Selecting Input Probability Distributions

Chapter 6

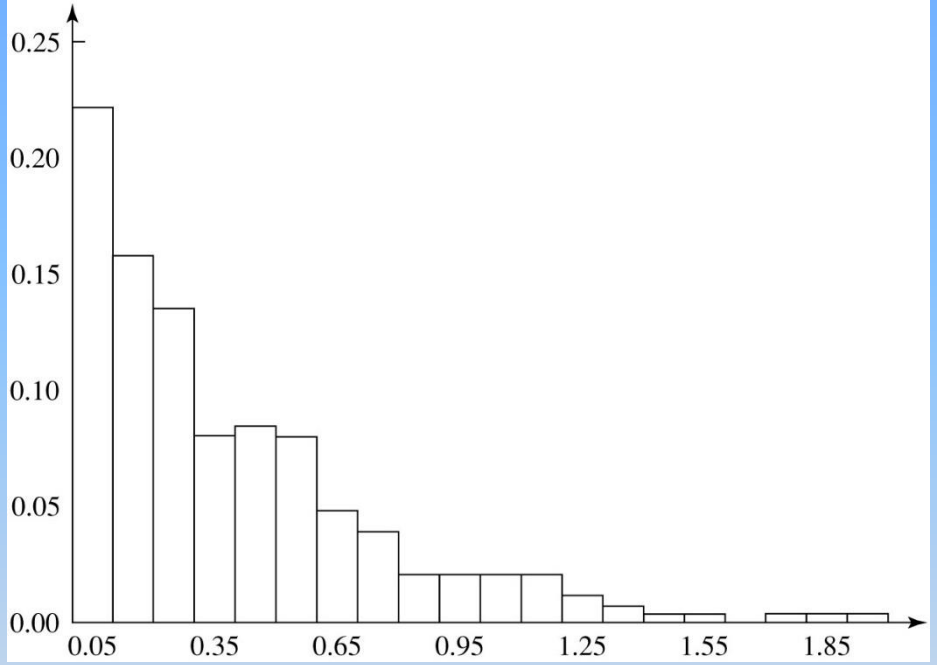
6.1 Introduction

- Almost all real-world systems contain randomness that needs to represented in models

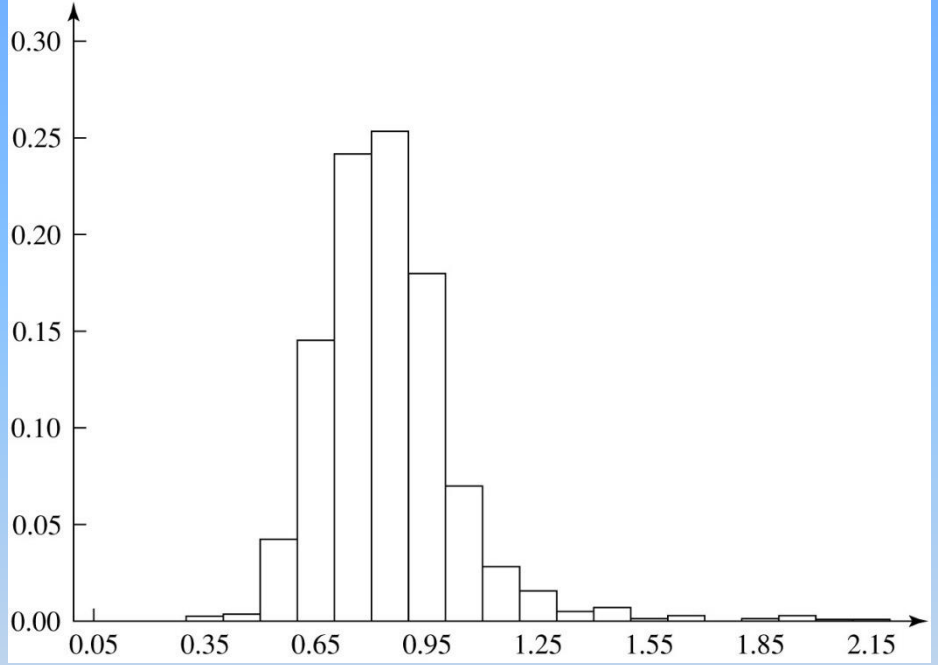
Type of system	Sources of randomness
Manufacturing	Processing times, machine times to failure, machine repair times
Defense-related	Arrival times and payloads of missiles or airplanes, outcome of an engagement, miss distances for munitions
Communications	Interarrival times of messages, message types, message lengths
Transportation	Ship-loading times, interarrival times of customers to a subway

Sources of randomness for common simulation applications

Simulation Data Set Examples (1)

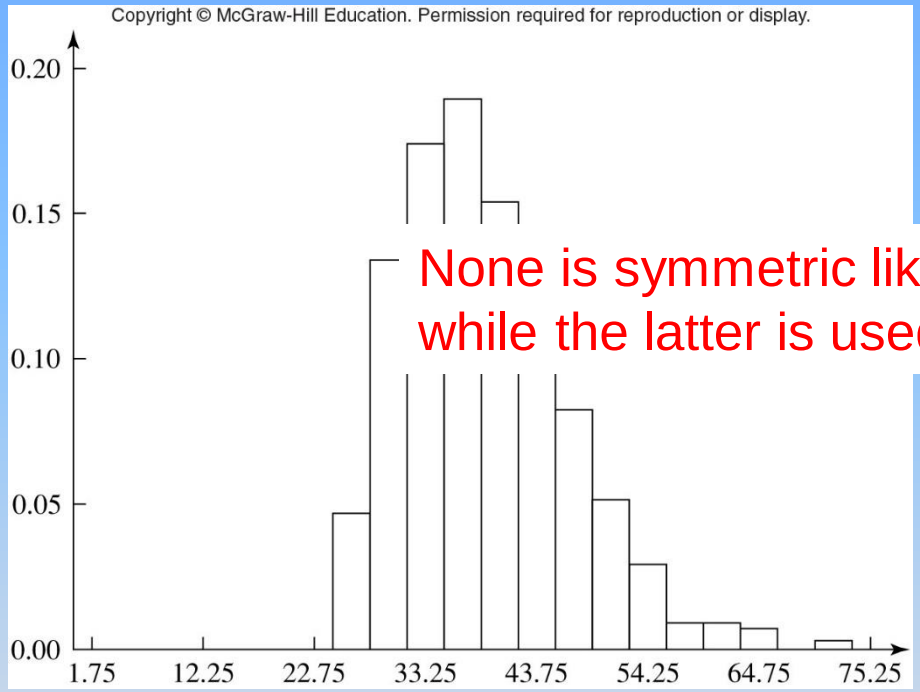


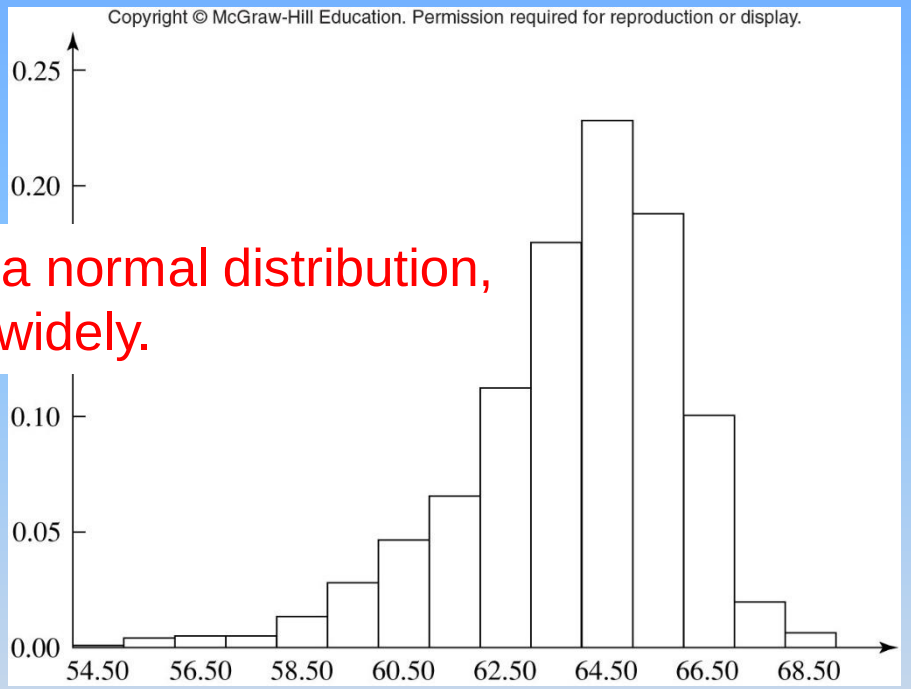
Interarrival times in minutes to a drive-up bank



Ship loading times in days

Simulation Data Set Examples (2)





None is symmetric like a normal distribution, while the latter is used widely.

Machine processing time for an automotive manufacturer
Longer right tail (positive skewness)
Minimum ≈ 25 minutes

Scaled number of yards of paper on 1000 large rolls of paper used to make facial or bathroom tissue
Longer left tail (negative skewness)

Impact of Using Incorrect Distributions

- Example:
  - A single-server queueing system has exponential interarrival times with a mean of 1 minute. Service time distribution is best fit by Weibull.

Service-time distribution	Average delay in queue	Average number in queue	Proportion of delays ≥20
Exponential	6.71	6.78	0.064
Gamma	4.54	4.60	0.019
Weibull	4.36	4.41	0.013
Lognormal	7.19	7.30	0.078
Normal	6.04	6.13	0.045

Has the same general shape as Weibull but has a "thicker" right tail

Processes for Selecting Distributions

- Use data values themselves directly in the simulation (trace-drive simulation)
  - Cons
    - Can only reproduce what has happened
    - Seldom enough data to run all the desired simulations
  - Pros
    - Good for use if modeling randomness is hard
    - Recommended for model validation

Processes for Selecting Distributions

- Define an empirical distribution function from the data values
  - Pros
    - Avoid the shortcomings of using real data values
  - Cons
    - Impossible to generate values outside the range of the observed data, if used in the usual way
    - Cumbersome to represent a large set of data values
      - 2n numbers stored for n data values

Processes for Selecting Distributions

- Fit a theoretical distribution to the data
  - Perform hypothesis test to determine the goodness of fit
  - Pros
    - "Smooth" data vs. irregular data from empirical distribution
    - Compact way of representing data values
    - Easy to change by tuning parameters
  - Cons
    - No fit for the observed data
      - Data are a mixture of multiple heterogeneous populations
      - Not enough data values
    - Arbitrarily large values can be generated
      - Truncate the distribution

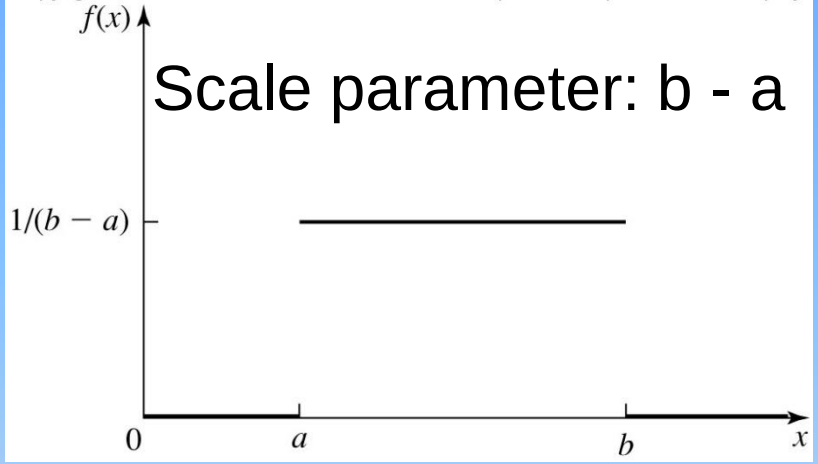
6.2 Useful Probability Distributions

- For a given family of continuous distributions
  - There are several ways to parameterize the probability distribution
- Three basic types of parameters
  - Location parameter
    - Midpoint or lower endpoint of the distribution's range
  - Scale parameter
    - Scale of the measurement of the values in the distribution's rnage
  - Shape parameter
    - Alters a distribution's properties more fundamentally
    - A distribution may have from 0 to 2 shape parameters

Useful Continuous Distributions (1)

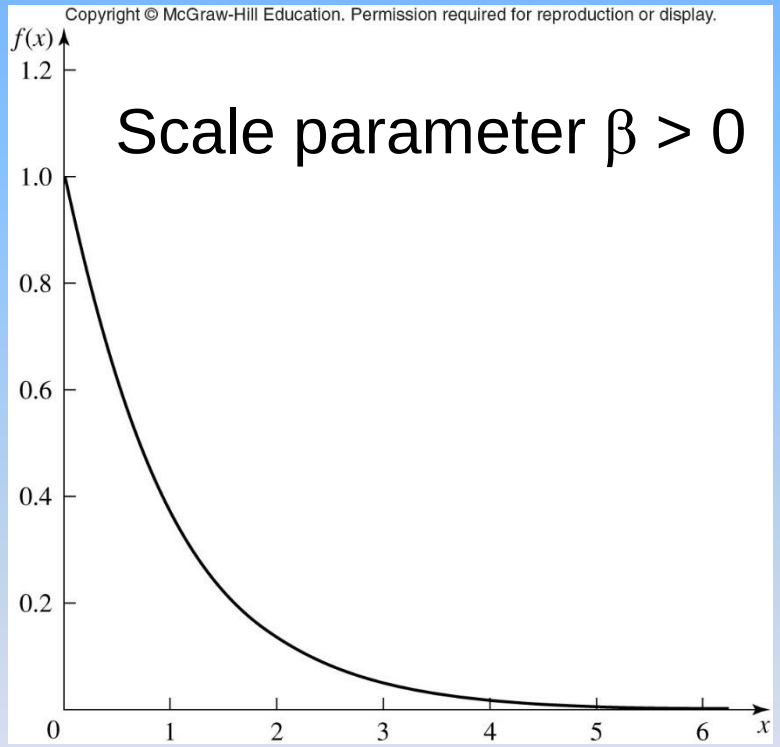
- Uniform U(a,b)
  - Used as a "first" model for a quantity felt to randomly varying between a and b,

    but little else is known
  - U(0,1) is essential in generating random values from all other distributions
  - U(0,1) is a special case of the beta distribution



