Fundamentals of Statistical Monitoring: The Good, Bad, & Ugly in Biosurveillance

Galit Shmuéli
Dept of Decision & Info Technologies
Robert H Smith School of Business
University of Maryland, College Park

Overview

- The main idea behind statistical monitoring
- Traditional monitoring tools
  - Control charts
  - Regression models
- Moving to pre-diagnostic data

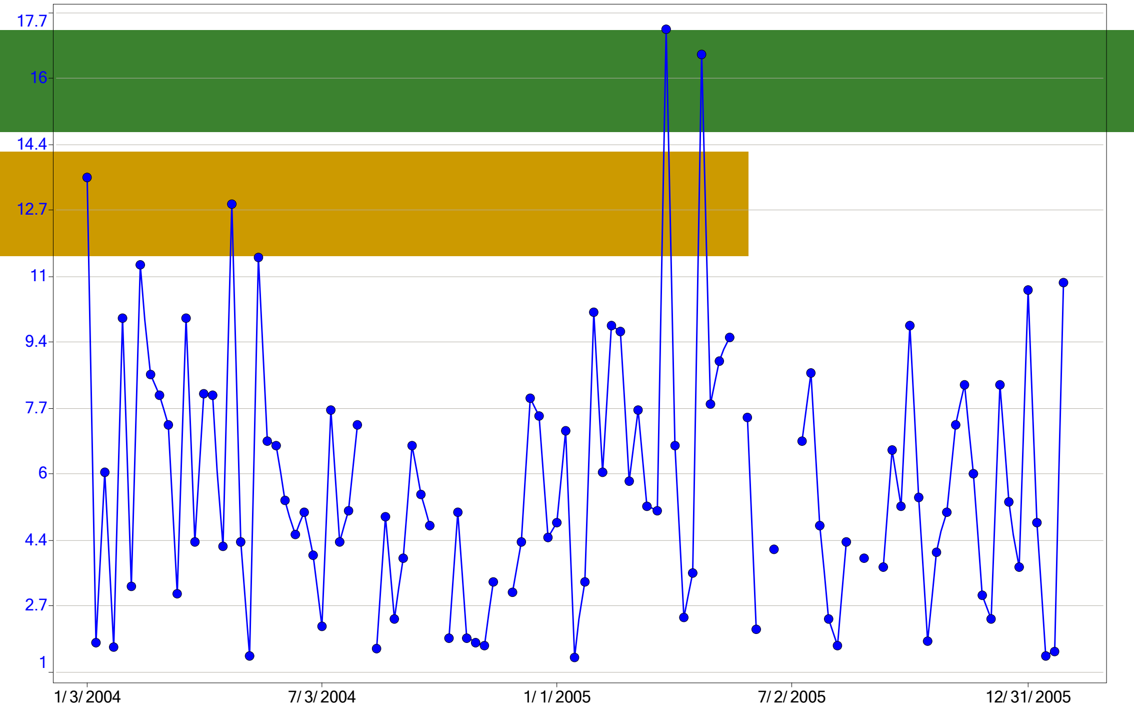
The main idea

- Monitor a stream of incoming data, and signal an alarm if there is indication of abnormality
- “Abnormality” – define normal

Any P&I outbreak(s) in Newark, NJ in this period (2004-2006)?

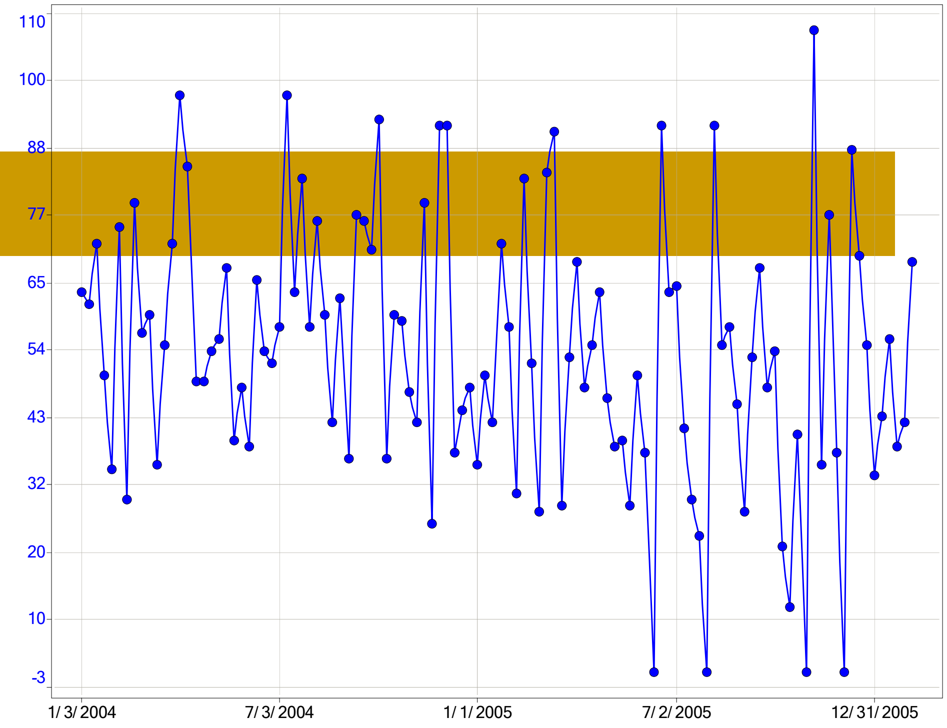
57%	1.	Yes
43%		No

Weekly % P&I deaths (relative to overall death)



Any outbreak(s) of Gonorrhea in Mass. in this period?

