MG26018 Simulation Modeling and Analysis
仿真建模与分析

Lecture 4: Input Modeling

SHEN Haihui 沈海辉

Sino-US Global Logistics Institute
Shanghai Jiao Tong University

⌂ shenhaihui.github.io/teaching/mg26018
✉ shenhaihui@sjtu.edu.cn

Fall 2019

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

董浩云航运与物流研究院
CY TUNG Institute of Maritime and Logistics
中美物流研究院
Sino-US Global Logistics Institute

Contents

Introduction

- Input models represent the uncertainty in real world.
- Input models provide the driving force for a simulation.
  - Queueing system: Distributions of interarrival time and service time.
  - Supply chain: Distributions of demand and lead time.
  - Financial risk management: Distributions of asset return.
- The quality of outputs is no better than the quality of inputs.
  - *"Garbage in, garbage out."*
- *"All models are wrong, but some are useful."* – George Box.
  - There is no "true" model for any stochastic input.
  - The best we can do is to obtain an approximation that yields reasonable and useful results.

Introduction

- Fundamental requirements for an input model:
  - can capture the physical properties of the system;
  - can be easily tuned to the situation at hand;
  - can be efficiently generated with certain random variate generation technique.
- Input modeling is sometimes more of an art than an engineering.
  - It nearly always requires the analysts to use their judgment as well as to apply appropriate statistical tools.
  - Since there is no “true” model, it is sensible to run the simulation with several plausible input models to see if the conclusions are robust or highly sensitive to the choices.

Introduction

- Typical steps for input modeling.
  1. Collect data from the real system.
  2. Identify a probability distribution family to represent the data.
     – based on physical characteristics of the process (consult domain experts for structural knowledge).
     – based on graphical examination of the data (examine the “shape” via, e.g., histogram).
  3. Fit the distribution to the data (determine values of the parameters).
     – method of moments (MoM)
     – maximum likelihood estimation (MLE)
  4. Evaluate the chosen distribution and parameters for goodness of fit.
     – graphical methods: histogram, quantile-quantile (Q-Q) plot.
     – statistical tests: chi-square (χ^2) test, Kolmogorov-Smirnov (K-S) test.
  5. If the fit is not good, select another candidate and go to Step 3, or use an empirical distribution.

Data Collection

- People often have the **false** impression that data are readily available, but it is one of the most *challenging* task in solving a real problem.
- **Never** trust data blindly!
    - A common mistake is to simply throw data into a software and ask for a “best” fit model.
    - Always take into account under what context (e.g., time, potential influence of other factors) the data was collected.

Data Collection

- The collected data can be
  - stale (out of date);
  - “dirty” (containing errors);
  - unexpected;
  - time-varying;
  - dependent.
- Sometimes the effort or cost to transform data into a usable form, or “clean” data, can be as significant as that required to obtain them.

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 they are being collected: check adequacy.
  - Combine homogeneous data sets, e.g., successive time periods, the same time period on successive days.
  - Be aware of data censoring (删失): some values exist but are not observed.

    – Example: customer may quit the queue due to excessive long waiting. How to find out the patience limit for those who don't experience long waiting and receive service?
  - Check for autocorrelation.
  - Collect input data, not output data.

    – Example: customer arrival times and service times are input, whereas waiting times are output.

Identifying Distribution

- When data are collected, we next want to select a *family* of input distributions.
    - **Assumption**: Data are *independent* and *identically distributed*!
- A family of distributions can be selected on the basis of
    - the context of the input variable;
    - the shape of the histogram.
- Do not ignore the physical characteristics of the process when selecting distributions.
    - Is the process naturally discrete or continuous valued?
    - Is it bounded or is there no natural bound?
- There are literally hundreds of probability distributions that have been created; many were created with some specific physical process in mind.

- Discrete Distributions:
  - **Bernoulli**: Models the outcome of a trial, where each trial has a probability p of success.
  - **binomial**: Models the number of successes in n trials, when the trials are independent with common success probability p.
    – *Example*: the number of defective computer chips found in n chips.
  - **negative binomial**: Models the number of trials required to achieve k successes.
    – *Example*: the number of computer chips that we must inspect to find 4 defective chips.
  - **Poisson**: Models the number of independent events that occur in a fixed amount of time or space.
    – *Example 1*: the number of customers that arrive to a store during 1 hour.
    – *Example 2*: the number of defects found in 30 square meters of sheet metal.

- Continuous Distributions:
  - **uniform**: Models the situation that an outcome is equally likely for every value in the range $[a, b]$.
  - **normal**: Models the distribution of a process that can be thought of as the sum of a number of component processes.

    – *Example*: the time to assemble a product that is the sum of the times required for each assembly operation.