Useful Continuous Distributions (2)

- Exponential expo(β)
  - Used for interarrival times of "customers" to a system that occur at a constant average rate (β), time to failure of a piece of equipment
  - A special case of both gamma and Weibull distributions
  - If $X_1, X_2, \ldots, X_n$ are IID expo(β), $X_1 + X_2 + \ldots + X_n \sim$ gamma(m, β), a.k.a m-Erlang(β)
  - The only continuous memoryless distribution
    - $P(X > t + s \mid X > t) = P(X > s)$



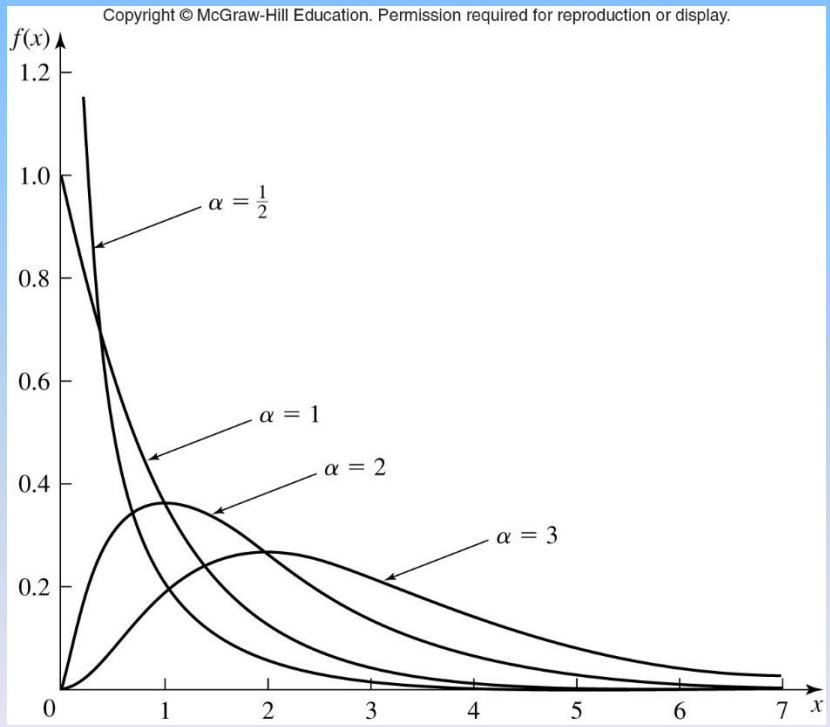
expo(1)

Useful Continuous Distributions (3)

- Gamma gamma(α,β)
  - Used for time to complete some task, e.g. customer server or machine repair
  - expo(β) $\equiv$ gamma(1,β)
  - X_i ~ gamma(α_i,β)
    - $X_1 + X_2 + \ldots + X_n$ ~ gamma($\alpha_1 + \alpha_2 + \ldots + \alpha_n$, β),
    - $X_1 / (X_1 + X_2)$ ~ beta(α_1 , α_2)
  - X ~ gamma(α_i,β)

    $\Leftrightarrow$ 1/X ~ PT5(α,1/β)

Shape param. $\alpha > 0$
Scale param. $\beta > 0$

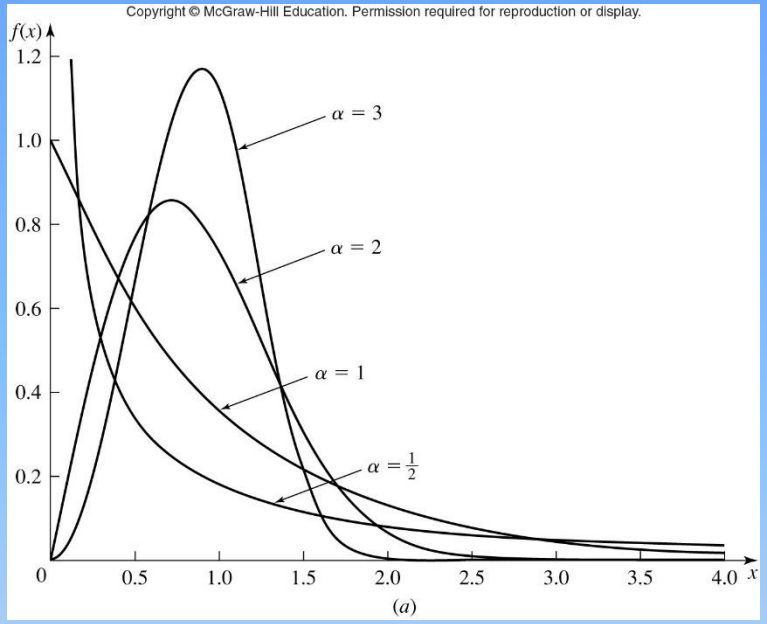


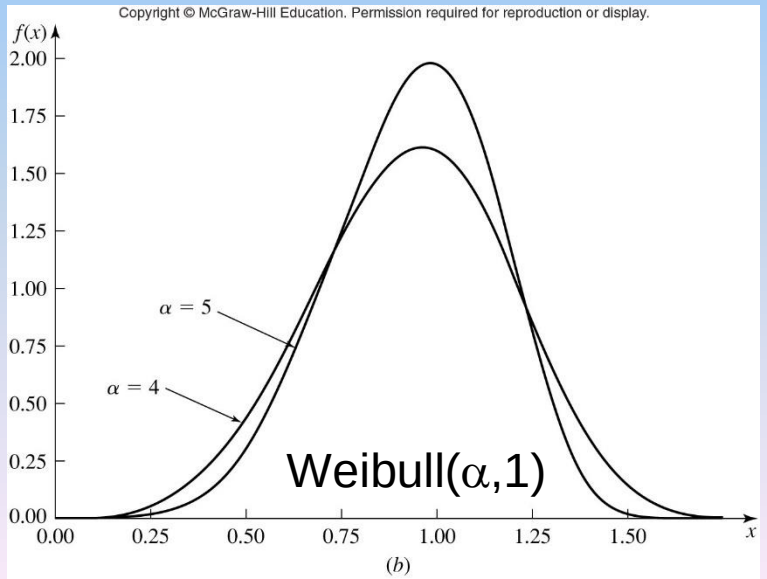
gamma(α,1)

Useful Continuous Distributions (4)

- Weibull(α,β)
  - Used for time to complete some task, time to failure of a piece of equipment, or as a rough model in the absence of data
  - expo(β) $\equiv$ Weibull (1,β)
  - X ~ Weibull(α,β) $\Leftrightarrow$ X^{α} ~ expo(β^{α})
  - α -> ∞, degenerate at β

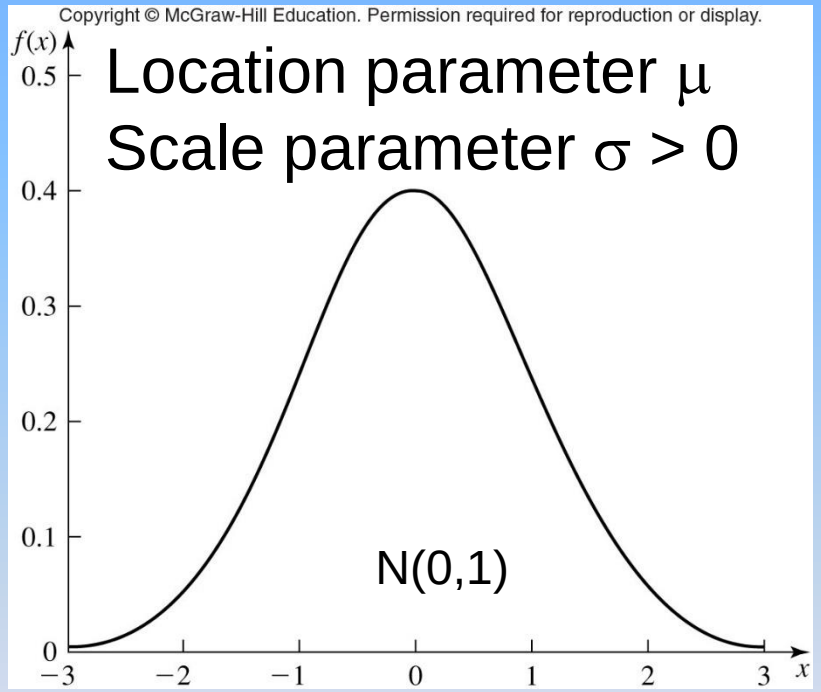
Shape param. $\alpha > 0$
Scale param. $\beta > 0$





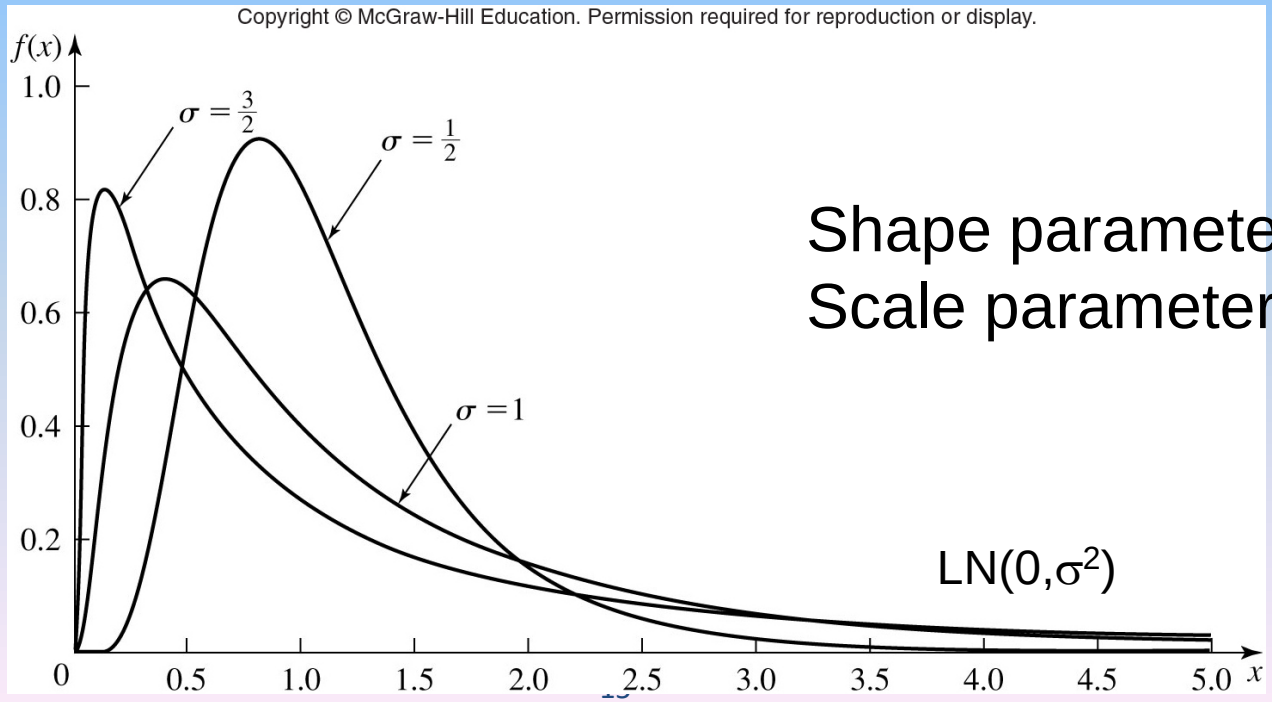
Useful Continuous Distributions (5)

- Normal $N(\mu,\sigma^2)$
  - Used for errors of various types (e.g. the impact point of a bomb), quantities that are the sum of a large number of other quantities
  - Two jointly distributed normal r.v. uncorrelated => independent
  - Sum of normal distributed r.v. also has normal distribution
  - If $X_1,X_2,\ldots,X_n \sim N(0,1)$, $X_1^2 + X_2^2 + \ldots + X_n^2 \sim$ gamma(n/2, 2)
  - $X \sim N(\mu,\sigma^2) \Rightarrow e^x \sim LN(\mu,\sigma^2)$
  - $\sigma \to 0$, degenerate at μ



Useful Continuous Distributions (6)

- Lognormal LN(μ,σ^2)
  - Used for time to perform a task, quantities that are the product of a large number of other quantities
  - X ~ LN(μ,σ^2) $\Leftrightarrow$ ln(x) ~ N(μ,σ^2)
  - σ -> 0, degenerate at e^{μ}



