CS626 Data Analysis and Simulation

Instructor: Peter Kemper

R 104A, phone 221-3462, email:kemper@cs.wm.edu

Office hours: Monday, Wednesday 2-4 pm

Today:

Stochastic Input Modeling

Reference: Law/Kelton, Simulation Modeling and Analysis, Ch 6.
NIST/SEMATECH e-Handbook of Statistical Methods,
http://www.itl.nist.gov/div898/handbook/

What is input modeling?

- Input modeling
  - Deriving a representation of the uncertainty or randomness in a stochastic simulation.
  - Common representations
    - Measurement data
    - Distributions derived from measurement data <-- focus of "Input modeling"
      - usually requires that samples are i.i.d and corresponding random variables in the simulation model are i.i.d
      - i.i.d. = independent and identically distributed
      - theoretical distributions
      - empirical distribution
    - Time-dependent stochastic process
    - Other stochastic processes
- Examples include
  - time to failure for a machining process;
  - demand per unit time for inventory of a product;
  - number of defective items in a shipment of goods;
  - times between arrivals of calls to a call center.

Representation by a single distribution

Given:

- Set of sample data for some real world phenomenon
  e.g. interarrival times of tasks for some computing node

Goal:

- Represent data by a single distribution which is used in a simulation study to draw interarrival times X_1, X_2, ... between tasks

Implies:

- Model uses random variables X_i for those times, X_i's are assumed to be independent and identically distributed

Is this a reasonable assumption?

Test question: Would $X_n,...,X_1$ tell you a different story, have a different meaning?

Check with

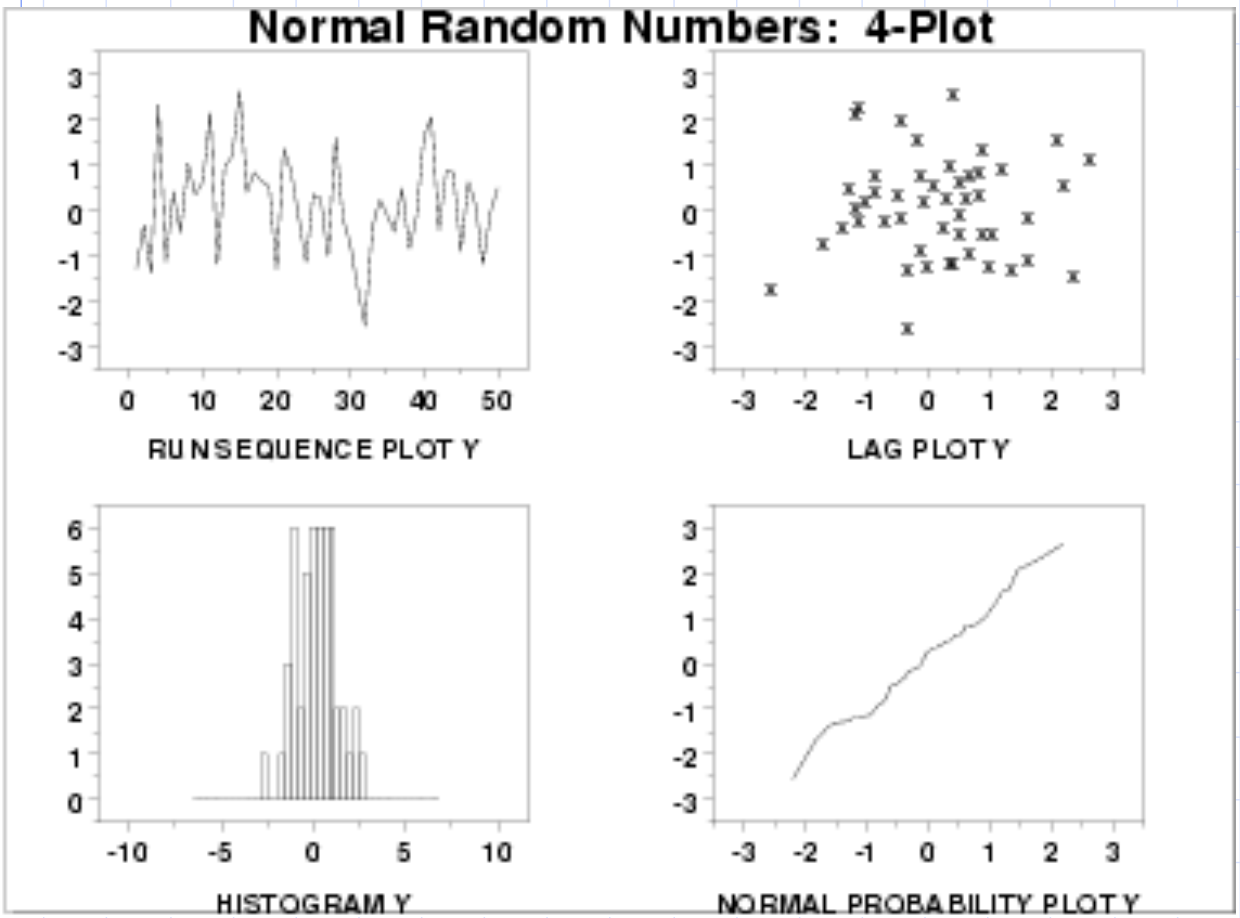
NIST/SEMATECH e-Handbook of Statistical Methods,
http://www.itl.nist.gov/div898/handbook/
in particular Chapter 1.2 EDA assumptions
http://www.itl.nist.gov/div898/handbook/eda/section2/eda2.htm

Exploratory Data Analysis (EDA): Assumptions

- Four typical assumptions for measurements processes: data from the process at hand "behave like":
  1. random drawings;
  2. from a fixed distribution;
  3. with the distribution having fixed location; and
  4. with the distribution having fixed variation.
- Fixed location:
  - response = deterministic component + random component
  - univariate case: response = constant + error
  - so fixed location is the unknown constant
  - can be extended to a function of many variables
  - effect: residuals (error) between measurement and response should behave like a univariate process with same assumed properties above
  - such that testing of underlying assumptions becomes a tool for the validation and quality of fit of the chosen model
- 4 assumptions hold => probabilistic predictability, process is “in statistical control”, can do predictions