    – *Caution*: normal distribution admits negative values, which could be impossible for some process.
  - **exponential**: Models the time between independent events, or a process time that is memoryless.

    – *Example 1*: the times between the arrivals from a large population of potential customers who act independently.

    – *Example 2*: the time to failure for a system that is memoryless or has constant failure rate over time.

    – *Note*: if the time between events is exponential, then the number of events in a fixed period of time is Poisson.
  - **Weibull**: Models the time to failure for components.

    – *Note*: the failure rate can be increasing, decreasing, or constant (reduce to exponential distribution).

- Continuous Distributions:
  - **Erlang**: Models the time that can be viewed as the sum of several exponentially distributed times.

    – *Example*: a computer network fails when a computer and two backup computers fail, and each has exponentially distributed time to failure.
    – *Note*: Erlang is a special case of gamma.
  - **gamma**: An extremely flexible distribution used to model *nonnegative* random variables.

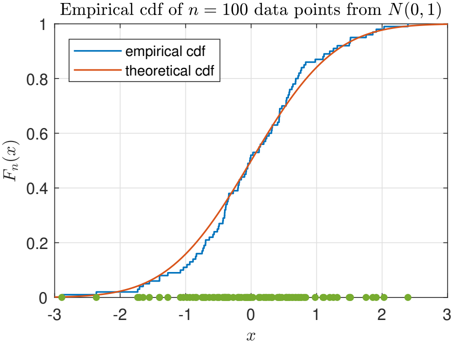
    – *Note*: can be shifted away from 0 by adding a constant.
  - **beta**: An extremely flexible distribution used to model *bounded* (originally in $[0, 1]$) random variables.

    – *Note*: can be shifted away from 0 by adding a constant; can cover a range different from $[0, 1]$ by multiplying by a constant.
  - **triangular**: Models a process for which only the minimum, most likely, and maximum values of the distribution are known.

    – *Example*: only the minimum, most likely, and maximum time required to test a product are known.

- **Empirical** Distribution: Often used when no theoretical distribution seems appropriate; its cdf is

$$F_n(x) = \frac{\text{number of points smaller than } x}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{x_i \le x\}}.$$



- Empirical distribution is discrete and its cdf is a step function.

Identifying Distribution

- **Histogram** describes frequency (or relative frequency, i.e., ratio) of data in different ranges.
  - Useful in determining the shape of the distribution from which the data have been sampled.
- For continuous data:
  - Corresponds to the pdf of a theoretical distribution.
  - In terms of the *shape*, not the exact *value*!
- For discrete data:
  - Corresponds to the pmf of a theoretical distribution.
  - In terms of both the *shape* and *value* (if the histogram uses relative frequency).
  - If there are few data points, it could be necessary to combine adjacent cells to eliminate the ragged appearance of the histogram.

Identifying Distribution

- Histogram's appearance heavily relies on how one partition the range of the data into intervals.
  - Intervals are too narrow: the histogram will be ragged (i.e., not smooth).
  - Intervals are too wide: the histogram will be coarse, or blocky, and its shape and other details will not show well.

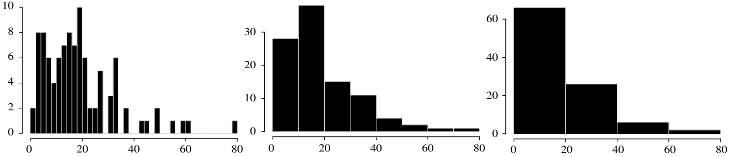


Figure: Ragged, Appropriate and Coarse Histograms (*from Banks et al. (2010)*)

- Choosing the number of intervals approximately equal to the square root of the sample size often works well in practice (Hines et al. 2002).

Distribution Fitting

- After a family of distributions has been selected, the next step is to determine the parameters of the distribution that can "best" fit the data.
  - Called distribution fitting, or parameter estimation.
- There are many different approaches and we discuss two simple ones:
  - method of moments (MoM)
  - maximum likelihood estimation (MLE)

- For a random variable X, its kth moment is defined as $\mathbb{E}[X^k]$.
- Let $X_1, \ldots, X_n$ be a random sample of X. The kth sample moment is defined as

$$m_k := \frac{X_1^k + \cdots + X_n^k}{n}.$$

- Suppose the considered distribution family has s unknown parameters.
  1. Analytically compute $\mathbb{E}[X^1], \ldots, \mathbb{E}[X^s]$, as functions of those parameters.
     - *Note*: the moments of common distributions are well-known.
  2. Compute $m_1, \ldots, m_s$ from the data.
  3. Solve $\mathbb{E}[X^k] = m_k,\ k = 1, \ldots, s$, for s unknown parameters.

