CHAPTER 6

Selecting Input Probability Distributions

6.1 Introduction

Part of modeling—what input probability distributions to use as input to simulation for:

Interarrival times

Service/machining times

Demand/batch sizes

Machine up/down times

Inappropriate input distribution(s) can lead to incorrect output, bad decisions

Usually, have observed data on input quantities—options for use:

Use	**Pros**	**Cons**
Trace-driven Use actual data values to drive simulation	Valid *vis à vis* real world Direct	Not generalizable
Empirical distribution Use data values to define a "connect-the-dots" distribution (several specific ways)	Fairly valid Simple Fairly direct	May limit range of generated variates (depending on form)
Fitted "standard" distribution Use data to fit a classical distribution (exponential, uniform, Poisson, etc.)	Generalizable—fills in "holes" in data	May not be valid May be difficult

6.2 Useful Probability Distributions

Many distributions exist, found useful for simulation input modeling

6.2.1 Parameterization of Continuous Distributions

Alternative ways to parameterize most distributions; not consistently done

Typically, parameters can be classified as one of:

- *Location parameter* g (also called *shift parameter*): specifies an abscissa (x axis) location point of a distribution’s range of values, often some kind of midpoint of the distribution.
    - Example: m for normal distribution
    - As g changes, distribution just shifts left or right without changing its spread or shape
    - If X has location parameter 0, then $X + g$ has location parameter g

- *Scale parameter* b: determines scale, or units of measurement, or spread, of a distribution.
    - Examples: s for normal distribution, b for exponential distribution
    - As b changes, the distribution is compressed or expanded without changing its shape
    - If X has scale parameter 1, then bX has scale parameter b

- *Shape parameter* a: determines, separately from location and scale, the basic form or shape of a distribution
    - Examples: normal and exponential distribution do not have shape parameter; a for gamma and Weibull distributions
    - May have more than one shape parameter (beta distribution has two shape parameters)
    - Change in shape parameter(s) alters distribution’s shape more fundamentally than changes in scale or location parameters

6.2.2 Continuous Distributions

Compendium of 13 continuous distributions

Possible applications
Density and distribution functions (where applicable)
Parameter definitions and ranges
Range of possible values
Mean, variance, mode
Maximum-likelihood estimator formula or method
General comments, including relationships to other distributions
Plots of densities

6.2.3 Discrete Distributions

Compendium of 6 discrete distributions, with similar information as for continuous distributions

6.2.4 Empirical Distributions

Use observed data themselves to specify directly an empirical distribution; maybe no standard distribution fits the data adequately

There are many different ways to specify empirical distributions, resulting in different distributions with different properties

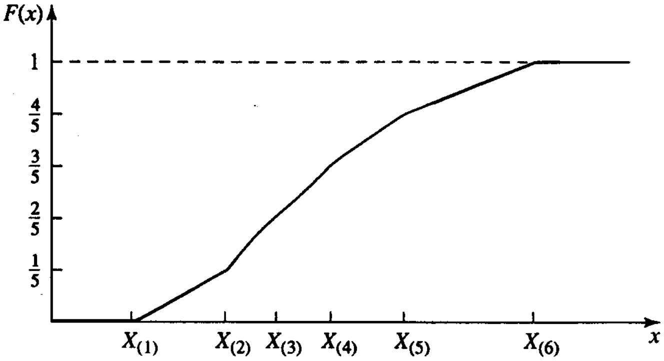
Continuous Empirical Distributions

If original individual data points are available (i.e., data are not grouped)

Sort data $X_1, X_2, \ldots, X_n$ into increasing order: $X_{(i)}$ is ith smallest

Define $F(X_{(i)}) = (i - 1)/(n - 1)$, approximately (for large n) the proportion of the data less than $X_{(i)}$, and interpolate linearly between observed data points:

$$F(x) = \begin{cases} 0 & \text{if } x < X_{(1)} \\ \dfrac{i-1}{n-1} + \dfrac{x - X_{(i)}}{(n-1)(X_{(i+1)} - X_{(i)})} & \text{if } X_{(i)} \le x < X_{(i+1)} \text{ for } i = 1, 2, \ldots, n-1 \\ 1 & \text{if } X_{(n)} \le x \end{cases}$$



Rises most steeply over regions where observations are dense, as desired.

Potential disadvantages

- Generated data will be within range of observed data
- Expected value of this distribution is not the sample mean

Other ways to define continuous empirical distributions, including putting an exponential tail on the right to make the range infinite on the right

If only grouped data are available

Don’t know individual data values, but counts of observations in adjacent intervals

Define empirical distribution function $G(x)$ with properties similar to $F(x)$ above for individual data points (details in text)

Discrete Empirical Distributions

If original individual data points are available (i.e., data are not grouped)

For each possible value x, define $p(x)$ = proportion of the data values that are equal to x

If only grouped data are available

Define a probability mass function such that the sum of the $p(x)$’s for the x’s in an interval is equal to the proportion of the data in that interval

Allocation of $p(x)$’s for x’s in an interval is arbitrary

6.3 Techniques for Assessing Sample Independence

Most methods to specify input distributions assume observed data X_1, X_2, ..., X_n are and independent (random) sample from some underlying distribution

If not, most methods are invalid

Need a way to check data empirically for independence

Heuristic plots vs. formal statistical tests for independence

Correlation plot: If data are observed in a time sequence, compute sample correlation $\hat{r}_j$ (see Sec. 4.4 for formula) and plot as a function of the lag j

If data are independent then the correlations should be near zero for all lags

Keep in mind that these are just estimates

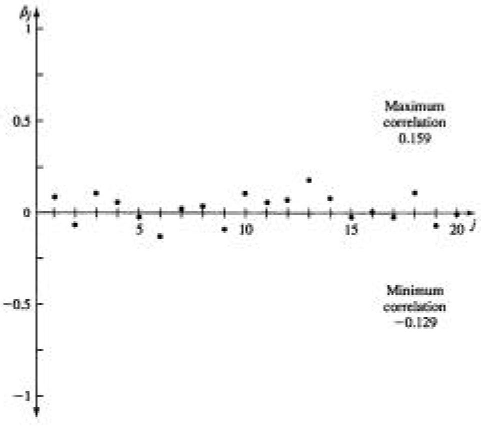
Scatter diagram: Plot pairs (X_i, X_{i+1})