EDA: Four techniques for testing assumptions

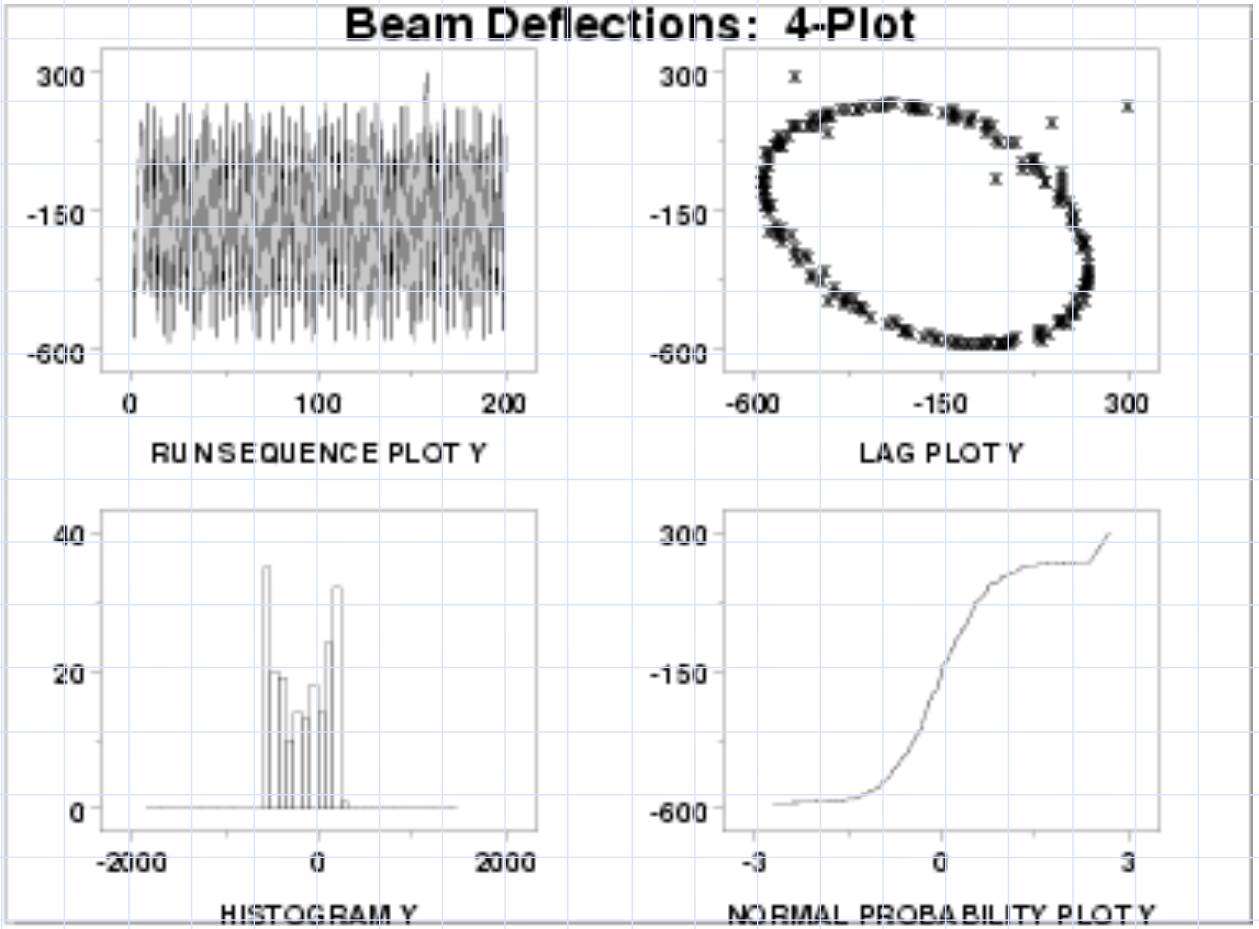
1. run sequence plot (Y_i versus i)
2. lag plot (Y_i versus Y_{i-1})
3. histogram (counts versus subgroups of Y)
4. normal probability plot (ordered Y versus theoretical ordered Y)



Example:
Process with
- fixed location
- fixed variation
- random distribution
- approx. normal
- no outliers

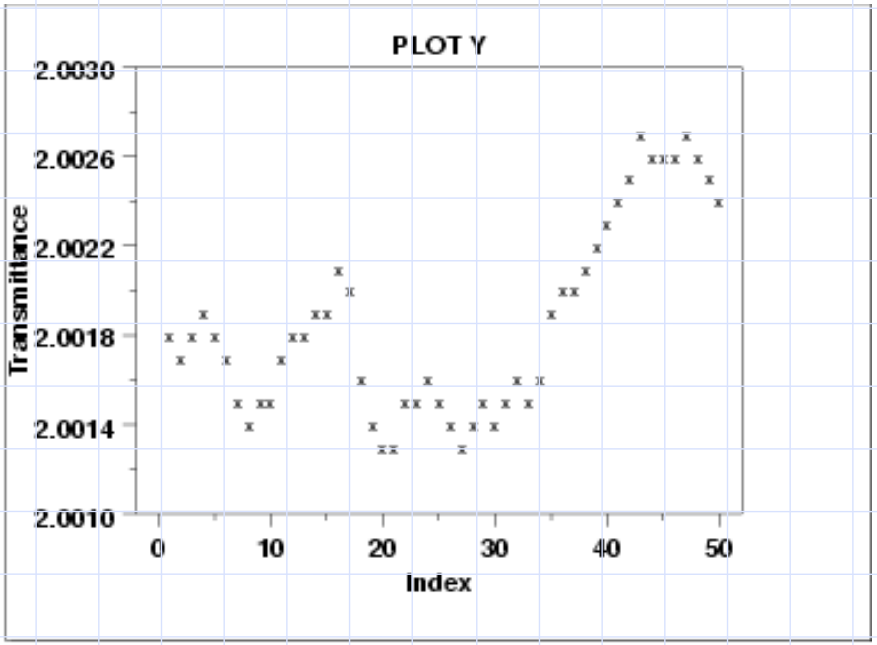
EDA: Four techniques for testing assumptions

1. run sequence plot (Y_i versus i)
2. lag plot (Y_i versus Y_{i-1})
3. histogram (counts versus subgroups of Y)
4. normal probability plot (ordered Y versus theoretical ordered Y)



Example:
Process with
- fixed location
- fixed variation
- non-random (oscillatory)
- non-normal U-shaped distribution
- several outliers

