CPSC 531: System Modeling and Simulation

Carey Williamson
Department of Computer Science
University of Calgary
Fall 2017

Motivational Quote

"If you can't measure it, you can't improve it."

- Peter Drucker

(Slightly Revised) Motivational Quote

model
"If you can't measure it, you can't ~~improve~~ it."

\- Peter Drucker

Simulation Input Analysis

- Input models are the driving force for many simulations
- Quality of the output depends on the quality of inputs
- There are four main steps for input model development:
  1. Collect data from the real system
  2. Identify a suitable probability distribution to represent the input process
  3. Choose parameters for the distribution
  4. Evaluate the goodness-of-fit for the chosen distribution and parameters

Data Collection

- Data collection is one of the biggest simulation tasks
- Beware of GIGO: Garbage-In-Garbage-Out
- Suggestions to facilitate data collection:
  - Analyze the data as it is being collected: check adequacy
  - Combine homogeneous data sets (e.g. successive time periods, or the same time period on successive days)
  - Be aware of inadvertent data censoring: quantities that are only partially observed versus observed in their entirety; gaps; outliers; risk of leaving out long processing times
  - Collect <u>input</u> data, not performance data (i.e., output)

Data Analysis Checklist (meta-level)

- Where did this data come from?
- How was it collected?
- What can it tell me?

- Do some exploratory data analysis (see next slide)

- Does this data make sense?
- Is it representative?
- What are the key properties?
- Does it resemble anything I've seen before?
- How best to model it?

Data Analysis Checklist (detailed-level)

- How much data do I have? (N)
- Is it discrete or continuous?
- What is the range for the data? (min, max)
- What is the central tendency? (mean, median, mode)
- How variable is it? (mean, variance, std dev, CV)
- What is the shape of the distribution? (histogram)
- Are there gaps, outliers, or anomalies? (tails)
- Is it time series data? (time series analysis)
- Is there correlation structure and/or periodicity?
- Other interesting phenomena? (scatter plot)

Identifying the Distribution

Non-Parametric Approach: does not care about the actual distribution or its parameters; simply (re-)generates observations from the empirically observed CDF for the distribution.

- less work for the modeler, but limited generative capability (e.g., variety; length; repetitive; preserves flaws in data)

Parametric Approach: tries to find a compact, concise, and parsimonious model that accurately represents the input data.

- more work, but potentially valuable model (parameterizable)

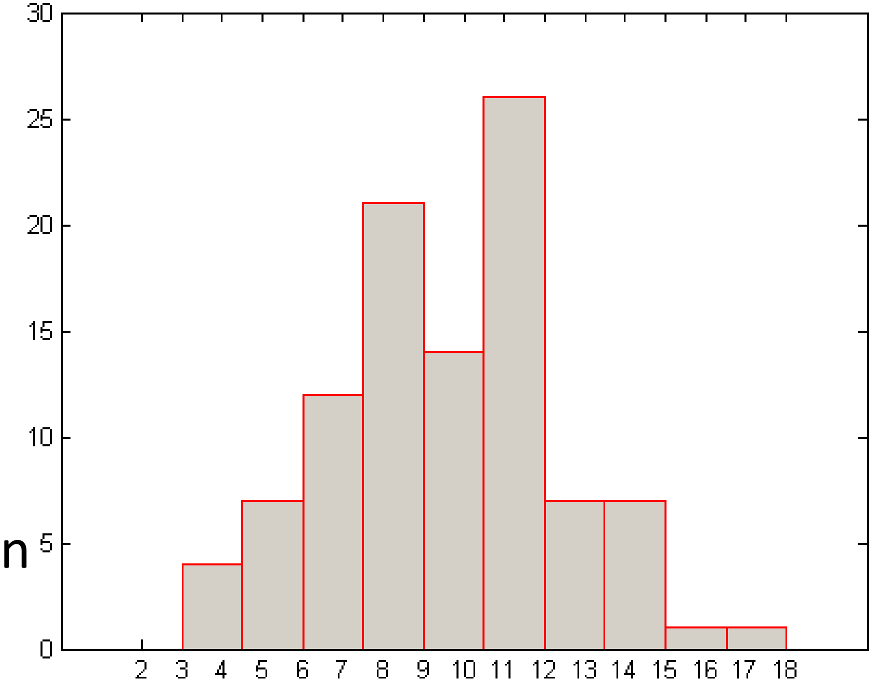
1. Histograms (visual/graphical approach)
2. Selecting families of distributions (logic/statistics)
3. Parameter estimation (statistical methods)
4. Goodness-of-fit tests (statistical/graphical methods)

Histograms (1 of 3)

- Histogram: A frequency distribution plot useful in determining the shape of a distribution
  - Divide the range of data into (typically equal) intervals or cells
  - Plot the frequency of each cell as a rectangle