If data are independent the pairs should be scattered randomly

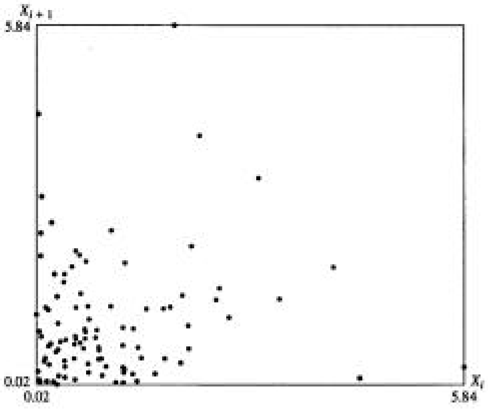
If data are positively (negatively) correlated the pairs will lie along a positively (negatively) sloping line

Independent draws from expo(1) distribution (independent by construction):

Correlation plot

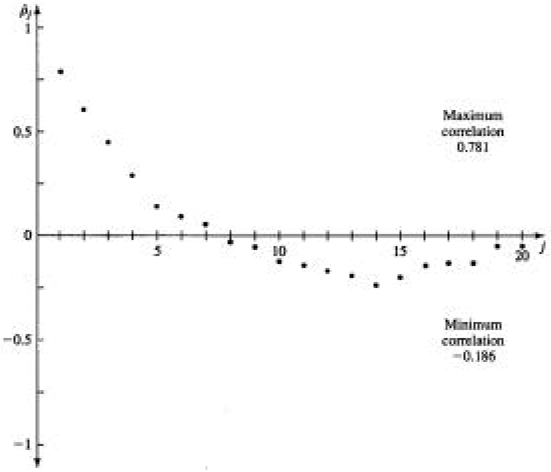


Scatter diagram

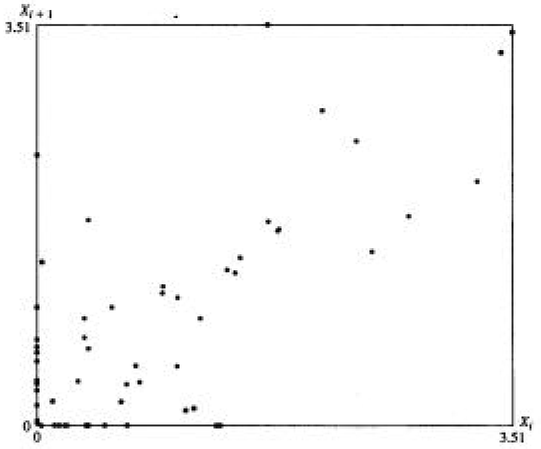


Delays in queue from *M*/*M*/1 queue with utilization factor r = 0.8 (positively correlated):

Correlation plot



Scatter diagram



Formal statistical tests for independence:

Nonparametric tests: rank von Neumann ratio

Runs tests

6.4 Activity I: Hypothesizing Families of Distributions

First, need to decide what *form* or *family* to use—exponential, gamma, or what?

Later, need to *estimate* parameters and *assess* goodness of fit

Sometimes have some *prior knowledge* of random variable's role in simulation

- Requires no data

- Use theoretical knowledge of random variable's role in simulation

- Seldom have enough prior knowledge to specify a distribution completely; exceptions:
  - Arrivals one-at-a-time, constant mean rate, independent: exponential interarrival times
  - Sum of many independent pieces: normal
  - Product of many independent pieces: lognormal

- Often use prior knowledge to *rule out* distributions on basis of *range*:
  - Service times: *not* normal (normal range always goes negative)

- Still should be supported by data (e.g., for parameter-value estimation)

6.4.1 Summary Statistics

Compare simple sample statistics with theoretical population versions for some distributions to get a hint

Bear in mind that we get only estimates subject to uncertainty

If sample mean $\bar{X}(n)$ and sample median $\hat{x}_{0.5}(n)$ are close, suggests a symmetric distribution

Coefficient of variation of a distribution: cv = s/m; estimate via $\widehat{\mathrm{cv}} = S(n) / \bar{X}(n)$; sometimes useful for discriminating between continuous distributions

cv < 1 suggests gamma or Weibull with shape parameter a < 1

cv = 1 suggests exponential

cv > 1 suggests gamma or Weibull with shape parameter a > 1

Lexis ratio of a distribution: $t = s^2/m$; estimate via $\hat{t} = S^2(n) / \bar{X}(n)$; sometimes useful for discriminating between discrete distributions

t < 1 suggests binomial

t = 1 suggests Poisson

t > 1 suggests negative binomial or geometric

Other summary statistics: range, skewness, kurtosis

6.4.2 Histograms

Continuous Data Set

Basically an unbiased estimate of $\Delta b\ f(x)$, where $f(x)$ is the true (unknown) underlying density of the observed data and Δb is a constant

Break range of data into k intervals of width Δb each

k, Δb are basically trial and error

One rule of thumb, *Sturges's rule*: $k = \lfloor 1 + \log_2 n \rfloor = \lfloor 1 + 3.332 \log_{10} n \rfloor$

Compute proportion h_j of data falling in jth interval; plot a constant of height h_j above the jth interval

Shape of plot should resemble density of underlying distribution; compare shape of histogram to density shapes in Sec. 6.2.2

Discrete Data Set

Basically an unbiased estimate of the (unknown) underlying probability mass function of the data

For each possible value x_j that can be assumed by the data, let h_j be the proportion of the data that are equal to x_j; plot a bar of height h_j above x_j

Shape of plot should resemble mass function of underlying distribution; compare shape of histogram to mass-function shapes in Sec. 6.2.3

Multimodal Data

Histogram might have multiple local modes, rather than just one; no single "standard" distribution adequately represents this

Possibility: data can be separated on some context-dependent basis (e.g., observed machine downtimes are classified as minor vs. major)

Separate data on this basis, fit separately, recombine as a mixture (details in text)

6.4.3 Quantile Summaries and Box Plots

Quantile Summaries

Numerical synopsis of sample quantiles useful for detecting whether underlying density or mass function is symmetric or skewed one way or the other