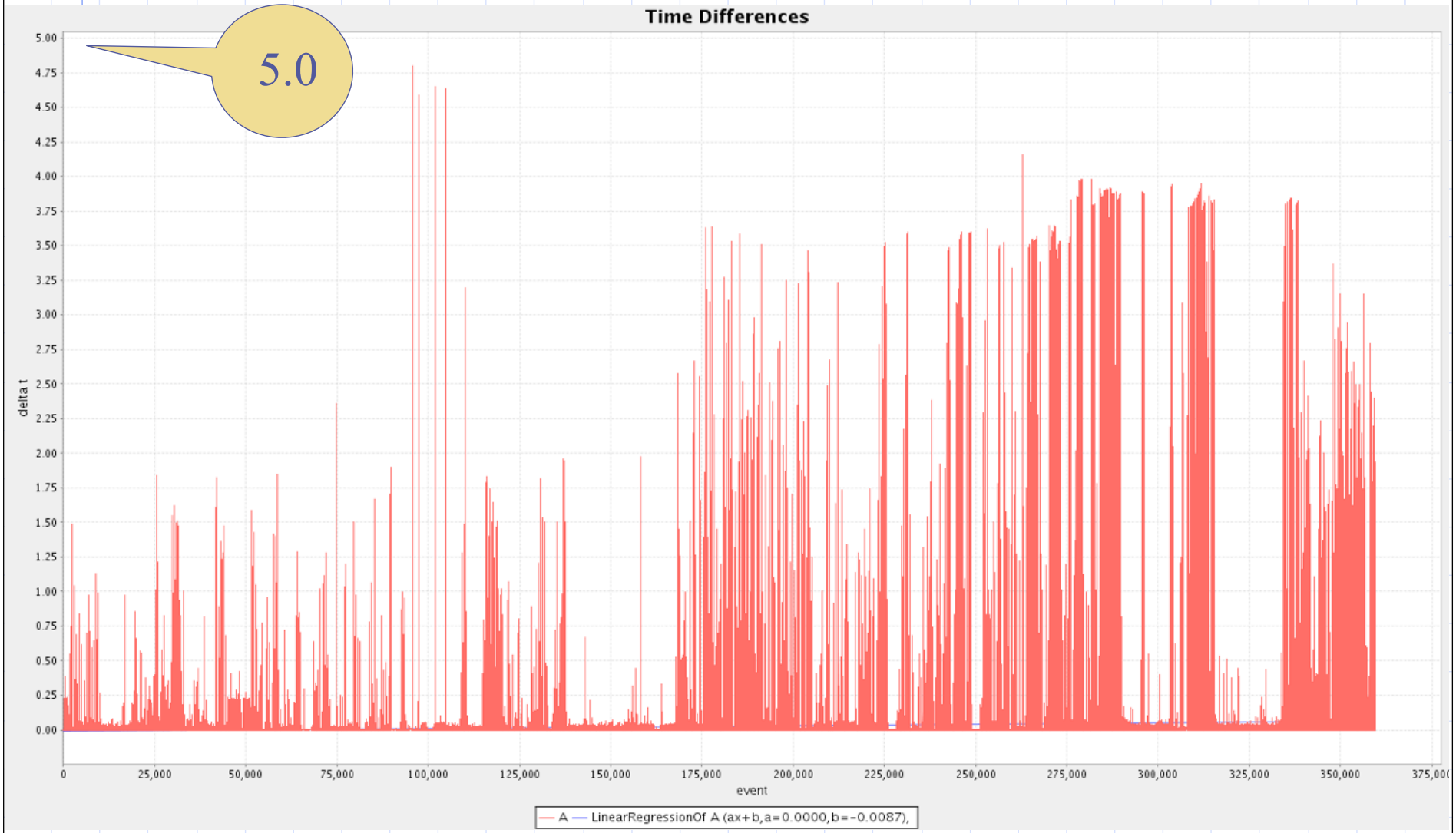
Run-Sequence Plot



- Purpose:
  - Check for Shifts in Location and Scale and Outliers
- Definition:
  - y(i) versus i
- Importance:
  - Check univariate assumptions: Y = constant + error where error is assumed to be random, from a fixed distribution with constant location and scale
  - More complex models $Y_i = f(X_1, ..., X_k) + E_i$ have same assumptions for the error term