- For discrete data:
  - Corresponds to the probability mass function

- For continuous data:
  - Corresponds to the probability density function



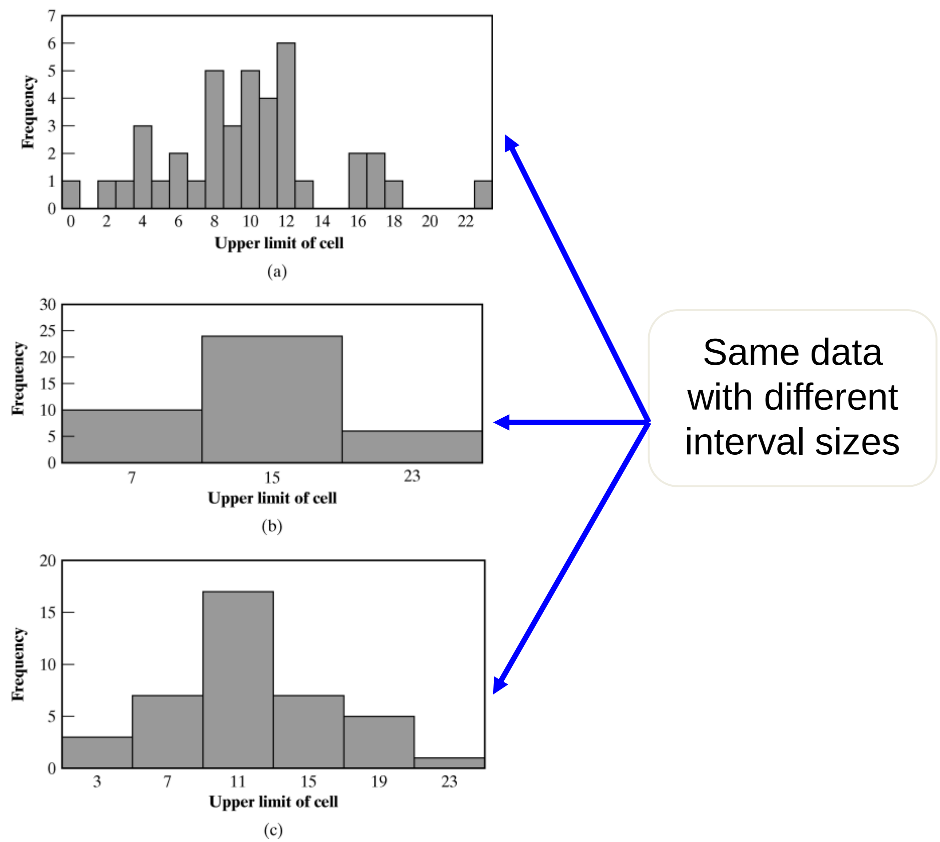
Histograms (2 of 3)

- The key problem is determining the cell size
  - Small cells: large variation in the number of observations per cell
  - Large cells: details of the distribution are completely lost
  - It is possible to reach very different conclusions about the distribution shape

- The cell size depends on:
  - The number of observations
  - The dispersion of the data

- Guideline:
  - The number of cells ≈ the square root of the sample size

- Example: It is possible to reach very different conclusions about the distribution shape by changing the cell size



Selecting the Family of Distributions (1 of 4)

- 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, Geometric, Poisson
  - Moderate to analyze: Normal, Log-Normal, Uniform
  - Harder to analyze: Beta, Gamma, Pareto, Weibull, Zipf

Selecting the Family of Distributions (2 of 4)

- Use the physical basis of the distribution as a guide
- Examples:
  - Binomial: number of successes in n trials
  - Poisson: number of independent events that occur in a fixed amount of time or space
  - Normal: distribution of a process that is the sum of a number of (smaller) component processes
  - Exponential: time between independent events, or a processing time duration that is memoryless
  - Discrete or continuous uniform: models the complete uncertainty about the distribution (other than its range)
  - Empirical: does not follow any theoretical distribution

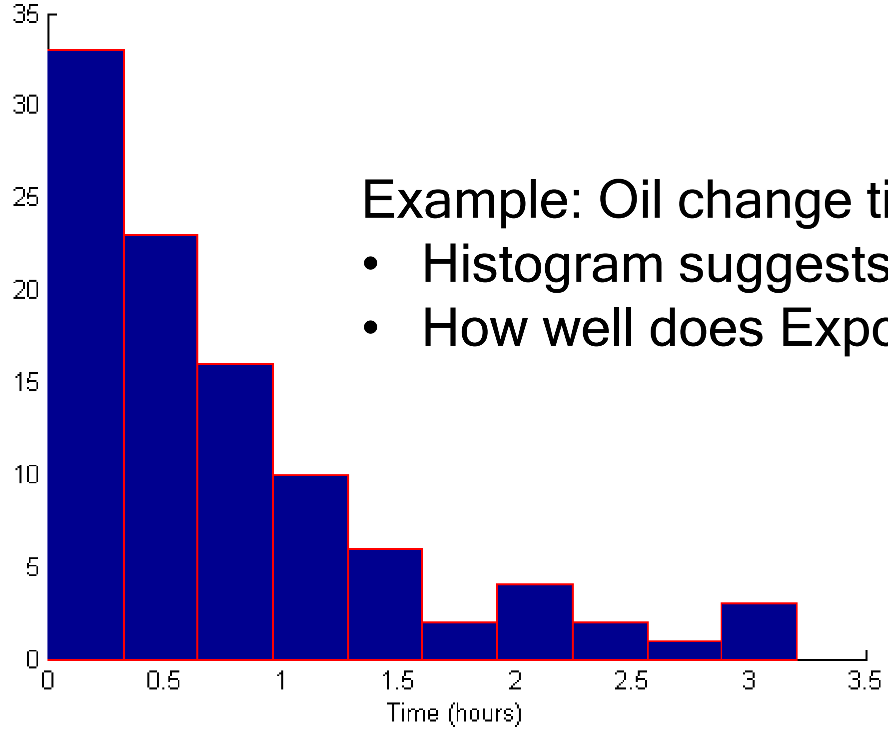
Selecting the Family of Distributions (3 of 4)

