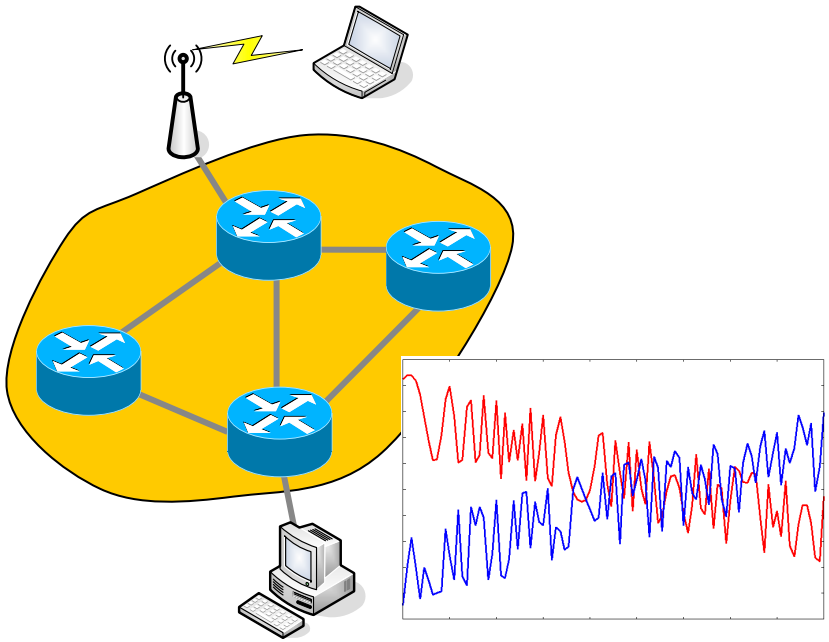
Chapter 9

Input Modeling



Contents

- Data Collection
- Identifying the Distribution with Data
- Parameter Estimation
- Goodness-of-Fit Tests
- Fitting a Nonstationary Poisson Process
- Selecting Input Models without Data
- Multivariate and Time-Series Input Data

Purpose & Overview

- Input models provide the driving force for a simulation model.
- The quality of the output is no better than the quality of inputs.
- In this chapter, we will discuss the 4 steps of input model development:

  (1) Collect data from the real system

  (2) Identify a probability distribution to represent the input process

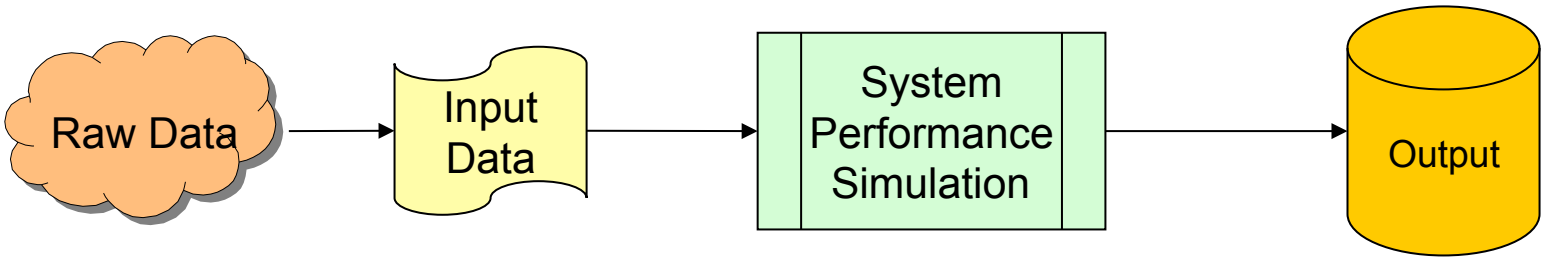
  (3) Choose parameters for the distribution

  (4) Evaluate the chosen distribution and parameters for goodness of fit

Data Collection

Data Collection

- One of the biggest tasks in solving a real problem
  - GIGO: Garbage-In-Garbage-Out



- Even when model structure is valid simulation results can be misleading, if the input data is
  - inaccurately collected
  - inappropriately analyzed
  - not representative of the environment

Data Collection

- Suggestions that may enhance and facilitate data collection:
  - **Plan ahead:** begin by a practice or pre-observing session, watch for unusual circumstances
  - **Analyze the data as it is being collected:** check adequacy
  - **Combine homogeneous data sets:** successive time periods, during the same time period on successive days
  - **Be aware of data censoring:** the quantity is not observed in its entirety, danger of leaving out long process times
  - **Check for relationship between variables** (scatter diagram)
  - **Check for autocorrelation**
  - **Collect input data, not performance data**

Identifying the Distribution

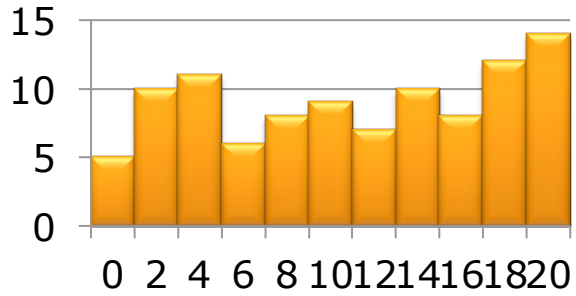
Histograms

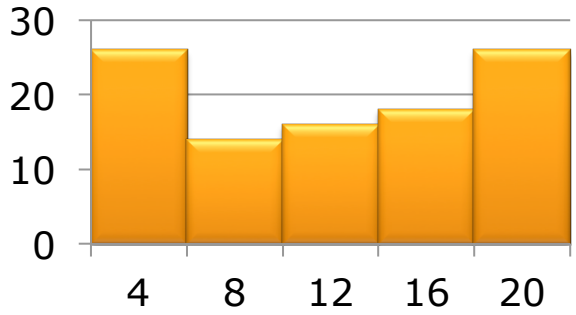
Histograms

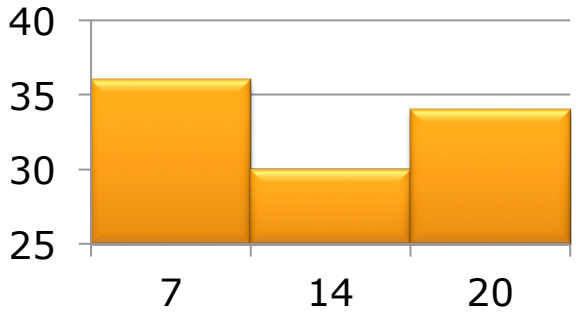
- A frequency distribution or histogram is useful in determining the shape of a distribution
- The number of class intervals depends on:
  - The number of observations
  - The dispersion of the data
  - Suggested number of intervals: the square root of the sample size
- For continuous data:
  - Corresponds to the probability density function (pdf) of a theoretical distribution
- For discrete data:
  - Corresponds to the probability mass function (pmf)
- If few data points are available
  - combine adjacent cells to eliminate the ragged appearance of the histogram

Histograms

- Same data with different interval sizes



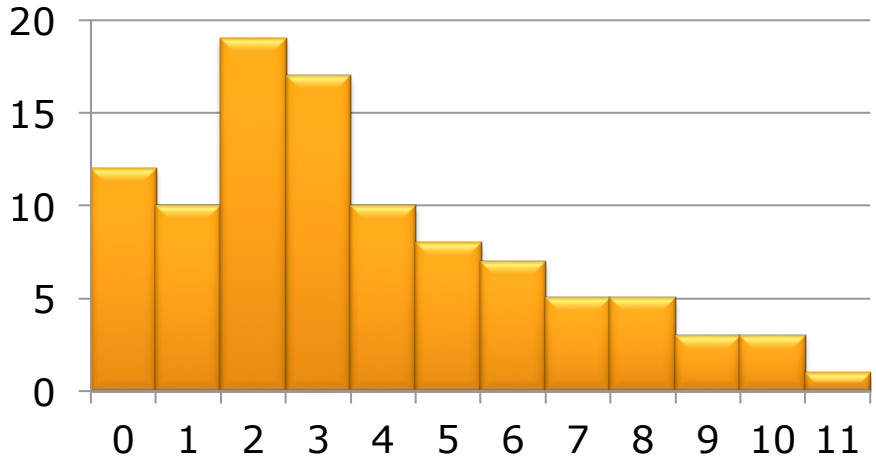




Histograms: Example

- Vehicle Arrival Example: Number of vehicles arriving at an intersection between 7 am and 7:05 am was monitored for 100 random workdays.
- There are ample data, so the histogram may have a cell for each possible value in the data range

