Input Distributions

Reading: Chapter 6 in Law

Peter J. Haas

CS 590M: Simulation
Spring Semester 2020

Overview

To specify a simulation model, we need to define clock-setting "input" sequences

Examples:

- Interarrival sequences
- Processing time sequences for a production system
- Asset-value sequence for a financial model

Even if we assume i.i.d. sequences for simplicity...

- What type of distribution should we use (gamma, Weibull, normal, exponential,...)?
- Given a type of probability distribution (i.e., a "distribution family") what parameter values should we use?

Two approaches: probability theory and historical data

Theoretical Justification for Normal Random Variables

Suppose that X can be expressed as a sum of random variables: $X = Y_1 + Y_2 + \cdots + Y_n$

In great generality, versions of CLT imply that $X \overset{\mathbf{D}}{\sim} N(\mu, \sigma^2)$ (approx.) for large n, where $\mu = E[X]$ and $\sigma^2 = \mathrm{Var}[X]$

- Y_i's need not be i.i.d., just not "too dependent" and not "too non-identical"

Q: Examples where CLT breaks down?

$X_1 = X_2 = \cdots$ (too dependent)

$X_1 \overset{D}{=} U[10^{10}, 10^{20}]$ and $X_i \overset{D}{=} U[10^{-20}, 10^{-10}]$ for $i > 1$ (too heterogeneous)

Moral:

If X is the sum of a large number of other random quantities, then X can be approx. modeled as a normal random variable.

Theoretical Justification for Lognormal Random Variables

Suppose that X can be expressed as a product of random variables: $X = Z_1 Z_2 \cdots Z_n$

- Example: Value of a financial asset $Z_2 = Z_1(1+r_1)$, $Z_3 = Z_2(1+r_2)$, ...

Then $\log(X) = Y_1 + Y_2 + \cdots + Y_n$, **where** $Y_i = \log(Z_i)$

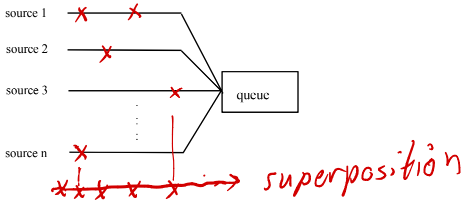
By prior discussion, $\log(X) \overset{\mathbf{D}}{\sim} N(\mu, \sigma^2)$ **(approx.) for large** n**, i.e.** $X \overset{\mathbf{D}}{\sim} \exp\big(N(0,1)\big)$**, so that** X **is approximately lognormally distributed**

Moral:

If X is the product of a large number of other random quantities, then X can be approx. modeled as a lognormal random variable.

Theoretical Justification for Poisson Arrival Process

Suppose that the arrival process is a superposition of arrivals from a variety of statistically independent sources



The Palm-Khintchine Theorem says that superposition of n i.i.d. sources looks $\approx$ Poisson as n becomes large

- Can relax i.i.d. assumption

Moral:

Poisson process often a reasonable model of arrival processes.

(But beware of long-range dependence)

Theoretical Justification for Weibull Distribution

Suppose that X can be expressed as a minimum of nonnegative random variables: $X = \min_{1 \leq i \leq n} Y_i$

- Example: Lifetime of a complex system

Extreme-value theory (Gnedenko's Theorem) says that if Y_i's are i.i.d. then X has approximately a Weibull distribution when n is large: $P(X \leq x) = 1 - e^{-(\lambda x)^{\alpha}}$

- α is the *shape* parameter and λ is the *scale* parameter

Moral:

If X is the the lifetime of a complicated component, then X can be approx. modeled as a Weibull random variable.

Controlling the Variability

Squared coefficient of variation: $\rho^2(X) = \frac{\text{Var}(X)}{(E[X])^2}$

Case 1: $X \overset{D}{\sim} \exp(\lambda)$

$\rho^2(X) = 1$

Case 2: $X \overset{D}{\sim} \text{gamma}(\lambda, \alpha)$ $\quad f_X(x) = \lambda e^{-\lambda x}(\lambda x)^{\alpha-1}/\Gamma(\alpha)$

$\rho^2(X) = \frac{\alpha/\lambda^2}{(\alpha/\lambda)^2} = \frac{1}{\alpha}$

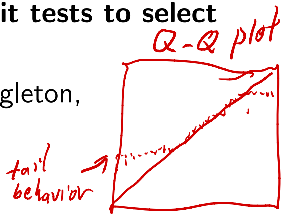
Three scenarios to play with:

- $\alpha > 1$: less variable than exponential
- $\alpha = 1$: exponential
- $\alpha < 1$: more variable than exponential

Goodness-of-Fit Software

GoF software applies a set of goodness-of-fit tests to select distribution family with "best fit" to data

- chi-square, Kolmogorov-Smirnov, Epps-Singleton, Anderson-Darling, . . .