- Remember the physical characteristics of the process
  - Is the process naturally discrete or continuous valued?
  - Is it bounded?
  - Is it symmetric, or is it skewed?
- No "true" distribution for any stochastic input process
- Goal: obtain a good approximation that captures the salient properties of the process (e.g., range, mean, variance, skew, tail behavior)

Selecting the Family of Distributions (4 of 4)

How to check if the chosen distribution is a good fit?

- Compare the shape of the pmf/pdf of the distribution with the histogram:
  - Problem: Difficult to visually compare probability curves
  - Solution: Use Quantile-Quantile plots



Example: Oil change time at MinitLube
- Histogram suggests "exponential" dist.
- How well does Exponential fit the data?

- Q-Q plot is a useful tool for evaluating distribution fit
  - It is easy to visually inspect since we look for a straight line

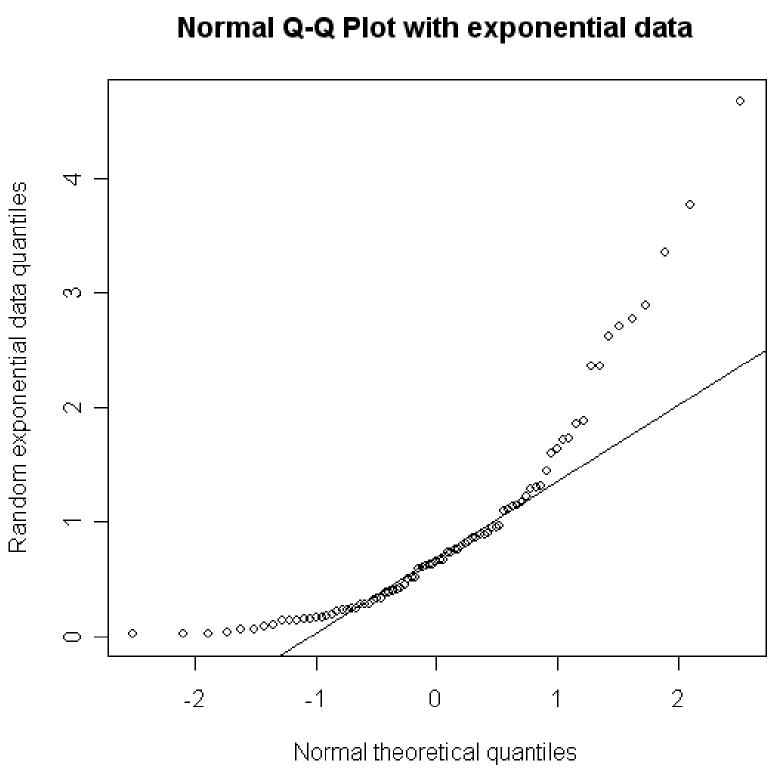
- If X is a random variable with CDF $F(x)$, then the q-quantile of X is given by x_q such that:

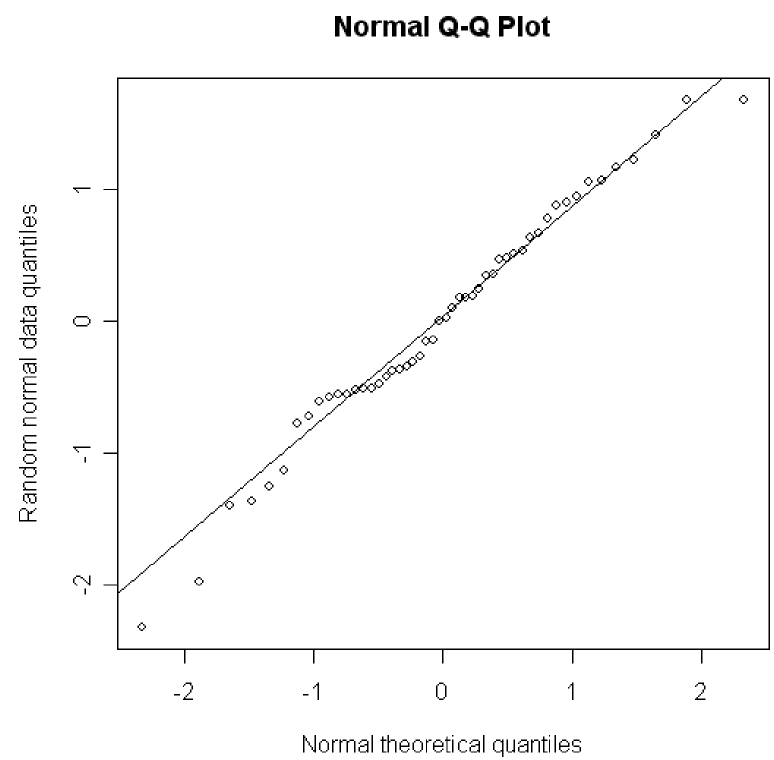
$$F(x_q) = \mathbb{P}(X \leq x_q) = q, \qquad 0 < q < 1$$