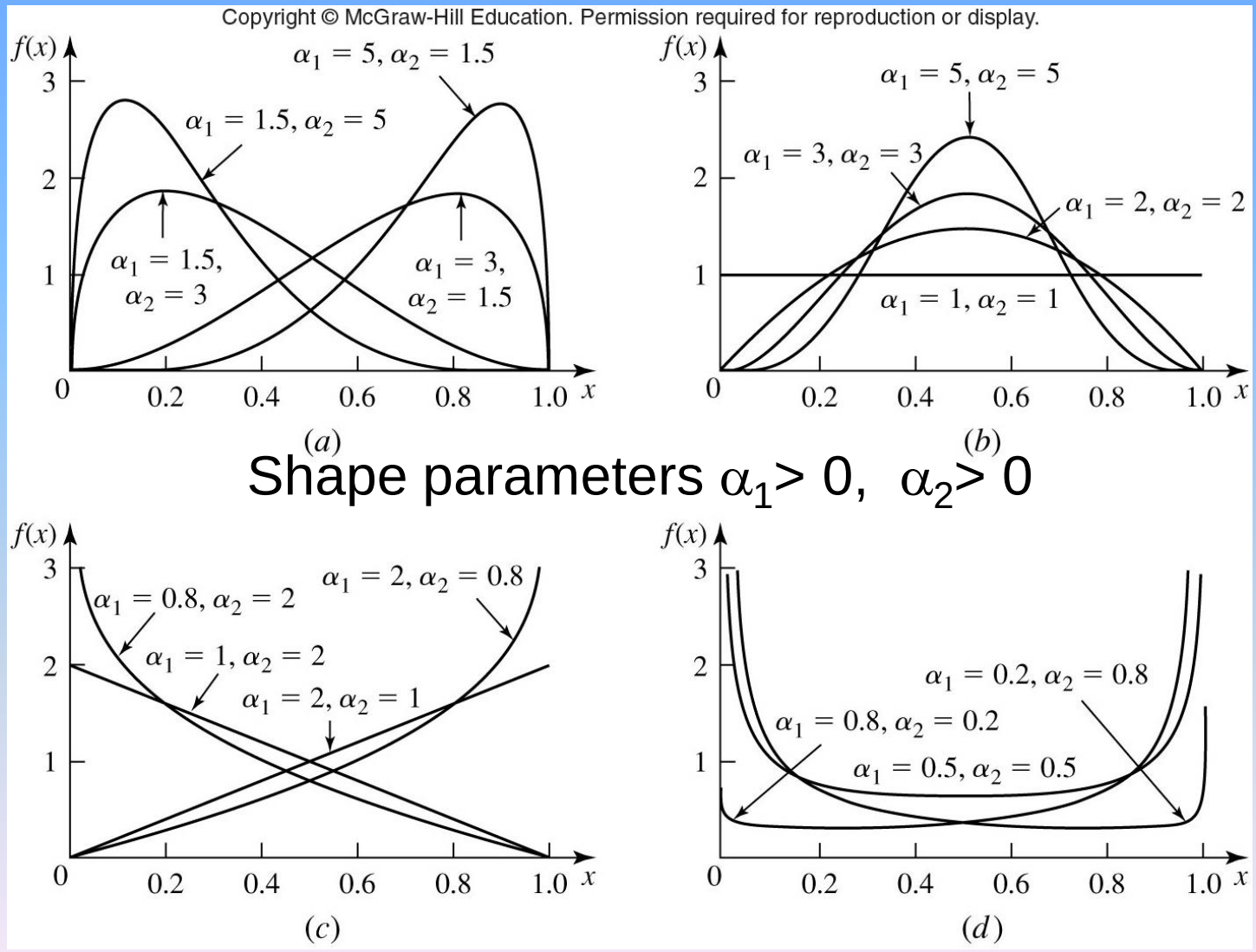
Shape parameter $\sigma > 0$
Scale parameter $e^{\mu} > 0$

LN(0,σ^2)

Useful Continuous Distributions (7)

- Beta beta(α_1, α_2)
  - Used for a random portion (e.g. the proportion of defective items in a shipment), time to complete a task, or a rough model in the absence of data
  - U(0,1) $\equiv$ beta(1,1)
  - $X_i \sim \text{gamma}(\alpha_i, \beta) \Rightarrow X_1 / (X_1 + X_2) \sim \text{beta}(\alpha_1, \alpha_2)$
  - X on [0,1] can be scaled to [a,b] by a + (b-a)X
  - $X \sim \text{beta}(\alpha_1, \alpha_2) \Leftrightarrow 1 - X \sim \text{beta}(\alpha_2, \alpha_1)$
    $\Leftrightarrow X/(1-X) \sim \text{PT6}(\alpha_1, \alpha_2, 1)$
  - Symmetric about x = ½ if and only if $\alpha_1 = \alpha_2$

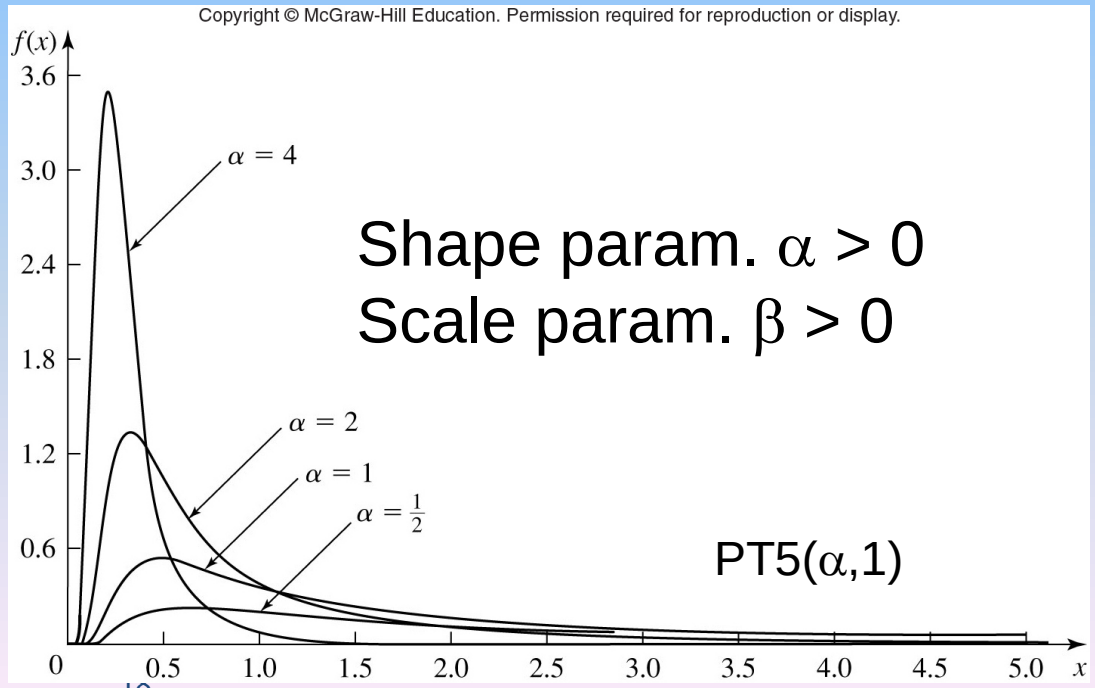
beta (α_1, α_2)



Shape parameters $\alpha_1 > 0$, $\alpha_2 > 0$

Useful Continuous Distributions (8)

- Person type V PT5(α,β)
  - Used for time to perform a task
  - Larger spike than lognormal close to x = 0
  - Inverted gamma distribution
    - X ~ PT5(α,1/β) $\Leftrightarrow$ 1/X ~ gamma(α_i,β)
  - Mean and variance exist only for certain values of the shape parameter

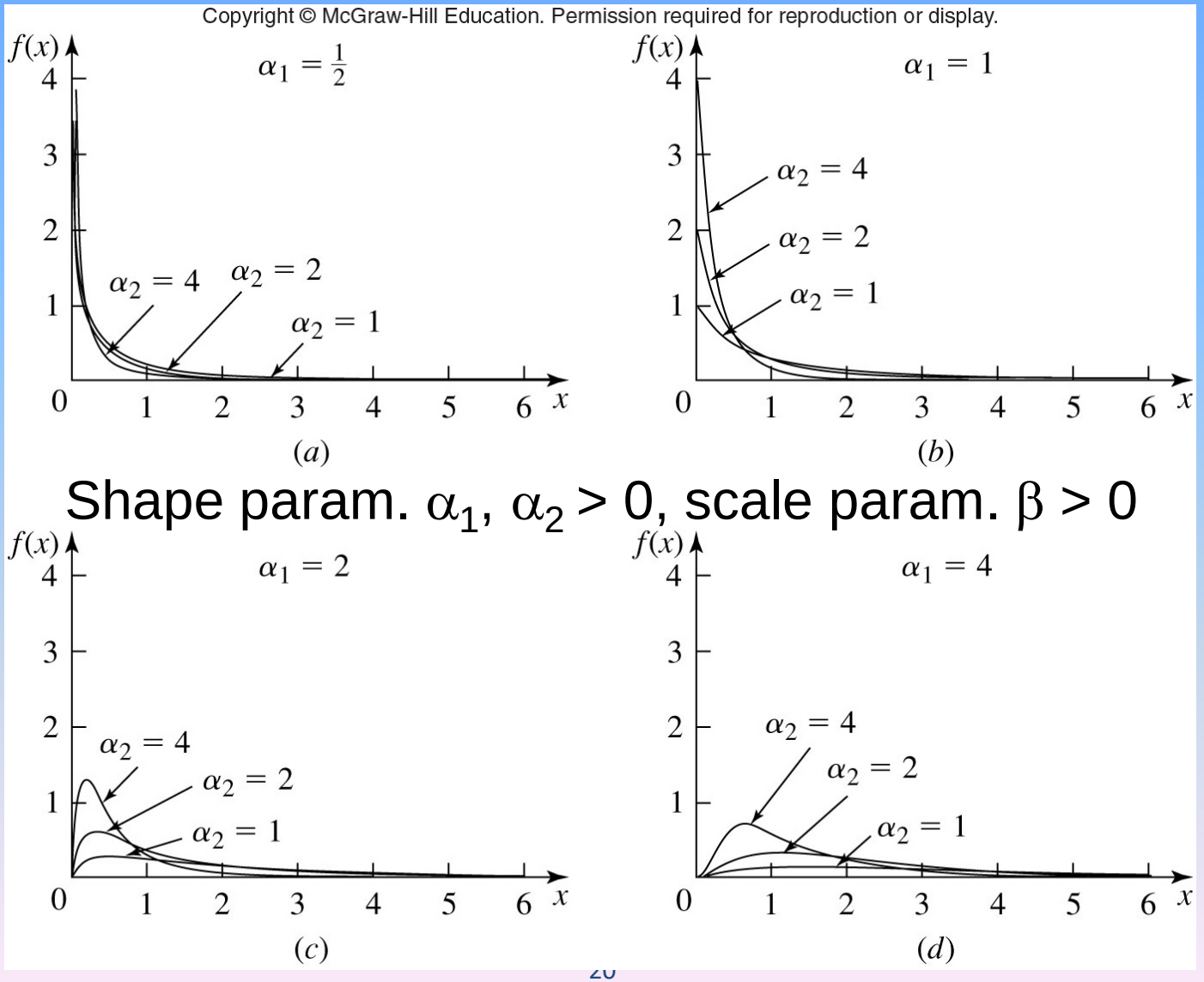


Useful Continuous Distributions (9)

- Person type VI PT6(α_1, α_2, β)
  - Used for time to perform a task
  - $X \sim PT6(\alpha_1, \alpha_2, 1) \Leftrightarrow X/(1+X) \sim beta(\alpha_1, \alpha_2)$
  - $X_1 \sim gamma(\alpha_1, \beta)$, $X_2 \sim gamma(\alpha_2, 1)$

    $\Rightarrow X_1 / X_2 \sim PT6(\alpha_1, \alpha_2, \beta)$
  - Mean and variance exist only for certain values of the shape parameter α_2