Weekly Gonorrhea counts in Mass. '04-'06



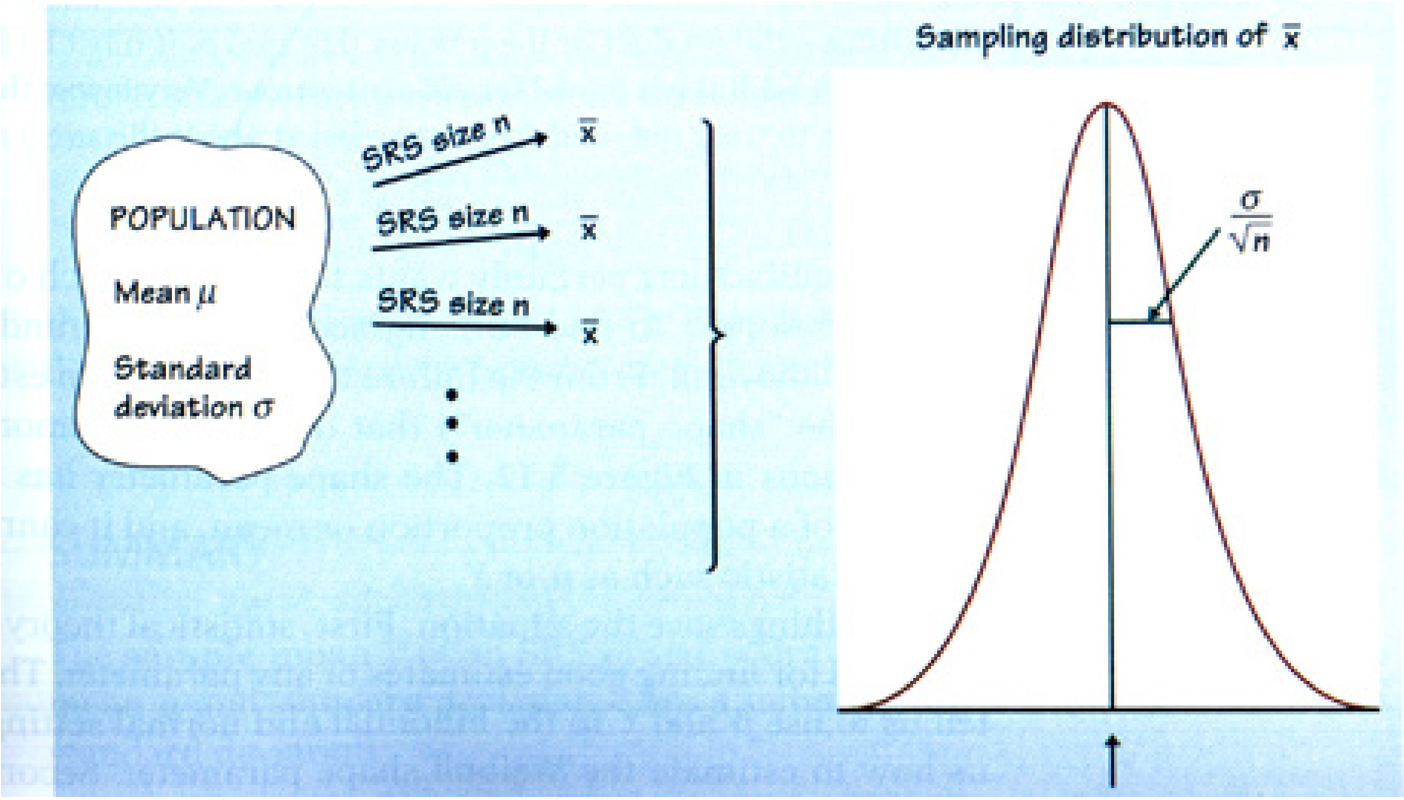
23%	1.	Yes
77%		No

Control charts: Shewhart charts

- Originally used to monitor a **process mean** in an industrial setting.
- Assumption: there is an “in-control” mean, and we want to detect when it goes “out-of-control”.
- Natural variability vs. “special cause”
- Method: draw a small random sample at repeated time intervals, and compare the **sample mean** to lower/upper thresholds.
- If the sample mean exceeds a threshold, then trigger an alarm and stop the process.

What is "normal"?

The mean ($\bar{X}$) should be Normal!



$$P\left(\mu - 3\frac{\sigma}{n} \le \bar{X} \le \mu + 3\frac{\sigma}{n}\right) = 0.9973$$

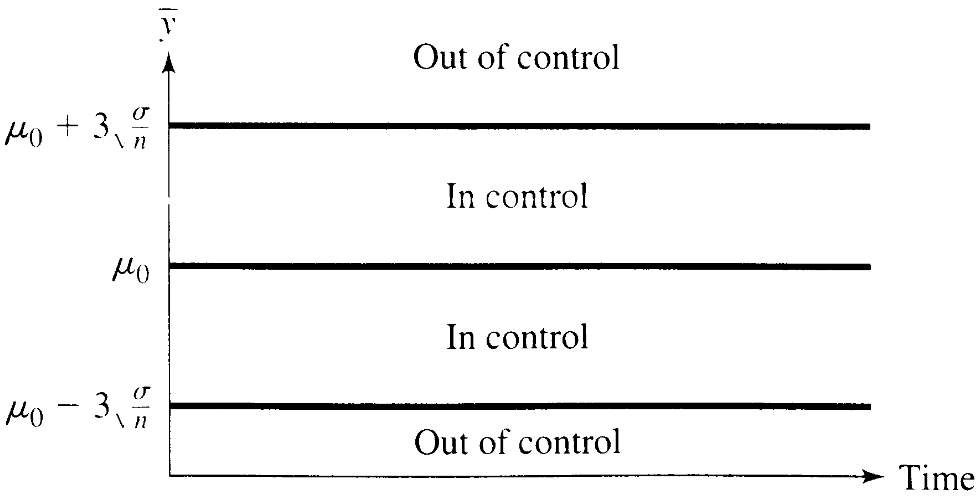
The X-bar chart (A Shewhart 3-sigma chart)

$$CL = \mu_0$$

$$LCL, UCL = CL \pm 3\sigma / \sqrt{n}$$

FIGURE 5.3

A Basic Illustration of a Control Chart



The thresholds take into account the variability of the sample mean around the process mean

Shewhart chart assumptions

- The statistic measured at time *t* is normally distributed
  - If a single measurement is taken every time unit – we assume the measurements are normally distributed. This is called an “i-chart”
  - If the statistic is a rate, you have a “p-chart”
- Samples taken at different time points are independent of each other

The X-bar chart: Example

- Data from Philips Semiconductors.
- 30 Samples of size *n=5* silicon wafers were taken every time unit.
- The thickness of each wafer was recorded, and the sample mean calculated.
- Target thickness = 244
- Standard deviation σ = 3.1

sample	X1	X2	X3	X4	X5		x-bar
1	240	243	250	253	248		246.8
2	238	242	245	251	247		244.6
3	239	242	246	250	248		245
4	235	237	246	249	246		242.6
5	240	241	246	247	249		244.6
6	240	243	244	248	245		244
7	240	243	244	249	246		244.4
8	245	250	250	247	248		248
9	238	240	245	248	246		243.4
10	240	242	246	249	248		245
11	240	243	246	250	248		245.4
12	241	245	243	247	245		244.2
13	247	245	255	250	249		249.2
14	237	239	243	247	246		242.4
15	242	244	245	248	245		244.8
16	237	239	242	247	245		242
17	242	244	246	251	248		246.2
18	243	245	247	252	249		247.2
19	243	245	248	251	250		247.4
20	244	246	246	250	246		246.4
21	241	239	244	250	246		244
22	242	245	248	251	249		247
23	242	245	248	243	246		244.8
24	241	244	245	249	247		245.2
25	236	239	241	246	242		240.8
26	243	246	247	252	247		247
27	241	243	245	248	246		244.6
28	239	240	242	243	244		241.6
29	239	240	250	252	250		246.2
30	241	243	249	255	253		248.2

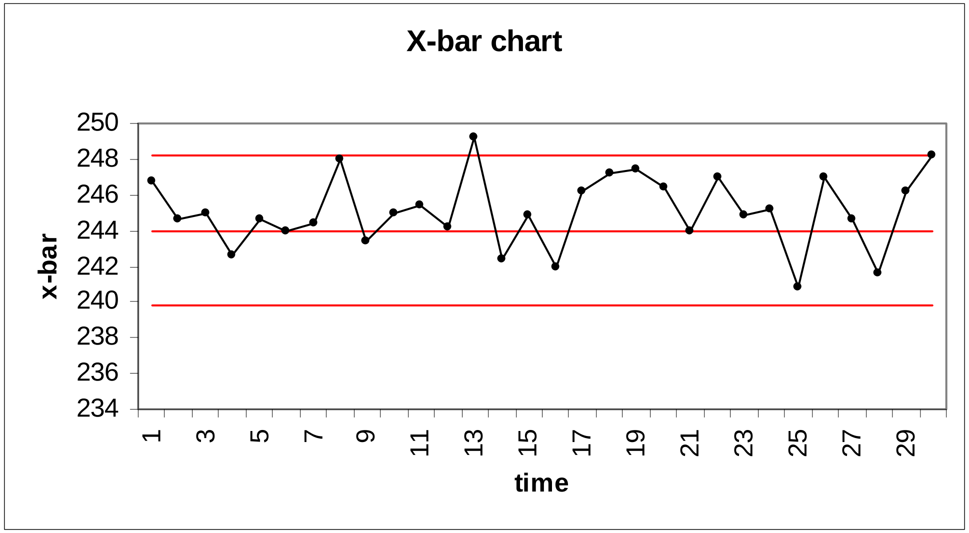
The X-bar chart: Example (cont.)