- Example 1: Suppose $X_1, \ldots, X_n$ are iid from $\mathrm{Gamma}(\alpha, \lambda)$ (in shape & rate parametrization).
  - Recall: $f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$, $\mathbb{E}[X] = \alpha/\lambda$, $\mathrm{Var}(X) = \alpha/\lambda^2$.

  Estimate α and λ using MoM.

Solution. The first two moments are

$$\begin{aligned} \mathbb{E}[X] &= \alpha/\lambda = m_1, \\ \mathbb{E}[X^2] &= \mathrm{Var}(X) + (\mathbb{E}[X])^2 = (\alpha + \alpha^2)/\lambda^2 = m_2. \end{aligned}$$

Solving two equations yields MoM estimators

$$\widehat{\alpha} = \frac{m_1^2}{m_2 - m_1^2}, \quad \widehat{\lambda} = \frac{m_1}{m_2 - m_1^2}. \quad \blacksquare$$

- Example 2: Suppose $X_1, \ldots, X_n$ are iid from $\mathrm{Exp}(\lambda)$. Estimate λ using MoM.

  Solution. The first moment is

  $$\mathbb{E}[X] = 1/\lambda = m_1.$$

  So the MoM estimator of λ is $\widehat{\lambda} = \frac{1}{m_1} = \frac{n}{X_1 + \cdots + X_n}$. ■

- Example 3: Suppose $X_1, \ldots, X_n$ are iid from $\mathcal{N}(\mu, \sigma^2)$. Estimate μ and σ^2 using MoM.

  Solution. The first two moments are

$$\begin{aligned}\mathbb{E}[X] &= \mu = m_1,\\ \mathbb{E}[X^2] &= \mathrm{Var}(X) + (\mathbb{E}[X])^2 = \sigma^2 + \mu^2 = m_2.\end{aligned}$$

  Solving two equations yields MoM estimators

$$\widehat{\mu} = m_1, \quad \widehat{\sigma}^2 = m_2 - m_1^2. \quad \blacksquare$$

- Remark: $\widehat{\mu} = \frac{\sum_{i=1}^n X_i}{n}$, and

$$\begin{aligned}\widehat{\sigma}^2 = \frac{X_1^2 + \cdots + X_n^2}{n} - \bar{X}^2 &= \frac{\sum_{i=1}^n X_i^2 - n\bar{X}^2}{n}\\ &= \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}. \quad \text{(why?)}\end{aligned}$$

- Many common distributions have no more than 2 parameters: $\mathrm{Ber}(p)$, $\mathrm{B}(n,p)$, $\mathrm{Neg\ B}(k,p)$, $\mathrm{Poisson}(\lambda)$, $\mathrm{Uniform}[a,b]$, $\mathcal{N}(\mu,\sigma^2)$, $\mathrm{Exp}(\lambda)$, $\mathrm{Weibull}(\alpha,\beta)$, $\mathrm{Erlang}(m,\beta)$, $\mathrm{Gamma}(\alpha,\lambda)$, $\mathrm{Beta}(\alpha,\beta)$.

- Instead of using MoM, another convenient way to estimate the parameters is using sample mean $\bar{X}$ and sample variance S^2:

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = m_1,$$
$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} = \frac{\sum_{i=1}^n X_i^2 - n\bar{X}^2}{n-1} = \frac{n}{n-1}(m_2 - m_1^2),$$

to solve $\mathbb{E}[X] = \bar{X}$, and $\mathrm{Var}(X) = S^2$ (if necessary).

- Note 1: Purpose of $n-1$ in S^2 is to ensure $\mathbb{E}[S^2] = \mathrm{Var}(X)$.

- Note 2: In original MoM, we solve $\mathrm{Var}(X) = m_2 - m_1^2$.

- Revisit Example 1: $\mathrm{Gamma}(\alpha, \lambda)$.
  Recall: using MoM, $\widehat{\alpha} = \frac{m_1^2}{m_2 - m_1^2}$, $\widehat{\lambda} = \frac{m_1}{m_2 - m_1^2}$.

  Solving $\mathbb{E}[X] = \alpha/\lambda = \bar{X}$ and $\mathrm{Var}(X) = \alpha/\lambda^2 = S^2$, yields
  $$\widetilde{\lambda} = \frac{\bar{X}}{S^2}, \quad \widetilde{\alpha} = \bar{X}\widetilde{\lambda} = \frac{\bar{X}^2}{S^2}. \quad \blacksquare$$
  Note: $\widetilde{\alpha} = \frac{n-1}{n}\frac{m_1^2}{m_2 - m_1^2} = \frac{n-1}{n}\widehat{\alpha}$, $\widetilde{\lambda} = \frac{n-1}{n}\frac{m_1}{m_2 - m_1^2} = \frac{n-1}{n}\widehat{\lambda}$.