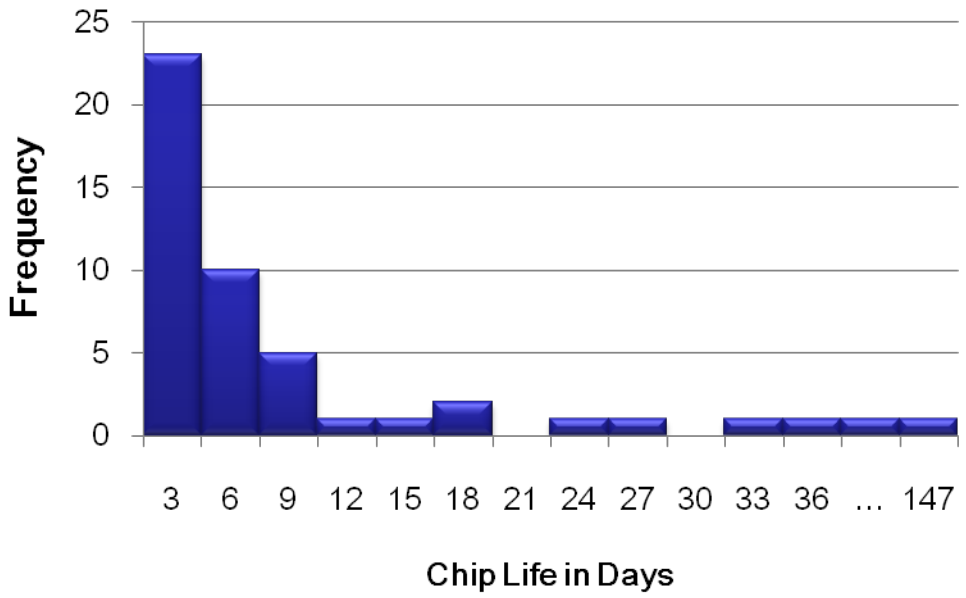
Arrivals per Period	Frequency
0	12
1	10
2	19
3	17
4	10
5	8
6	7
7	5
8	5
9	3
10	3
11	1



Histograms: Example

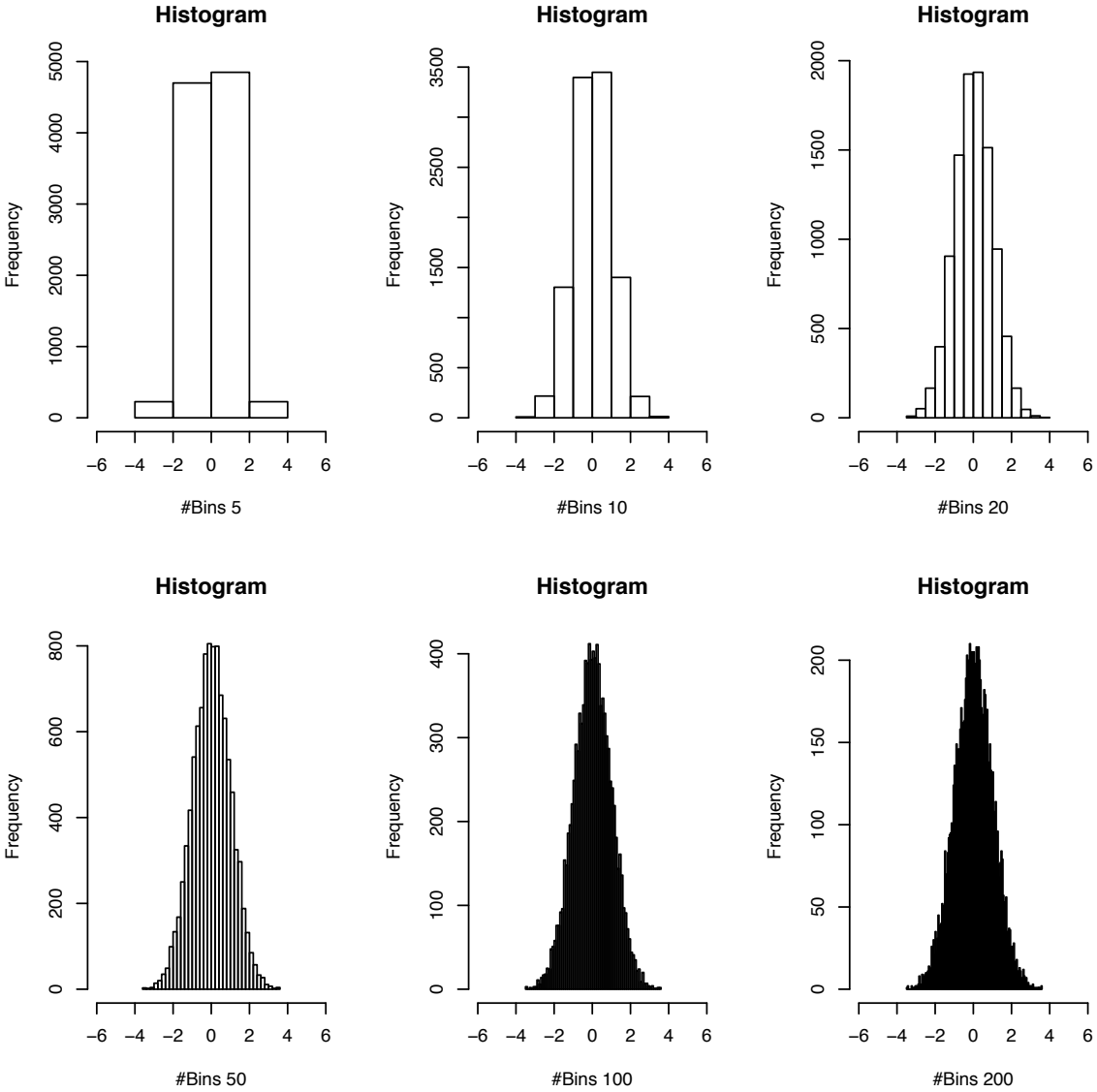
- Life tests were performed on electronic components at 1.5 times the nominal voltage, and their lifetime was recorded

Component Life	Frequency
0 ≤ x < 3	23
3 ≤ x < 6	10
6 ≤ x < 9	5
9 ≤ x < 12	1
12 ≤ x < 15	1
...	
42 ≤ x < 45	1
...	
144 ≤ x < 147	1



Histograms: Example

- Sample size 10000
- Histograms with different numbers of bins



Histograms: Example

Stanford University Mobile Activity Traces (SUMATRA)

- Target community: cellular network research community
- Traces contain mobility as well as connection information

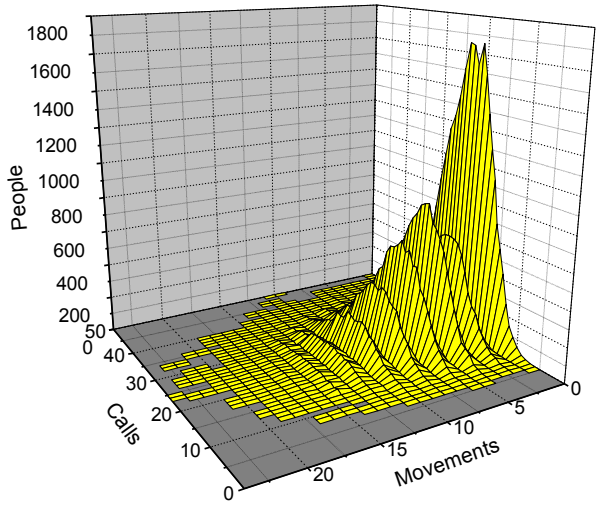
- Available traces
  - SULAWESI (S.U. Local Area Wireless Environment Signaling Information)
  - BALI (Bay Area Location Information)