$$CL = 244$$

$$LCL, UCL = 244 \pm 3 \times 3.1/\sqrt{5}$$

$$LCL = 239.84$$

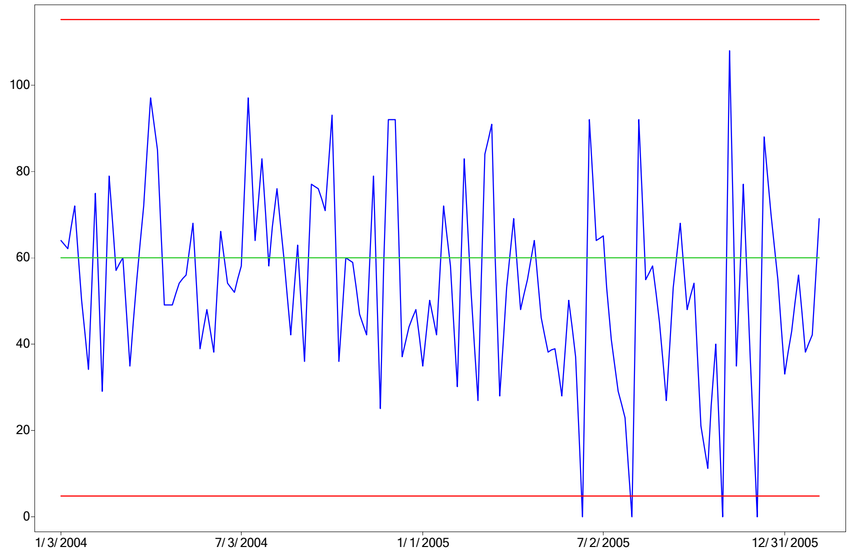
$$UCL = 248.16$$



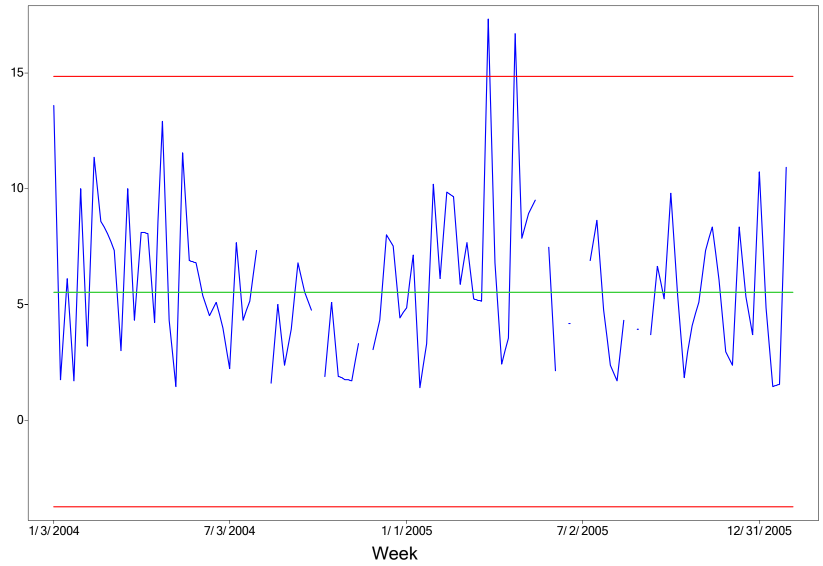
Shewhart chart for weekly data

- Use “stable” period to estimate mean and std for thresholds (used 2004)

Gonorrhea in Mass.



% P&I Deaths in Newark, NJ



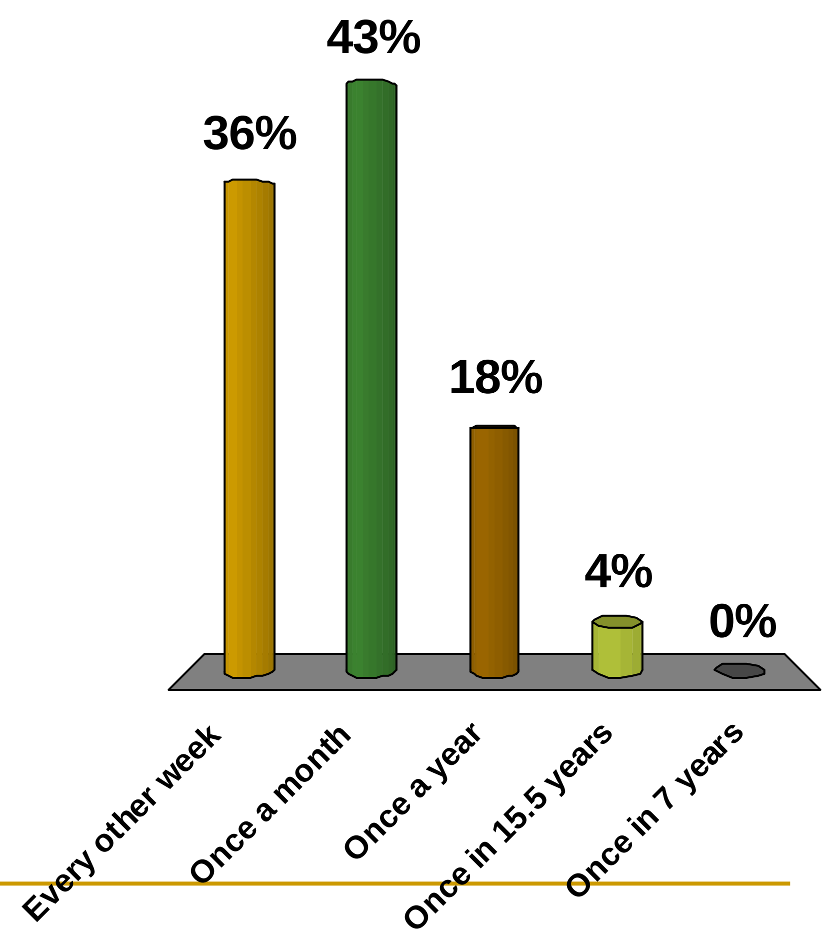
When will a Shewhart signal an alarm?