- Revisit Example 2: $\mathrm{Exp}(\lambda)$. No difference.

- Revisit Example 3: $\mathcal{N}(\mu, \sigma^2)$.
  Recall: $\widehat{\mu} = m_1 = \bar{X}$, $\widehat{\sigma}^2 = m_2 - m_1^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$.

  Letting $\mathbb{E}[X] = \bar{X}$ and $\mathrm{Var}(X) = S^2$, we have
  $$\widetilde{\mu} = \bar{X}, \quad \widetilde{\sigma}_2 = S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}. \quad \blacksquare$$

- MoM is often “quick and dirty” and is not using all the information contained within the data efficiently.
- Maximum Likelihood Estimation (MLE), by contrast, is known to be as efficient as possible.
- MLE says that the parameters should take values under which the observed data are mostly likely to occur.
- Sometimes, both MoM and MLE yield the same estimator.

- Revisit Example 2: Suppose $x_1, \ldots, x_n$ are iid observations from $\mathrm{Exp}(\lambda)$. Estimate λ using MLE.

  Solution. The pdf is $f(x) = \lambda e^{-\lambda x}$, $x \geq 0$, $\lambda > 0$. So the *likelihood* of observing the above data is

  $$L(\lambda) := \prod_{i=1}^{n} f(x_i) = \lambda^n e^{-\lambda(x_1+\cdots+x_n)}.$$

  We want to solve λ that maximizes $L(\lambda)$. To make it easier, we consider to maximize the *log likelihood*, which is equivalent:

  $$\ln(L(\lambda)) = n\ln(\lambda) - \lambda(x_1 + \cdots + x_n).$$

  Taking its derivative w.r.t. λ and setting it to zero gives the solution $\lambda^* = \frac{n}{x_1+\cdots+x_n}$. (Check it is indeed the global maximizer!) ■

- Remarks:
  - If $X_1, \ldots, X_n$ haven't been observed, $\lambda^* = n/(X_1 + \cdots + X_n)$.
  - The estimator is the same as in MoM.

- Revisit Example 1: Suppose $x_1, \dots, x_n$ are iid observations from $\mathrm{Gamma}(\alpha, \lambda)$. Estimate α and λ using MLE.

  Solution. The pdf is $f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x},\ x > 0,\ \alpha > 0, \lambda > 0.$ So the *log likelihood* of observing the above data is

$$\begin{aligned}\ln(L(\alpha,\lambda)) &= \sum_{i=1}^n \ln(f(x_i)) \\ &= n[\alpha \ln(\lambda) - \ln(\Gamma(\alpha))] + (\alpha - 1)\sum_{i=1}^n \ln(x_i) - \lambda \sum_{i=1}^n x_i.\end{aligned}$$

  To maximize $\ln(L(\alpha,\lambda))$, notice that for any value of α, the global maximizer of λ is that satisfying $\frac{\partial \ln(L(\alpha,\lambda))}{\partial \lambda} = n\alpha/\lambda - \sum_{i=1}^n x_i = 0$, which is $\lambda^*(\alpha) = \alpha/\bar{x}$. (Check this!)

  Then we need to find α is that maximizes $\ln(L(\alpha, \lambda^*(\alpha)))$. Unfortunately, this can only be done numerically. ■

- For discrete distributions, replace the pdf with pmf.

Goodness of Fit

- After a family of distributions has been selected and the parameters are determined to "best" fit the data, the next step is to evaluate how good the fitting is.
- If the goodness of fit is not good, select another candidate and repeat the previous processes, or use an empirical distribution.
- There are two types of approaches:
  - Graphical methods: histogram against fitted pdf/pmf, quantile-quantile (Q-Q) plot, etc.
  - Statistical tests: chi-square (χ^2) test, Kolmogorov-Smirnov (K-S) test, etc.
- Try more than one plot/test before making conclusion.

- Compare the shape of **histogram** of data (plotted in the same way as before) against the fitted pdf or pmf.
- For better comparison, one may consider to aline the histogram and pdf/pmf:
    - Use relative frequency (i.e., ratio) for histogram.
    - For continuous data, one may consider to re-scale the vertical axis of histogram or pdf to make them aligned.
    - Commercial softwares usually take care of that by default.