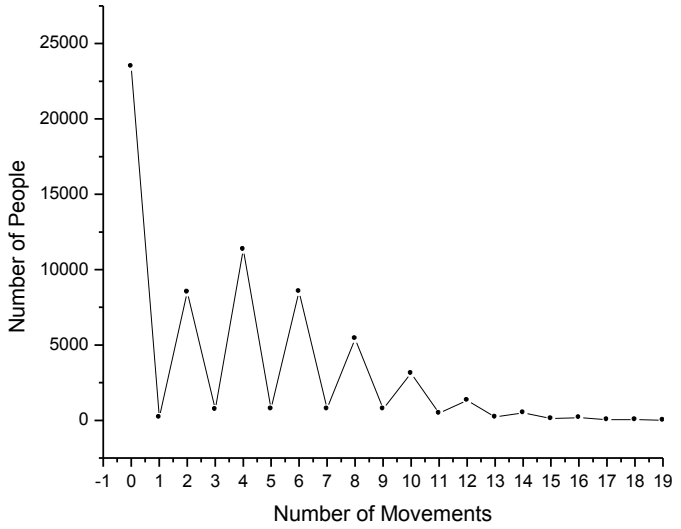
- BALI Characteristics
  - San Francisco Bay Area
  - Trace length: 24 hour
  - Number of cells: 90
  - Persons per cell: 1100
  - Persons at all: 99.000
  - Active persons: 66.550
  - Move events: 243.951
  - Call events: 1.570.807

- Question: How to transform the BALI information so that it is usable with a network simulator, e.g., ns-2?
  - Node number as well as connection number is too high for ns-2

Histograms: Example

- Analysis of the BALI Trace
  - Goal: Reduce the amount of data by identifying user groups
- User group
  - Between 2 local minima
  - Communication characteristic is kept in the group
  - A user represents a group
- Groups with different mobility characteristics
  - Intra- and inter group communication
- Interesting characteristic
  - Number of people with odd number movements is negligible!



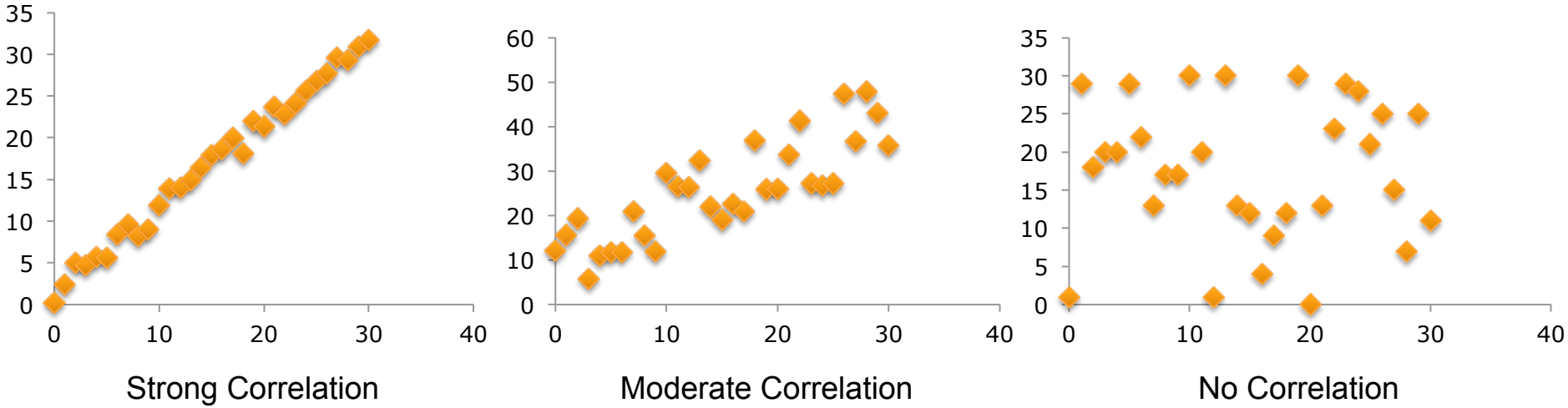


Identifying the Distribution

Scatter diagrams

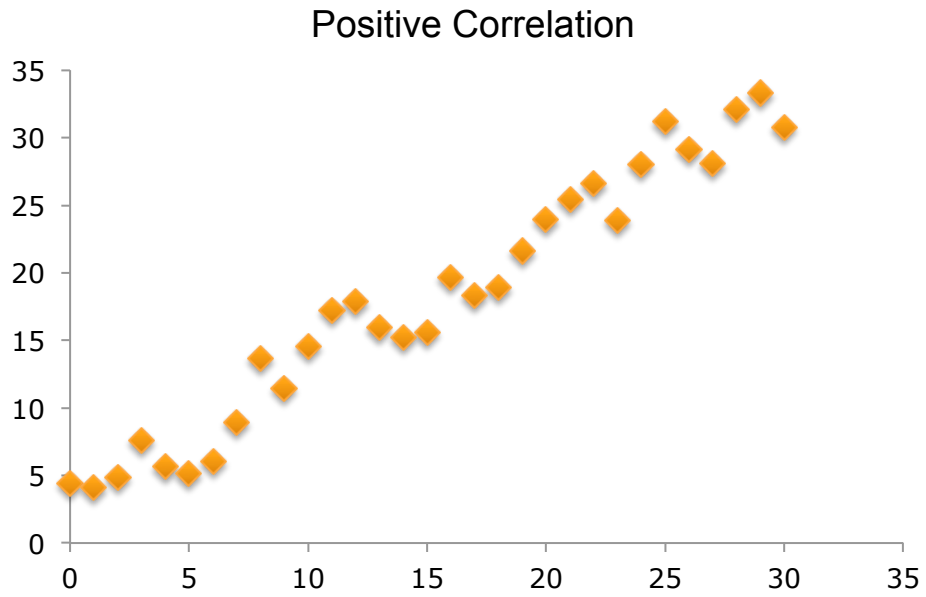
Scatter Diagrams

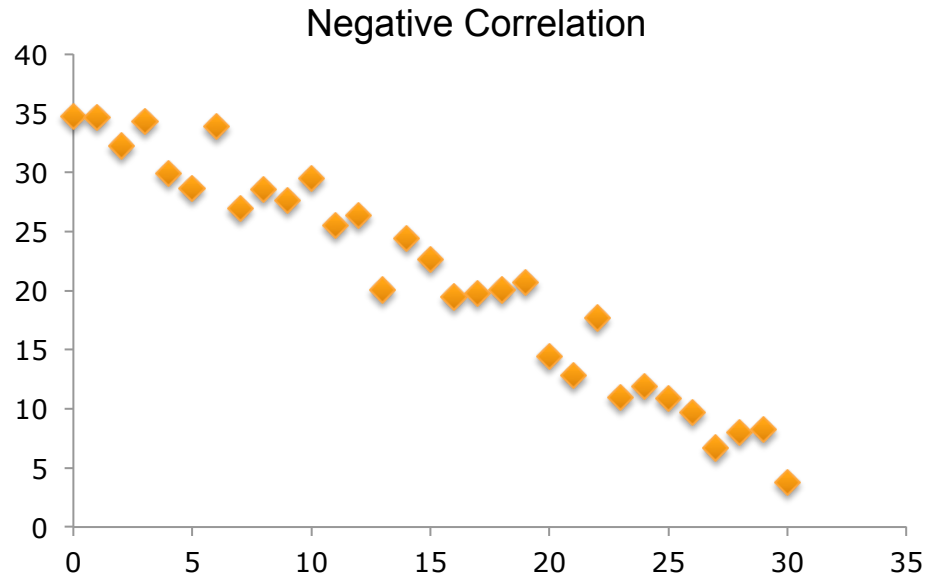
- A scatter diagram is a quality tool that can show the relationship between paired data
  - Random Variable X = Data 1
  - Random Variable Y = Data 2
  - Draw random variable X on the x-axis and Y on the y-axis