- Probability that a point exceeds the limits, when the process mean shifts by *k* std:

k	P(Alarm)
0	.0027
1	.0228
-1	.0228
2	.1587
3	.5000

How often should we expect a false alarm with a Shewhart chart? (with weekly data)

1. Every other week
2. Once a month
3. Once a year
4. Once in 15.5 years
5. Once in 7 years



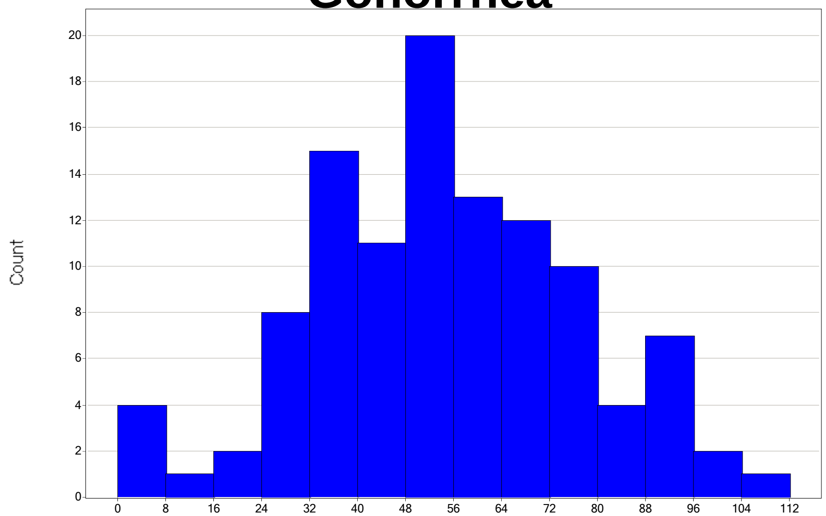
1/0.0027 = 370 weeks ≅ 7 years

Catch #1: How to set LCL, UCL?

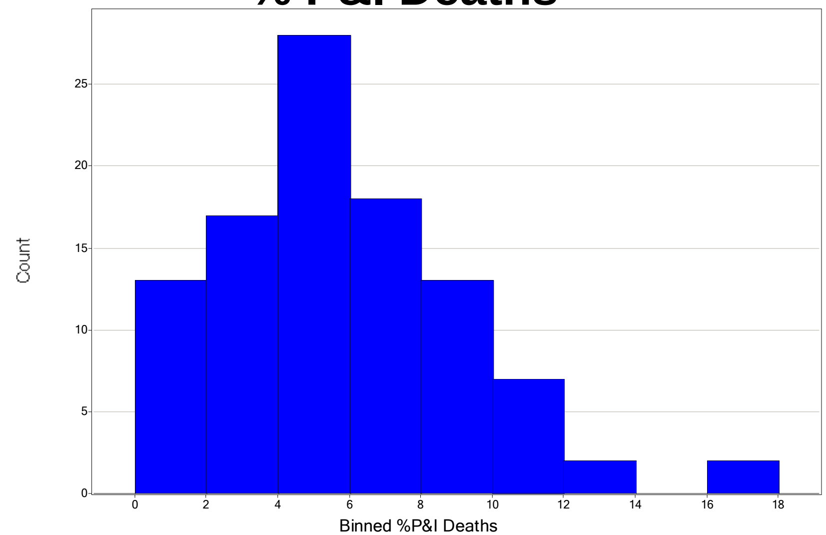
- Best: underlying domain knowledge
  - “Rate of Gonorrhea in population above X considered outbreak”
  - “Number of weekly cases above X…”
- In the absence, use historical data
  - To estimate of population parameter
  - Make sure the historic period has no outbreaks!
  - How to determine?
- The bad: lack of gold standards

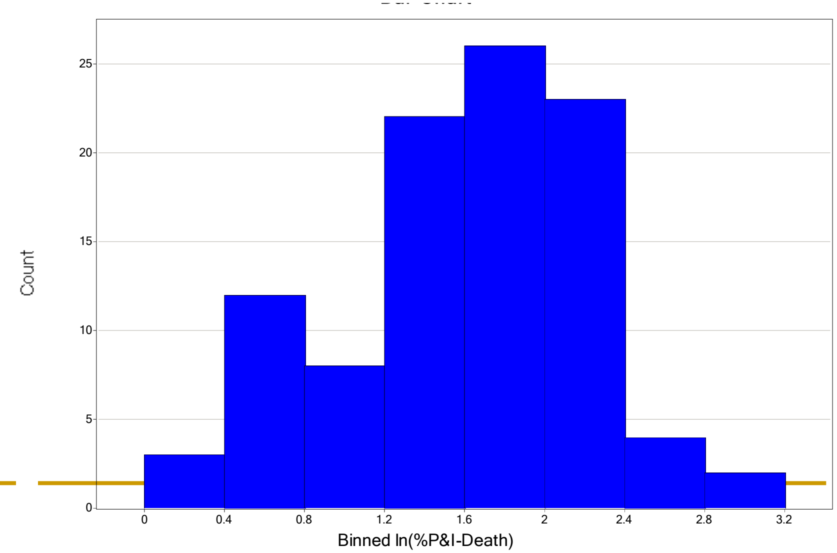
Catch #2: are the data normal?

Gonorrhea



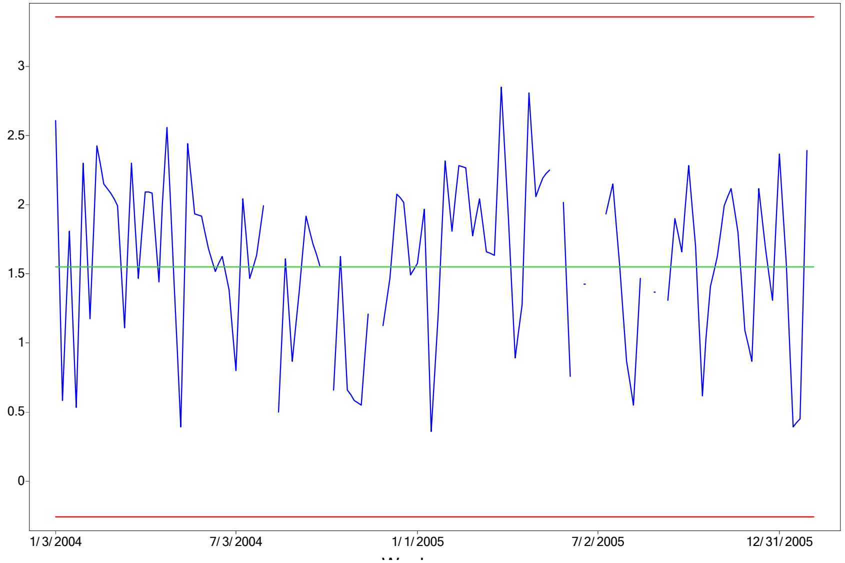
% P&I Deaths



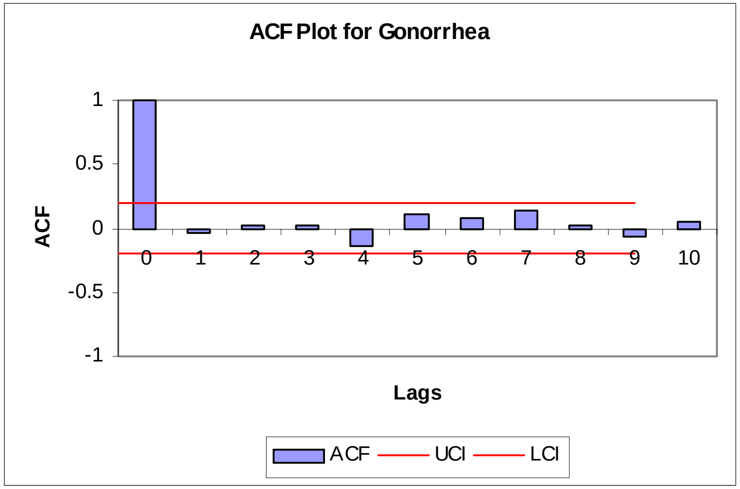


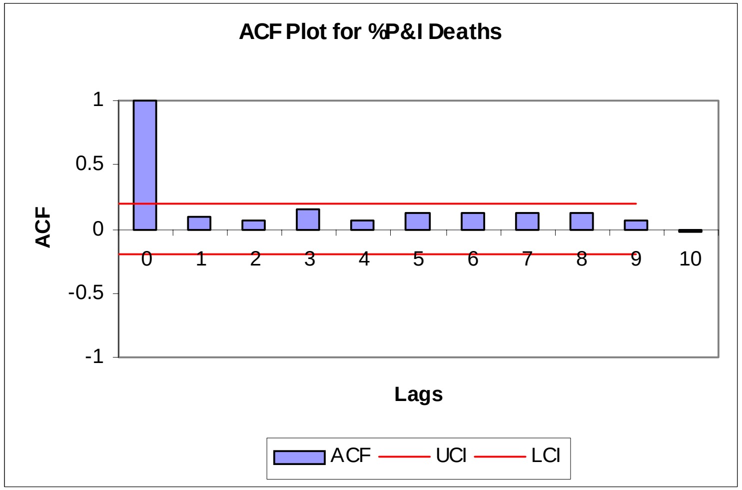
- If not, two tricks:
  - Transform the data (right skew -> take log)
  - Use a more suitable Shewhart chart