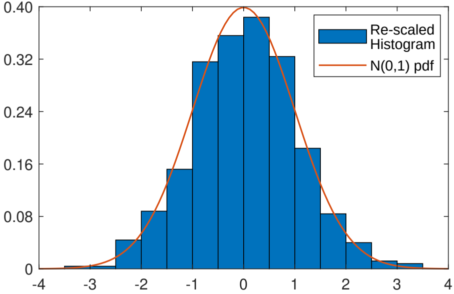


Figure: Example of Re-scaled Histogram vs. Fitted pdf

- **Quantile-Quantile** (Q-Q) **plot** compares the quantiles of the data against those of the fitted distribution.
- The q-quantile of X is that value γ such that $\mathbb{P}(X \leq \gamma) = F(\gamma) = q$, for $0 < q < 1$. When $F(x)$ has an inverse, we can write $\gamma = F^{-1}(q)$.
  - Median: 50% quantile.
  - In financial risk management, quantile of the profit-and-loss of a portfolio is also called Value-at-Risk (VaR).

- To make Q-Q plots, given the data $\{x_1, \ldots, x_n\}$ and the fitted distribution with cdf $F(x)$:
  - Order the observations from the smallest to the largest, and rename them as $y_1 \leq y_2 \leq \cdots \leq y_n$.
  - y_j is an estimate of the $(j - 0.5)/n$ quantile of X which generates the data.
  - For $X \sim F(x)$, its $(j - 0.5)/n$ quantile is $F^{-1}\left(\frac{j-0.5}{n}\right)$.
  - Q-Q plot displays $y_1, \ldots, y_n$ vs. $F^{-1}\left(\frac{1-0.5}{n}\right), \ldots, F^{-1}\left(\frac{n-0.5}{n}\right)$.
  - If the data is indeed generated from distribution $F(x)$, then

    $$y_j \approx F^{-1}\left(\frac{j - 0.5}{n}\right),$$

    so the plot will be approximately a straight line with slop 1.

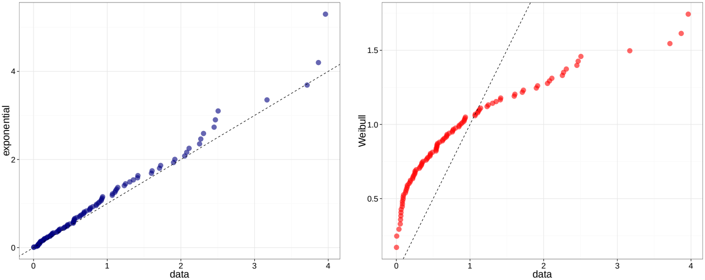


Figure: Examples of Q-Q Plot (*from* ZHANG Xiaowei)

- The observed values will never fall exactly on a straight line
- The ordered values are not independent because they are ranked. Hence, if one point lies above the line, it is likely that the next one will too.
- The values at the extremes have a much higher variance than those in the middle. So greater discrepancies can be acceptable at the extremes; linearity in the middle is much more important.

Goodness of Fit

- Graphical methods *qualitatively* measure the fitting goodness of a certain distribution with cdf $F(x)$ to the data.

- Goodness-of-fit tests are statistical hypothesis tests that *quantitatively* measure the fitting goodness.
    - Whether or not the observations $x_1, \ldots, x_n$ are an independent sample from a certain distribution with cdf $F(x)$?

- A hypothesis test is a data-based rule to decide between the null hypothesis (H_0) and the alternative hypothesis (H_1).
    - The basic idea is to assume H_0 is true, and then check if the data provide enough evidence to conclude that H_0 is not true. If yes, we **reject** H_0; otherwise, we **fail to reject** H_0.

Decision Truth	reject H_0	fail to reject H_0
H_0 is true	type I error	correct
H_1 is true	correct	type II error

Goodness of Fit

- A hypothesis test only directly controls the type I error.
  - A test with the same type I error probability but smaller type II error probability is better (*more powerful*).
  - The **level of significance** (显著水平), α, means that $\mathbb{P}(\text{type I error}) \leq \alpha$.

- The **test statistic** (检验统计量) is a statistic computed from the data.

- The p-**value** is the probability that we would observe the same value of the computed test statistic or an even more extreme value, **given** H_0 **is true**.

- We will reject H_0 if
  - p-value is smaller than some specified α, or, equivalently,
  - the computed test statistic falls in certain range (called *rejection region*), which is determined by α.

- For goodness-of-fit tests, the basic question is, "is it reasonable, statistically speaking, to assume that the observations $x_1, \ldots, x_n$ are an independent sample from the specified distribution?"
- H_0: The data come from the specified distribution
  H_1: The data do not come from the specified distribution
- Logic: Assume H_0 is true, is it likely to observe the data at hand? If the likelihood is very small (i.e., p-value is very small), then H_0 is unlikely to be true (reject H_0); otherwise, there is no enough evidence to reject H_0.