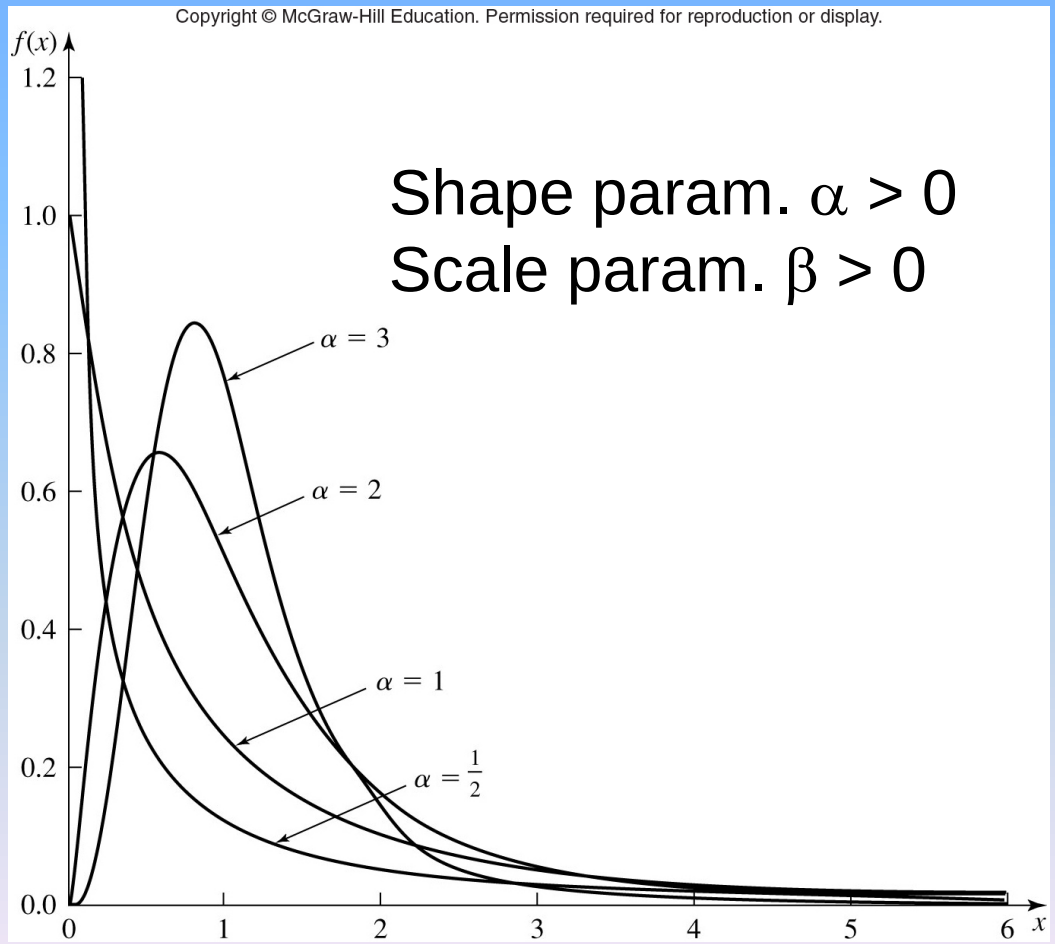
PT6(α_1, α_2, 1)



Shape param. α_1, $\alpha_2 > 0$, scale param. $\beta > 0$

Useful Continuous Distributions (10)

- Log-logistic LL(α,β)
  - Used for time to perform a task



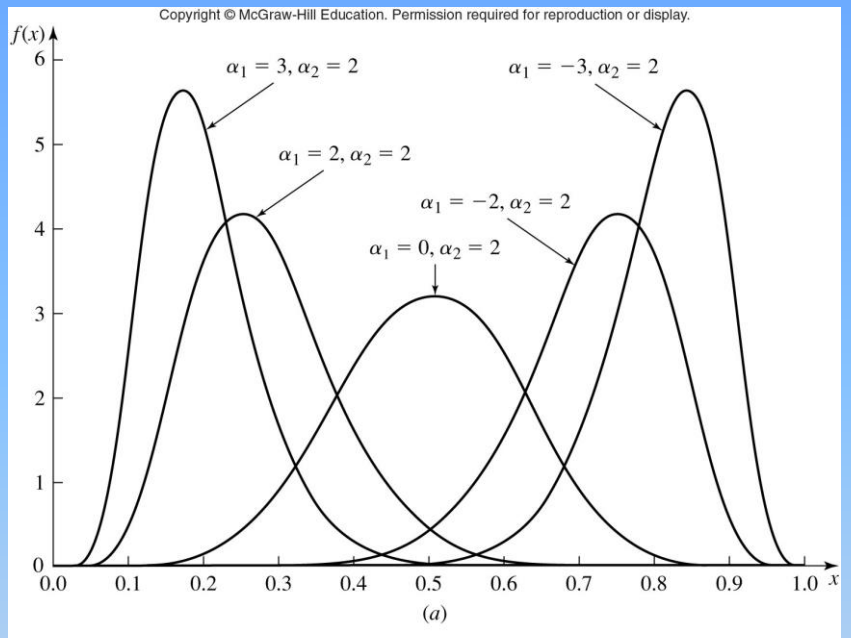
LL(α,1)

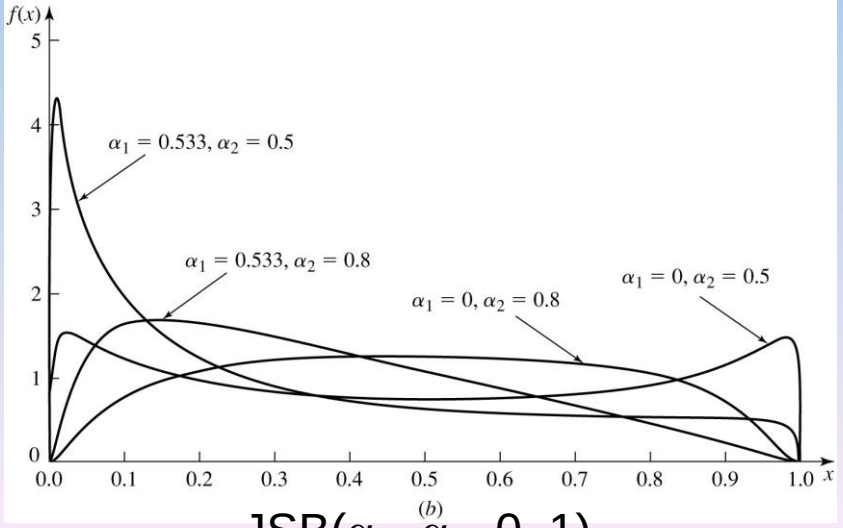
Useful Continuous Distributions (11)

- Johnson S_B JSB(α_1, α_2, a, b)
- X ~ JSB(α_1, α_2, a, b)

  $\Leftrightarrow \alpha_1 + \alpha_2 ln\left(\frac{X-a}{b-X}\right)$ ~ N(0,1)
- pdf skewed to left/symmetric/right for α_1 > 0, = 0, < 0

Location param. a
Scale param. b – a > 0
Shape param. α_2 > 0, α_1





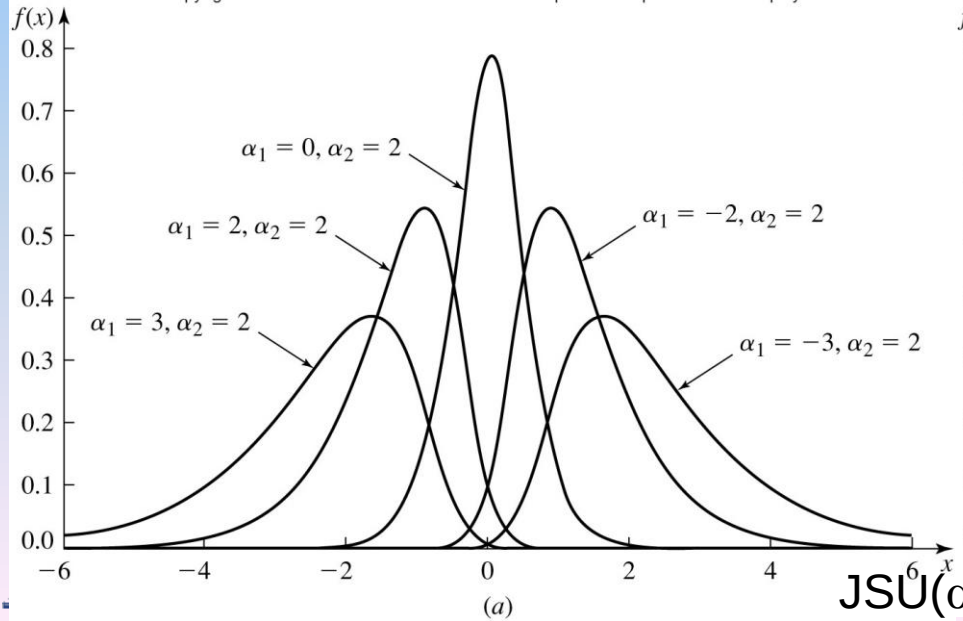
JSB(α_1, α_2, 0, 1)

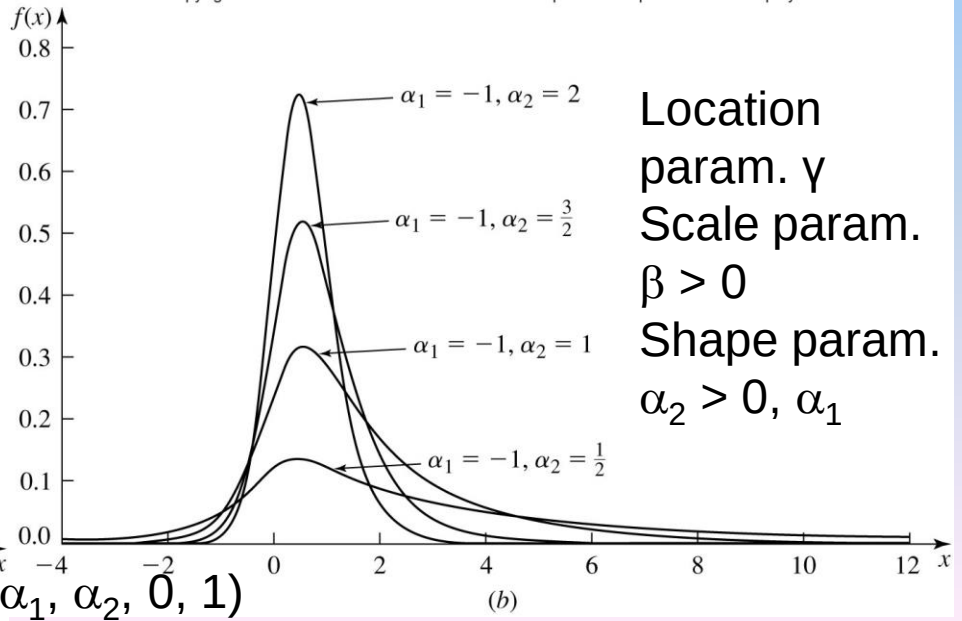
Useful Continuous Distributions (12)

- Johnson S_U JSU(α_1, α_2, γ, β)
- X ~ JSU(α_1, α_2, γ, β)

$$\Leftrightarrow \alpha_1 + \alpha_2 ln\left(\frac{X-\gamma}{\beta} + \sqrt{\left(\frac{X-\gamma}{\beta}\right)^2 + 1}\right) \sim N(0,1)$$

- pdf skewed to left/symmetric/right for $\alpha_1 > 0, = 0, < 0$





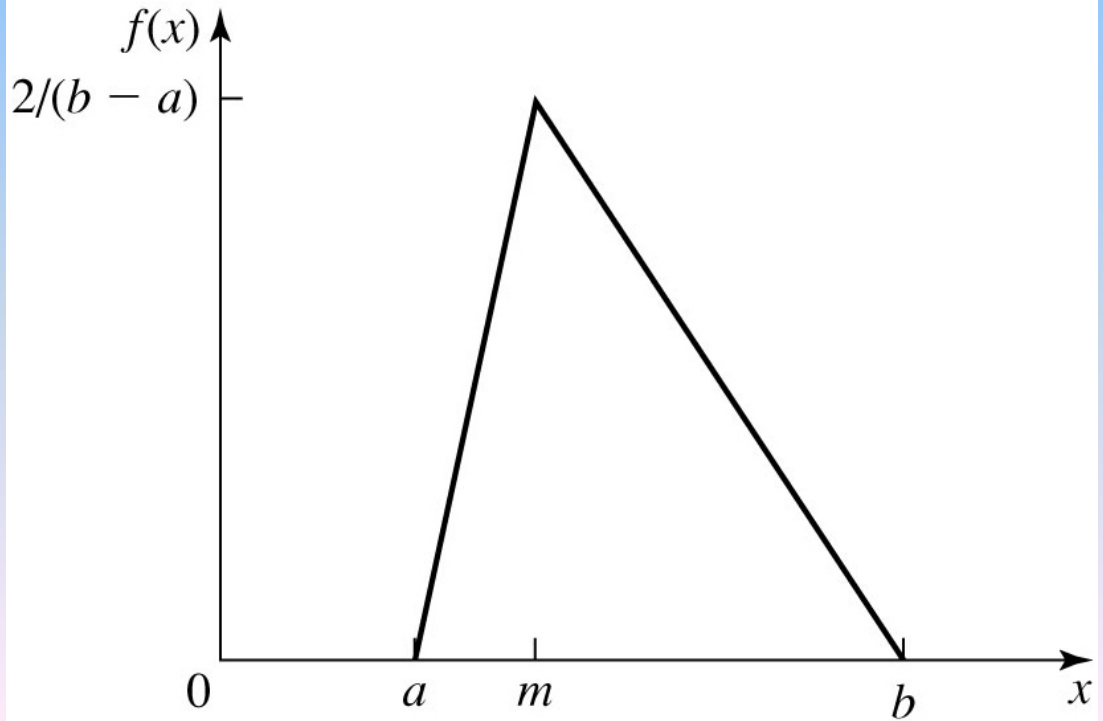
Location param. γ
Scale param. $\beta > 0$
Shape param. $\alpha_2 > 0$, α_1

JSU(α_1, α_2, 0, 1)

Useful Continuous Distributions (13)

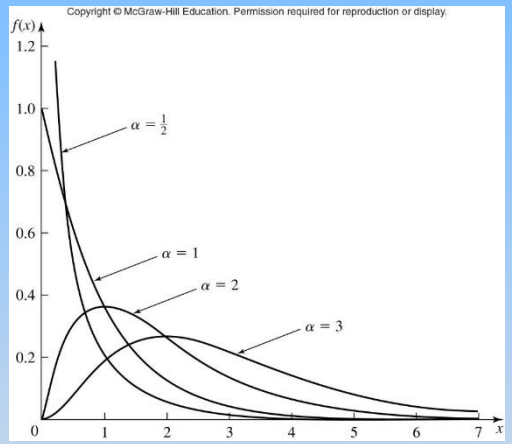
- Triangular triang(a,b,m)
  - Used as a rough model in the absence of data
  - m -> b: right triangular, m -> a: left triangular
  - triang(a,b,m) (m -> a or m-> b) are special cases of the beta distribution