Scatter Diagrams

- Linear relationship
  - Correlation: Measures how well data line up
  - Slope: Measures the steepness of the data
  - Direction
  - Y intercept



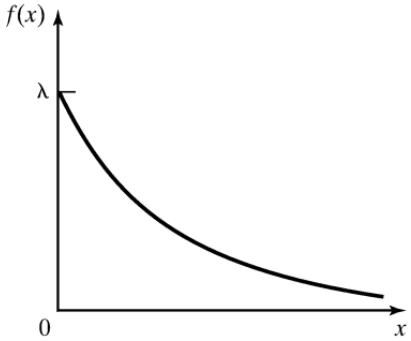


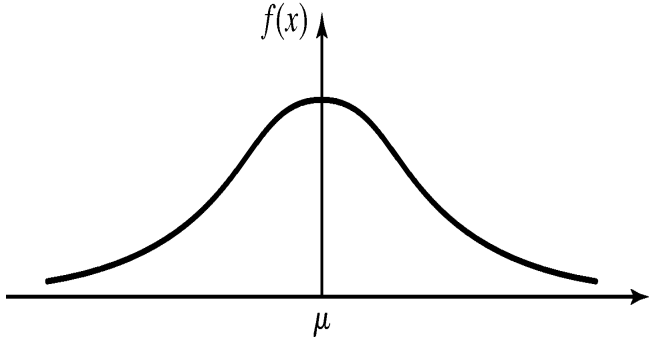
Identifying the Distribution

Selecting the Family of Distributions

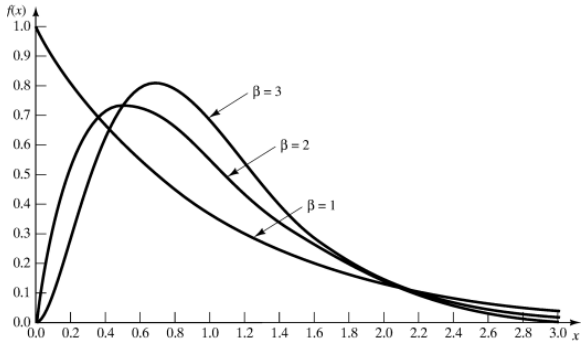
Selecting the Family of Distributions

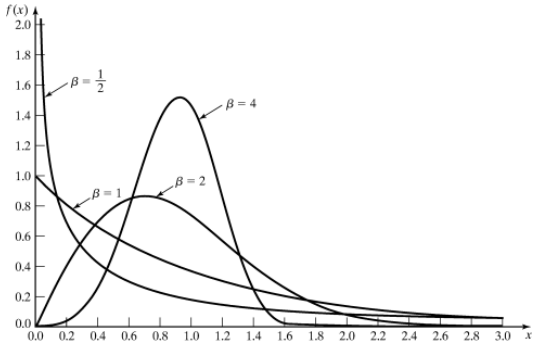
- A family of distributions is selected based on:
  - The context of the input variable
  - Shape of the histogram
- Frequently encountered distributions:
  - Easier to analyze: Exponential, Normal, and Poisson





  - Difficult to analyze: Beta, Gamma, and Weibull





Selecting the Family of Distributions

- Use the physical basis of the distribution as a guide, e.g.:
  - **Binomial**: Number of successes in n trials
  - **Negative binomial and geometric**: Number of trials to achieve k successes
  - **Poisson**: Number of independent events that occur in a fix amount of time or space
  - **Normal**: Distribution of a process that is the sum of a number of component processes
  - **Lognormal**: Distribution of a process that is the product of a number of component processes
  - **Exponential**: Time between independent events, or a process time that is memoryless
  - **Weibull**: Time to failure for components
  - **Discrete or continuous uniform**: Models complete uncertainty
  - **Triangular**: A process for which only the minimum, most likely, and maximum values are known
  - **Empirical**: Resamples from the actual data collected

Selecting the Family of Distributions

- Remember the physical characteristics of the process
  - Is the process naturally discrete or continuous valued?
  - Is it bound?
  - Value range?
    - Only positive values
    - Only negative values
    - Interval of [-a:b]
- No “true” distribution for any stochastic input process
- Goal: obtain a good approximation

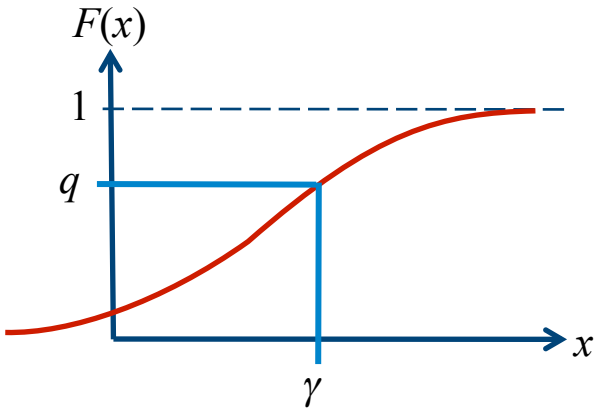
Identifying the Distribution

Quantile-Quantile Plots

Quantile-Quantile Plots

- *Q-Q* plot is a useful tool for evaluating distribution fit
- If X is a random variable with CDF F, then the q-quantile of X is the γ such that

$$F(\gamma) = P(X \le \gamma) = q \ , \quad \text{for } 0 < q < 1$$



  - When F has an inverse, $\gamma = F^{-1}(q)$

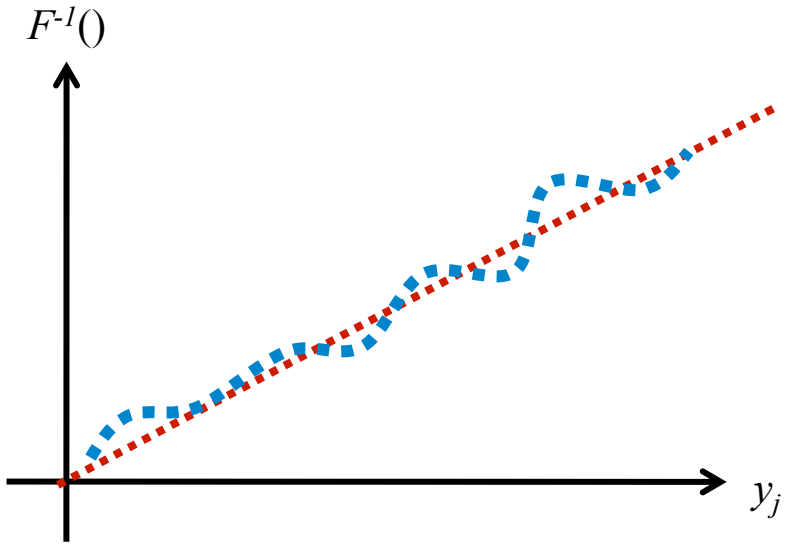
- Let $\{x_i, i = 1,2, \ldots, n\}$ be a sample of data from X and $\{y_j, j = 1,2, \ldots, n\}$ be this sample in ascending order:

$$y_j \text{ is approximately } F^{-1}\left(\frac{j-0.5}{n}\right)$$

  - where j is the ranking or order number

Quantile-Quantile Plots

- The plot of y_j versus $F^{-1}((j - 0.5) / n)$ is
  - Approximately a straight line if F is a member of an appropriate family of distributions
  - The line has slope 1 if F is a member of an appropriate family of distributions with appropriate parameter values