Goodness of Fit

- The **chi-square test** (χ^2 test, 卡方检验) is a more formal comparison of a histogram with the fitted pdf $f(x)$ or pmf $p(x)$.
- The procedure of chi-square test:
  1. First divide the entire range of the fitted distribution into k adjacent intervals, $[a_0, a_1), [a_1, a_2), \ldots, [a_{k-1}, a_k)$.
  2. Define

$$\begin{aligned} O_i &:= \textbf{actual} \text{ number of data points in } [a_{i-1}, a_i), \\ E_i &:= \textbf{expected} \text{ number of points in } [a_{i-1}, a_i) \text{ for fitted dist.} \\ &= n \times \mathbb{P}(a_{i-1} \leq X < a_i) \\ &= n \int_{a_{i-1}}^{a_i} f(x) \mathrm{d}x \quad \text{or} \quad n \sum_{a_{i-1} \leq x_j < a_i} p(x_j). \end{aligned}$$

  3. Compute the test statistic $R := \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$.
  4. Reject H_0 if R is too large.

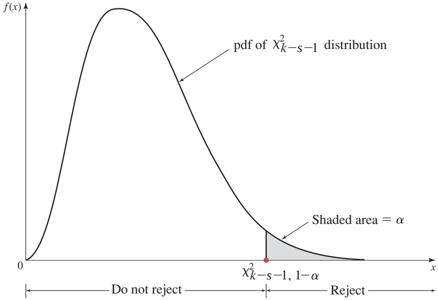
     – Reason: A large value of R indicates a poor fit, whereas a small value indicates a good fit.

     – Question: How large is too large? (i.e., what is the rejection region?)

Goodness of Fit

- View the test statistic R as a random variable.
  - Since we assume the collected data is one observed random sample from some unknown distribution, if we conduct the study multiple times, the values of the statistics will be different because the collected data will be different.
  - For current data at hand, we have already observed the value of R, which is denoted as r.
- So, the p-value for this hypothesis test is $\mathbb{P}(R \geq r)$.
- For the fitted distribution, suppose $s \geq 0$ parameters are unknown and estimated via MLE.
- If H_0 is true, then R *approximately* follows the chi-square distribution with $k - s - 1$ degrees of freedom (i.e., χ^2_{k-s-1} distribution) when sample size n is large.

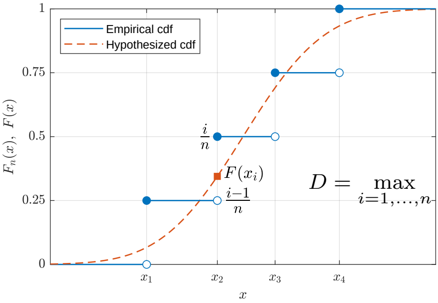
- The p-value $= \mathbb{P}(R \geq r) = \int_r^\infty f(x)\mathrm{d}x$, where $f(x)$ is the pdf of χ^2_{k-s-1} distribution.
- If we have selected some significance level α (i.e., we want to control $\mathbb{P}(\text{type I error})$ below α), then we will **reject** H_0 if
  - p-value $< \alpha$, or, equivalently,
  - $r > \chi^2_{k-s-1,1-\alpha}$, where $\chi^2_{k-s-1,1-\alpha}$ is the $(1-\alpha)$-quantile of χ^2_{k-s-1} distribution (as shown in the following figure).



- Advantage of chi-square test:
  - It can be applied to any hypothesized distribution, which makes it widely use.
- Disadvantage of chi-square test:
  - It is valid only in an asymptotic sense (large n).
  - **Major drawback**: The validity and power of chi-square test are affected by the number and size of the chosen intervals, while there is no clear prescription for such selection.
- In the absence of a definitive guideline for choosing the intervals, it's usually recommended to make E_i equal (or approximately equal) and no less than 5, for all intervals.

Goodness of Fit

- The Kolmogorov-Smirnov test (柯尔莫哥洛夫–斯米尔诺夫检验) formally compares the empirical cdf $F_n(x)$ with the cdf of the hypothesized distribution, $F(x)$.



$$D = \max_{i=1,\dots,n} \left\{ \left(\frac{i}{n} - F(x_i)\right) \bigvee \left(F(x_i) - \frac{i-1}{n}\right) \right\}.$$

Note: $x_1, \dots, x_n$ are the **sorted** data points.

- The test statistic is $D := \sup_x |F_n(x) - F(x)|$.
  - D is the largest deviation between the empirical cdf and the hypothesized cdf.
  - Since the empirical cdf is a step function, to compute D, it suffices to evaluate $|F_n(x) - F(x)|$ only at the "jump" points.

- The procedure of K-S test:
  1. Compute the test statistic D.
  2. Reject H_0 if D is too large.
     – Reason: A large value of D indicates a poor fit, whereas a small value indicates a good fit.