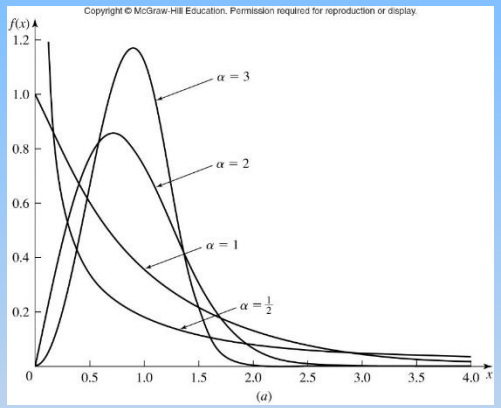
Location param. a
Scale param. b – a > 0
Shape param. m

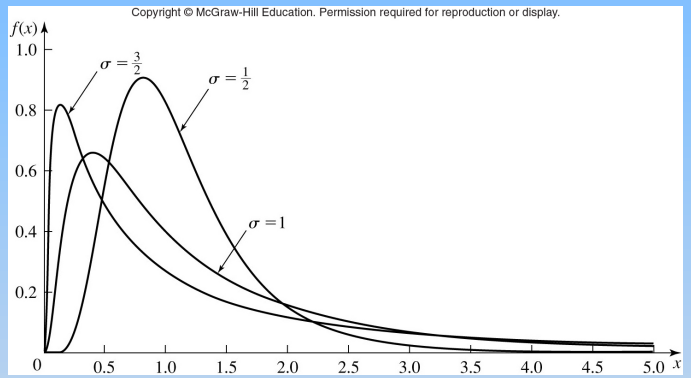


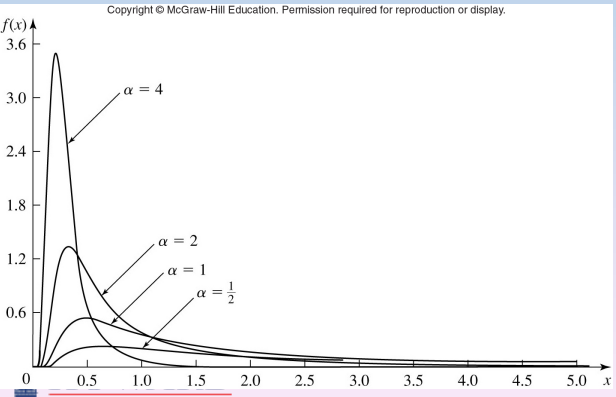
Continuous Distribution Use Summary

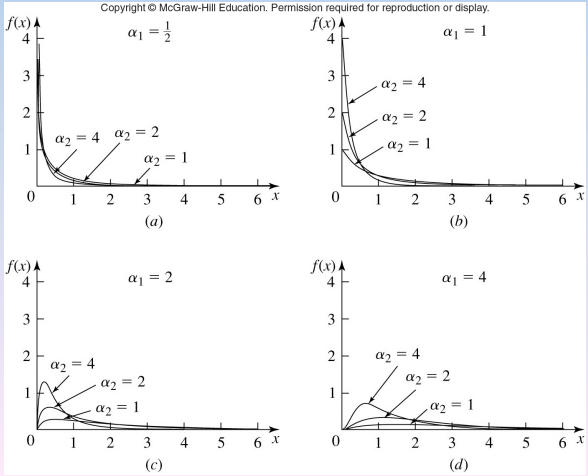
- Time to complete a tasks
  - Gamma, Weibull, lognormal, Pearson type V, Pearson type VI, Log-logistic

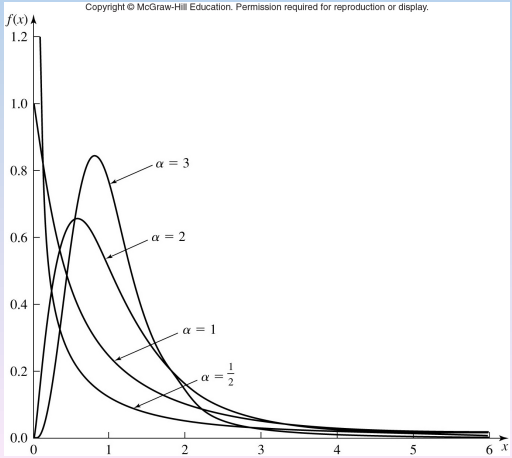






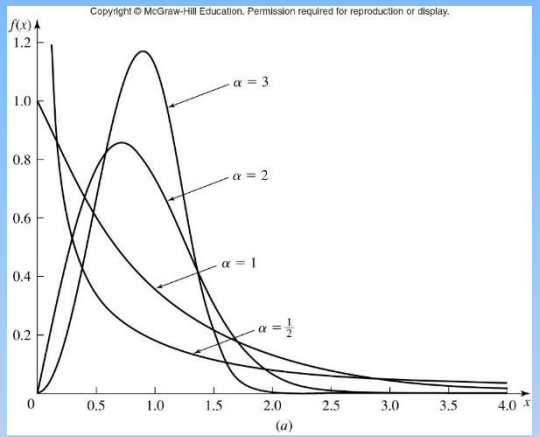


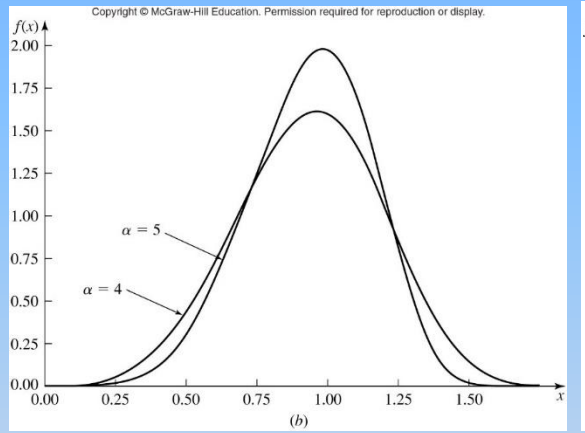


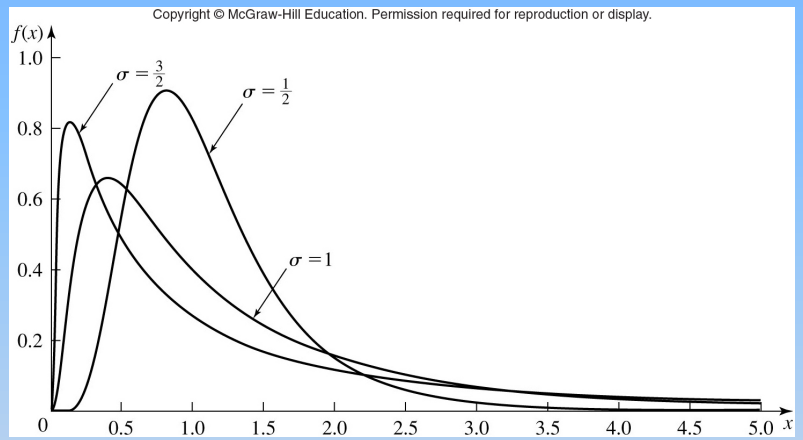


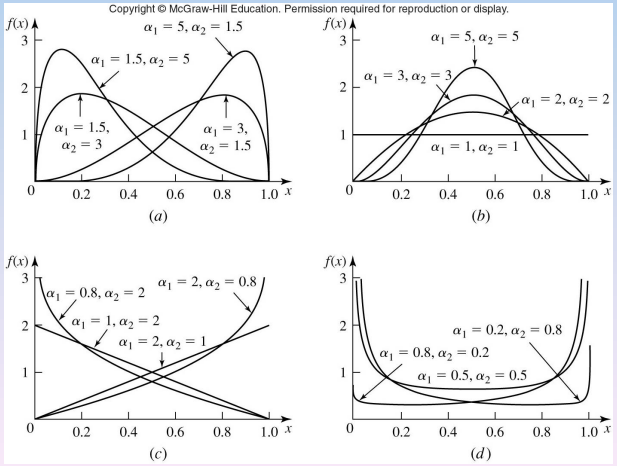
Continuous Distribution Use Summary

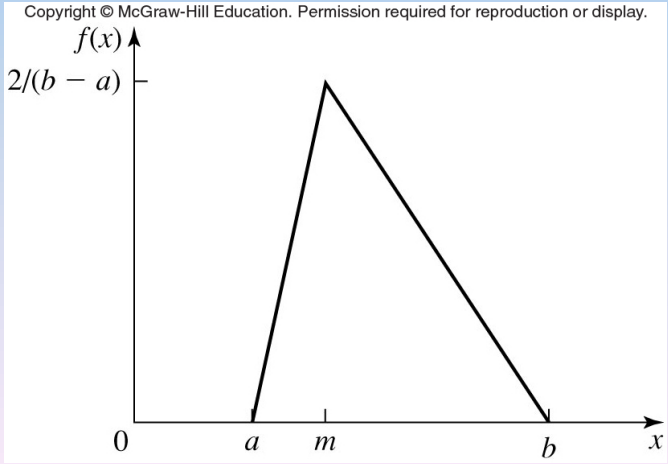
- Rough model in the absence of data
  - Weibull, lognormal, beta, triangular





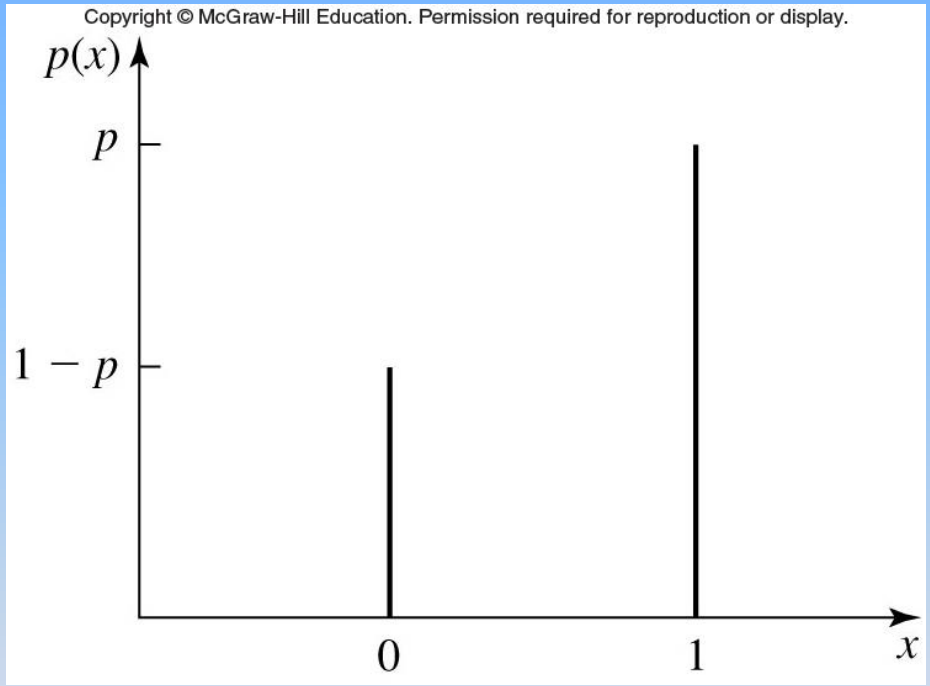






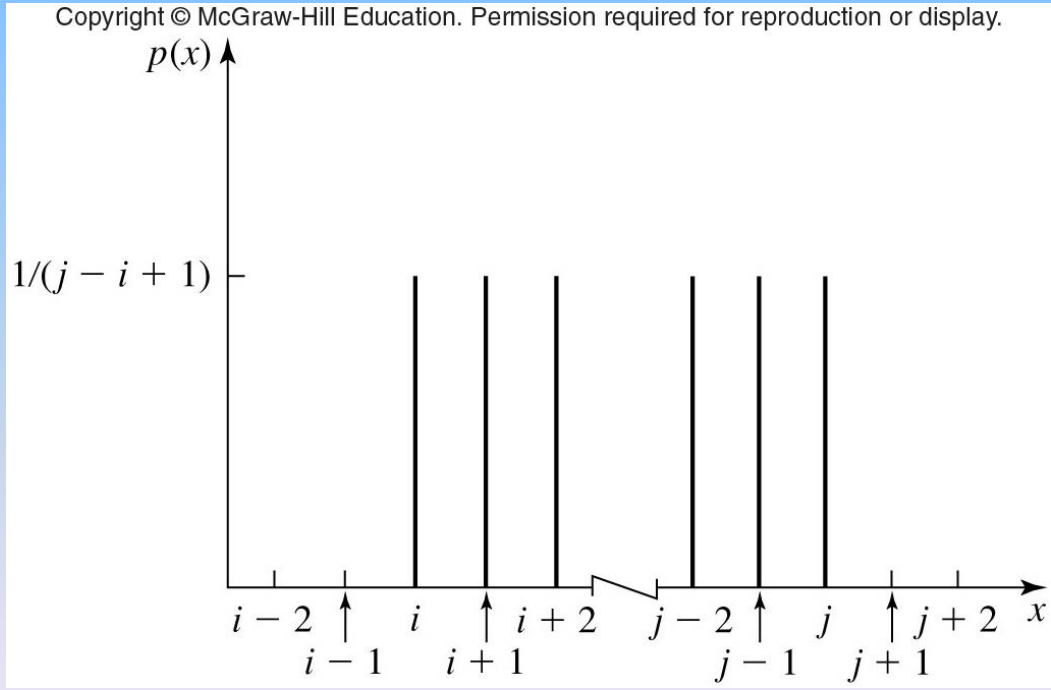
Useful Discrete Distributions (1)