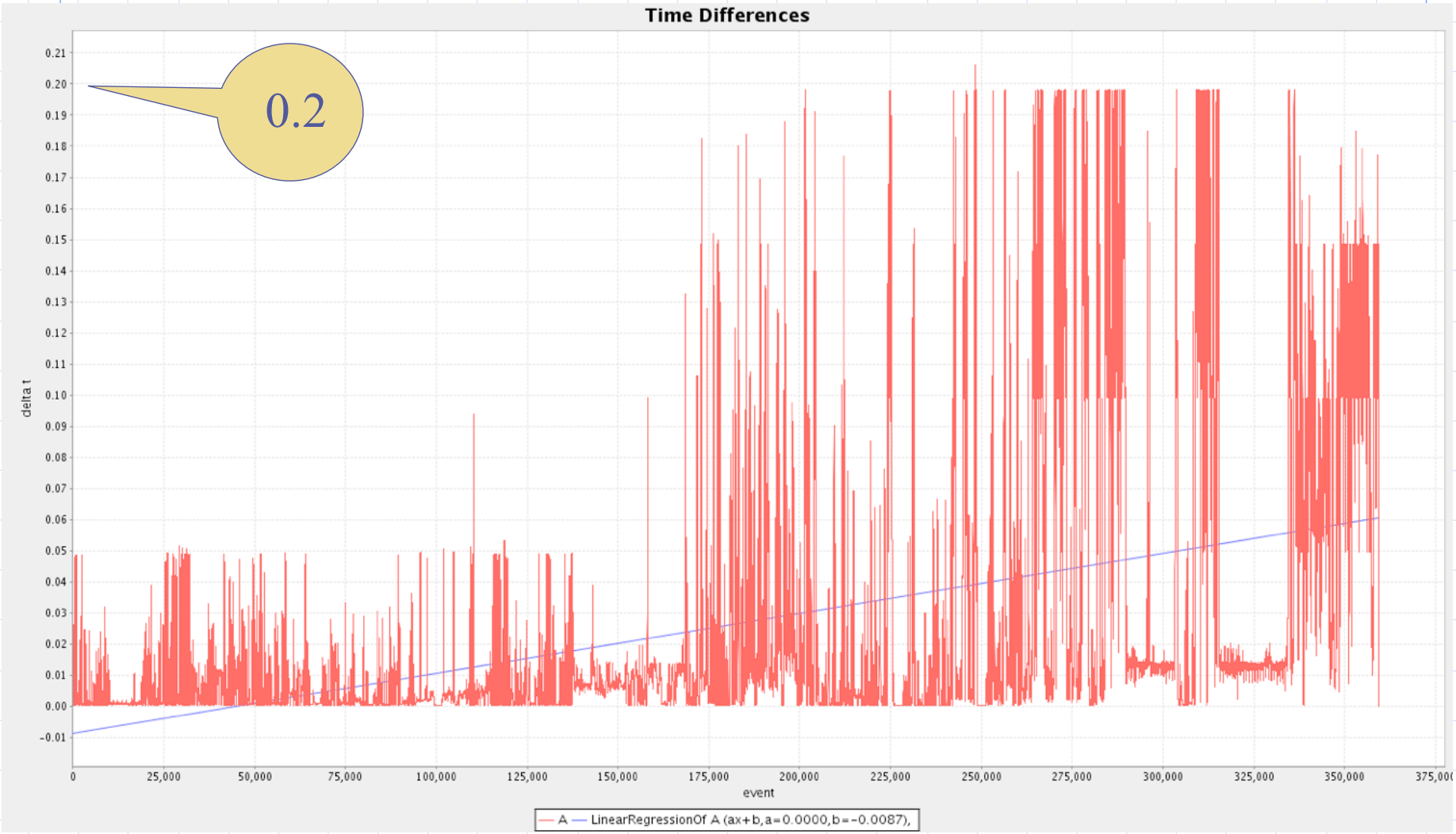
Run-Sequence Plot

- Scalability is an issue: 360000 interarrival times



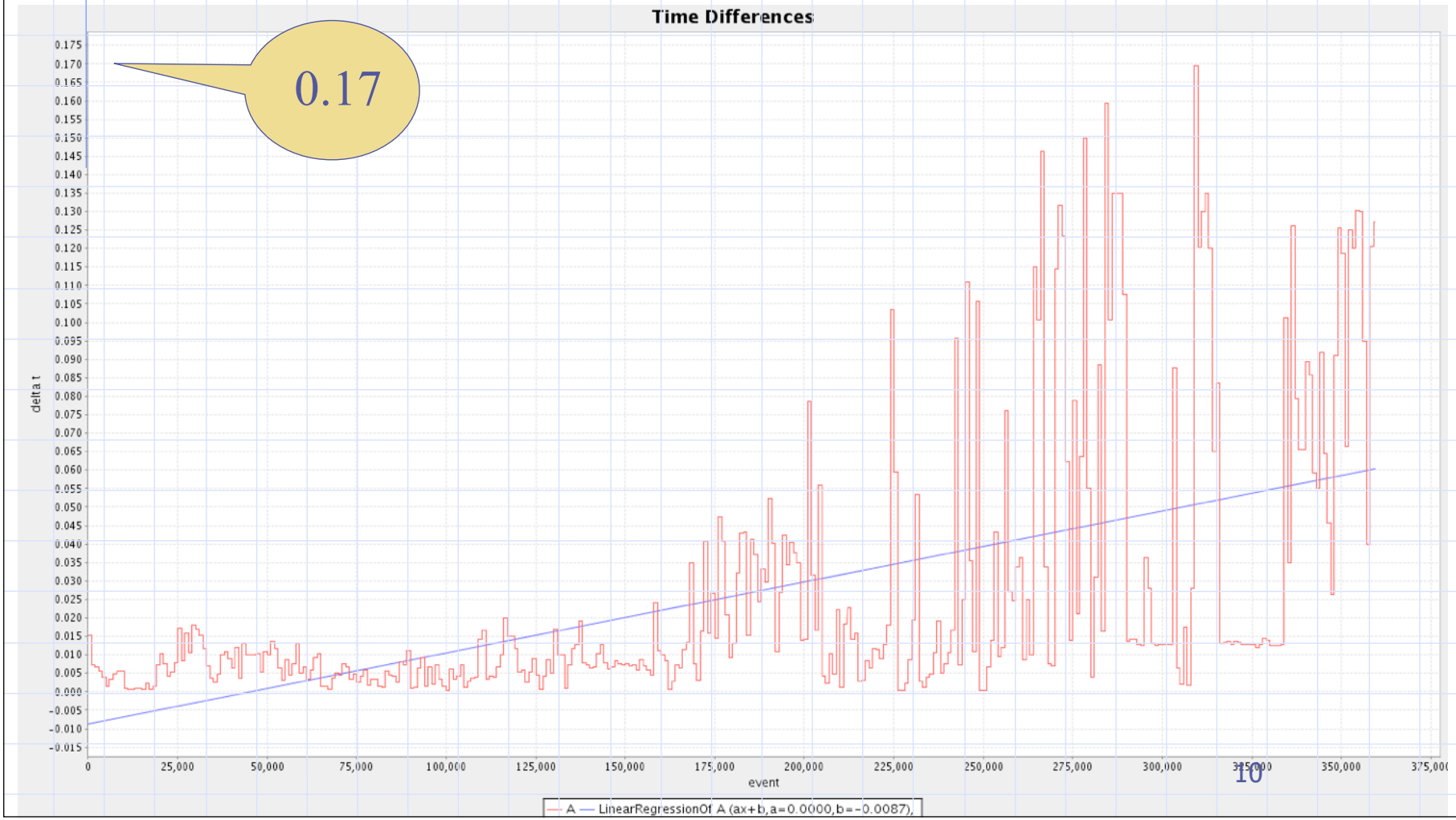
Run-Sequence Plot

- Smoothing emphasizes location, trend
- Moving averages $y_i = (x_{i-50} + x_{i-49} + \ldots + x_{i+50})/101$

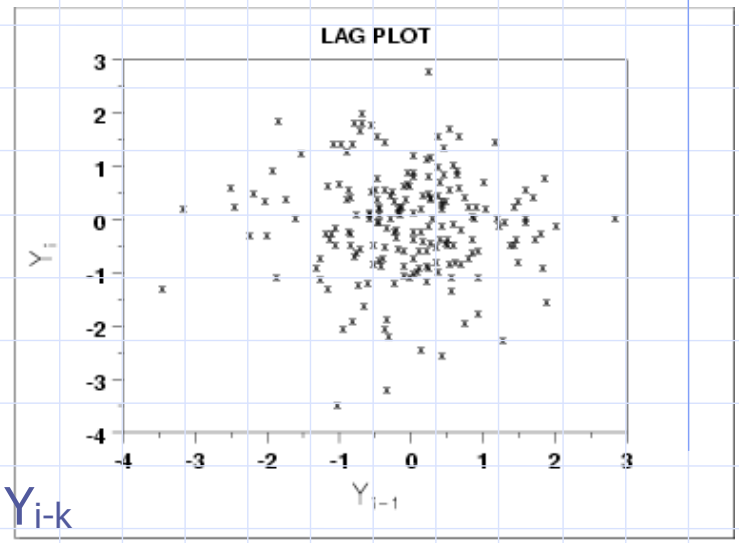


Run-Sequence Plot

- Smoothing emphasizes location, trend
- Batch means: $y_i = (x_i + x_{i+1} + \ldots + x_{i+999})/1000$

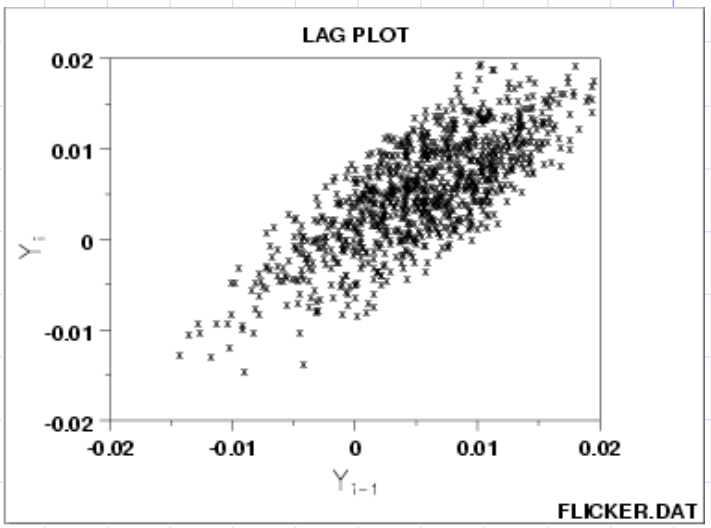


Lag Plot



- Purpose:
  - check for randomness in a time series
  - more precisely: correlation
- Definition:
  - Plot of lag k is a plot of values Y_i versus Y_{i-k}
  - Most commonly considered k=1
- Example: Random Data
  - Observations:
    - random
    - no autocorrelation of lag 1
    - no outliers
  - based on absence of structure
  - for given Y_{i-1} on cannot infer position of Y_i
  - such non-association is an indication of randomness

Lag 1 Plot Example for Moderate Autocorrelation



Conclusions from plot

- Data are from an underlying autoregressive model with moderate positive autocorrelation of lag 1
- Data contain no outliers

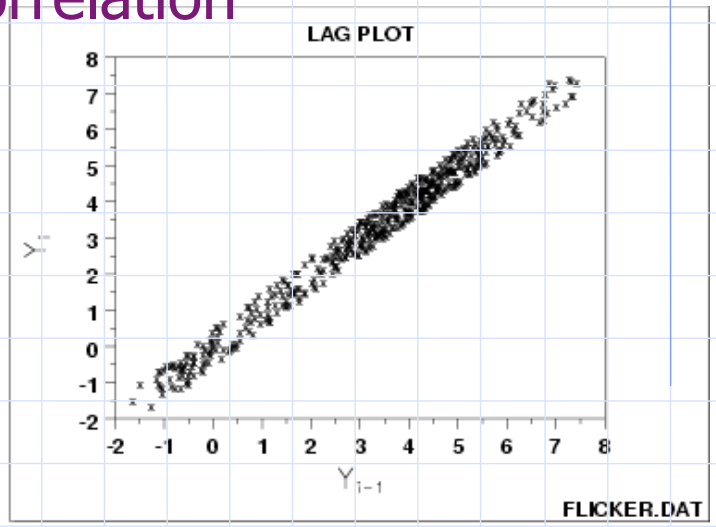
Discussion

- Points tend to cluster along diagonal -> positive autocorrelation
- Noise -> moderate correlation
- Range of Y_i somewhat predictable if Y_{i-1} is known.