Shewhart chart for transformed data



Catch #3: are the counts correlated?





- Compute autocorrelation at lag 1,2,…
- If autocorrelated at a low lag, need time-series model
- If autocorrelated at constant multiples then there is seasonality

Shewhart Charts – useful for biosurveillance?

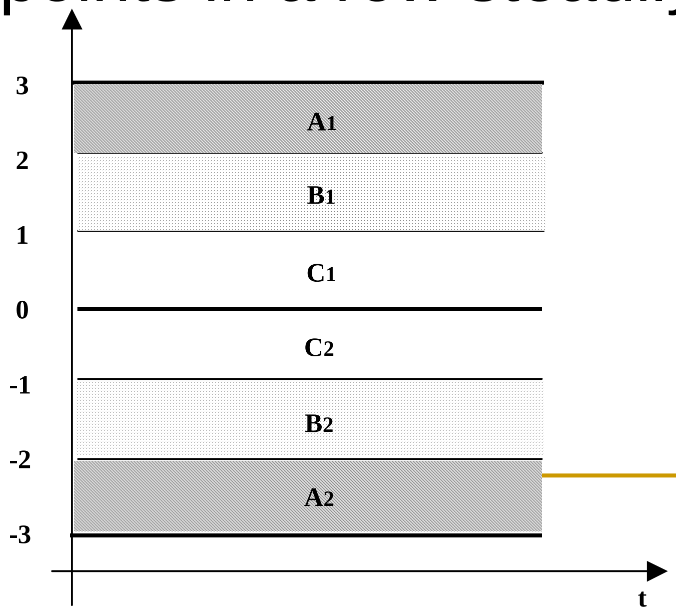
- The good:
  - When assumptions are satisfied, these charts are good at quickly detecting large spikes/dips
  - Very simple
- The bad:
  - Outbreak that manifests as smaller, consistent increases will go undetected
  - Hard in some cases to determine "normal period"
- The ugly: Assumptions are often violated. Even more so with pre-diagnostic data.

Detecting small or other types of changes

- Method 1: make the Shewhart more sensitive
- Method 2: use a different chart altogether

Shewhart chart with extra alarming rules

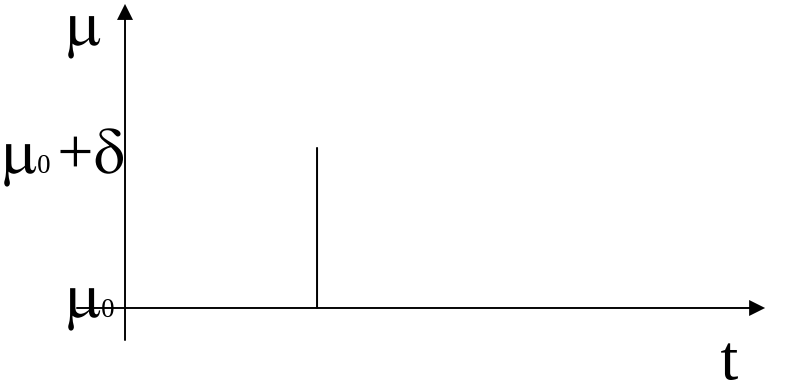
- Western Electric Rules (1956) -- Signal if (in addition to exceeding LCL,UCL):
  - 8 consecutive points are on one side of the CL
  - 2 of 3 consecutive points are in zone A
  - 6 points in a row steadily increasing/decreasing



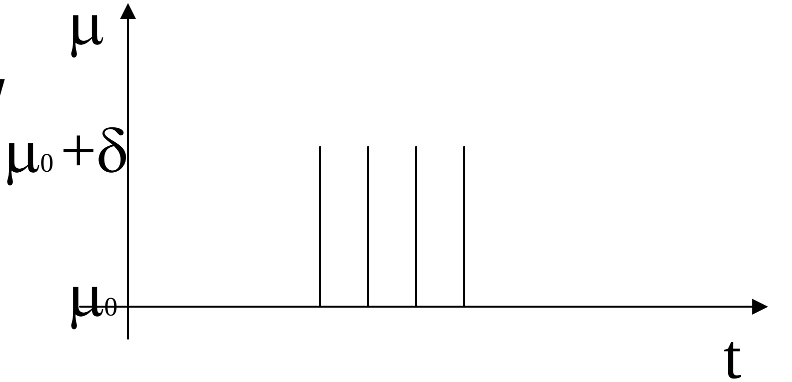
- Increases false alarms
- Choose only relevant rules
- Don't run all rules together

Detecting a shift with a known pattern

- Shewhart charts:

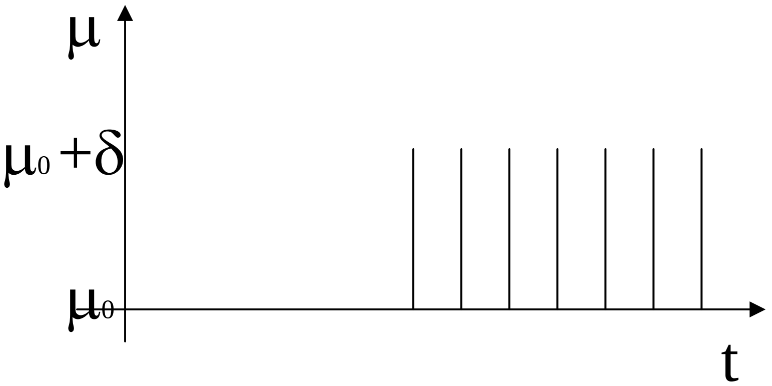


- Moving Average charts (with window of 4):



Detecting a shift with a known pattern – cont.