- When $F(x)$ has an inverse, then $x_q = F^{-1}(q)$

Quantile-Quantile Plots (2 of 8)

- x_q^S: empirical q-quantile from the sample
- x_q^M: theoretical q-quantile from the model
- Q-Q plot: plot x_q^S versus x_q^M as a scatterplot of points





Quantile-Quantile Plots (3 of 8)

- X: a random variable with CDF $F(x)$
- $\{X_i, i = 1, \dots, n\}$: a sample of X consisting of n observations
- Define $F_n(x)$: empirical CDF of X,

$$F_n(x) = \frac{\text{number of } X_i' s \le x}{n}$$

- $\{X_{(j)}, j = 1, \dots, n\}$: observations ordered from smallest to largest

$$X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$$

- It follows that

$$F_n(x) = \frac{j}{n}$$

where j is the rank or order of x, i.e., x is the j-th value among X_i's.

Quantile-Quantile Plots (4 of 8)

- Problem:
  - For finite value $x = X_{(n)}$, we have $F_n^{-1}(1) = X_{(n)}$
  - But from the model we generally have: $F^{-1}(1) = \infty$
  - How to resolve this mismatch?

- Solution: slightly modify the empirical distribution

$$\tilde{F}_n(X_{(j)}) = F_n(X_{(j)}) - \frac{0.5}{n} = \frac{j - 0.5}{n}$$

- Therefore,

$$\tilde{F}_n^{-1}\left(\frac{j - 0.5}{n}\right) = X_{(j)}$$

- and, thus,

empirical $\left(\frac{j-0.5}{n}\right)$ $-\text{quantile of X} = X_{(j)}$

Quantile-Quantile Plots (5 of 8)

- $F(x)$: the CDF fitted to the observed data, i.e., the model

- Q-Q plot: plotting empirical quantiles vs. model quantiles
  - $\left(\frac{j-0.5}{n}\right)$-quantiles for $j = 1, \ldots, n$
    - Empirical quantile = $X_{(j)}$
    - Model quantile = $F^{-1}\left(\frac{j-0.5}{n}\right)$

- Q-Q plot features:
  - 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

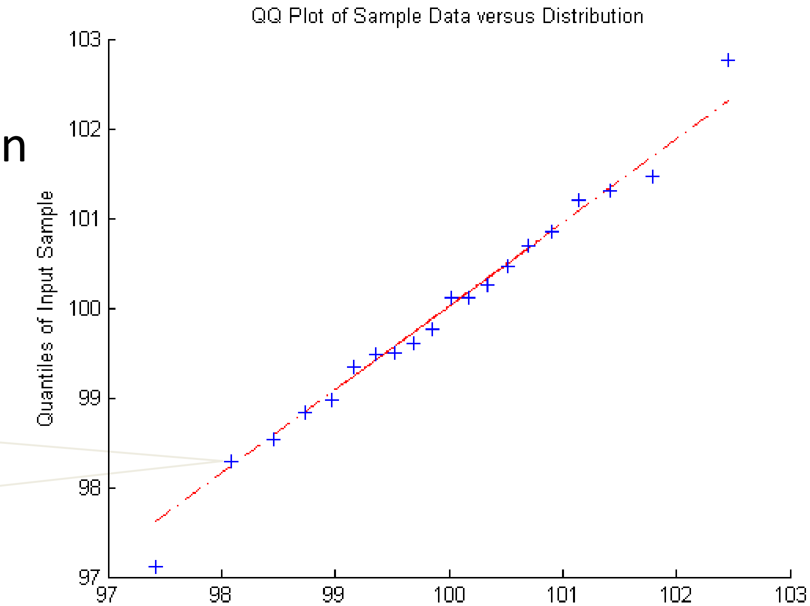
Quantile-Quantile Plots (6 of 8)

- Example: Check whether the door installation times follow a normal distribution.
  - The observations are ordered from smallest to largest:

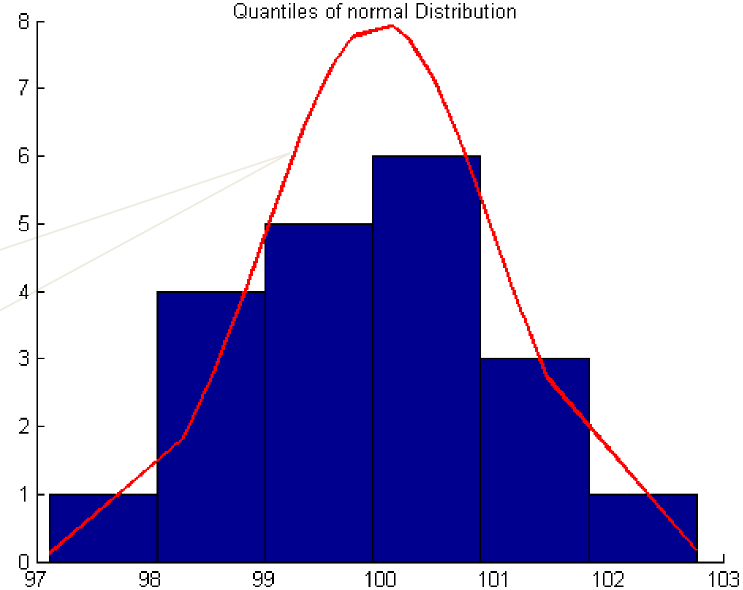
j	value	j	value	j	value	j	value
1	97.12	6	99.34	11	100.11	16	100.85
2	98.28	7	99.50	12	100.11	17	101.21
3	98.54	8	99.51	13	100.25	18	101.30
4	98.84	9	99.60	14	100.47	19	101.47
5	98.97	10	99.77	15	100.69	20	102.77