Definition of quantiles: Suppose the CDF $F(x)$ is continuous and strictly increasing whenever $0 < F(x) < 1$, and let q be strictly between 0 and 1. Then the *q-quantile* of $F(x)$ is the number x_q such that $F(x_q) = q$. If F^{-1} is the inverse of F, then $x_q = F^{-1}(q)$

$q = 0.5$: median

$q = 0.25$ or 0.75: quartiles

$q = 0.125, 0.875$: octiles

$q = 0, 1$: extremes

Quantile summary: List median, average of quartiles, average of octiles, and avg. of extremes

If distribution is symmetric, then median, avg. of quartiles, avg. of octiles, and avg. of extremes should be approximately equal

If distribution is skewed right, then

median < avg. of quartiles < avg. of octiles < avg. of extremes

If distribution is skewed left, then

median > avg. of quartiles > avg. of octiles > avg. of extremes

Box Plots

Graphical display of quantile summary

On horizontal axis, plot median, extremes, octiles, and a box ending at quartiles

Symmetry or asymmetry of plot indicates symmetry or skewness of distribution

Hypothesizing a Family of Distributions: Example with Continuous Data

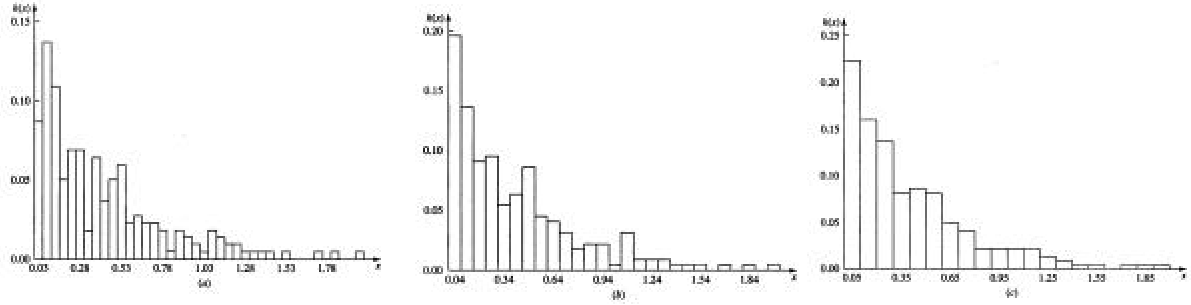
Sample of n = 219 interarrival times of cars to a drive-up bank over a 90-minute peak-load period

Number of cars arriving in each of the six 15-minute periods was approximately equal, suggesting stationarity of arrival rate

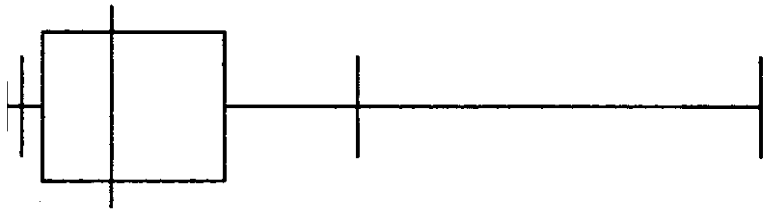
Sample mean = 0.399 (all times in minutes) > median = 0.270, skewness = +1.458, all suggesting right skewness

cv = 0.953, close to 1, suggesting exponential

Histograms (for different choices of interval width Δb) suggest exponential:



Box plot is consistent with exponential:



Hypothesizing a Family of Distributions: Example with Discrete Data

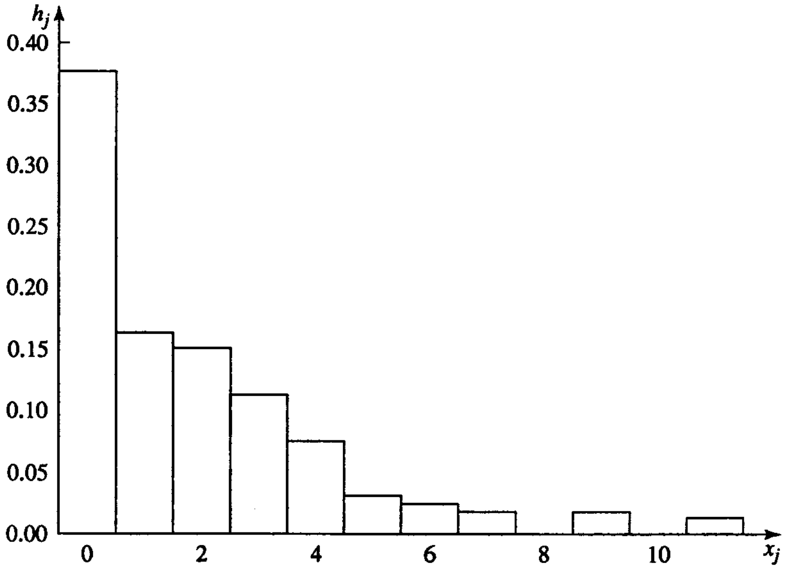
Sample of n = 156 observations on number of items demanded per week from an inventory over a three-year period

Range 0 through 11

Sample mean = 1.891 > median = 1.00, skewness = +1.655, all suggesting right skewness

Lexis ratio = 5.285/1.891 = 2.795 > 1, suggesting negative binomial or geometric (special case of negative binomial)

Histogram suggests geometric:



6.5 Activity II: Estimation of Parameters

Have: Hypothesized distribution

Need: Numerical estimates of its parameter(s)—this constitutes the “fit”

Many methods to estimate distribution parameters

- Method of moments
- Unbiased
- Least squares
- Maximum likelihood (MLE)

In some sense, MLE is the preferred method for our purposes

- Good statistical properties
- Somewhat justifies chi-square goodness-of-fit test
- Intuitive
- Allows estimates of error in the parameters—sensitivity analysis

Idea for MLEs:

- Have observed sample X_1, X_2, ..., X_n
- Came from some true (unknown) parameter(s) of the distribution form
- Pick the parameter(s) that make it most likely that you *would* get what you *did* get (or *close* to what you got in the continuous case)
- An *optimization* (mathematical-programming) problem, often messy

MLEs for Discrete Distributions

Have hypothesized family with PMF $p_q\,(x_j)$ =$P_q(X = x_j)$

Single (for now) unknown parameter q to be estimated

For any trial value of q, the probability of getting the already-observed sample is

$$\begin{aligned} P(\text{Getting } X_1, X_2, ..., X_n) &= P(X_1)P(X_2)\cdots P(X_n) \\ &= P(X = X_1)P(X = X_2)\cdots P(X = X_n) \\ &= \underbrace{p_q(X_1)\,p_q(X_2)\cdots p_q(X_n)}_{\text{Likelihood function} L(q)} \end{aligned}$$

Task: Find the (legal) value of q that makes $L(q)$ as big as it can be

How?: Differential calculus, take logarithm (turns products into sums), nonlinear programming methods, tabled values, staring at it, ...