- CuSum charts:



- EWMA charts:

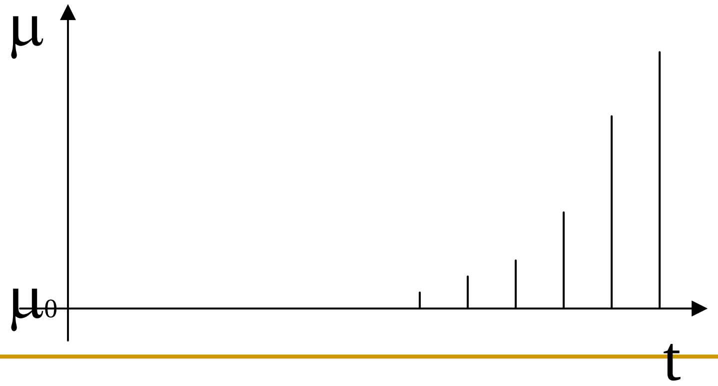


Chart assumptions

- Target mean is constant
- The statistic measured at time t is normally distributed
- Samples taken at different times are independent of each other

The Moving-Average (MA) chart for single daily counts

- Points on the plot are averages of sliding window:

$$MA_t = (X_t + X_{t-1} + ... + X_{t-b+1}) / b$$

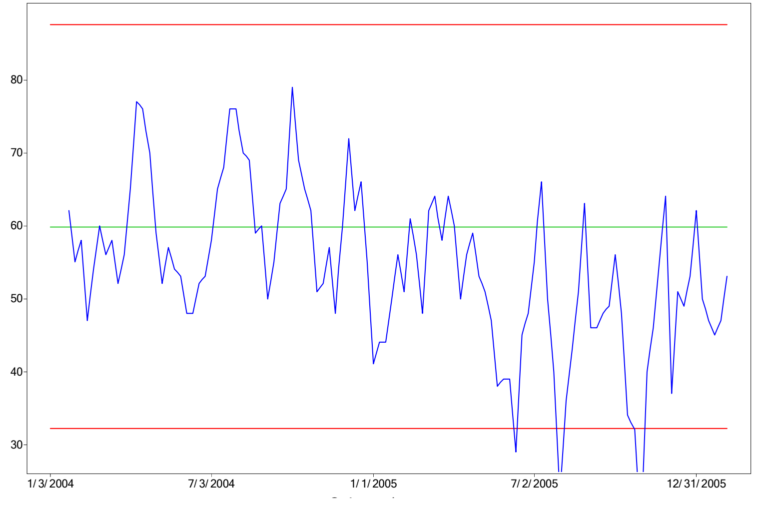
- Control limits:

$$CL = \mu_0$$

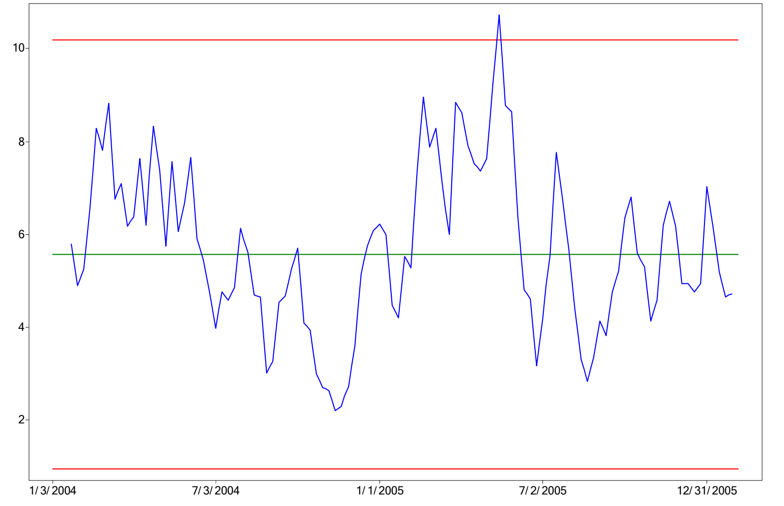
$$LCL, UCL = CL \pm 3\frac{\sigma}{\sqrt{b}}$$

Moving Average chart (b=4 weeks)

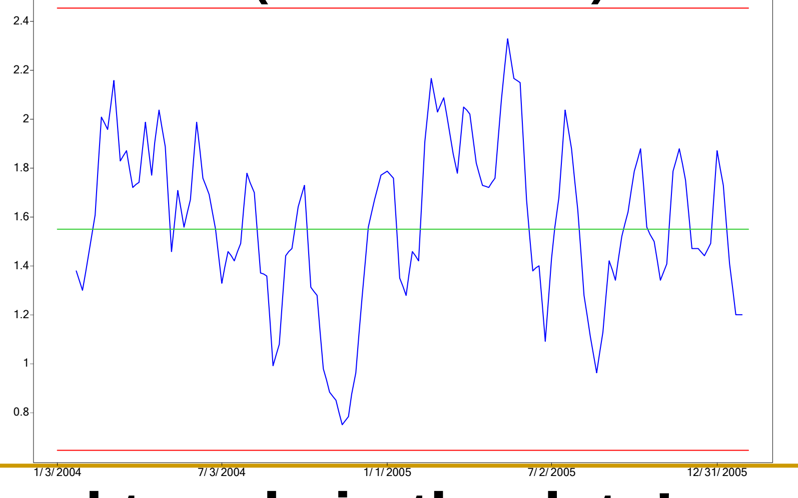
Gonorrhea



% P&I Deaths



LOG(% P&I Deaths)



Good way to SEE patterns and trends in the data!

The Cumulative Sum (CuSum) chart

- On day t,
  - Compute deviation of count from target $X_t - \left(\mu_0 + \frac{\delta}{2}\right)$
  - Accumulate the deviations until time t
  - Restart the counter if it goes below zero

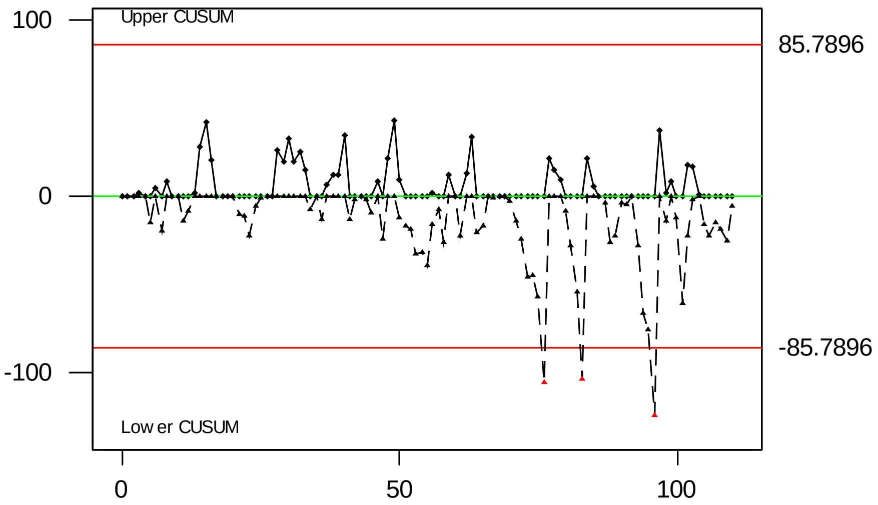
$$S_t^+ = \max\left\{0,\ S_{t-1}^+ + X_t - \left(\mu_0 + \frac{\delta}{2}\right)\right\}$$