Suggestion in NIST (since it is for time series analysis)

- Go for an autoregressive model with $Y_i = A_0 + A_1 * Y_{i-1} + E_i$
- which is a linear regression from the lag plot since Y_i, Y_{i-1} are the axes

Lag 1 Plot Example for Strong Autocorrelation



Conclusions from plot

- Data are from an underlying autoregressive model with strong positive autocorrelation of lag 1
- Data contain no outliers

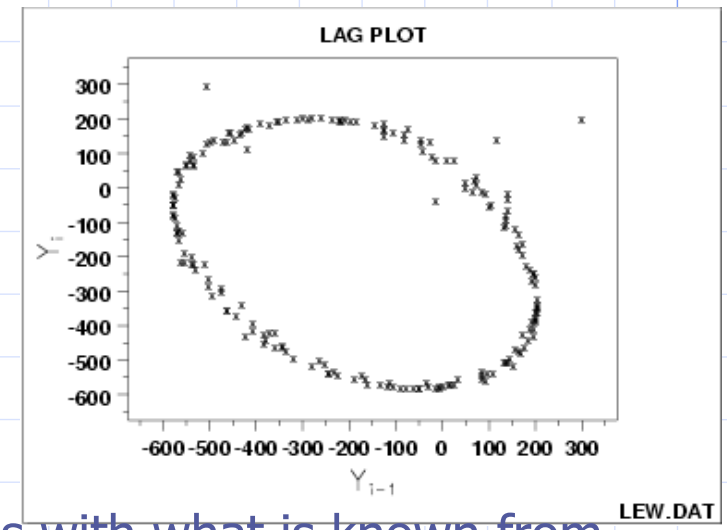
Discussion

- Points tightly clustered along diagonal -> positive autocorrelation
- Little noise -> strong correlation
- Range of Y_i well predictable if Y_{i-1} is known.

Suggestion in NIST (since it is for time series analysis)

- Go for an autoregressive model with $Y_i = A_0 + A_1 * Y_{i-1} + E_i$
- which is a linear regression from the lag plot since Y_i, Y_{i-1} are the axes
- Reexamine system for an explanation, due to phenomenon under study, drifting in environment, contamination from data acquisition system ... ????

Lag 1 Plot Example for Sinusoidal Model and Outliers



Conclusions from plot

- Data come from an underlying single-cycle sinusoidal model
- Data contain three outliers

Discussion

- Tight elliptical clustering of points, matches with what is known from sinusoidal models
- Aside: shows several outliers that need consideration

Suggestion in NIST (since it is for time series analysis)

- Go for a model with $Y_i = C + \alpha \sin(2\pi\omega t_i + \phi) + E_i$
- with amplitude α, frequency ω, phase ϕ
- can be fitted with standard non-linear least squares to estimate coefficients

Covariance and Correlation

X stochastic variable

$f_X(x)$ density function

$$F_X(x) = P(X \le x) = \int_{-\infty}^{x} f_X(y)dy \quad \text{distribution function}$$

$$E(X) = \int_{-\infty}^{\infty} y f_X(y)dy \quad \text{expected value}$$

$$E(cX) = cE(X)$$

$$E(X + Y) = E(X) + E(Y)$$

$$\text{independent} : P(X = x, Y = y) = P(X = x)\,P(Y = y), E(XY) = E(X)E(Y)$$

$$\sigma_X^2 = \text{var}(X) = E((X - E(X))^2)$$

$$\text{var}(aX + b) = a^2\,\text{var}(X)$$

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2\,\text{cov}(X, Y)$$

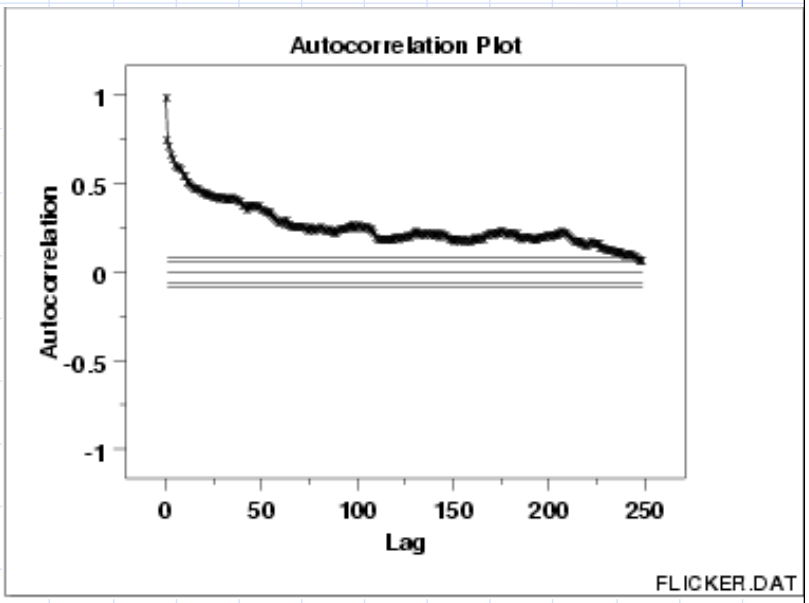
$$\text{covariance} : \text{cov}(X, Y) = E((X - E(X))(Y - E(Y)))$$

$$\text{correlation} : \frac{\text{cov}(X, Y)}{\sqrt{\sigma_X^2 \sigma_Y^2}}$$

$$\text{independent} : \text{cov}(X, Y) = 0$$

Autocovariance and Autocorrelation:
for series $X_1,\ldots,X_n$, consider covariance cov(X,Y) with $X=X_i$, $Y=X_{i+k}$
k: lag k

Autocorrelation Plot



Purpose: check randomness

- if random autocorrelations should be near zero for any and all time-lag separations.
- If non-random, then one or more of the autocorrelations will be significantly non-zero.

Observation

- rather high degree of correlations, hence not random
- horizontal lines around zero indicate thresholds for noise

Definition: r(h) vs h

- vertical axis: autocorrelation coefficient
- horizontal axis: time lag h (h = 1,2,3, ...)
- Note: range [-1,+1]

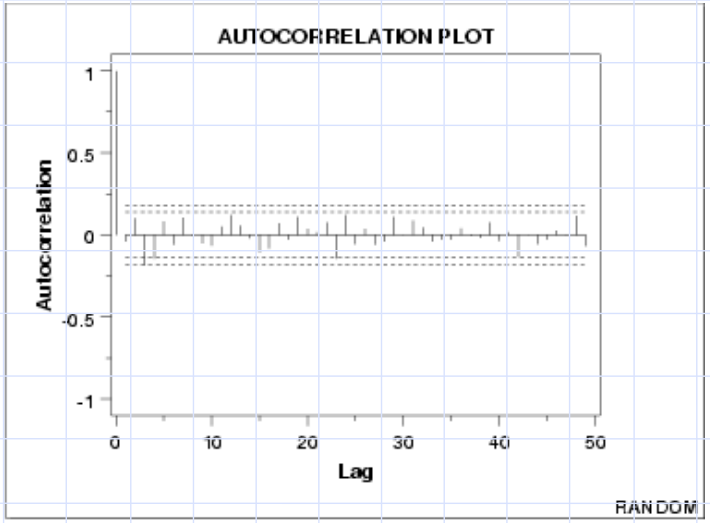
$$R_h = C_h / C_0$$

$$C_h = \frac{1}{N} \sum_{t=1}^{N-h} (Y_t - \bar{Y})(Y_{t+h} - \bar{Y})$$

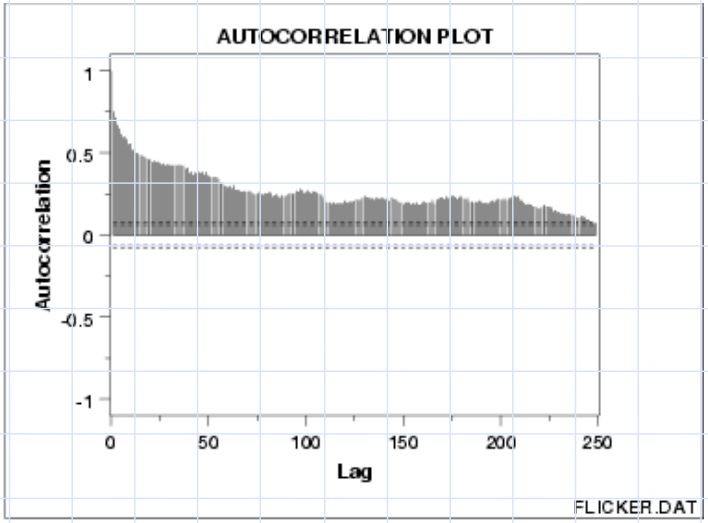
$$C_0 = \frac{\sum_{t=1}^{N} (Y_t - \bar{Y})^2}{N}$$