  - $X_{(j)}$'s are plotted versus $F^{-1}\left(\frac{j-0.5}{n}\right)$ where F is the normal CDF with sample mean (99.93 sec) and sample STD (1.29 sec)

- 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 scaled by the number of observation, that is $20 \times f(x)$

Quantile-Quantile Plots (8 of 8)

- Consider the following while evaluating the linearity of a Q-Q plot:
  - The observed values never fall exactly on a straight line
  - Variation of the extremes is higher than the middle.
  - Linearity of the points in the middle of the plot (the main body of the distribution) is more important.

Parameter Estimation (1 of 4)

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 variance are:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}, \qquad s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} = \frac{\sum_{i=1}^{n} X_i^2 - n\,\bar{X}^2}{n-1}$$

- If the data are discrete and have been grouped into a frequency distribution with k distinct values:

$$\bar{X} = \frac{\sum_{j=1}^{k} f_j X_j}{n},$$

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

where f_j is the observed frequency of value X_j

Parameter Estimation (3 of 4)

- Vehicle Arrival Example: number of vehicles arriving at an intersection between $7{:}00$ am and $7{:}05$ am was monitored for 100 random workdays.

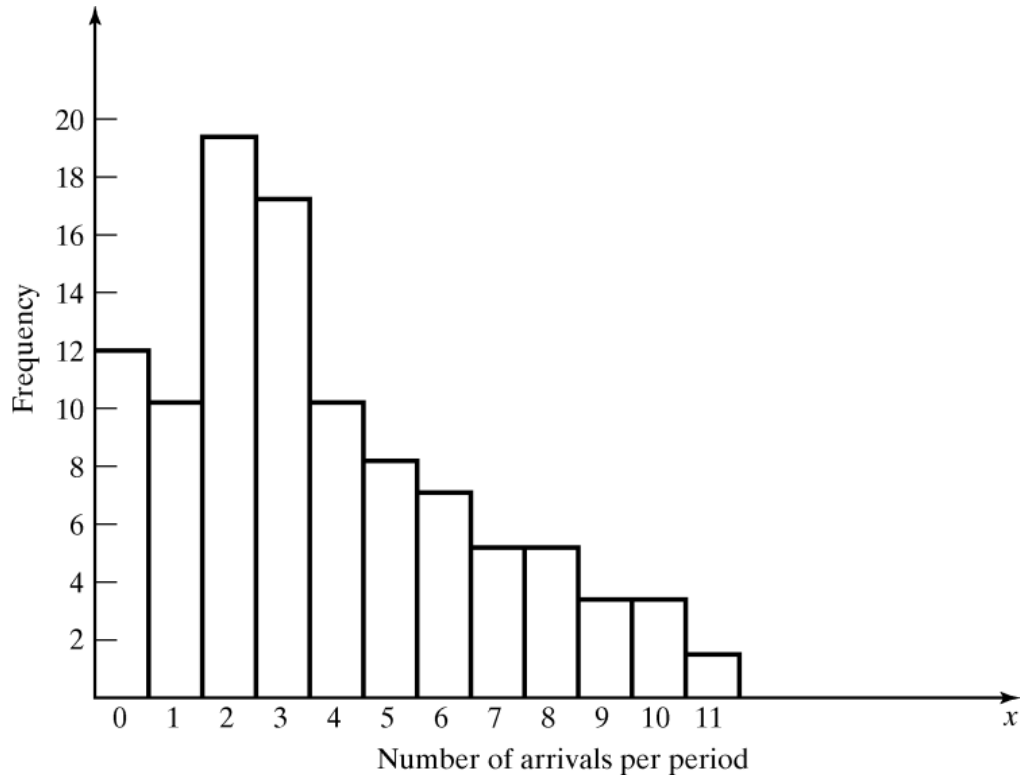
$$n = 100$$
$$\sum_{j=1}^{k} f_j X_j = 364$$
$$\sum_{j=1}^{k} f_j X_j^2 = 2080$$

# Arrivals (X_j)	Frequency (f_j)
0	12
1	10
2	19
3	17
4	10
5	8
6	7
7	5
8	5
9	3
10	3
11	1

  - The sample mean and variance are

$$\bar{X} = \frac{364}{100} = 3.64$$
$$s^2 = \frac{2080-100*(3.64)^2}{99} = 7.63$$

- The histogram suggests X is a Poisson distribution
  - However, the sample mean is not equal to sample variance
  - Reason: each estimator is a random variable (not perfect)



Goodness-of-Fit Tests (1 of 2)

- Conduct hypothesis testing on input data distribution using well-known statistical tests, such as:
  - Chi-square test
  - Kolmogorov-Smirnov test

- Note: you don't always get a single unique correct distributional result for any real application:
  - 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 (2 of 2)

Objective: to determine how well a (theoretical) statistical model fits a given set of empirical observations (sample)

- Vehicle Arrival Example:
  - The histogram suggests X might be a Poisson distribution
  - Hypothesis:
    X has a Poisson distribution with rate 3.64
  - How can we test the hypothesis?