MLEs for Continuous Distributions

Change “getting” above to “getting close to” for motivation (see Prob. 6.26)

Wind up just replacing PMF p_q by density f_q and proceed the same way

MLEs for Multiple-Parameter Distributions

Same idea, but have optimization problem in dimensionality of number of parameters to be estimated

MLEs and Confidence Intervals on Distribution Parameters

Have MLE estimate $\hat{q}$ of q

Would also like a confidence interval on q for sensitivity analysis of simulation output to parameter

Asymptotic normality property of MLEs:

$$\frac{\hat{q}-q}{\sqrt{d(\hat{q})/n}} \xrightarrow{D} \text{N}(0,1) \text{ as } n \to \infty \text{, where } d(q) = -\frac{n}{E\left[\frac{d^2}{dq^2} \ln L(q)\right]}$$

Thus, by the usual confidence-interval manipulations, an approximate 100(1–a)% confidence interval for q is

$$\hat{q} \pm z_{1-a/2}\sqrt{\frac{d(\hat{q})}{n}}$$

where $z_{1-\alpha/2}$ is the 1 – a/2 critical point of N(0, 1)

Use in simulation:

Question: Is the estimate $\hat{q}$ of q good enough?

Approach:

Get c.i. on q as above

Run simulation with input parameter set at left, then right end

If simulation output changes significantly, then need better $\hat{q}$

If not, this $\hat{q}$ is good enough

Example of Continuous MLE: Interarrival-Time Data for Drive-Up Bank

Hypothesized exponential family: density function is $f_{b}(x) = \begin{cases} \frac{1}{b} e^{-x/b} \text{ if } x > 0 \\ \text{Otherwise} \end{cases}$

Likelihood function is

$$L(b) = \left(\frac{1}{b} e^{-X_1/b} \right)\left(\frac{1}{b} e^{-X_2/b} \right) \cdots \left(\frac{1}{b} e^{-X_n/b} \right) = b^{-n} \exp\left(-\frac{1}{b} \sum_{i=1}^{n} X_i \right)$$

Want value of b that maximizes $L(b)$ over all $b > 0$

Equivalent (and easier) to maximize the *log-likelihood function* $l(b) = \ln L(b)$ since ln is a monotonically increasing function

In this case, $l(b) = -n \ln b - \frac{1}{b} \sum_{i=1}^{n} X_i$, which can be maximized by simple differential calculus:

Set $\frac{dl}{db} = \frac{-n}{b} + \frac{1}{b^2} \sum_{i=1}^{n} X_i = 0$ and solve for $b = \frac{\sum_{i=1}^{n} X_i}{n} = \overline{X}(n)$

Check second-order sufficient conditions for a maximizer:

$\frac{d^2 l}{db^2} = \frac{n}{b^2} - \frac{2}{b^3} \sum_{i=1}^{n} X_i$, which is negative when $b = \overline{X}(n)$ since the X_i's are positive

Thus, the MLE is $\hat{b} = \overline{X}(n) = 0.399$ from the observed sample of n = 219 points

Example of Discrete MLE: Demand-Size Data from Inventory

Hypothesized geometric family: mass function is $p_p(x) = p(1-p)^x$ for $x = 0,1,2,\ldots$

Likelihood function is $L(p) = p^n(1-p)^{\sum_{i=1}^n X_i}$

In this case, log-likelihood function is $l(p) = n \ln p + \sum_{i=1}^n X_i \ln(1-p)$, which can be maximized by simple differential calculus:

Set $\dfrac{dl}{dp} = \dfrac{n}{p} - \dfrac{\sum_{i=1}^n X_i}{1-p} = 0$ and solve for $p = \dfrac{1}{\overline{X}(n)+1}$

Check second-order sufficient conditions for a maximizer:

$\dfrac{d^2 l}{dp^2} = -\dfrac{n}{p^2} - \dfrac{\sum_{i=1}^n X_i}{(1-p)^2}$, which is negative for any valid p

So MLE is $\hat{p} = \dfrac{1}{1.891+1} = 0.346$ from the observed sample of n = 156 points

Confidence interval for true p:

$$E\left(\frac{d^2 l}{dp^2}\right) = -\frac{n}{p^2} - \frac{\sum_{i=1}^n E(X_i)}{(1-p)^2} = -\frac{n}{p^2} - \frac{n(1-p)/p}{(1-p)^2} = -\frac{n}{p^2(1-p)}$$

Thus, $d(p) = p^2(1-p)$ and for large n, an approximate 90% confidence interval for p is

$$\hat{p} \pm 1.645\sqrt{\frac{\hat{p}^2(1-\hat{p})}{n}}$$

$$0.346 \pm 1.645\sqrt{\frac{0.346^2(1-0.346)}{156}}$$

$$0.346 \pm 0.037$$

$$[0.309,\ 0.383]$$

6.6 Activity III: Determining How Representative the Fitted Distributions Are

Have: Hypothesized family, have estimated parameters
Question: Does the fitted distribution agree with the observed data?
Approaches: Heuristic and formal statistical hypothesis tests