- For current data at hand, we have already observed the value of D, which is denoted as d. **Reject** H_0 if
  - p-value $= \mathbb{P}(D \geq d) < \alpha$, or equivalently,
  - $d > d_{n,1-\alpha}$, where $d_{n,1-\alpha}$ is the $(1-\alpha)$-quantile of D.

- Unfortunately, the distribution of D (thus the $d_{n,1-\alpha}$ and p-value) depends on how the hypothesized distribution $F(x)$ was specified:

Case 1: No parameter of $F(x)$ is estimated in any way;

Case 2: $F(x)$ is cdf of distribution such as normal, exponential, or Weibull, and parameters are estimated via MLE (except for normal σ^2, which is estimated by S^2).

- Advantage of K-S test:
  - It does not require us to group the data in any way, so no information is lost and no troublesome selection is faced.
  - It is valid (exactly) for any sample size, whereas chi-square test is valid only in an asymptotic sense.
  - It tends to be more powerful than chi-square test.
- Disadvantage of K-S test:
  - Its range of applicability is more limited than that for chi-square test.
  - When applicable, its computation of p-value and rejection region is usually complicated.
- K-S test is relatively more convenient to be used in a case where the hypothesized distribution is continuous and no parameter is estimated. For example:
  - Test random number generators.
  - Test a Poisson process (more details later).

- Comments on p-value:
  - p-value can be viewed as a measure of fit: a large p-value tends to indicate a good fit, while a small p-value suggests a poor fit.
  - We could try several families of distributions and select the one with the largest p-value.
  - **However**, p-value is just a summary measure. It says little or nothing about where the lack of fit occurs (body? left tail? right tail?).
  - Different statistical tests may give different p-values.
  - Whether or not you reject H_0 also depends on the significance level α chosen by yourself.

- Comments on general goodness-of-fit tests:
  - If very little data are available, then a goodness-of-fit test is unlikely to reject any candidate distribution.

    – No enough evidence to reject H_0.
  - If a lot of data are available, then a goodness-of-fit test is likely to reject all candidate distributions.

    – H_0 is virtually never exactly true, and even a tiny departure from the hypothesized distribution will be detected for large n.
  - Do not have blind faith in goodness-of-fit tests!

    – Failing to reject a candidate distribution should be taken as only **one piece of evidence** in favor of that choice.

    – Rejecting a candidate distribution should be taken as only **one piece of evidence** against the choice.

- Graphical Methods vs. Statistical Tests
  - Graphical methods *qualitatively* measure the fitting goodness, while statistical tests *quantitatively* measure the fitting goodness.
  - Statistical tests measure the lack of fit by summary statistics, while graphical methods show where the lack of fit occurs (body, left tail, right tail) and allow users to decide whether it is important.
  - Statistical tests may accept the fit, but plots may suggest otherwise, especially when the number of observations is small.
- Always combine statistical test results with graphical analysis.
- When no model fits the data satisfactorily, we may end up with the empirical distribution.

Goodness of Fit

- Many softwares do have a “best fit” option (or button).
  - It recommends the “best” distribution in its library based on summary measure like the p-value (and perhaps other factors such as discrete or continuous, bounded or unbounded).
- Always keep the following in mind when using such an option:
  - The software might know nothing about the physical basis of the data.
  - Automated best-fit procedures tend to choose the more flexible distributions (gamma over Erlang, Weibull over exponential). But, close conformance to the data does not always lead to the most appropriate input model (overfitting).
  - The limitation of summary measure like p-value.
  - View the automated distribution selection as one suggestion, inspect it using graphical methods, and remember that *the final choice is yours*.

Goodness of Fit

- All the graphical methods and statistical tests can be used to check the uniformity of a random number generator (RNG).
  1. Generate a sequences of numbers (as many as you want) using the RNG.
  2. Check if $\mathrm{Uniform}[0,1]$ fits the data well enough.

- Poisson-Process Test
  - Suppose we observe an arrival process for a time interval $[0,T]$, where T is a constant decided before we start our observation.
  - We see n arrivals during $[0,T]$ with arrival times $s_1, s_2, \ldots, s_n$, and want to check if Poisson process is a good model for it.
  - Method 1: Test if an exponential distribution can fit the data $\{s_1, s_2 - s_1, \ldots, s_n - s_{n-1}\}$ well.
  - Method 2: Test if $\mathrm{Uniform}[0,T]$ can fit the data $\{s_1, \ldots, s_n\}$ well. (**Why?**)

    – Given $N(T) = n$, the n arrival times $S_1, \ldots, S_n$ have the same distribution as n independent RVs from $\mathrm{Uniform}[0,T]$ that are **sorted**.

An Illustrative Example

- Suppose we want to build a statistical model for the life time (i.e., time to failure) of a electronic component at 1.5 times the nominal voltage.