GoF software must be used with caution

- Low power: Poor discrimination between different distributions
- Sequential testing: Test properties mathematically ill-defined
- Discourages sensitivity analysis: Unwary users stop with "best"
- Can obscure non-i.i.d. features: e.g., trends, autocorrelation
- Fails on big data: All test fail on real-world datasets
- Over-reliance on summary statistics: Should also plot data
  - Ex: Q-Q plots [Law, p. 339-344] better indicate departures from candidate distribution (lower/upper, heavy/light tails)

Feature Matching to Data

Pragmatic approach: match key features in the data

Ex: Use gamma dist'n, match first two empirical moments

- Hence match the empirical coefficient of variation (see below)
- Note: nothing about "process physics" implies gamma, we use it for it's convenence and flexibility in modeling a range of variability

Can use even more flexible distribution families

- Ex: Johnson translation system (4 parameters)
- Ex: Generalized lambda distribution (4 parameters)
- Useful when extreme values not important to simulation [Nelson (2013), pp. 113–116]

Estimating Parameters: Maximum Likelihood Method

Ex: Geometric distribution

- Number of failures before first success in Bernoulli trials, success probability $= p$
- $P\{X = k\} = (1-p)^k p$ for k ≥ 0 $\quad X_1, X_2, X_3, X_4$
- How to estimate p given four observations of X? $X = 3, 5, 2, 8$
- Given a value of p, likelihood for obs1: $(1-p)^3 p$ obs2: $(1-p)^5 p$ obs3: $(1-p)^2 p$ obs4: $(1-p)^8 p$
- Joint likelihood $L(p)$ for all four observations: $(1-p)^{18} p^4$
- Choose estimator $\hat{p}$ to maximize $L(p)$ ("best explains data")
- Equivalently, maximize $\tilde{L}(p) = \log\big(L(p)\big)$: $18 \cdot \log(1-p) + 4 \log(p)$

$\tilde{L}'(p) = 0 \Rightarrow \hat{p} = \frac{4}{22} = \frac{1}{1+\bar{X}_4}$ $\quad \bar{X}_4 = \text{avg}(X_1, X_2, X_3, X_4)$

Maximum Likelihood, Continued

pdf: $f(x) = \lambda e^{-\lambda x}$

Ex: Poisson arrival process to a queue

- Exponential interarrivals: 3.0, 1.0, 4.0, 3.0, 8.0
- Goal: estimate λ
- For a continuous dist'n, likelihood of an observation = pdf [Law, Problem 6.26]
- Given a value of λ, likelihoods for observations: $\lambda e^{-3\lambda}, \ldots, \lambda e^{-8\lambda}$
- Joint likelihood: $L(\lambda) = \lambda^5 e^{-19\lambda}$
- Joint log-likelihood: $\tilde{L}(\lambda) = 5 \log(\lambda) - 19\lambda$
- Estimate $\hat{\lambda} =$ $\tilde{L}'(\lambda) = 0$

$$\frac{5}{\hat{\lambda}} = 19 \Rightarrow \hat{\lambda} = 5/19$$

Q: Why is this a reasonable estimator? $X \sim \exp(\lambda)$

$$\hat{\lambda} = \frac{1}{\bar{X}_5} \qquad E[X] = \frac{1}{\lambda} \rightarrow \lambda = \frac{1}{E[X]}$$

Maximum Likelihood, Continued

General setup (continuous i.i.d. case)

- Given $X_1, \ldots, X_n$ i.i.d. samples from pdf $f(\,\cdot\,; \alpha_1, \ldots, \alpha_k)$
- MLE's $\hat{\alpha}_1, \ldots, \hat{\alpha}_k$ maximize the likelihood function

$$L_n(\alpha_1, \ldots, \alpha_k) = \prod_{i=1}^{n} f(X_i; \alpha_1, \ldots, \alpha_k)$$

or, equivalently, the log-likelihood function

$$\tilde{L}_n(\alpha_1, \ldots, \alpha_k) = \sum_{i=1}^{n} \log\big(f(X_i; \alpha_1, \ldots, \alpha_k)\big)$$