6.6.1 Heuristic Procedures

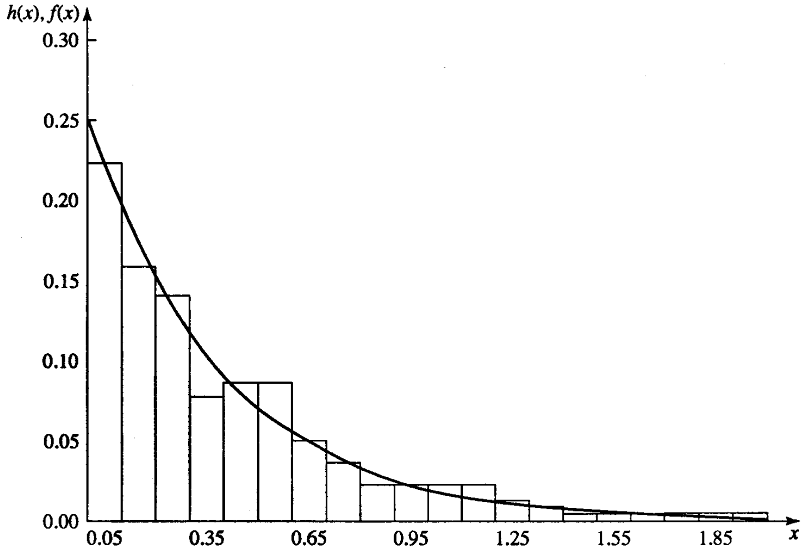
Density/Histogram Overplots and Frequency Comparisons

Continuous Data

Density/histogram overplot:

Plot $\Delta b\ \hat{f}(x)$ over the histogram $h(x)$; look for similarities (recall that the area under $h(x)$ is Δb and $\hat{f}$ is the density of the fitted distribution)

Interarrival-time data for drive-up bank and fitted exponential:



Frequency comparison

Histogram intervals interval $[b_{j-1}, b_j]$ for $j = 1, 2, ..., k$, each of width Δb

Let h_j = the *observed* proportion of data in jth interval

Let $r_j = \int_{b_{j-1}}^{b_j} \hat{f}(x)\, dx$, the *expected* proportion of data in jth interval if the fitted distribution is correct

Plot h_j and r_j together, look for similarities

Discrete Data

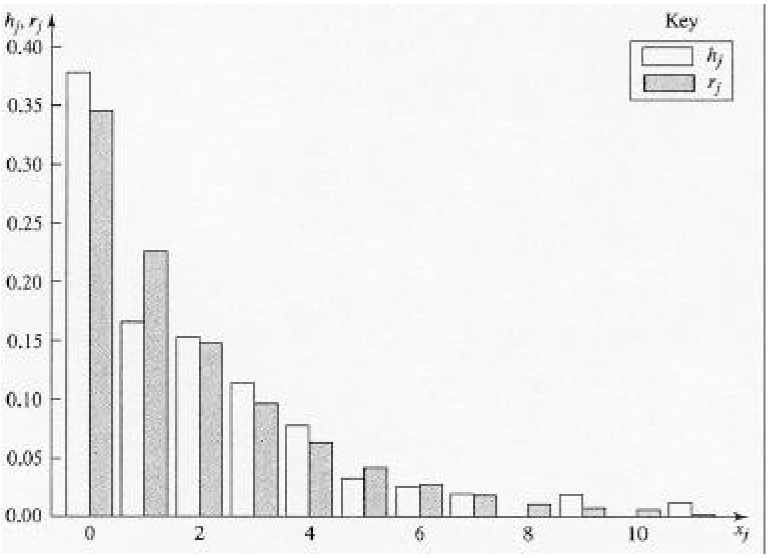
Frequency comparison

Let h_j = the observed proportion of data that are equal to the jth possible value x_j

Let $r_j = \hat{p}(x_j)$, the expected proportio n of the data equal to x_j if the fitted probability mass function $\hat{p}$ is correct

Plot h_j and r_j together, look for similarities

Demand-size data for inventory and fitted geometric:



Distribution Function Differences Plots

Above density/histogram overplots are comparisons of *individual* probabilities of fitted distribution with observed *individual* probabilities

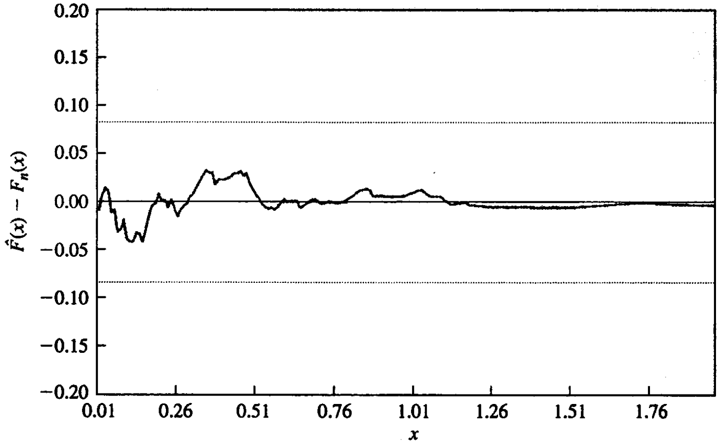
Instead of individual probabilities, could compare *cumulative* probabilities via fitted CDF $\hat{F}(x)$ against a (new) empirical CDF

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

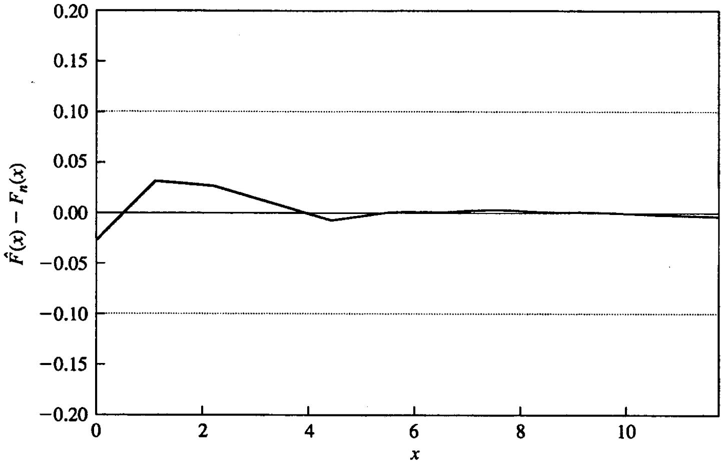
Could plot $\hat{F}(x)$ with $F_n(x)$ and look for similarities, but it is harder to see such similarities for cumulative than for individual probabilities