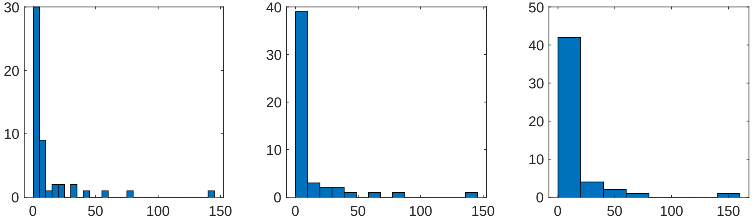
1. Data Collection.
   - Perform life tests on a random sample ($n = 50$) of electronic components and record their lifetime, in days:

79.919	3.081	0.062	1.961	5.845
3.027	6.505	0.021	0.013	0.123
6.769	59.899	1.192	34.760	5.009
18.387	0.141	43.565	24.420	0.433
144.695	2.663	17.967	0.091	9.003
0.941	0.878	3.371	2.157	7.579
0.624	5.380	3.148	7.078	23.960
0.590	1.928	0.300	0.002	0.543
7.004	31.764	1.005	1.147	0.219
3.217	14.382	1.008	2.336	4.562

An Illustrative Example

❷ Identifying Distribution.

- Lifetime, although recorded to three-decimal-place accuracy, is a positive continuous variable.
- For this life time, naturally, exponential and Weibull are considered.
- Plot the histogram.

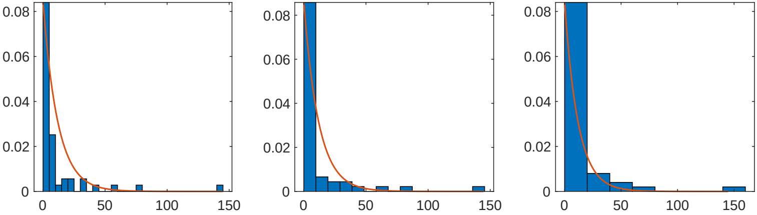


- We decide to first try exponential distribution family $\mathrm{Exp}(\lambda)$.

An Illustrative Example

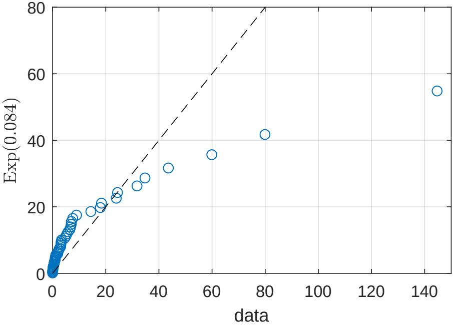
3. Distribution Fitting.
   - Recall Example 2, MoM (or its variation) and MLE yield the same estimator for λ, which is $\widehat{\lambda} = \frac{n}{X_1+\cdots+X_n}$.
   - Plug the data in, and the estimate of λ is 0.084.

4. Goodness of Fit.
   - Re-scaled histogram vs. pdf of $\mathrm{Exp}(0.084)$.



An Illustrative Example

4. Goodness of Fit.
   - Q-Q plot.

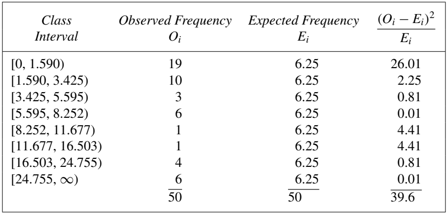


An Illustrative Example

4 Goodness of Fit.

- Chi-square test (H_0: The data come from $\mathrm{Exp}(0.084)$). Number of estimated parameters is $s = 1$.

  Choose intervals (make E_i equal).



Class Interval	*Observed Frequency* O_i	*Expected Frequency* E_i	$\frac{(O_i - E_i)^2}{E_i}$
[0, 1.590)	19	6.25	26.01
[1.590, 3.425)	10	6.25	2.25
[3.425, 5.595)	3	6.25	0.81
[5.595, 8.252)	6	6.25	0.01
[8.252, 11.677)	1	6.25	4.41
[11.677, 16.503)	1	6.25	4.41
[16.503, 24.755)	4	6.25	0.81
[24.755, ∞)	6	6.25	0.01
	50	50	39.6

  Number of intervals is $k = 8$.

  Compute test statistic $r = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = 39.6$.

  Note that $R \sim \chi^2_{k-s-1}$ distribution $= \chi^2_6$ distribution.

  So, p-value $= \mathbb{P}(R \geq r) = \mathbb{P}(R \geq 39.6) = 5 \times 10^{-7}$.

  Hence, at almost any practical level of significance, e.g., $\alpha = 0.05$, $\alpha = 0.01$, we will reject H_0.