Autocorrelation plots

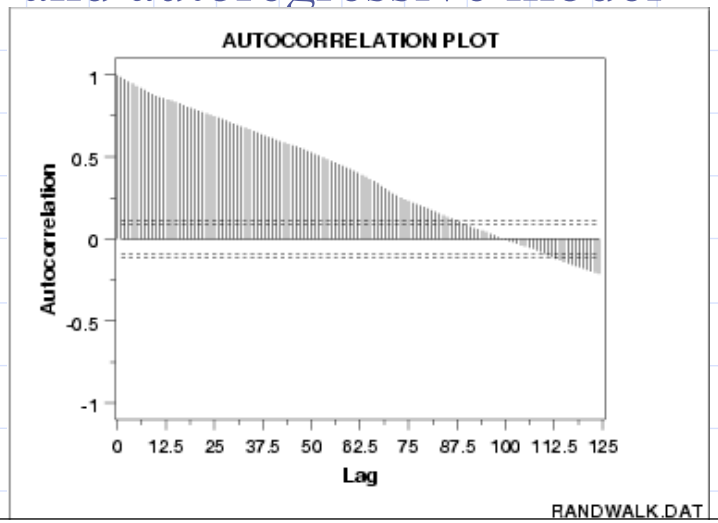
Random data



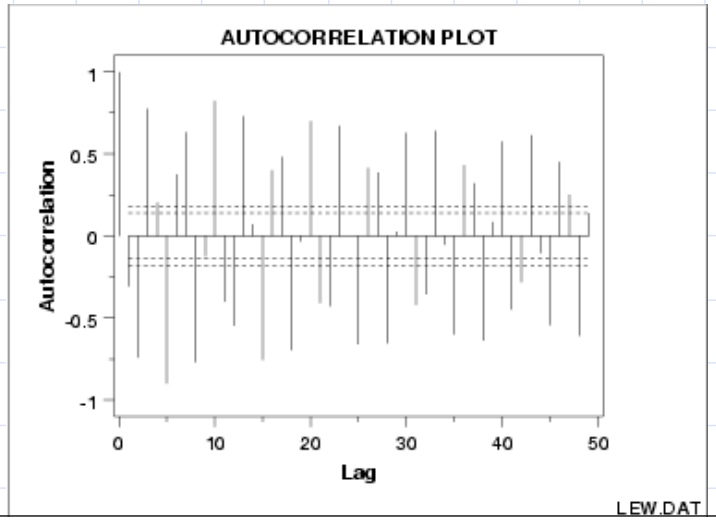
Moderate positive autocorrelation



Strong autocorrelation and autoregressive model

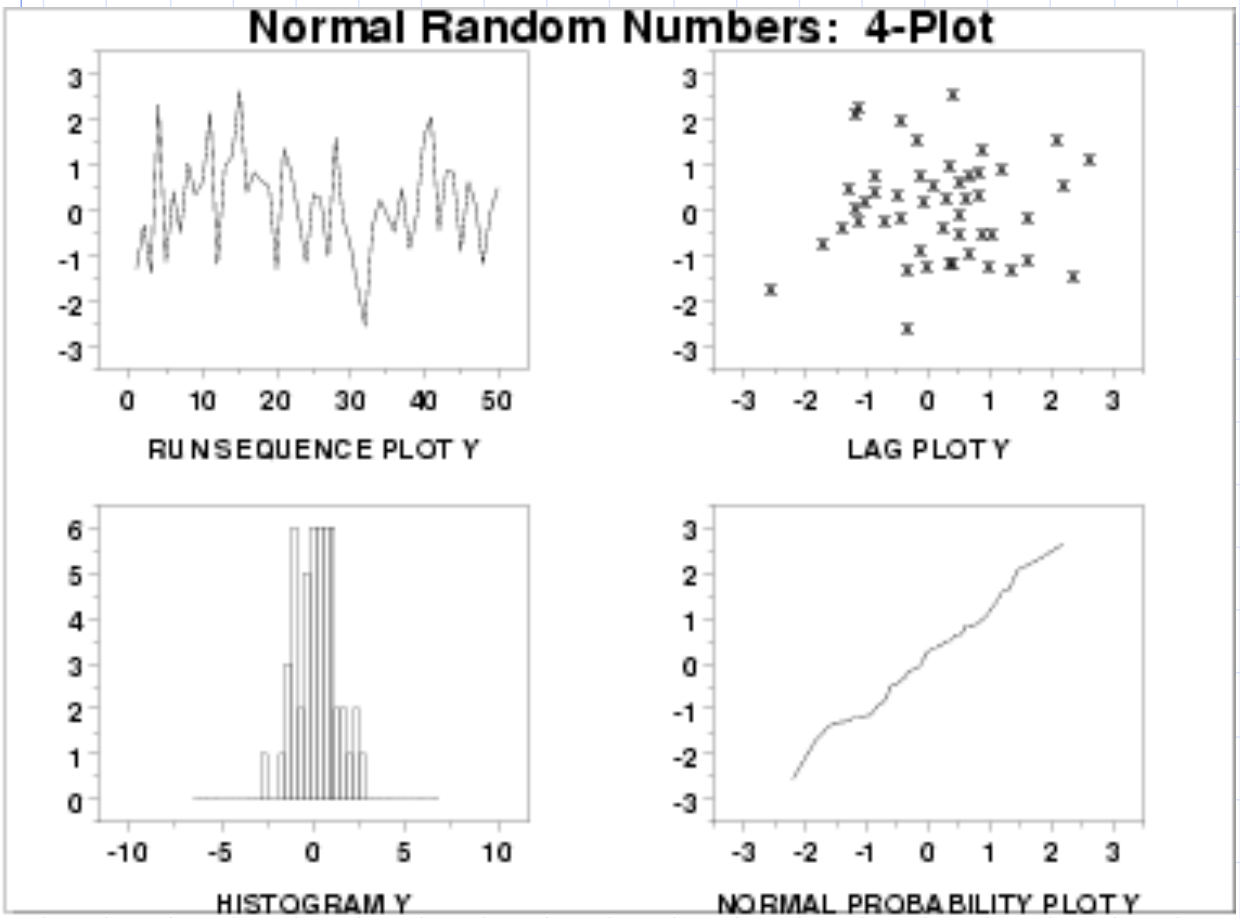


Sinusoidal model



Recall: Four techniques for testing assumptions

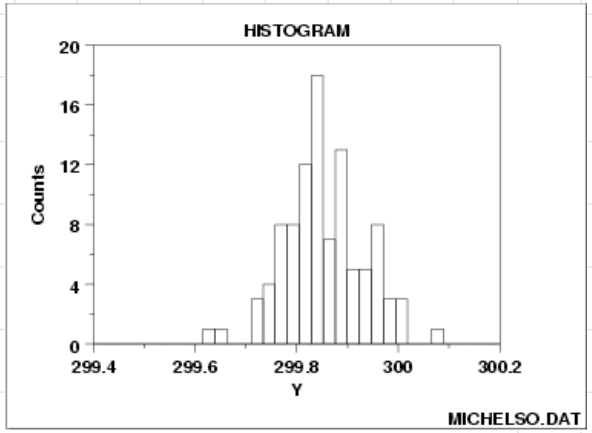
1. run sequence plot (Y_i versus i)
2. lag plot (Y_i versus Y_{i-1})
3. histogram (counts versus subgroups of Y)
4. normal probability plot (ordered Y versus theoretical ordered Y)