Alternatively, plot $\hat{F}(x) - F_n(x)$ against the range of x values and look for closeness to a flat horizontal line at height 0

Interarrival-time data for drive-up bank and fitted exponential:



Demand-size data for inventory and fitted geometric:



Probability Plots

Another class of ways to compare CDF of fitted distribution with an empirical directly from the data

Sort data into increasing order: $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ (called the *order statistics* of the data)

Another empirical CDF definition, defined only at the order statistics: $\tilde{F}_n\left(X_{(i)}\right)$ is the observed proportion of data $\leq X_{(i)}$, which is i/n (adjust to $(i - 0.5)/n$ since it's inconvenient to hit 0 or 1)

If $F(x)$ is the true (unknown) CDF of the data then $F(x) = P(X \leq x)$ for any x, so taking $x = X_{(i)}$, $F(X_{(i)}) = P(X \leq X_{(i)})$, which is estimated by $(i - 0.5)/n$

Thus, we should have $F(X_{(i)}) \approx (i - 0.5)/n$, for all $i = 1, 2, \ldots, n$

P-P Plot: If the fitted distribution (with CDF $\hat{F}$) is correct, i.e. close to the true unknown F, we should have

$$\hat{F}(X_{(i)}) \approx (i - 0.5)/n, \text{ for all } i = 1, 2, \ldots, n$$

so plotting the pairs $\left((i - 0.5)/n, \hat{F}(X_{(i)})\right)$, for all $i = 1, 2, \ldots, n$ should result in an approximately straight line from (0, 0) to (1, 1) if $\hat{F}$ is correct

Valid for both continuous and discrete data

Sensitive to misfits in the center of the range of the distribution

Q-Q Plot: Taking $\hat{F}^{-1}$ across the above,

$$\left(\hat{F}^{-1}((i - 0.5)/n), X_{(i)}\right), \text{ for all } i = 1, 2, \ldots, n$$

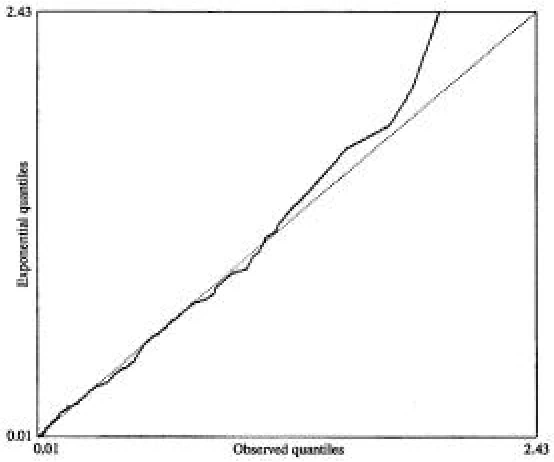
so plotting the pairs $\left((i - 0.5)/n, \hat{F}(X_{(i)})\right)$, for all $i = 1, 2, \ldots, n$ should result in an approximately straight line from $(X_{(1)}, X_{(1)})$ to $(X_{(n)}, X_{(n)})$ if $\hat{F}$ is correct

Valid only for continuous data

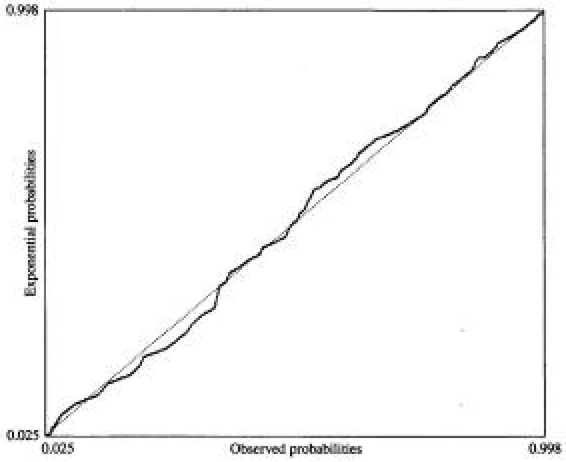
Depending on the form of the fitted distribution, there may not be a closed-form formula for $\hat{F}^{-1}$

Sensitive to misfits in the tails of the distributions

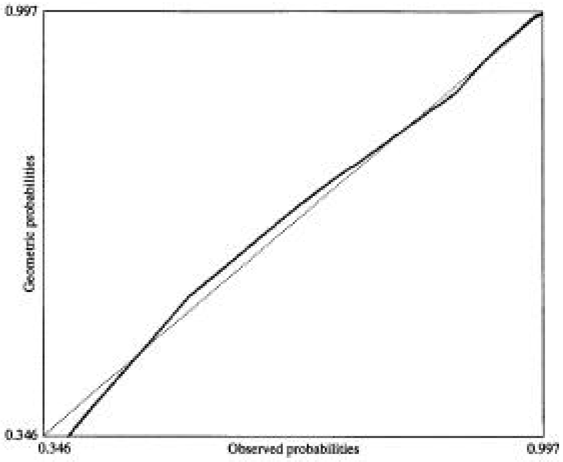
Q-Q plot of interarrival-time data for fitted exponential distribution:



P-P plot of interarrival-time data for fitted exponential distribution:



P-P plot of demand-size data for fitted geometric distribution:



6.6.2 Goodness-of-Fit Tests

Formal statistical hypothesis tests for

H$_0$: The observed data X_1, X_2, ..., X_n are IID random variables with distribution function $\hat{F}$

Caution: Failure to reject H$_0$ does not constitute “proof” that the fit is good

Power of some goodness-of-fit tests is low, particularly for small sample size n

Also, large n creates high power, so tests will nearly always reject H$_0$

Keep in mind that null hypotheses are seldom *literally* true, and we are looking for an “adequate” fit of the distribution

Chi-Square Tests

Very old (Karl Pearson, 1900), and general (continuous or discrete data)

Formalization of frequency comparisons

Divide range of data into *k* intervals, *not* necessarily of equal width:

$[a_0, a_1), [a_1, a_2), ..., [a_{k-1}, a_k]$

a_0 could be $-\infty$ or a_k could be $+\infty$

Compare actual amount of observed data in each interval with what the fitted distribution would predict

Let N_j = the number of observed data points in the *j*th interval

Let p_j = the expected proportion of the data in the *j*th interval if the fitted distribution were literally true:

$$p_j = \begin{cases} \int_{a_{j-1}}^{a_j} \hat{f}(x)\, dx \text{ for continuous} \\ \sum_{a_{j-1} \le x \le x a_j} \hat{p}(x) \text{ for discrete} \end{cases}$$

Thus, $n\,p_j$ = expected (under fitted distribution) number of points in the *j*th interval

If fitted distribution is correct, would expect that $N_j \approx n\,p_j$

Test statistic: $c^2 = \sum_{j=1}^{k} \frac{(N_j - np_j)^2}{np_j}$

Under H_0: Fitted distribution is correct, χ^2 has (approximately—see book for details) a chi-square distribution with $k - 1$ d.f.

Reject H_0 at level a if χ^2 > upper critical value

Advantages: Completely general

Asymptotically valid (as $n \to \infty$) *if* MLEs were used

Drawback: Arbitrary choice of intervals (can affect test conclusion)

Conventional advice:

Want $n\,p_j \ge 5$ or so for all but a couple of *j*’s

Pick intervals such that the p_j’s are close to each other

Chi-square test for exponential distribution fitted to interarrival-time data:

Chose $k = 20$ intervals so that $p_j = 1/20 = 0.05$ for each interval (see book for details on how the endpoints were chosen ... involved inverting the exponential CDF and taking $a_{20} = +\infty$)

Thus, $np_j = (219)\ (0.05) = 10.95$ for each interval

Counted observed frequencies N_j, computed test statistic $c^2 = 22.188$

Use d.f. $= k - 1 = 19$; upper 0.10 critical level is $c^2_{19,0.90} = 27.204$

Since test statistic does not exceed the critical level, do not reject H_0

Chi-square test for geometric distribution fitted to demand-size data:

Since data are discrete, cannot choose intervals so that the p_j's are exactly equal to each other

Chose $k = 3$ intervals (classes) {0}, {1, 2}, and {3, 4, ...}

Got $np_1 = 53.960$, $np_2 = 58.382$, and $np_3 = 43.658$

Counted observed frequencies N_j, computed test statistic $c^2 = 1.930$

Use d.f. $= k - 1 = 2$; upper 0.10 critical level is $c^2_{2,0.90} = 4.605$

Since test statistic does not exceed the critical level, do not reject H_0

Kolmogorov-Smirnov Tests

Advantages with respect to chi-square tests:

- No arbitrary choices like intervals
- Exactly valid for any (finite) n

Disadvantage with respect to chi-square tests:

- Not as general

A kind of a formalization of probability plots

- Compare empirical CDF from data against fitted CDF

Yet another version of empirical distribution function:

- $F_n(x)$ = proportion of the X_i data that are $\leq x$ (piecewise linear step function)
- On the other hand, we have the fitted CDF $\hat{F}(x)$
- In a perfect world, $F_n(x) = \hat{F}(x)$ for all x
- The worst (vertical) discrepancy is $D_n = \sup_x \left| F_n(x) - \hat{F}(x) \right|$
  - (“sup” instead of “max” because it may not be attained for any x)
- Computing D_n (must be careful; sometimes stated incorrectly):

$$D_n^+ = \max_{i=1,2,\ldots,n} \left(\frac{i}{n} - \hat{F}\left(X_{(i)}\right) \right)$$

$$D_n^- = \max_{i=1,2,\ldots,n} \left(\hat{F}\left(X_{(i)}\right) - \frac{i-1}{n} \right)$$

$$D_n = \max\left\{ D_n^+, D_n^- \right\}$$

- Reject H_0: The fitted distribution is correct if D_n is too big
  - There are several different kinds of tables depending on the form and specification of the hypothesized distribution (see book for details and example)

Anderson-Darling Tests

As in K-S test, look at vertical discrepancies between $\hat{F}(x)$ and $F_n(x)$

Difference: K-S weights differences the same for each x

- Sometimes more interested in getting accuracy in (right) tail
  - Queueing applications
  - P-K formula depends on *variance* of service-time RV
- A-D applies increasing weight on differences toward tails
- A-D more sensitive (powerful) than K-S in tail discrepancies

Define the weight function $y(x) = \dfrac{1}{\hat{F}(x)\left[1 - \hat{F}(x)\right]}$

Note that $Y(x)$ is smallest (= 4) in the middle (median) where $\hat{F}(x) = 1/2$ and largest (→ ∞) in either tail

Test statistic is

$$A_n^2 = n\int_{-\infty}^{\infty} \left[F_n(x) - \hat{F}(x)\right]^2 y(x)\hat{f}(x)\,dx$$

$$= -\frac{\sum_{i=1}^{n}(2i-1)\left[\ln \hat{F}(X_{(i)}) + \ln\left(1 - \hat{F}(X_{(n-i+1)})\right)\right]}{n} - n \quad \text{computatio nally}$$

Reject H_0: The fitted distribution is correct if A_n^2 is too big

There are several different kinds of tables depending on the form and specification of the hypothesized distribution (see book for details and example)

Poisson-Process Tests

Common situation in simulation: modeling an *event process* over time

Arrivals of customers or jobs
Breakdowns of machines
Accidents

Popular (and realistic) model: *Poisson process* at rate l

Equivalent definitions:

1. Number of events in $(t_1, t_2]$ ~ Poisson with mean $l(t_2 - t_1)$
2. Time between successive events ~ exponential with mean $1/l$
3. Distribution of events over a fixed period of time is uniform

Use second or third definitions to develop test for observed data coming from a Poisson process:

2. Test for inter-event times' being exponential (chi-square, K-S, A-D, ...)
3. Test for placement of events' over time being uniform

See book for details and example

6.7 The ExpertFit Software and an Extended Example

Need software assistance to carry out the above calculations

Standard statistical-analysis packages do not suffice

- Often too oriented to normal-theory and related distributions
- Need wider variety of “nonstandard” distributions to achieve and adequate fit
- Difficult calculations like inverting non-closed-form CDFs, computation of critical values and *p*-values for tests