- Bernoulli(p)
  - Used to generate some other discrete r.v.
- If X_i ~ Bernoulli(p), $X_1 + X_2 + \ldots + X_n$ ~ Binomial(n, p)
- A special case of Binomial(1, p)
- Number of failures before the first success ~ geom(p)



Useful Discrete Distributions (2)

- Discrete Uniform DU(I,j)
  - DU(0,1) is the same as Bernoulli(0.5)

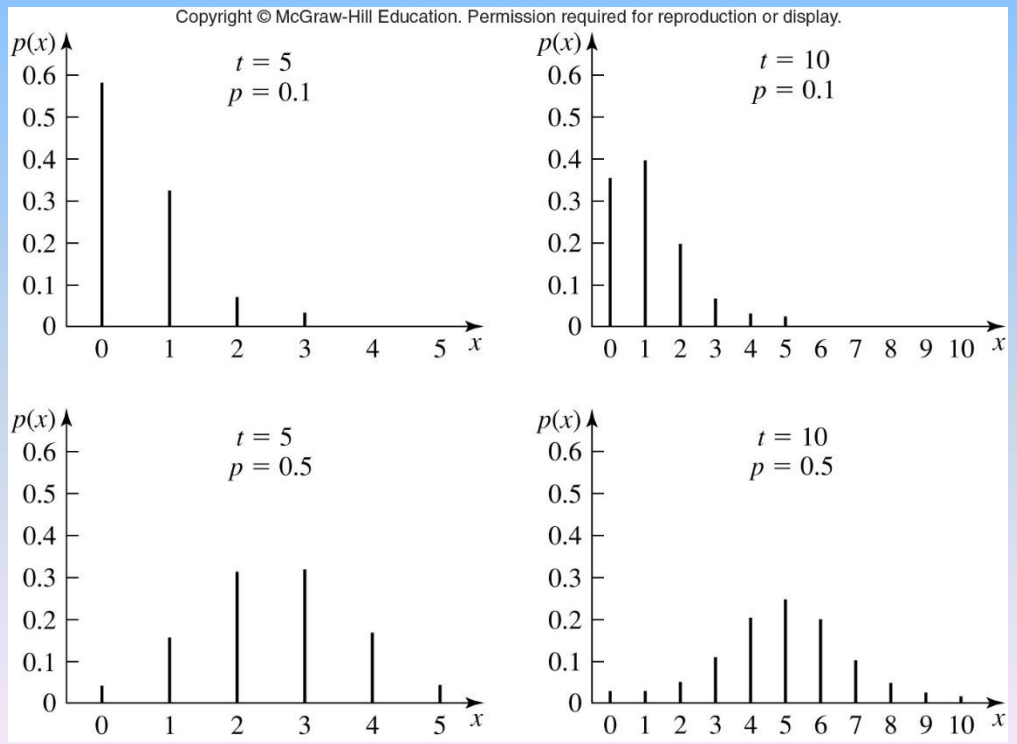


Useful Discrete Distributions (3)

- Binomial bin(t,p)
  - Number of successes in t independent Bernoulli trials with probability p of success
- If $X_i \sim bin(t_i,p)$, $X_1 + X_2 + \ldots + X_n \sim bin(t_1+t_2+\ldots+t_n, p)$
- $p = 0.5 \Leftrightarrow$

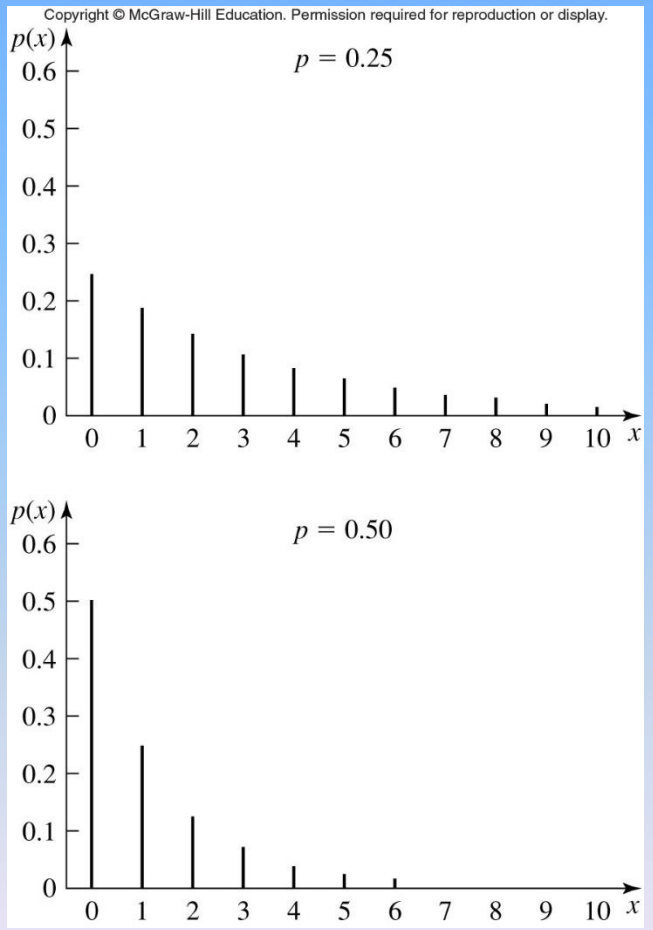
  bin(t,p) is symmetric
- $X \sim bin(t,p)$

  $\Leftrightarrow t - x \sim bin(t, 1-p)$



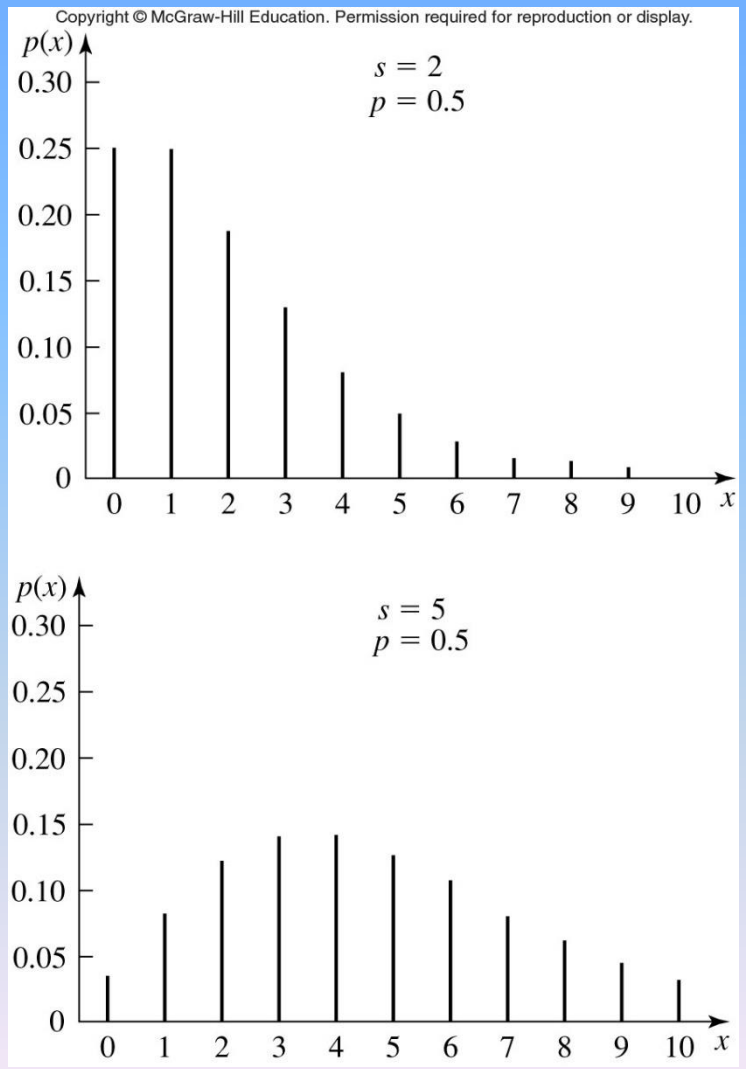
Useful Discrete Distributions (4)

- Geometric geom(p)
  - Number of failures before the first success in a sequence of independent Bernoulli trials with probability p of success
  - Discrete analog of the expo. distribution, also memoryless
  - Model number of items inspected before seeing the first defective item, number of items in a batch of random size, number of items demanded from an inventory
  - Y_i ~ Bernoulli(p), X = min{i:Y_i=1} = 1, X ~ geom(p)
  - X_i ~ geom(p), $X_1 + X_2 + \ldots + X_n$ ~ negbin(n,p)



Useful Discrete Distributions (5)

- Negative binomial negbin(s,p)
  - Number of failures before the *s*-th success in Bernoulli trials
  - Model number of good items inspected before seeing the s-th defection, # items in a batch of random size, # items demanded from an inventory
  - $X_i \sim \text{negbin}(s_i, p)$, $X_1 + X_2 + \ldots + X_n \sim \text{negbin}(s_1+s_2 + \ldots + s_n,p)$

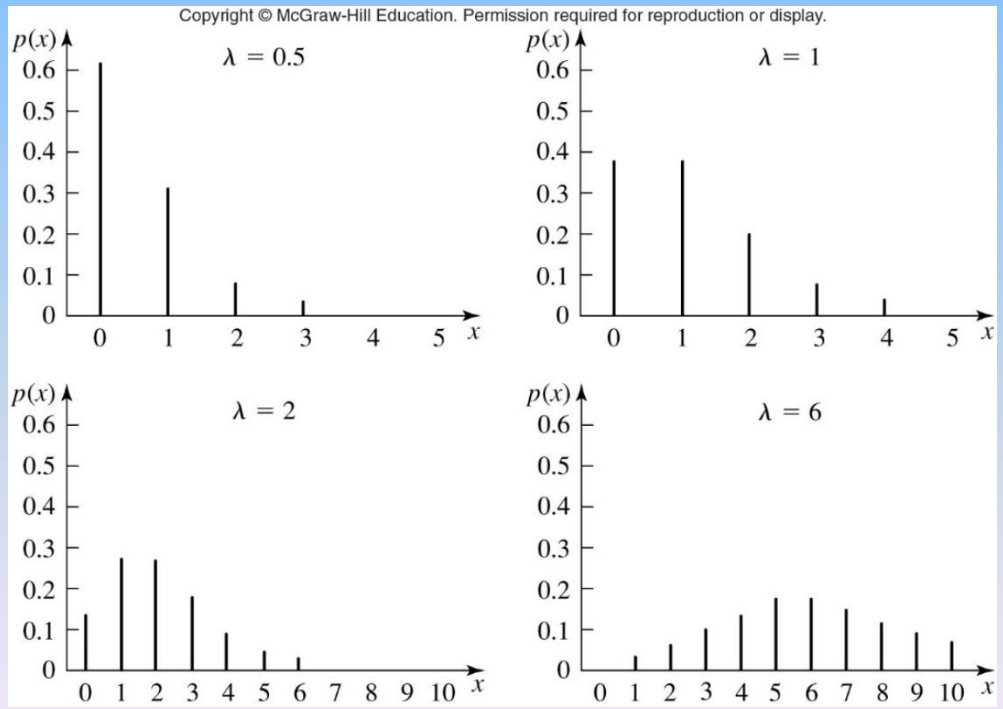


Useful Discrete Distributions (6)

- Poisson(λ)
  - Number of events in an interval of time when the events are occurring at a constant rate
  - Model # items in a batch of random size, # items demanded from an inventory
  - Y_i IID, $X = max\{i: \sum_{j=1}^{i} Y_j \leq 1\}$,

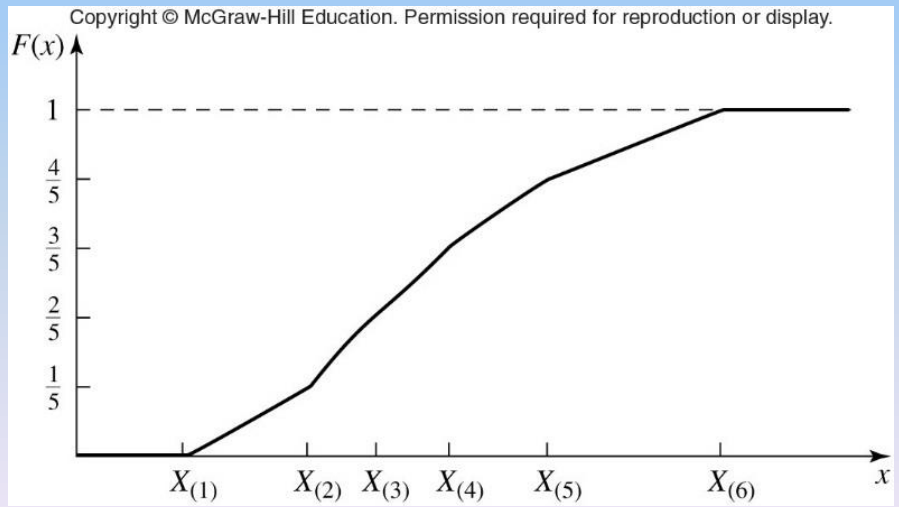
    X ~ Possion(λ) $\Leftrightarrow$ Y_i ~ expo($1/\lambda$),
  - Y_i IID, $X' = max\{i: \sum_{j=1}^{i} Y_j \leq \lambda\}$,