Quantile-Quantile Plots: Example

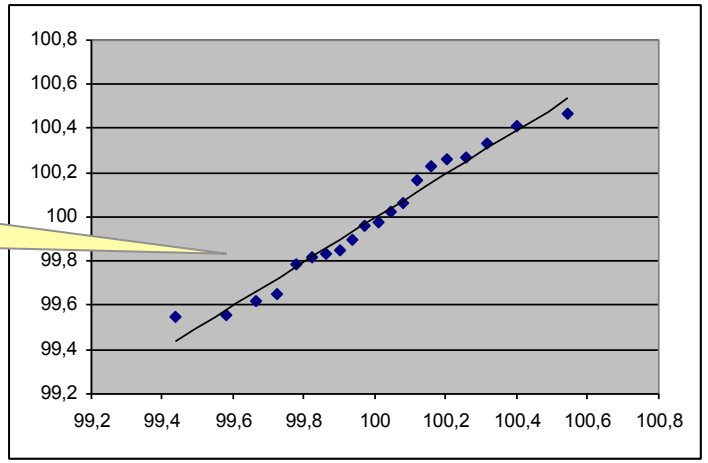
- Example: Door installation times of a robot follows a normal distribution.
  - The observations are ordered from the smallest to the largest
  - y_j are plotted versus $F^{-1}((j - 0.5)/n)$ where F has a normal distribution with the sample mean (99.99 sec) and sample variance (0.2832^2 sec^2)

j	Value
1	99,55
2	99,56
3	99,62
4	99,65
5	99,79
6	99,98
7	100,02
8	100,06
9	100,17
10	100,23
11	100,26
12	100,27
13	100,33
14	100,41
15	100,47

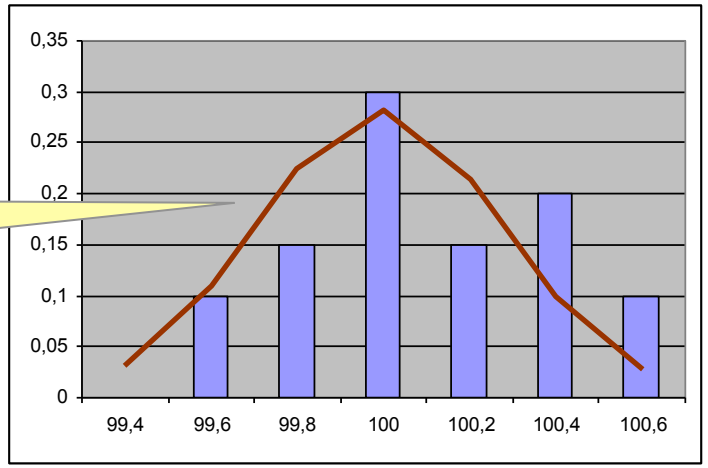
Quantile-Quantile Plots: Example

- Example (continued): Check whether the door installation times follow a normal distribution.

Straight line, supporting the hypothesis of a normal distribution

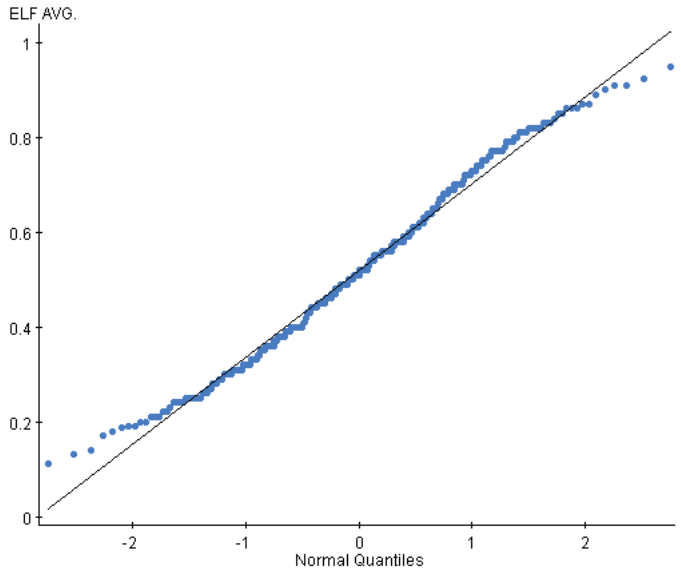


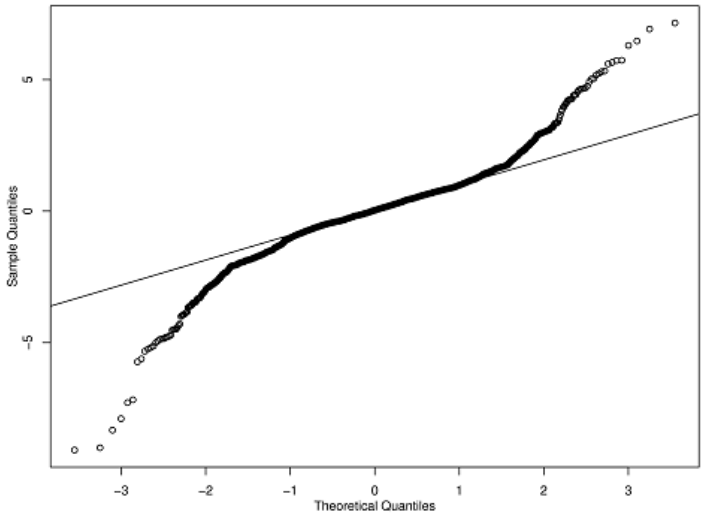
Superimposed density function of the normal distribution



Quantile-Quantile Plots

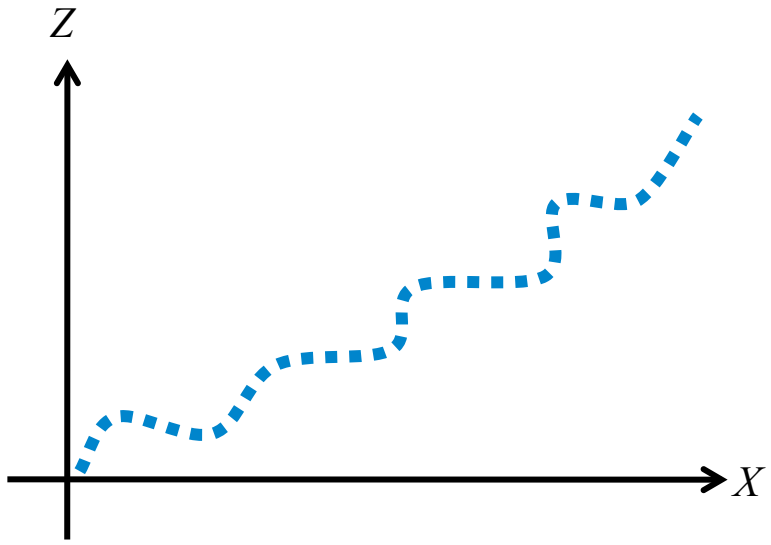
- Consider the following while evaluating the linearity of a Q-Q plot:
  - The observed values never fall exactly on a straight line
  - The ordered values are ranked and hence not independent, unlikely for the points to be scattered about the line
  - Variance of the extremes is higher than the middle. Linearity of the points in the middle of the plot is more important.