Chi-Square Test (1 of 11)

Intuition:

- It establishes whether an observed frequency distribution differs from a model distribution
  - Model distribution refers to the hypothesized distribution with the estimated parameters
  - Can be used for both discrete and continuous random variables
  - Valid for large sample sizes

- If the difference between the distributions is smaller than a critical value, the model distribution fits the observed data well, otherwise, it does not.

Chi-Square Test (2 of 11)

Concepts:

- Null hypothesis H_0:
  The observed random variable X conforms to the model distribution

- Alternative hypothesis H_1:
  The observed random variable X does not conform to the model distribution

- Test statistic χ^2:
  The measure of the difference between sample data and the model distribution

- Significance level α:
  The probability of rejecting the null hypothesis when the null hypothesis is true. Common values are 0.05 and 0.01.

Chi-Square Test (3 of 11)

Approach:

- Arrange the n observations into a set of k intervals or cells, where interval i is given by $[a_{i-1}, a_i)$
  - Suggestion: set the interval length such that at least 5 observations fall in each interval

- 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	$n^{1/2}$ to n/5

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

Test Statistic:

- O_i: the number of observations X_j that fall in interval i
- E_i: the expected number of observations in interval i if taking n samples from the model distribution:
  - Continuous model with fitted PDF $f(x)$:

$$E_i = n \cdot \int_{a_{i-1}}^{a_i} f(x)dx$$

  - Discrete model with fitted PMF $p(x)$:

$$E_i = n \cdot \sum_{a_{i-1} \leq x < a_i} p(x)$$

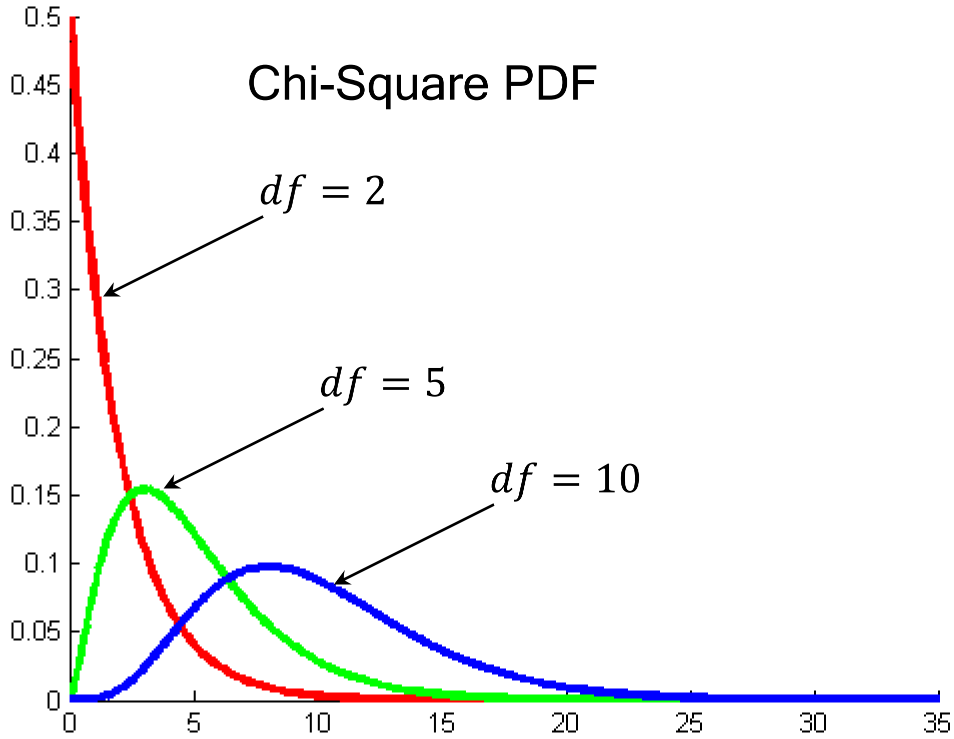
Test Statistic:

- Test statistic χ^2 is defined as

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

- χ^2 approximately follows the chi-square distribution with $k - s - 1$ degrees of freedom
  - k: the number of intervals
  - s: the number of parameters of the model (i.e., hypothesized distribution) estimated by the sample statistics
    - Uniform: $s = 0$
    - Poisson, Exponential, Bernoulli, Geometric: $s = 1$
    - Normal, Binomial: $s = 2$

- The distribution is not symmetric
- Minimum value is 0
- Mean = degrees of freedom



Chi-Square Test (7 of 11)

Intuition:

- χ^2 measures the normalized squared difference between the frequency distribution of the sample data and hypothesized model

- A large χ^2 provides evidence that the model is not a good fit for the sample data:
  - If the difference is greater than a critical value then reject the null hypothesis
  - Question: what is an appropriate critical value?
  - Answer: it is pre-specified by the modeler.