    X' ~ Possion(λ) $\Leftrightarrow$ Y_i ~ expo(1)
  - X_i ~ Poisson(λ_i), $X_1 + X_2 + \ldots + X_n$ ~ Possion ($\lambda_1 + \lambda_2 + \ldots + \lambda_n$)



Empirical Distributions (1)

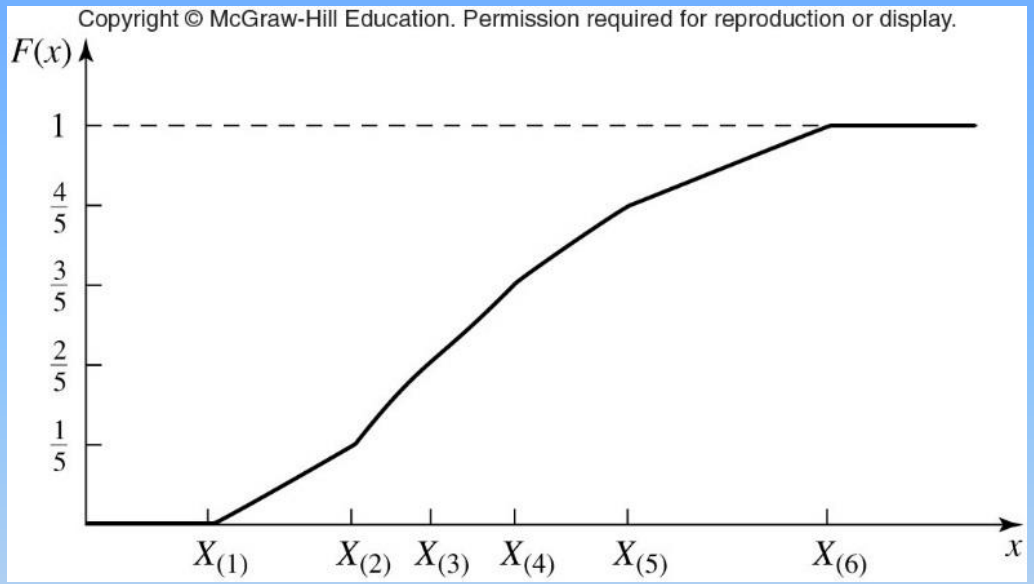
- For continuous r.v.
  - Have individual original sample values
  - Sort the samples into increasing order
  - $$F(x) = \begin{cases} 0 & \text{if } x < X_{(i)} \\ \frac{i-1}{n-1} + \frac{x - X_{(i)}}{(n-1)(X_{(i+1)} - X_{(i)})} & \text{if } X_{(i)} \le x < X_{(i+1)} \text{ for } i = 1,2,\dots,n-1 \\ 1 & \text{if } X_{(n)} \le x \end{cases}$$
  - Look at the r.v. as discrete uniform
  - Use uniform dist. within each interval
  - Con
    - Random values generated are within $X_{(1)}$ and $X_{(n)}$
    - The mean of F(x) $\neq \bar{X}(n)$



Continuous, piecewise-linear empirical distribution function from original data

Empirical Distributions

- For continuous r.v. (cont'd)
  - Have group data
    - The number of samples in each of several specified intervals $[a_0,a_1)$, $[a_1,a_2)$, …, $[a_{k-1}, a_k)$
    - Total number: n
    - *j*th interval contains n_j observations

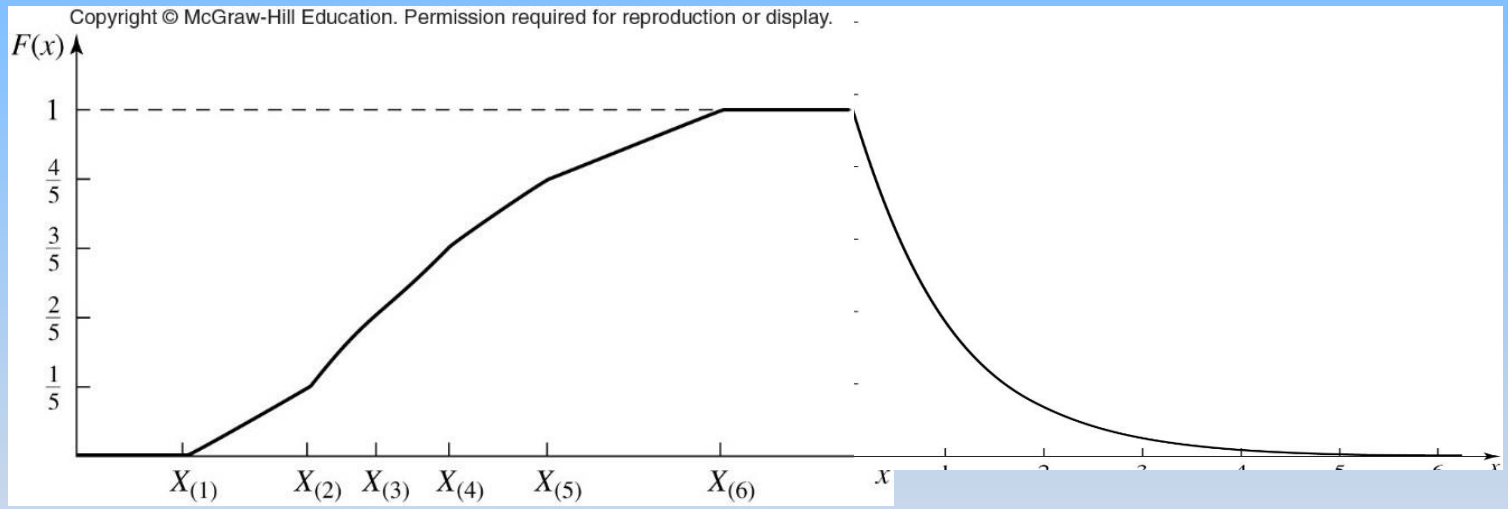


  - $G(a_0) = 0$, $G(a_j) = (n_1+n_2+\ldots+n_j)/n$ (for $0 < j \leq k$)
  - $F(x) =$

$$\begin{cases} 0 & \text{if } x < a_0 \\ G(a_{j-1}) + \frac{x-a_{j-1}}{a_j-a_{j-1}}[G(a_j) - G(a_{j-1})] & \text{if } a_{j-1} \leq x < a_j \text{ for } 0 < j \leq k \\ 1 & \text{if } a_n \leq x \end{cases}$$

Solve the Bound Problem

- Can append an expo. dist. to the right side



Empirical Distributions (3)

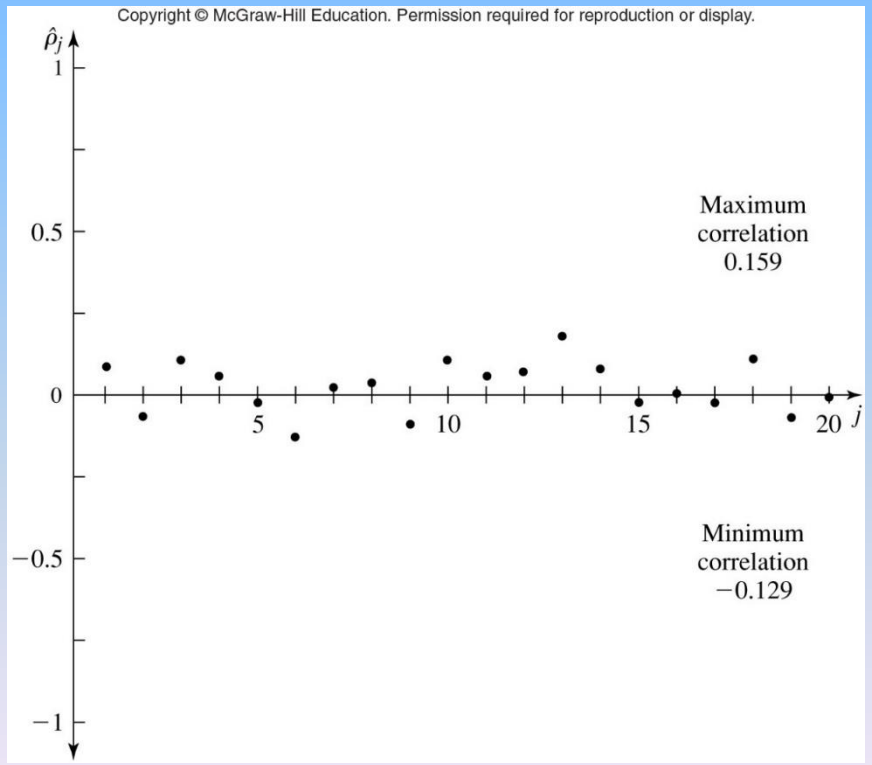
- For discrete r.v.
  - Have individual original sample values
    - For each possible value x, an empirical mass function is the proportion of the samples of that value
  - Have group data
    - Define a mass function such that the sum of the p(x)'s over all possible values of x in an interval is equal to the proportion of the samples in that interval
    - Individual p(x)'s can be allocated arbitrarily

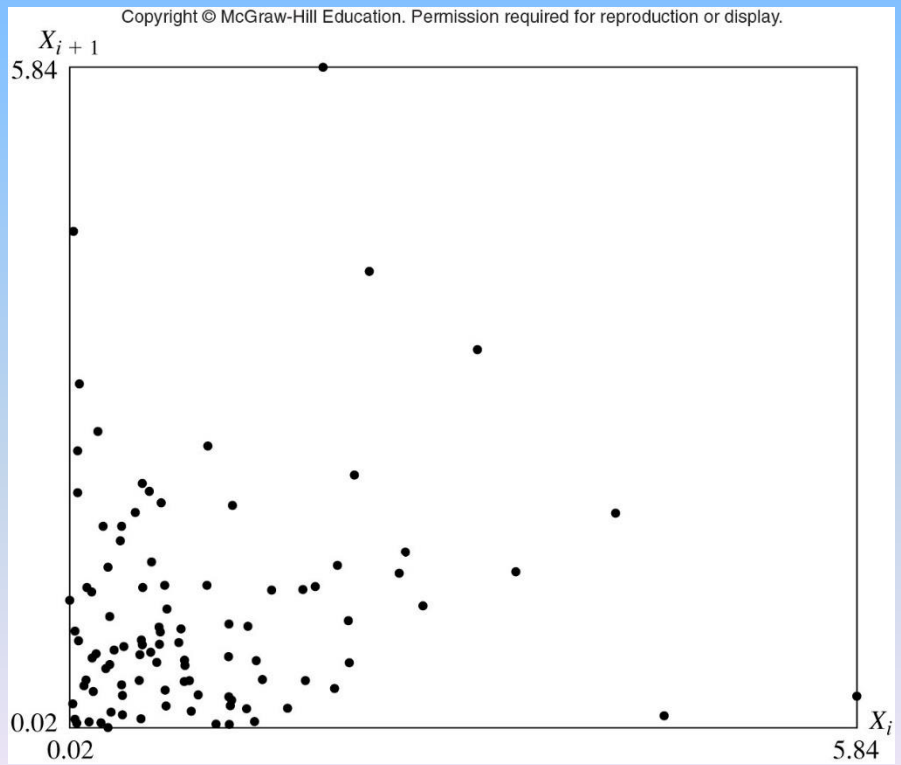
6.3 Techniques for Verifying Sample Independence

- A key assumption by many statistical techniques
  - Individual observations represent independent samples from an underlying distribution
- An example of non-independent data
  - Hourly samples of temperature from a specific city, starting at noon
    - Adjacent sample values will be positively correlated