- For discrete case use pmf instead of pdf

Maximum Likelihood, Continued

Maximizing the likelihood function

- Simple case: solve

$$\frac{\partial \tilde{L}_n(\hat{\alpha}_1, \ldots, \hat{\alpha}_k)}{\partial \alpha_i} = 0, \qquad \text{for } i = 1, 2, \ldots, k$$

- Harder cases: maximum occurs on boundary, constraints (Kuhn-Tucker conditions)
- Hardest cases (typical in practice): solve numerically

Why bother?

- MLEs maximize asymptotic statistical efficiency
- For large n, MLEs "squeeze maximal info from sample" (i.e., smallest variance ⇒ narrowest confidence intervals)

Estimating Parameters: Method of Moments

Idea: equate k sample moments to k true moments & solve

Example: Exponential distribution

- $k = 1$ and $E[X] = 1/\lambda$
- So equate first moments:

$$\bar{X}_n = 1/\hat{\lambda} \quad \Rightarrow \quad \hat{\lambda} = 1/\bar{X}_n$$

Example: Gamma distribution

- $k = 2$, $E[X] = \alpha/\lambda$, and $\mathrm{Var}[X] = \alpha/\lambda^2$
- equate moments:

$$\bar{X}_n = \hat{\alpha}/\hat{\lambda} \text{ and } s_n^2 = \hat{\alpha}/\hat{\lambda}^2 \quad \Rightarrow \quad \hat{\alpha} = \bar{X}_n^2/s_n^2 \text{ and } \hat{\lambda} = \bar{X}_n/s_n^2$$

Can match other stats, e.g., quantiles [Nelson 2013, p. 115]

Bayesian Parameter Estimation

View unknown parameter θ as a random variable with prior distribution $f(\theta)$

- Encapsulates prior knowledge about θ before seeing data
- Ex: If we know that $\theta \in [5, 10]$ set $f = \text{Uniform}(5,10)$

After seeing data Y, use Bayes' Rule to compute posterior

$$f(\theta \mid Y) = f(Y \mid \theta)\, f(\theta)/c$$

where $f(Y \mid \theta)$ is the likelihood of Y under θ and c is normalizing constant

Estimate θ by mean (or mode) of posterior

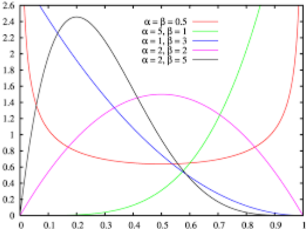
- $f(\theta \mid Y)$ usually very complex
- Use Markov Chain Monte Carlo (MCMC) to estimate mean—see HW 2

Bayesian Parameter Estimation, Continued

Simple example (see refresher handout)

- Goal: Estimate success probability for Bernoulli distribution
- Assume a Beta(α, β) prior on θ
- Observe n Bernoulli trials with Y successes
- $Y \mid \theta$ has Binomial(n, θ) likelihood distribution
- Posterior, given $Y = y$, is Beta$(\alpha + y, \beta + n - y)$ (Beta is "conjugate prior" to binomial)
- $\hat{\theta}$ = mean of posterior $= \dfrac{\alpha + y}{\alpha + \beta + n}$

Note: MLE is a special case of Bayesian estimate (uniform prior + MAP)



Other Approaches to Input Distributions

Trace-driven simulation: Use actual measured values

- Can't generalize simulation results beyond data
- Can't assess variability
- Hard to do "what if' and sensitivity analyses
- Privacy concerns
- Bootstrapping can help [Efron and Tibshirani book]

Empirical Distributions

- $\hat{F}_n(x) = (1/n)(\#\text{ obs} \leq x)$ or use histogram [Law 6.4.2]
- Truncation effect: problems with tail probabilities
- No smoothing $\Rightarrow$ sensitive to data anomalies

Modified Empirical Distributions

- Smoothed empirical distribution with exponential tail [Bratley et al., pp. 131–132]

- Bezier distributions [Law, Sec. 6.9]

Other Approaches to Input Distributions, Continued

Generative neural networks
(current research with Cen Wang)

- Good for situations with lots of historical data
- Automated distribution fitting
- Can deal with multimodal, non-i.i.d. arrival-process distributions
- Automatic generation of samples via fast matrix multiplications