ExpertFit package is tailored to these needs

Other packages exist, sometimes packaged with simulation-modeling software

See book for details on ExpertFit and an extended, in-depth example

6.8 Shifted and Truncated Distributions

Shifted Distributions

Many standard distributions have range $[0, \infty)$

- Exponential, gamma, Weibull, lognormal PT5, PT6, log-logistic

But in some situations we'd like the range to be $[\gamma, \infty)$ for some parameter $\gamma > 0$

- A service time cannot physically be arbitrarily close to 0; there is some absolute positive minimum γ for the service time

Can *shift* one of the above distributions up (to the right) by γ

- Replace x in their density definitions by $x - \gamma$ (including in the definition of the ranges)

Introduces a new parameter γ that must be estimated from the data

- Depending on the distribution form, this may be relatively easy (e.g., exponential) or very challenging (e.g., global MLEs are ill-defined for gamma, Weibull, lognormal)
- See book for details and example

Truncated Distributions

Data are well-fitted by a distribution with range $[0, \infty)$ but physical situation dictates that no value can exceed some finite constant b

Need to truncate the distribution above b, to make effective range $[0, b]$

Really a variate-generation issue: covered in Chap. 8

6.9 Bézier Distributions

Can approximate the underlying CDF $F(x)$ arbitrarily closely by a *Bézier distribution* (related to Bézier curves used in drawing)

Specify control points for distribution

Can fit an optimally fitting Bézier distribution, or use specialized software to drag control points around visually with a mouse to achieve a visually acceptable fit

This is an alternative to simpler empirical distributions, useful when no standard distribution adequately fits the observed data

6.10 Specifying Multivariate Distributions, Correlations, and Stochastic Processes

Assumption so far: Want to generate independent, identically distributed (IID) random variables (RVs) for input to drive the simulation

Sometimes have correlation between RVs in reality

A = interarrival time of a job from an upstream process

S = service time of job at the station being modeled

Perhaps a large A means that the job is "large," taking a lot of time upstream—then it probably will take a lot of time here too (S large), i.e., $\text{Cor}(A, S) > 0$

Ignoring this correlation can lead to serious errors in output validity

Need ways to estimate this dependence, and (later) generate it in the simulation

There are several different specific situations and goals

6.10.1 Specifying Multivariate Distributions

Some of the model's input RVs together form a jointly distributed random vector

Must specify the joint distribution form and estimate its parameters

Correlations between the RVs is then determined by the joint distribution form

This is an ambitious goal, in terms of both methods for specification, observed-data requirements, and later variate-generation methods

At present, limited to several specific special cases (see book for details): multivariate normal, multivariate lognormal, multivariate Johnson, and bivariate Bézier

6.10.2 Specifying Arbitrary Marginal Distributions and Correlations

Less ambitious than specifying the joint distribution, but affords greater flexibility

Allow for possible correlation between input RVs, but fit their univariate (marginal) distributions separately

Must specify the univariate marginal distributions (earlier methods) and estimate the correlations (fairly easy)

Does not in general uniquely specify (control) the joint distribution

Except in multivariate normal case, specifying the marginal distributions and all the correlations does not uniquely specify the joint distribution

Must take care that the correlations are compatible with the marginal distributions

Marginal distributions place constraints on what correlation structure is theoretically possible

How to generate this structure for input to the simulation? (Chap. 8)

6.10.3 Specifying Stochastic Processes

Have an input stochastic process $\{X_1, X_2, ...\}$ where the X_i's have the same distribution, but there is a correlation structure for them at various lags

e.g., X_i is the size of the *i*th incoming message in a communications system, and it could be that large messages tend to be followed by other large messages (or the reverse)

Can regard this as an infinite-dimensional random vector for input

Some specific models (see book for details): AR, ARMA, gamma processes, EAR, TES, ARTA

6.11 Selecting a Distribution in the Absence of Data

No data? (it happens)

Must rely to some extent on subjective information (guesses)

Ask “expert” for:

min, max ⇒ uniform distribution
min, max, mode ⇒ triangular distribution
min, max, mode, mean ⇒ beta distribution

See book for details and example

Must do sensitivity analysis

Change input distributions, see if output changes appreciably

6.12 Models of Arrival Processes

Want probabilistic model of event process happening over time
Common application: arrival process

As in distributions, need to specify form, estimate parameters

Three common models:

6.12.1 Poisson Processes

Three “behavioral” assumptions:

1. Events occur one at a time
2. Number of events in a time interval is independent of the past
3. Expected rate of events is constant over time

Fitting: Fit exponential to interevent times via MLE

Testing: Saw above

6.12.2 Nonstationary Poisson Process

Drop behavioral assumption 3 above (keep 1, one-at-a-time events)

Allow for expected rate of events to vary with time: replace arrival-rate constant l with a function $l\,(t)$, where t = time

Number of events in $(t_1, t_2]$ ~ Poisson with mean $\int_{t_1}^{t_2} l\,(t)\,dt$

Estimation of rate function

Assume rate function is constant over subintervals of time
Must specify subintervals thought to be appropriate
Must be careful to keep the units straight

Other methods exist (see book for discussion and references)

6.12.3 Batch Arrivals

Drop behavioral assumption 1 above

Allow number of events arriving to be a discrete RV, independent of event-time process

Fitting
Fit distribution to interevent times via MLE
Fit a discrete RV to observed “group” sizes

Testing
Separately for interevent times, group sizes

6.13 Assessing the Homogeneity of Different Data Sets

Sometimes have different data sets on related but separate processes

- Have service-time observations for ten different days
- Can the ten data sets be merged?
- In other words, is the underlying distribution the same for each day?

Advantages of merging (if it turns out to be justified)

- Larger sample size, so get better specification of *the* input distribution
- Just one specification problem rather than several
- Just one distribution from which to generate in the simulation model

Want to test

- H_0: All the population distribution functions are identical
- vs.
- H_1: At least one of the populations tends to yield larger observations than at least one of the other populations

Formal statistical test for doing so: *Kruskal-Wallis test*, which is a nonparametric test based on the ranks of the data sets (see book for details)