Quantile-Quantile Plots

- *Q-Q* plot can also be used to check homogeneity
  - It can be used to check whether a single distribution can represent two sample sets
  - Given two random variables
    - X and $x_1, x_2, \ldots, x_n$
    - Z and $z_1, z_2, \ldots, z_n$
  - Plotting the **ordered** values of X and Z against each other reveals approximately a straight line if X and Z are well represented by the same distribution



Parameter Estimation

Parameter Estimation

- Parameter Estimation: Next step after selecting a family of distributions
- If observations in a sample of size n are $X_1, X_2, \ldots, X_n$ (discrete or continuous), the sample mean and sample variance are:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad S^2 = \frac{\left(\sum_{i=1}^{n} X_i^2\right) - n\bar{X}^2}{n-1}$$

- If the data are discrete and have been grouped in a frequency distribution:

$$\bar{X} = \frac{\sum_{j=1}^{n} f_j X_j}{n} \qquad S^2 = \frac{\left(\sum_{j=1}^{n} f_j X_j^2\right) - n\bar{X}^2}{n-1}$$

  - where f_j is the observed frequency of value X_j

Parameter Estimation

- When raw data are unavailable (data are grouped into class intervals), the approximate sample mean and variance are:

$$\bar{X} = \frac{\sum_{j=1}^{c} f_j m_j}{n} \qquad S^2 = \frac{\left(\sum_{j=1}^{n} f_j m_j^2\right) - n\bar{X}^2}{n-1}$$

  - f_j is the observed frequency in the j-th class interval
  - m_j is the midpoint of the j-th interval
  - c is the number of class intervals

- A parameter is an unknown constant, but an estimator is a statistic.

Parameter Estimation: Example

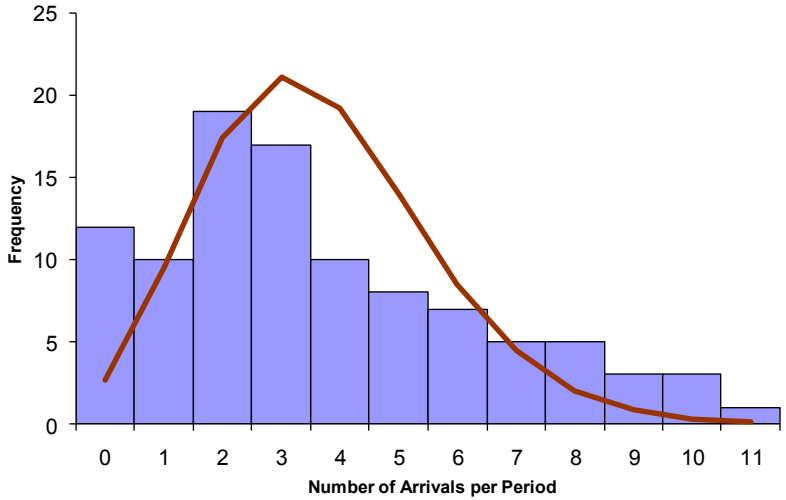
- Vehicle Arrival Example (continued): Table in the histogram of the example on Slide 10 can be analyzed to obtain:

$$n = 100, f_1 = 12, X_1 = 0, f_2 = 10, X_2 = 1, \ldots \quad \text{and} \quad \sum_{j=1}^{k} f_j X_j = 364, \text{ and } \sum_{j=1}^{k} f_j X_j^2 = 2080$$

  - The sample mean and variance are

$$\bar{X} = \frac{364}{100} = 3.64$$

$$S^2 = \frac{2080 - 100 \cdot (3.64)^2}{99} = 7.63$$



  - The histogram suggests X to have a Poisson distribution
    - However, note that sample mean is not equal to sample variance.
    - Theoretically: Poisson with parameter λ ➡ $\mu = \sigma^2 = \lambda$
    - Reason: each estimator is a random variable, it is not perfect.

Parameter Estimation

- Maximum-Likelihood Estimators (MLE)
  - Discrete distribution with one parameter θ ➡ $p_\theta(x)$
  - Given iid sample $X_1, X_2, \ldots, X_n$
  - Likelihood function $L(\theta)$ is defined as

$$L(\theta) = p_\theta(X_1)\, p_\theta(X_2) \ldots p_\theta(X_n)$$

  - MLE of the unknown θ is θ' given by θ that maximizes $L(\theta)$ ➡ $L(\theta') \geq L(\theta)$ for all values of θ

Parameter Estimation

- Maximum-Likelihood Estimators (MLE)
- Suggested estimators for distributions often used in simulation

Distribution	Parameter	Estimator
Poisson	α	$\hat{\alpha} = \overline{X}$
Exponential	λ	$\hat{\lambda} = \frac{1}{\overline{X}}$
Gamma	β, θ	$\hat{\theta} = \frac{1}{\overline{X}}$
Normal	μ, σ^2	$\hat{\mu} = \overline{X}, \hat{\sigma}^2 = S^2$
Lognormal	μ, σ^2	$\hat{\mu} = \overline{X}, \hat{\sigma}^2 = S^2$

After taking *ln* of data.

Parameter Estimation

- Maximum Likelihood example exponential distribution

Goodness-of-Fit Tests

Goodness-of-Fit Tests

- Conduct hypothesis testing on input data distribution using
  - Kolmogorov-Smirnov test
  - Chi-square test

- No single correct distribution in a real application exists
  - If very little data are available, it is unlikely to reject any candidate distributions
  - If a lot of data are available, it is likely to reject all candidate distributions

Goodness-of-Fit Tests

- Be aware of mistakes in decision finding
  - Type I Error: α
    - Error of first kind, False positive
    - Reject H_0 when it is true
  - Type II Error: β
    - Error of second kind, False negative
    - Retain H_0 when it is not true

Statistical Decision	State of the null hypothesis	
	H_0 True	H_0 False
Accept H_0	Correct	Type II Error Incorrectly accept H_0 False negative
Reject H_0	Type I Error Incorrectly reject H_0 False positive	Correct

Chi-Square Test

- Intuition: comparing the histogram of the data to the shape of the candidate density or mass function
- Valid for large sample sizes when parameters are estimated by maximum-likelihood
- Arrange the n observations into a set of k class intervals
- The test statistic is:

Observed frequency in the i-th class

Expected Frequency
$E_i = n \times p_i$
where p_i is the theoretical prob. of the i-th interval.
Suggested Minimum = 5

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

- χ_0^2 approximately follows the Chi-square distribution with k-s-1degrees of freedom
- s = number of parameters of the hypothesized distribution estimated by the sample statistics.

Chi-Square Test

- The hypothesis of a Chi-square test is
  - H_0: The random variable, X, conforms to the distributional assumption with the parameter(s) given by the estimate(s).
  - H_1: The random variable X does not conform.

$$\text{Test result} \begin{cases} \chi_0^2 \le \chi^2_{\alpha,k-s-1} & \text{Accept } H_0 \\ \chi_0^2 > \chi^2_{\alpha,k-s-1} & \text{Reject } H_0 \end{cases}$$

- If the distribution tested is discrete and combining adjacent cells is not required (so that E_i > minimum requirement):
  - Each value of the random variable should be a class interval, unless combining is necessary, and

$$p_i = p(x_i) = P(X = x_i)$$

Chi-Square Test

- If the distribution tested is continuous:

$$p_i = \int_{a_{i-1}}^{a_i} f(x)\,dx = F(a_i) - F(a_{i-1})$$

  - where a_{i-1} and a_i are the endpoints of the i-th class interval
  - $f(x)$ is the assumed PDF, $F(x)$ is the assumed CDF
  - Recommended number of class intervals (k):

Sample size (n)	Number of class intervals (k)
20	Do not use the chi-square test
50	5 to 10
100	10 to 20
> 100	$\sqrt{n}$ to $\frac{n}{5}$

  - Caution: Different grouping of data (i.e., k) can affect the hypothesis testing result.

Chi-Square Test: Example

- Vehicle Arrival Example (continued):

  H_0: the random variable is Poisson distributed.