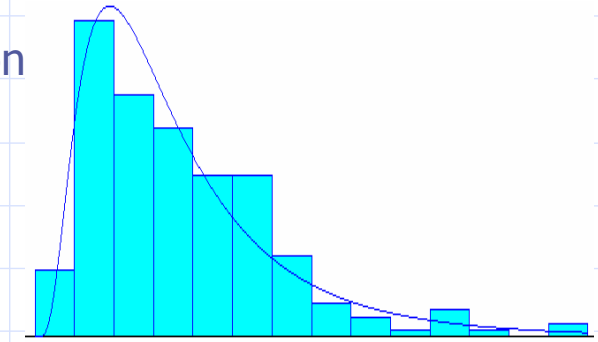


Example:
Process with
fixed location
fixed variation
random
distribution
approx. normal
no outliers

Histogram

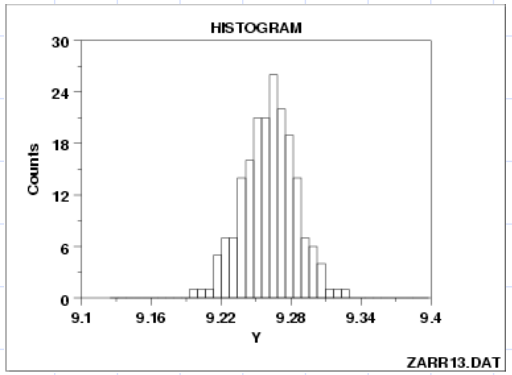
- Purpose: summarize univariate data set
  - Shows center (location), spread (scale), skewness, presence of outliers, presence of multiple modes in data
- Definition:
  - Most common form: split range of data into equal-sized bins (classes), count number of points in each bin.
    - Vertical axis: frequency
    - Horizontal axis: response variable
  - also:
    - cumulative histograms,
    - relative histograms
      - normalized by number of points
      - normalized s.th. total area is 1
      - useful if plotted with cont distribution