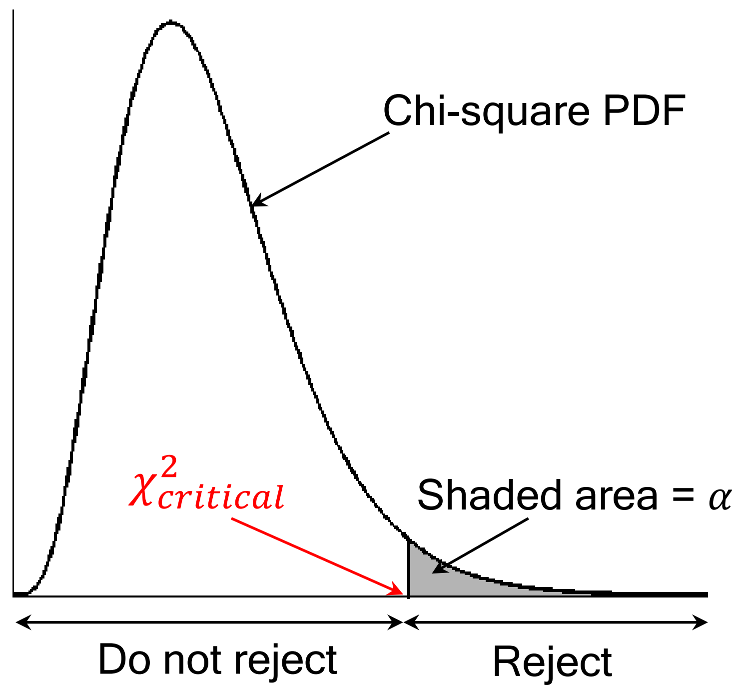
Critical Value:

- For significance level α, the critical value $\chi^2_{critical}$ is defined such that:

$$\mathbb{P}\left(\chi^2_{k-s-1} \geq \chi^2_{critical}\right) = \alpha$$

Chi-Square distributed random variable with $k-s-1$ degrees of freedom.

- $\chi^2_{critical} = \chi^2_{k-s-1,1-\alpha}$ the $(1-\alpha)$-quantile of chi-square distribution with $k-s-1$ degrees of freedom

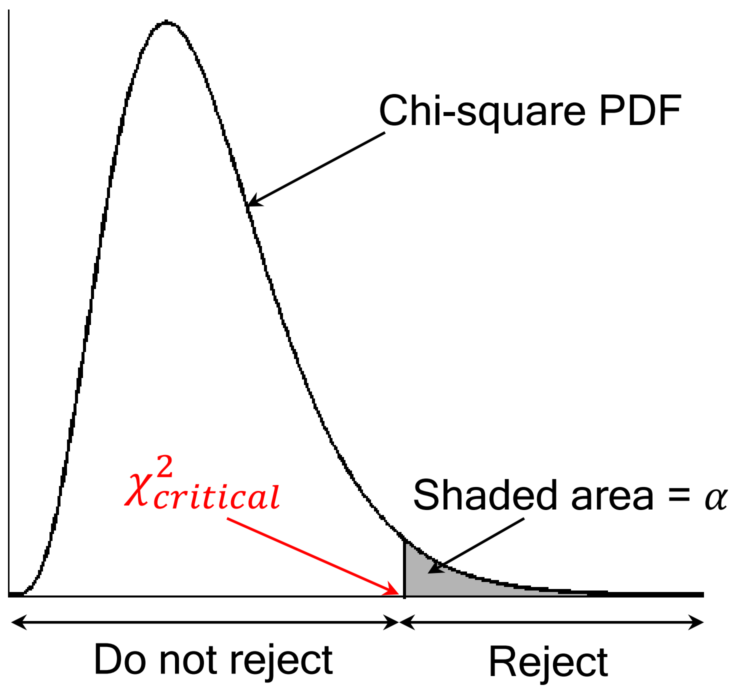


Chi-Square Test (9 of 11)