  - Signal if $S_t^+ > h\sigma$

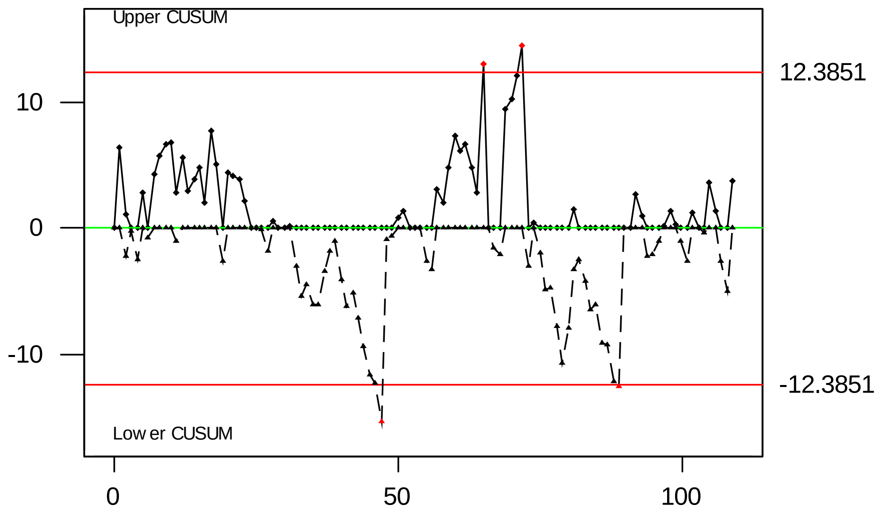
  - Can construct Cusum for detecting decrease

CuSum with (h=4, δ=1)

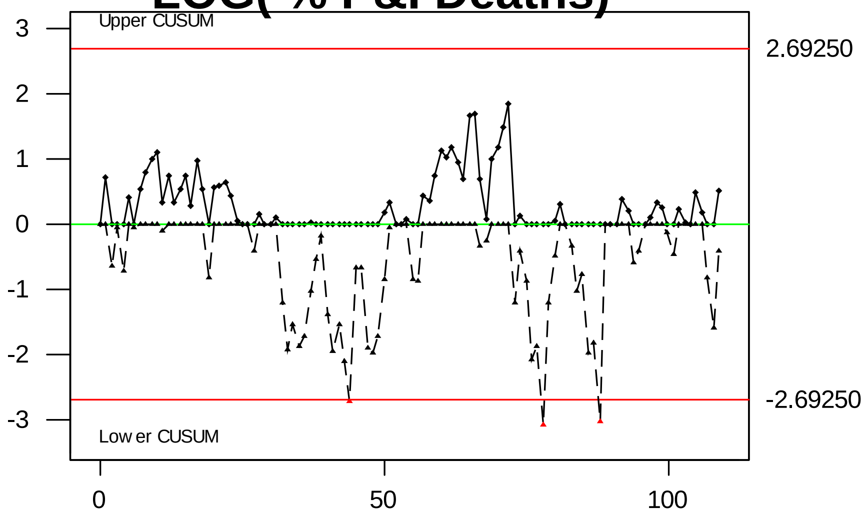
Gonorrhea



% P&I Deaths



LOG(% P&I Deaths)



Missing values? Zero them?

Exponentially Weighted Moving-Average (EWMA) chart

- Points on the plot:

$$\tilde{X}_t = (1-\theta)\left(X_t + \theta X_{t-1} + \theta^2 X_{t-2} + \cdots\right) = (1-\theta)X_t + \theta\tilde{X}_{t-1}$$

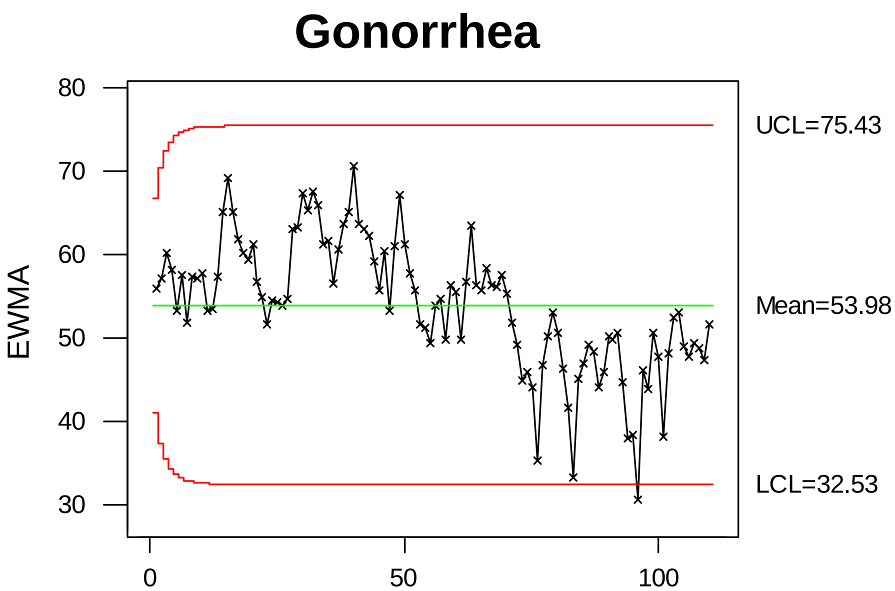
$$0 < \theta < 1$$

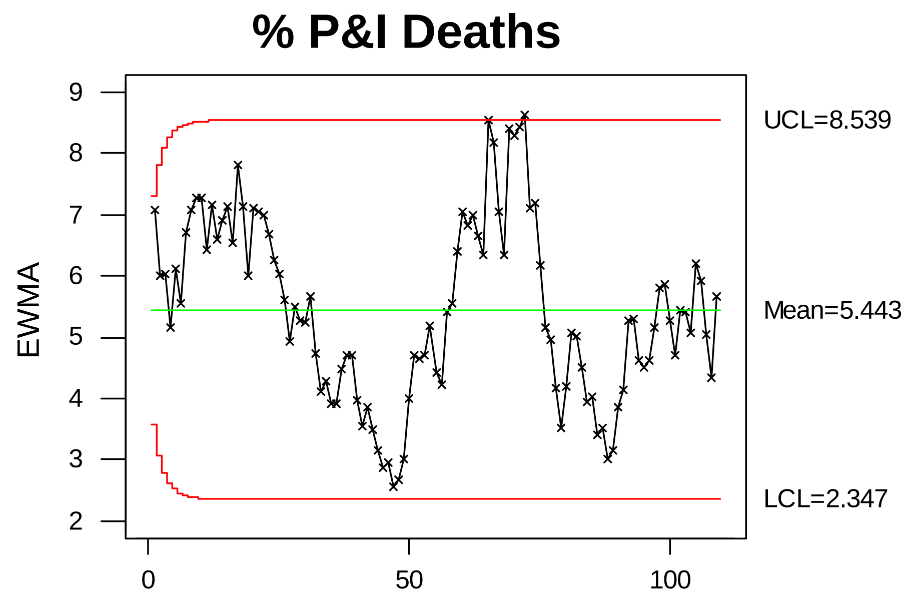
- Control limits:

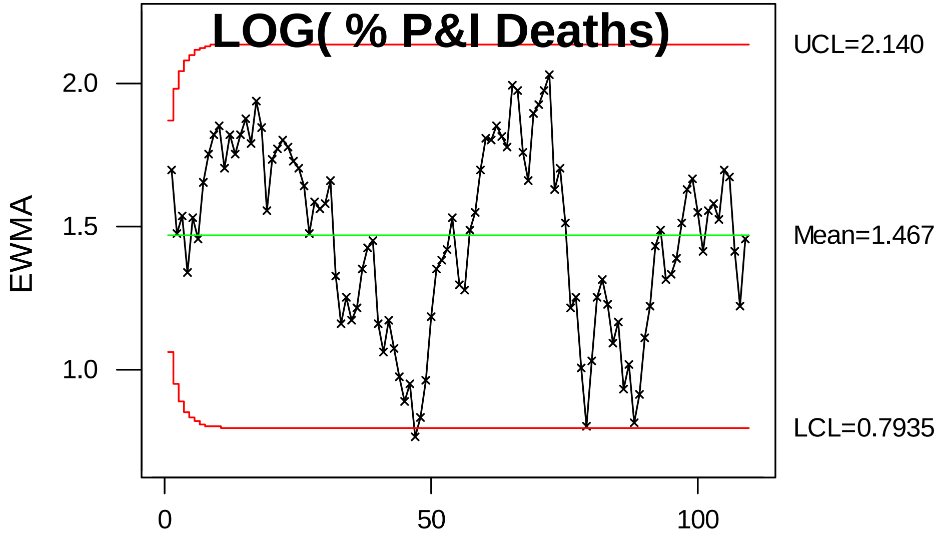
$$CL = \mu_0$$

$$LCL, UCL = CL \pm 3\sigma\sqrt{\frac{1-\theta}{1+\theta}}$$

EWMA charts for weekly data







Regression models for removing seasonality and trend

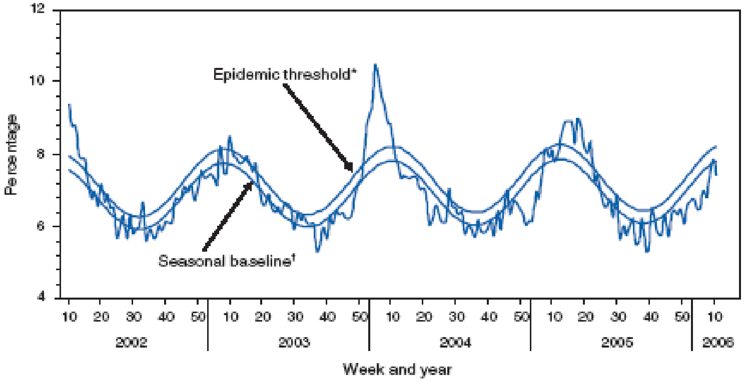
- Control charts assume no trend, no seasonality
- Regression models
  - Exp trend + multiplicative quarterly seasonality

$$\log(y_t) = \alpha + \beta_1 Q_1 + \beta_2 Q_2 + \beta_3 Q_3 + \beta t + \varepsilon_t$$

  - Sinusoidal (CDC model for %P&I deaths, annual cycle)

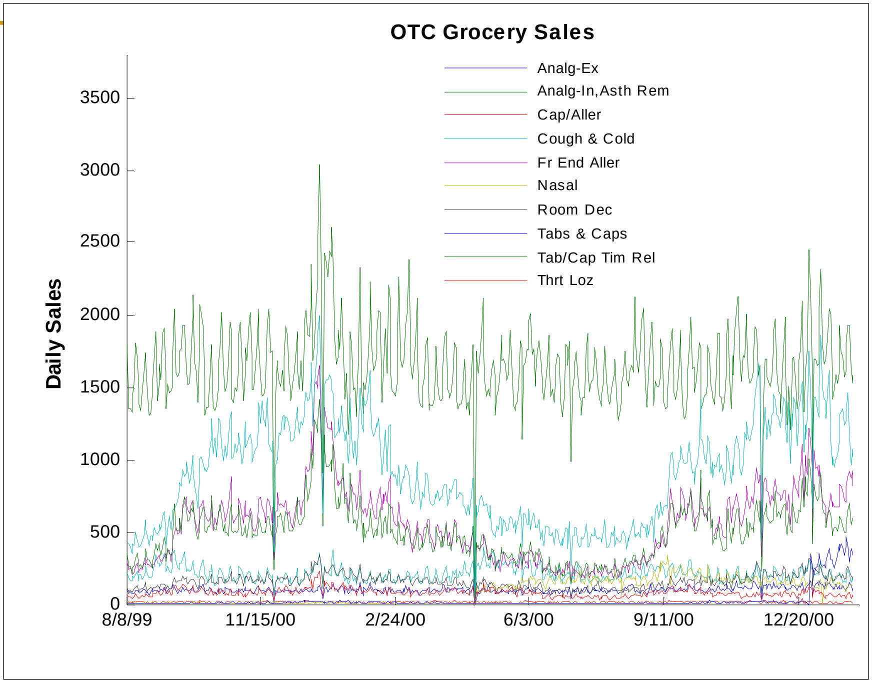
$$y_t = \alpha + \beta_1 Cos(t/365.25 + \beta_2) + \beta t + \varepsilon_t$$

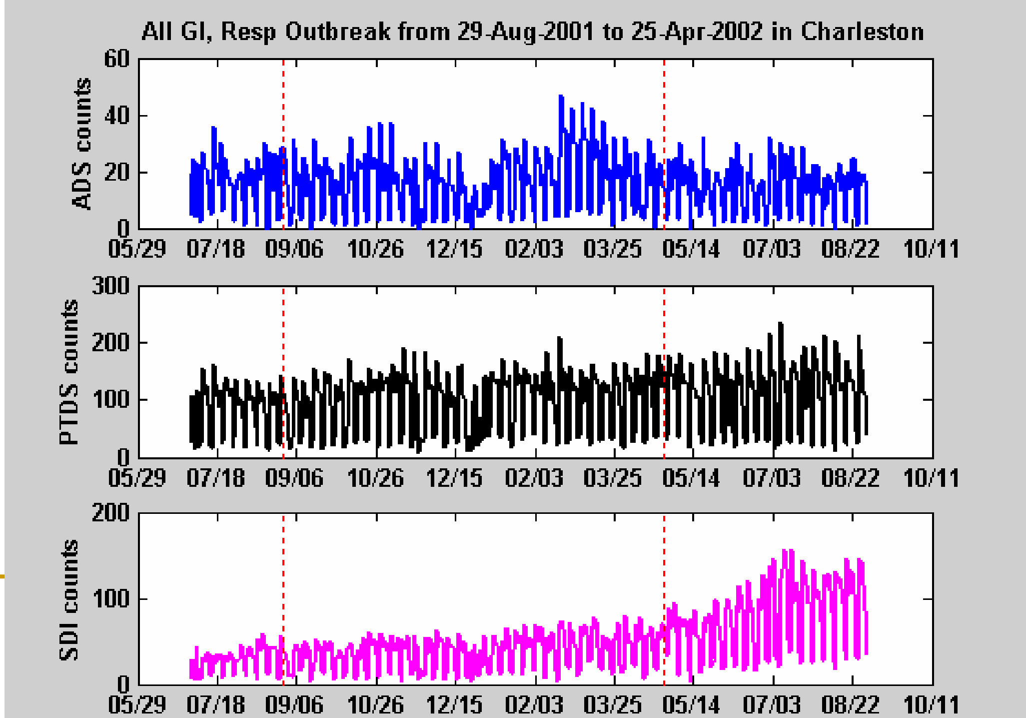
  - Can stratify by adding predictors
  - Use RESIDUALS in control chart
  - The ugly:
    - What if pattern changes?
    - Autocorrelation



Pre-diagnostic data: A whole new ball game

- Daily data
- Day-of-week effect
- Some series seasonal
- Non-stationary, local
- Vastly different across/within sources
- Correlate with other irrelevant variables
- Missing data (school absences on holidays)
- Infected by provider issues
- Low vs. high counts
- Lack of **domain knowledge**





What's the moral?

1. Preprocess series before applying control charts — 3%
2. Pre-diagnostic data require different tools/treatment than traditional data — 8%
3. I need a refresher statistics course — 3%
4. 1&2 — 33%
5. All of the above — 53%