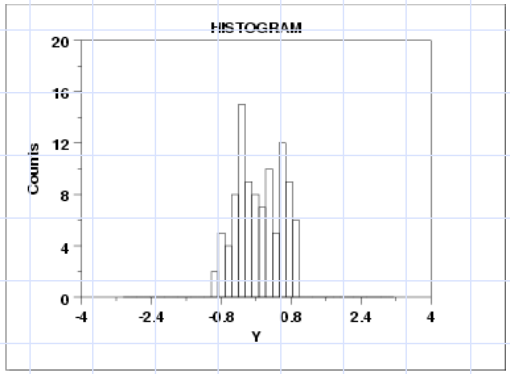
Histogram examples

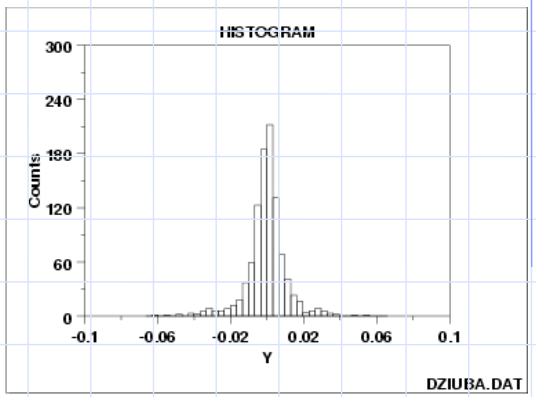
Normal



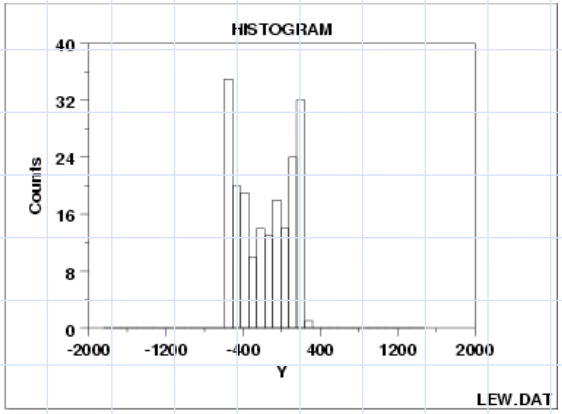
Symmetric, Non-Normal

Short-tailed vs Long-tailed

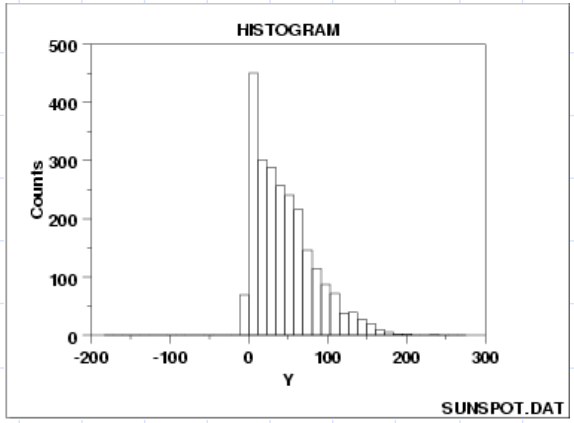




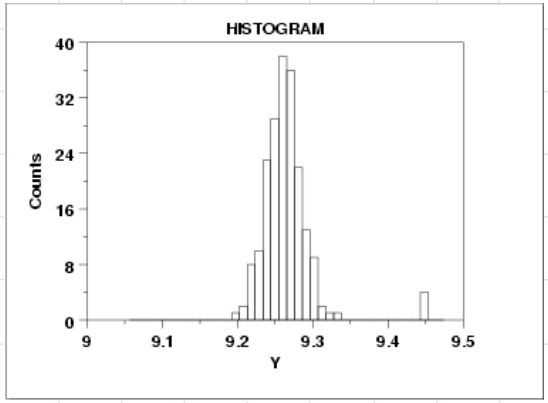
Symmetric, bimodal



Skewed Right

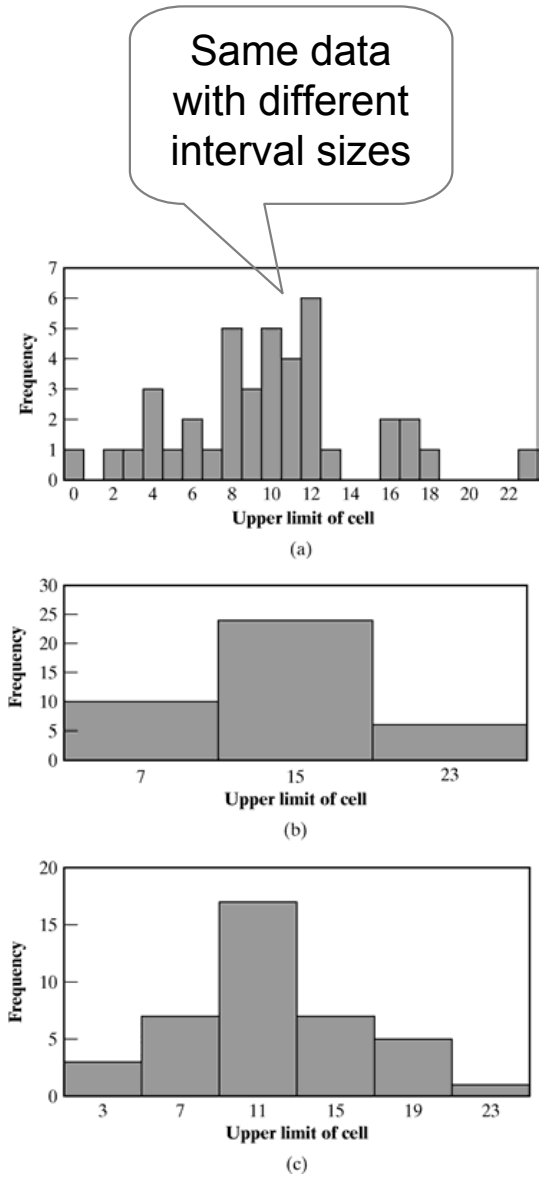


Symmetric with outlier



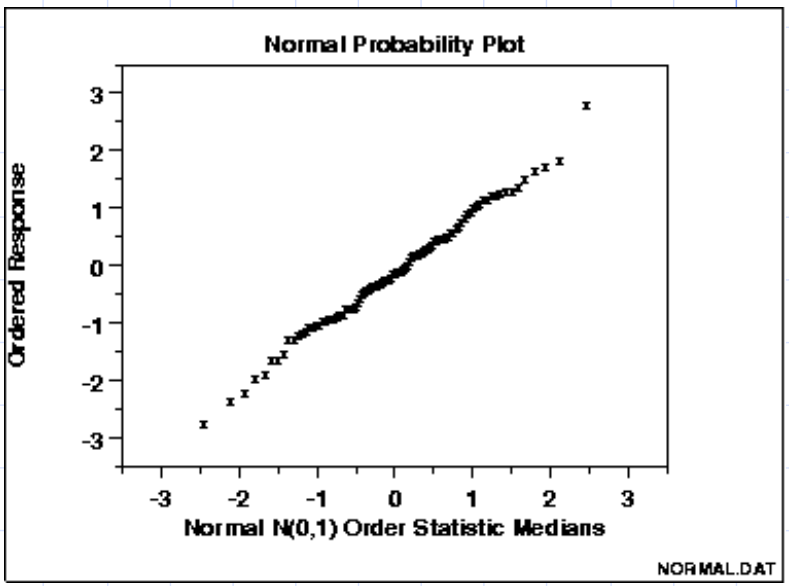
Histogram: Sensitive to selection of Bins

- Number of class intervals depends on:
  - The number of observations
  - The dispersion of the data
- If few data points are available:
  - combine adjacent cells to eliminate the ragged appearance of the histogram
- Note: Visual impression varies a lot with the number of bins selected.
- Recommendation:
  - Square root of the sample size
  - Try a range of values
  - (Law/Kelton) Pick value such that
    - irregularities are smoothed out,
    - curve characteristics are pronounced



Normal Probability Plot

- Purpose: check if data is approximately Normally distributed
- Special case of Probability Plot
  - needs Percent Point Function of particular distribution (here: normal)
  - needs uniform order statistics medians (known)
- Definition
  - Ordered Response values vs normal order statistic medians N(i)=G(U(i)) where U(i) are uniform order statistic medians, G is percent point function of the normal distribution.
- Observation:
  - linear relationship -> match!



Definition of U(i)

$U(i) = 1 - U(n)$ for $i = 1$

$U(i) = (i - 0.3175)/(n + 0.365)$ for $i = 2, 3, ..., n-1$

$U(i) = 0.5^{(1/n)}$ for $i = n$

Note:

1) intercept and slope estimates of the fitted line are also estimates for the location and scale parameters of the distribution.

2) correlation coefficient can be computed and used to test if this is a match.

3) Percent point function is inverse of cdf.

Interpretation of 4-Plot

- Fixed Location:
  If the fixed location assumption holds, then the run sequence plot will be flat and non-drifting.
- Fixed Variation:
  If the fixed variation assumption holds, then the vertical spread in the run sequence plot will be the approximately the same over the entire horizontal axis.
- Randomness:
  If the randomness assumption holds, then the lag plot will be structureless and random.
- Fixed Distribution:
  If the fixed distribution assumption holds, in particular if the fixed normal distribution holds, then
  - the histogram will be bell-shaped, and
  - the normal probability plot will be linear.

Overview of fitting with data

- Check if key assumptions hold (i.i.d)
- Select one or more candidate distributions
  - based on physical characteristics of the process and
  - graphical examination of the data.
- Fit the distribution to the data
  - determine values for its unknown parameters.
- Check the fit to the data
  - via statistical tests and
  - via graphical analysis.
- If the distribution does not fit,
  - select another candidate and repeat the process, or
  - use an empirical distribution.

Parameter estimates

- Common methods for parameter estimation are
  - maximum likelihood,
  - method of moments, and
  - least squares.
- While the method matters, the variability in the data often overwhelms the differences in the estimators.
- Decide what parameter estimates to use with goodness-of-fit tests and graphical comparisons.
- Remember:
  - There is no "true distribution" just waiting to be found!

Method of Moments

- The method of moments equates sample moments to parameter estimates.
- When moment methods are available, they have the advantage of simplicity.
- The disadvantage is that they are often not available and they do not have the desirable optimality properties of maximum likelihood and least squares estimators.
- The primary use of moment estimates is as starting values for the more precise maximum likelihood and least squares estimates.
- Example: Normal distribution $N(\mu,\sigma^2)$
  - use mean estimate from sample data for μ
  - use var estimate from sample data for σ^2

Maximum Likelihood

- Idea: determine the parameters that maximize the probability (likelihood) of the sample data
- Consider a joint density function

$$f(x_1, x_2, \ldots, x_n \mid \theta) = f(x_1|\theta) \cdot f(x_2|\theta) \cdots f(x_n|\theta).$$

  - $x_1,\ldots,x_n$ are samples, θ is the parameter of f whose best fitting value we want to determine
  - We want to optimize the likelihood function

$$\mathcal{L}(\theta \mid x_1, \ldots, x_n) = f(x_1, x_2, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

  - resp. the log-likelihood function

$$\ln \mathcal{L}(\theta \mid x_1, \ldots, x_n) = \sum_{i=1}^{n} \ln f(x_i|\theta), \qquad \hat{\ell} = \frac{1}{n} \ln \mathcal{L}.$$

  - for the maximum likelihood estimator (MLE)

$$\hat{\theta}_{\text{mle}} = \arg\max_{\theta \in \Theta} \hat{\ell}(\theta \mid x_1, \ldots, x_n).$$

  - where the sample values are fixed, θ can be freely chosen

Maximum Likelihood

- For many theoretical distributions, the MLE is known
  - see Law/Kelton, Chapter 6, tables of distribution characteristics
  - if not known in general, MLE can be determined for a given sample set numerically as a solution of the optimization problem
- MLE has attractive asymptotic properties
  - consistency, asymptotic normality, efficiency
  - asymptotic: for large sample sizes
- But also:
  - Maximum likelihood estimates can be heavily biased for small samples. The optimality properties may not apply for small samples.
- Example: Normal distribution $N(\mu,\sigma^2)$
  - MLE estimate for μ: same as mean estimate from sample data
  - MLE estimate for σ^2 : $S^2(n)(n-1)/n$ for var estimate $S^2(n)$

Least Squares Estimate

- Idea: Minimize the sum of the squared residuals where residuals measure the difference between the value of a function and the given data.
- Does not enjoy the same favorable properties as MLE, thus Law/Kelton focuses on MLE only.

Overview of fitting with data

- Check if key assumptions hold (i.i.d)
- Select one or more candidate distributions
  - based on physical characteristics of the process and
  - graphical examination of the data.
- Fit the distribution to the data
  - determine values for its unknown parameters.
- Check the fit to the data
  - via statistical tests and
  - via graphical analysis.
- If the distribution does not fit,
  - select another candidate and repeat the process, or
  - use an empirical distribution.