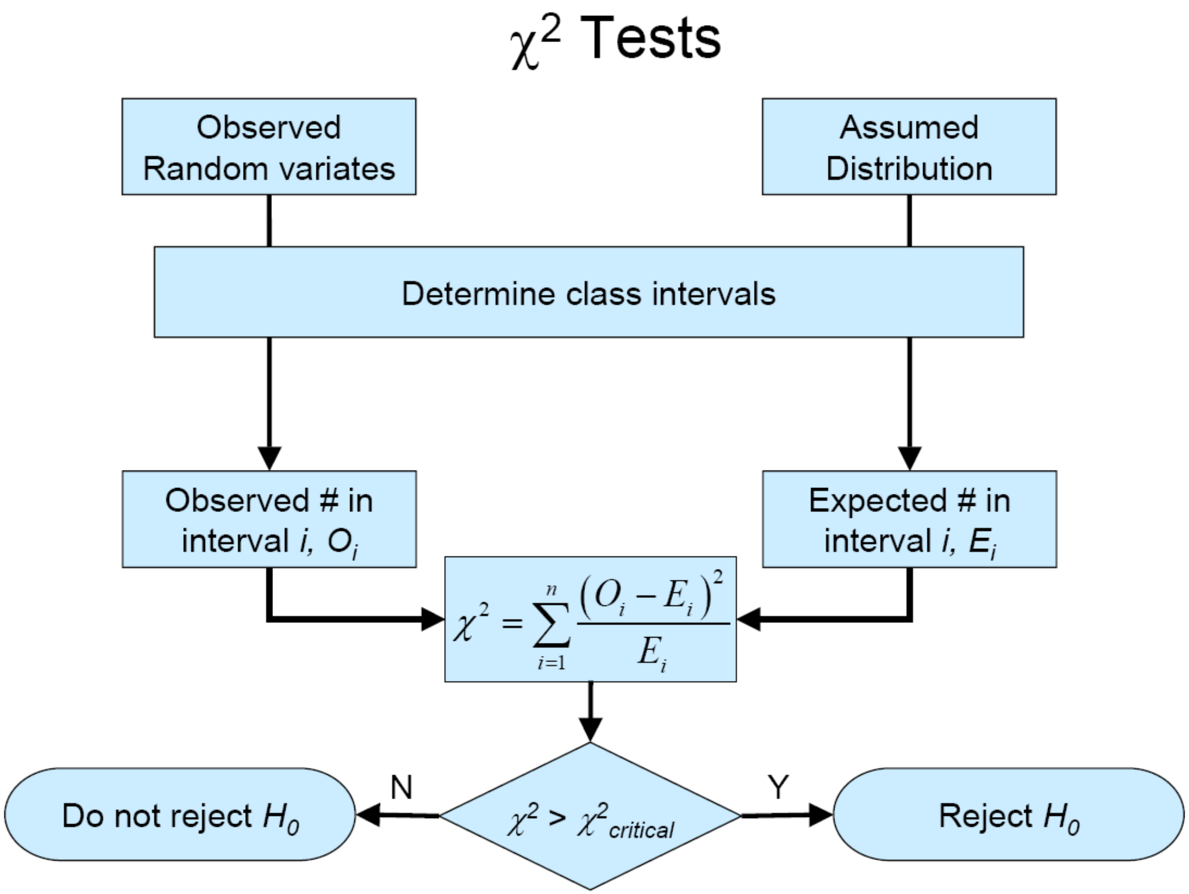
- We say that the null hypothesis H_0 is rejected at the significance level α, if:

$$\chi^2 > \chi^2_{k-s-1,1-\alpha}$$

- Interpretation:
  - The test statistic can be as large as the critical value
  - If the test statistic is greater than the critical value then, the null hypothesis is rejected
  - If the test statistic is not greater than the critical value then, the null hypothesis can not be rejected



Chi-Square Test (10 of 11)



Chi-Square Test (11 of 11)

- Vehicle Arrival Example (continued):

  H_0: the random variable is Poisson distributed (with $\lambda = 3.64$).

  H_1: the random variable is not Poisson distributed.

0	12	2.6	7.87
1	10	9.6	
2	19	17.4	0.15
3	17	21.1	0.83
4	10	19.2	4.41
5	8	14.0	2.57
6	7	8.5	0.26
7	5	4.4	11.63
8	5	2.0	
9	3	0.8	
10	3	0.3	
>11	1	0.1	
	100	100.0	27.72

$$E_i = np(x) = n\frac{e^{-\lambda}\lambda^x}{x!}$$

Combined because of min E_i

— Degrees of freedom is $k - s - 1 = 7 - 1 - 1 = 5$, hence, the hypothesis is rejected at the 0.05 level of significance:

$$\chi^2 = 27.72 > \chi^2_{0.95,5} = 11.1$$

Kolmogorov-Smirnov Test

- Intuition:
  - Formalizes the idea behind examining a Q-Q plot
  - The test compares the CDF of the hypothesized distribution with the empirical CDF of the sample observations based on the maximum distance between two cumulative distribution functions.
- A more powerful test that is particularly useful when:
  - Sample sizes are small
  - No parameters have been estimated from the data

Selecting Model without Data (1 of 2)

- 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 that 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.

Selecting Model without Data (2 of 2)

- Example: Production planning simulation.
  - Input of sales volume of various products is required, salesperson of product XYZ says that:
    - No fewer than $1,000$ units and no more than $5,000$ units will be sold.
    - Given her experience, she believes there is a 90% chance of selling more than $2,000$ units, a 25% chance of selling more than $3,000$ units, and only a 1% chance of selling more than $4,000$ units.
  - Translating these information into a cumulative probability of being less than or equal to those goals for simulation input:

i	**Interval (Sales)**	**Cumulative Frequency, c_i**
1	$1000 \le x \le 2000$	0.10
2	$2000 < x \le 3000$	0.75
3	$3000 < x \le 4000$	0.99
4	$4000 < x \le 5000$	1.00

Multivariate and Time-Series Models

- So far, we have considered:
  - Single variate models for independent input parameters

- To model correlation among input parameters
  - Multivariate models
  - Time-series models