  H_1: the random variable is not Poisson distributed.

x_i	Observed Frequency, O_i		Expected Frequency, E_i		$(O_i - E_i)^2/E_i$
0	12	22	2.6	12.2	7.87
1	10		9.6		
2	19		17.4		0.15
3	17		21.1		0.8
4	19		19.2		4.41
5	6		14.0		2.57
6	7		8.5		0.26
7	5	17	4.4	7.6	11.62
8	5		2.0		
9	3		0.8		
10	3		0.3		
> 11	1		0.1		
	100		100.0		27.68

$$E_i = n \cdot p(x) = n \cdot \frac{e^{-\alpha}\alpha^x}{x!}$$

Combined because of the assumption of $\min E_i = 5$, *e.g.*,

$E_1 = 2.6 < 5$, *hence combine with* E_2

- Degree of freedom is $k\text{-}s\text{-}1 = 7\text{-}1\text{-}1 = 5$, hence, the hypothesis is rejected at the $\alpha = 0.05$ level of significance.

$$\chi_0^2 = 27.68 > \chi_{0.05,5}^2 = 11.1$$

Kolmogorov-Smirnov Test

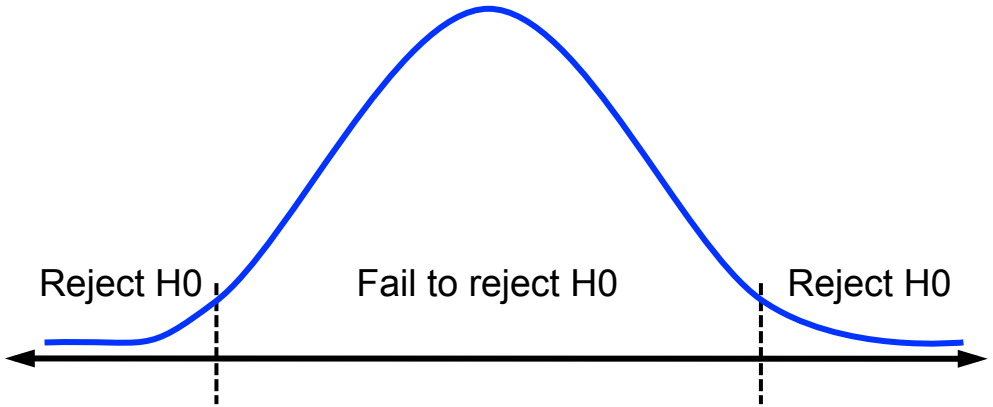
- Intuition: formalize the idea behind examining a Q-Q plot
- Recall
  - The test compares the continuous CDF, $F(x)$, of the hypothesized distribution with the empirical CDF, $SN(x)$, of the N sample observations.
  - Based on the maximum difference statistic:

$$D = max| F(x) - SN(x) |$$

- A more powerful test, particularly useful when:
  - Sample sizes are small
  - No parameters have been estimated from the data
- When parameter estimates have been made:
  - Critical values are biased, too large.
  - More conservative, i.e., smaller Type I error than specified.

p-Values and "Best Fits"

- Hypothesis testing requires a significance level
  - Significance level (α) is the probability of falsely rejecting H_0
  - Common significance levels
    - $\alpha = 0.1$
    - $\alpha = 0.05$
    - $\alpha = 0.01$
- Be aware that significance level does not tell anything about the subject of the test!
- Generalization of the significance level: p-value



p-Values and "Best Fits"

- *p-value* for the test statistics
  - The significance level at which one would just reject H_0 for the given test statistic value.
  - A measure of fit, the larger the better
  - Large p-value: good fit
  - Small p-value: poor fit

- Vehicle Arrival Example (cont.):
  - H_0: data is Poisson
  - Test statistics: $\chi_0^2 = 27.68$, with 5 degrees of freedom
  - The *p-value* $F(5, 27.68) = 0.00004$, meaning we would reject H_0 with 0.00004 significance level, hence Poisson is a poor fit.

p-Values and "Best Fits"

- Many software use *p-value* as the ranking measure to automatically determine the "best fit".

- Things to be cautious about:
  - Software may not know about the physical basis of the data, distribution families it suggests may be inappropriate.
  - Close conformance to the data does not always lead to the most appropriate input model.
  - *p-value* does not say much about where the lack of fit occurs

- Recommended: always inspect the automatic selection using graphical methods.

Fitting a Non-stationary Poisson Process

Fitting a Non-stationary Poisson Process

- Fitting a NSPP to arrival data is difficult, possible approaches:
  - Fit a very flexible model with lots of parameters
  - Approximate constant arrival rate over some basic interval of time, but vary it from time interval to time interval.
- Suppose we need to model arrivals over time $[0, T]$, our approach is the most appropriate when we can:
  - Observe the time period repeatedly
  - Count arrivals / record arrival times
  - Divide the time period into k equal intervals of length $\Delta t = T/k$
  - Over n periods of observation let C_{ij} be the number of arrivals during the i-th interval on the j-th period

Fitting a Non-stationary Poisson Process

- The estimated arrival rate during the i-th time period $(i\text{-}1)\,\Delta t < t \le i\,\Delta t$ is:

$$\hat{\lambda}(t) = \frac{1}{n\Delta t}\sum_{j=1}^{n} C_{ij}$$

  - n = Number of observation periods
  - Δt = Time interval length
  - C_{ij} = Number of arrivals during the i-th time interval on the j-th observation period
- Example: Divide a 10-hour business day [8am,6pm] into equal intervals $k = 20$ whose length $\Delta t = ½$, and observe over n=3 days

Time Period	Number of Arrivals			Estimated Arrival Rate (arrivals/hr)
	Day 1	Day 2	Day 3	
8:00 - 8:30	12	14	10	24
8:30 - 9:00	23	26	32	54
9:00 - 9:30	27	18	32	52
9:30 - 10:00	20	13	12	30

For instance,
1/3(0.5)*(23+26+32)
= 54 arrivals/hour

Selecting Models without Data

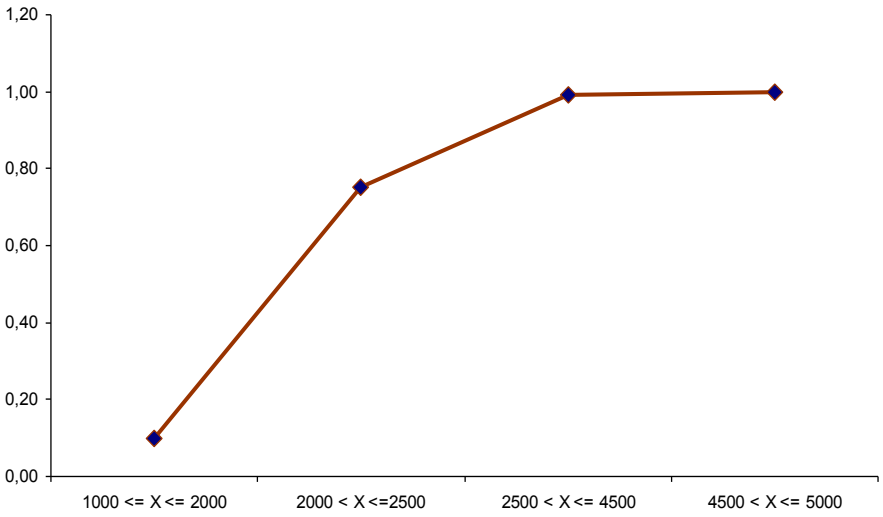
Selecting Models without Data

- If data is not available, some possible sources to obtain information about the process are:
  - **Engineering data:** often product or process has performance ratings provided by the manufacturer or company rules specify time or production standards.
  - **Expert option:** people who are experienced with the process or similar processes, often, they can provide optimistic, pessimistic and most-likely times, and they may know the variability as well.
  - **Physical or conventional limitations:** physical limits on performance, limits or bounds that narrow the range of the input process.
  - **The nature of the process**.
- The uniform, triangular, and beta distributions are often used as input models.
  - Speed of a vehicle?

