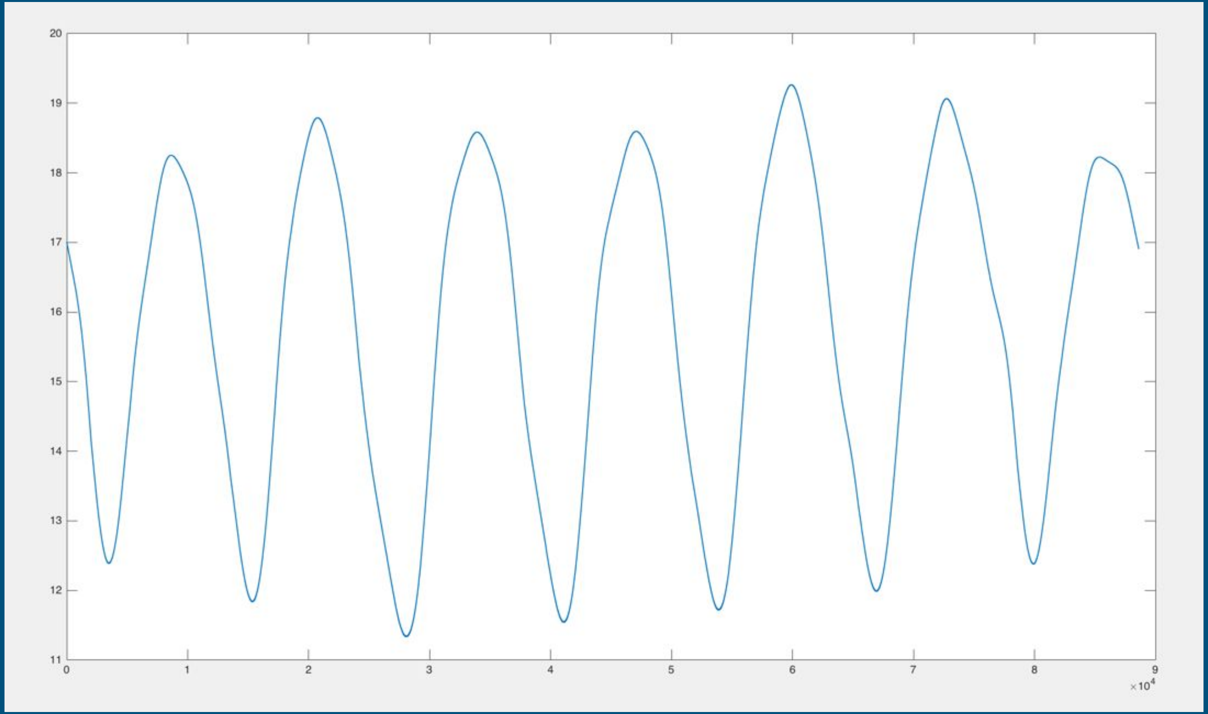
Specify

Initial cond., $m(0)$.

Service rate, $\boldsymbol{\mu}$.

Function, λ(t).

Hit "enter."



Recovering $m(t)$ by solving a differential eqn.

Do I recommend this?

- Not if you have access to historical estimates of $m(t)$

Why?

- Lots of variation in historical estimates of λ(t).
- Differential equation requires specifying initial condition, and mean service time.
- *DiffEq solution is* _*nearly, but not perfectly*_ *periodic with period of 1 week, even if* λ(t) *is*.
    - ^ This actually comments on the validity of the 1-week period assumption made at the beginning of this presentation.

Summary

- Queueing theory can make strong predictions about systems *without* stationary distributions.
- Real-world emergency services systems (e.g. ambulances for SFFD) can be modeled as simple $[M/G/\infty](t)$ queues (maybe "transit factor" is important?).
- Parameters of non-stationary stochastic systems can be modeled with deterministic differential equations, and we can solve these differential equations numerically.
  - This differential equations approach could be useful for systems-design work currently handled by simulation.

References

"Server Staffing to Meeting Time-Varying Demand"

Jennings, Massey, Whit (1996).

"Coping with Time-Varying Demand When Setting Staffing Requirements for a Service System"

Green, Kolesar, Whitt (2007).

Discusses application areas / practical concerns moreso than the 1996 paper.

Both of the papers above also discuss [M/M/s + M](t) queues (i.e. queues with "impatient customers" / customer abandonment).

Thank you!

Questions?