Selecting Models without Data

- Example: Production planning simulation.
  - Input of sales volume of various products is required, salesperson of product XYZ says that:
    - No fewer than 1000 units and no more than 5000 units will be sold.
    - Given her experience, she believes there is a 90% chance of selling more than 2000 units, a 25% chance of selling more than 2500 units, and only a 1% chance of selling more than 4500 units.
  - Translating these information into a cumulative probability of being less than or equal to those goals for simulation input:

i	Interval (Sales)	PDF	Cumulative Frequency, *ci*
1	$1000 \leq X \leq 2000$	0.1	0.10
2	$2000 < X \leq 2500$	0.65	0.75
3	$2500 < X \leq 4500$	0.24	0.99
4	$4500 < X \leq 5000$	0.01	1.00



Multivariate and Time-Series Input Models

Multivariate and Time-Series Input Models

- The random variable discussed until now were considered to be independent of any other variables within the context of the problem
  - However, variables may be related
  - If they appear as input, the relationship should be investigated and taken into consideration
- Multivariate input models
  - Fixed, finite number of random variables $X_1, X_2, \ldots, X_k$
  - For example, lead time and annual demand for an inventory model
  - An increase in demand results in lead time increase, hence variables are dependent.
- Time-series input models
  - Infinite sequence of random variables, e.g., $X_1, X_2, X_3, \ldots$
  - For example, time between arrivals of orders to buy and sell stocks
  - Buy and sell orders tend to arrive in bursts, hence, times between arrivals are dependent.

Covariance and Correlation

- Consider a model that describes relationship between X_1 and X_2:

$$(X_1 - \mu_1) = \beta(X_2 - \mu_2) + \varepsilon$$

ε is a random variable with mean 0 and is independent of X_2

  - $\beta = 0$, X_1 and X_2 are statistically independent
  - $\beta > 0$, X_1 and X_2 tend to be above or below their means together
  - $\beta < 0$, X_1 and X_2 tend to be on opposite sides of their means

- Covariance between X_1 and X_2:

$$\text{cov}(X_1, X_2) = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E(X_1 X_2) - \mu_1 \mu_2$$

- Covariance between X_1 and X_2:

  - where $\text{cov}(X_1, X_2) \begin{cases} = 0 \\ < 0 \\ > 0 \end{cases} \Rightarrow \beta \begin{cases} = 0 \\ < 0 \\ > 0 \end{cases} \qquad -\infty < \text{cov}(X_1, X_2) < \infty$

Covariance and Correlation

- Correlation between X_1 and X_2 (values between -1 and 1):

$$\rho = \text{corr}(X_1, X_2) = \frac{\text{cov}(X_1, X_2)}{\sigma_1 \sigma_2}$$

- where $$corr(X_1, X_2) \begin{cases} = 0 \\ < 0 \\ > 0 \end{cases} \Rightarrow \beta \begin{cases} = 0 \\ < 0 \\ > 0 \end{cases}$$

- The closer ρ is to -1 or 1, the stronger the linear relationship is between X_1 and X_2.

Time-Series

- A time series is a sequence of random variables $X_1, X_2, X_3, \ldots$ which are identically distributed (same mean and variance) but dependent.
  - $cov(X_t, X_{t+h})$ is the lag-h autocovariance
  - $corr(X_t, X_{t+h})$ is the lag-h autocorrelation
  - If the autocovariance value depends only on h and not on t, the time series is **covariance stationary**
  - For covariance stationary time series, the shorthand for lag-h is used

$$\rho_h = corr(X_t, X_{t+h})$$

- Notice
  - autocorrelation measures the dependence between random variables that are separated by h-1 others in the time series

Multivariate Input Models

- If X_1 and X_2 are normally distributed, dependence between them can be modeled by the bivariate normal distribution with μ_1, μ_2, σ_1^2, σ_2^2 and correlation ρ
  - To estimate μ_1, μ_2, σ_1^2, σ_2^2, see "Parameter Estimation"
  - To estimate ρ, suppose we have n independent and identically distributed pairs (X_{11}, X_{21}), (X_{12}, X_{22}), … (X_{1n}, X_{2n}),

  - Then the sample covariance is

$$\hat{\text{cov}}(X_1, X_2) = \frac{1}{n-1}\sum_{j=1}^{n}(X_{1j} - \bar{X}_1)(X_{2j} - \bar{X}_2)$$

  - The sample correlation is

$$\hat{\rho} = \frac{\hat{\text{cov}}(X_1, X_2)}{\hat{\sigma}_1\hat{\sigma}_2}$$

Sample deviation

Multivariate Input Models: Example

- Let X_1 the average lead time to deliver and X_2 the annual demand for a product.
- Data for 10 years is available.

Lead Time (X_1)	Demand (X_2)
6,5	103
4,3	83
6,9	116
6,0	97
6,9	112
6,9	104
5,8	106
7,3	109
4,5	92
6,3	96

$$\overline{X}_1 = 6.14, \quad \overline{\sigma}_1 = 1.02$$
$$\overline{X}_2 = 101.8, \quad \overline{\sigma}_2 = 9.93$$

$$\hat{\text{cov}}_{\text{sample}} = 8.66$$ ← Covariance

$$\hat{\rho} = \frac{8.66}{1.02 \times 9.93} = 0.86$$

- Lead time and demand are strongly dependent.
  - Before accepting this model, lead time and demand should be checked individually to see whether they are represented well by normal distribution.

Time-Series Input Models

- If $X_1, X_2, X_3, \ldots$ is a sequence of identically distributed, but dependent and covariance-stationary random variables, then we can represent the process as follows:
  - Autoregressive order-1 model, **AR(1)**
  - Exponential autoregressive order-1 model, **EAR(1)**

- Both have the characteristics that:

$$\rho_h = corr(X_t, X_{t+h}) = \rho^h, \quad \text{for } h = 1,2,\ldots$$

  - Lag-h autocorrelation decreases geometrically as the lag increases, hence, observations far apart in time are nearly independent

Time-Series Input Models: Autoregressive order-1 model AR(1)

- Consider the time-series model:

$$X_t = \mu + \phi(X_{t-1} - \mu) + \varepsilon_t, \quad \text{for } t = 2,3,\ldots$$

where $\varepsilon_2, \varepsilon_3, \ldots$ are i.i.d. normally distributed with $\mu_\varepsilon = 0$ and variance σ_ε^2

- If initial value X_1 is chosen appropriately, then
  - $X_1, X_2, \ldots$ are normally distributed with *mean* = μ, and *variance* = $\sigma^2/(1-\phi^2)$
  - Autocorrelation $\rho_h = \phi^h$

- To estimate $\phi, \mu, \sigma_\varepsilon^2$:

$$\hat{\mu} = \overline{X}, \qquad \hat{\sigma}_\varepsilon^2 = \hat{\sigma}^2(1-\hat{\phi}^2), \qquad \hat{\phi} = \frac{\hat{\text{cov}}(X_t, X_{t+1})}{\hat{\sigma}^2}$$

where $\hat{\text{cov}}(X_t, X_{t+1})$ is the *lag*-1 autocovariance

Time-Series Input Models: Exponential AR(1) model EAR(1)

- Consider the time-series model:

$$X_t = \begin{cases} \phi X_{t-1}, & \text{with probability } \phi \\ \phi X_{t-1} + \varepsilon_t, & \text{with probability } 1\text{-}\phi \end{cases} \qquad \text{for } t = 2,3,\ldots$$

where $\varepsilon_2, \varepsilon_3, \ldots$ are i.i.d. exponentially distributed with $\mu_\varepsilon = 1/\lambda$, and $0 \le \phi < 1$

- If X_1 is chosen appropriately, then
  - $X_1, X_2, \ldots$ are exponentially distributed with *mean* $= 1/\lambda$
  - Autocorrelation $\rho_h = \phi^h$, and only positive correlation is allowed.
- To estimate ϕ, λ:

$$\hat{\lambda} = \frac{1}{\overline{X}}, \qquad \hat{\phi} = \hat{\rho} = \frac{\hat{\text{cov}}(X_t, X_{t+1})}{\hat{\sigma}^2}$$

where $\hat{\text{cov}}(X_t, X_{t+1})$ is the *lag*-1 autocovariance

Summary

- In this chapter, we described the 4 steps in developing input data models:
  (1) Collecting the raw data
  (2) Identifying the underlying statistical distribution
  (3) Estimating the parameters
  (4) Testing for goodness of fit