Techniques for Verifying Sample Independence

- Graphical techniques
  - Correlation plot, scatter diagram

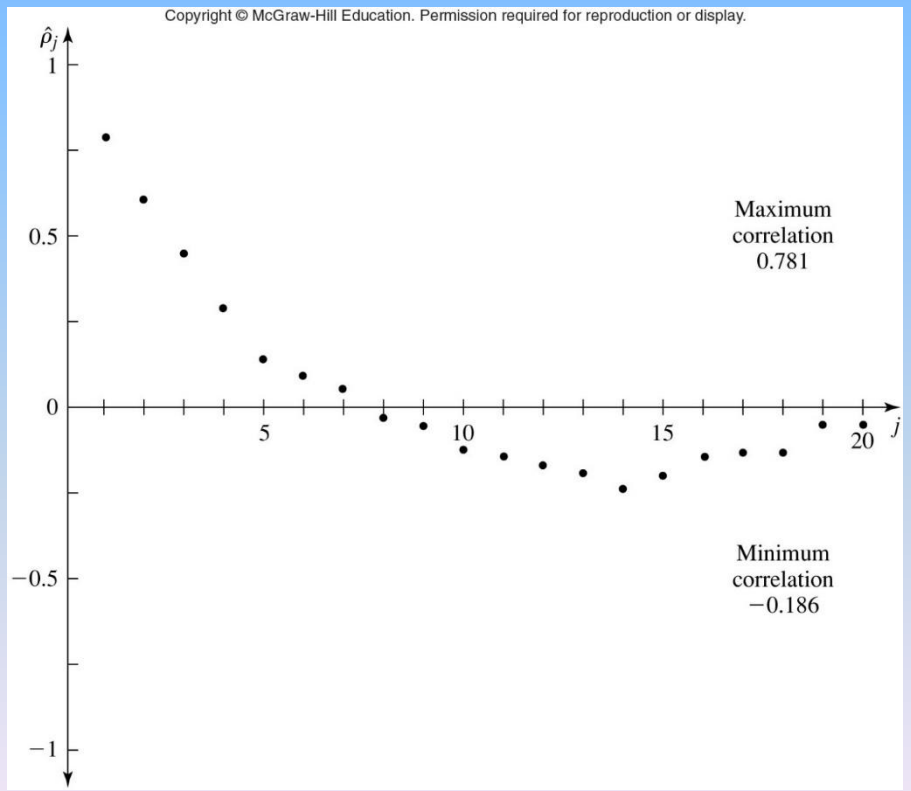


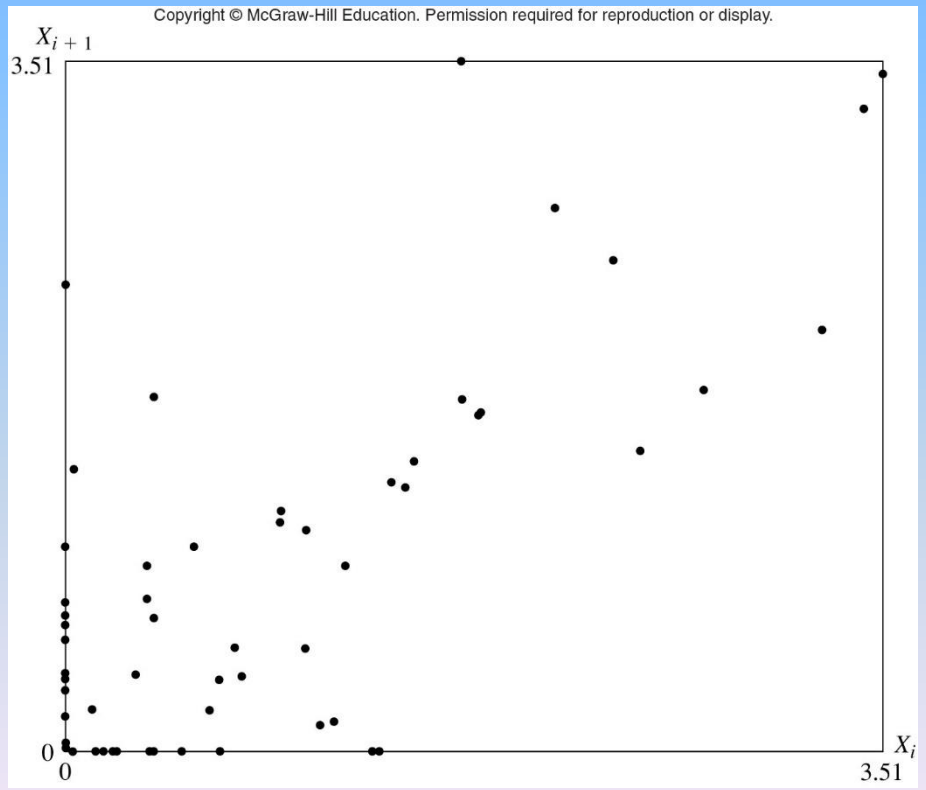


100 independent observations from an expo. dist. with a mean of 1

Techniques for Verifying Sample Independence

- Graphical techniques (cont'd)
  - Correlation plot, scatter diagram





100 delays in queue from M/M/1 with $\rho = 0.8$

Techniques for Verifying Sample Independence (cont'd)

- Rank von Neumann test
  - Requires there be no “ties” (equal values) in the data
  - This requirement generally will not be met for discrete data
- Run the test
  - Sec. 7.4.1

6.4 Activity I: Hypothesizing Families of Distributions

- Process of determining appropriate general families of distributions, based on their shape
  - Without concern for parameter values
- Prior knowledge can sometimes be used
  - E.g. customers arrive one at a time at a constant rate => IID expo.
  - E.g. Service time can't use normal dist. as it must be positive
  - E.g. Proportion of defective items can't be gamma as proportion must be in [0, 1]
- In practice, hypothesizing a distribution family is somewhat less structured

Summary Statistics (1)

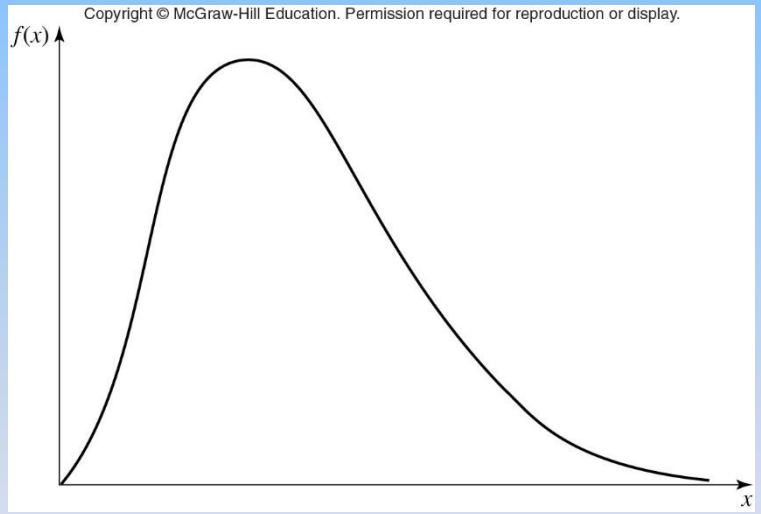
Function	Sample estimate (summary statistic)	Continuous (C) or discrete (D)	Comments
Minimum, maximum	$X_{(1)}, X_{(n)}$	C, D	$[X_{(1)}, X_{(n)}]$ is a rough estimate of the range
Mean μ	$\overline{X}(n)$	C, D	Measure of central tendency
Median $x_{0.5}$	$\hat{x}_{0.5}(n) = \begin{cases} X_{((n+1)/2)} & \text{if } n \text{ is odd} \\ [X_{(n/2)} + X_{((n/2)+1)}]/2 & \text{if } n \text{ is even} \end{cases}$	C, D	Alternative measure of central tendency
Variance σ^2	$S^2(n)$	C, D	Measure of variability
Coefficient of variation, $\text{cv} = \frac{\sqrt{\sigma^2}}{\mu}$	$\widehat{\text{cv}}(n) = \frac{\sqrt{S^2(n)}}{\overline{X}(n)}$	C	Alternative measure of variability
Lexis ratio, $\tau = \frac{\sigma^2}{\mu}$	$\hat{\tau}(n) = \frac{S^2(n)}{\overline{X}(n)}$	D	Alternative measure of variability
Skewness, $\nu = \frac{E[(X-\mu)^3]}{(\sigma^2)^{3/2}}$	$\hat{\nu}(n) = \frac{n^2}{(n-1)(n-2)} \frac{\sum_{i=1}^{n} [X_i - \overline{X}(n)]^3/n}{[S^2(n)]^{3/2}}$	C, D	Measure of symmetry

Summary Statistics (2)

- Can be used to suggest an appropriate distribution family
- If $\bar{X}(n) \approx \hat{X}_{0.5}(n)$, the underlying distribution may be symmetric
- $\widehat{cv}\ (n) \approx 1$, the underlying distribution may be exponential
- For gamma or Weibull dist., $\widehat{cv}\ (n) > 1 / \approx 1 / < 1$ for shape parameter α < 1 / =1 / >1

Summary Statistics (3)

- $\widehat{cv}\ (n) > 1$, the underlying distribution has the following shape, it'd better be modelled by lognormal
  - Lognormal has this shape and cv can be any positive value
  - Gamma and Weibull has this shape when α > 1 and cv < 1



- cv not useful for other distributions

Summary Statistics (4)

- Lexis ratio = 1, < 1, > 1, distribution may be Poisson, binomial, and negative binomial respectively
- Estimated skewness can be used to ascertain the shape of the underlying distribution
  - > 0, skewed to the right
    - Many distributions in practice are
  - < 0, skewed to the left

Histogram for Continuous Data Set (1)

- A graphical estimate of the plot of the PDF corresponding to data
- Split the range of data values to multiple disjoint adjacent intervals $[b_{i-1}, b_i)$ $(i = 1…k)$ of the same width
  - May need to remove extremely large or small values
  - Height of the bar of an interval is the proportion of the data values in the interval
- Compare the basis of shape
  - Ignore location and scale

Histogram for Continuous Data Set (2)

- $h(x) = \begin{cases} 0 & \text{if } x < b_0 \\ h_j & \text{if } b_{j-1} \le x < b_j \\ 0 & \text{if } b_k \le x \end{cases}$
- $P\left(b_{j-1} \le X < b_j\right) = \int_{b_{j-1}}^{b\_j} f(x)dx = \Delta b f(y)$, $y \in \left(b_{j-1}, b_j\right)$
- $h(y) \approx \Delta b f(y), \therefore h(y) \propto f(y)$

Histogram for Continuous Data Set (3)

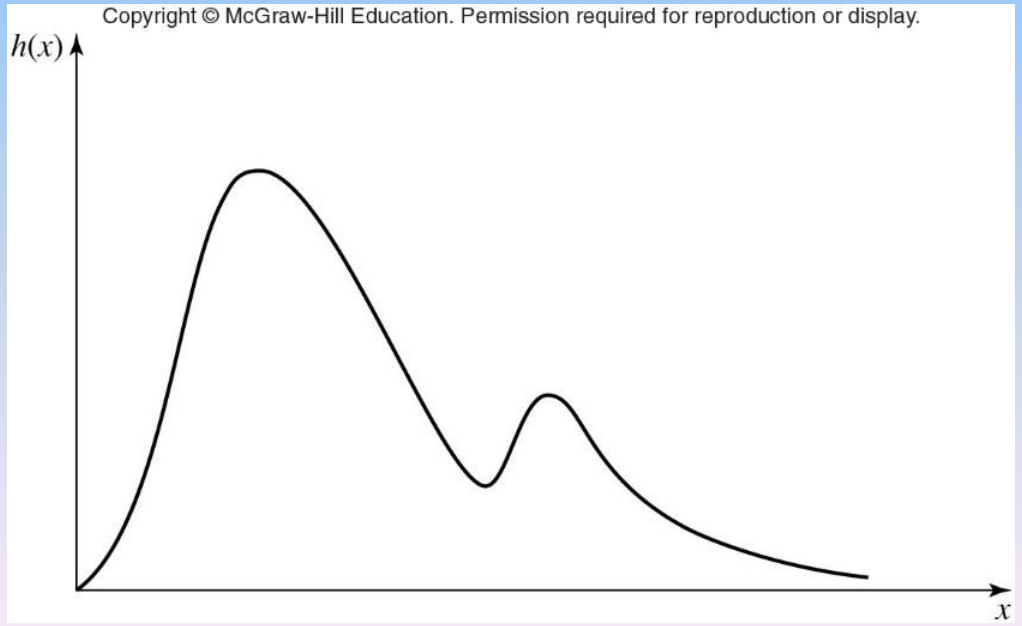
- How many intervals?
  - No definite guide
  - Struge's rule: $k = \lfloor 1 + \log_2 n \rfloor$
    - Not very useful
  - Try multiple values and choose the smallest one giving a "smooth" histogram
    - Too many intervals => h_j's vary too much
    - Too few intervals => underlying density is masked

Histogram for Discrete Data Set (1)

- No need for intervals
- Plot vertical bars of height h_j vs x_j
  - For each possible value x_j, h_j is the proportion of the sample values = x_j
  - h_j is an unbiased estimator of $p(x_j)$ ($p(x)$ is the true PMF)

Histogram with Several Local Modes

- Split the data into two cases
  - p_j being the proportion of observations for case j
- Overall PDF
  - $f(x) = p_1 f_1(x) + p_2 f_2(x)$



Quantile Summaries

- Useful for determining whether the distribution is symmetric or skewed right or left
- q-quantile of F(x): x_q
  - $0 < F(x) < 1$, continuous and strictly increasing
  - For $0 < q < 1$, $F(x_q) = q$
- If the underlying distribution is symmetric, the four midpoints should be about the same

Estimates of quantiles

Quantile	Depth	Sample value(s)		Midpoint
Median	$i = (n + 1)/2$	$X_{(i)}$		$X_{(i)}$
Quartiles	$j = (\lfloor i \rfloor + 1)/2$	$X_{(j)}$	$X_{(n-j+1)}$	$[X_{(j)} + X_{(n-j+1)}]/2$
Octiles	$k = (\lfloor j \rfloor + 1)/2$	$X_{(k)}$	$X_{(n-k+1)}$	$[X_{(k)} + X_{(n-k+1)}]/2$
Extremes	1	$X_{(1)}$	$X_{(n)}$	$[X_{(1)} + X_{(n)}]/2$

Drive-up Banking Example (1)

- 220 car arrived during 90 minutes and thus 219 interarrival times
- Hypothesize the distribution family of interarrival times
  - Cars arrive one at a time -> independent
  - Number of cars arriving in every 15 minutes are about the same
  - Exponential interarrival times

Drive-up Banking Example (2)

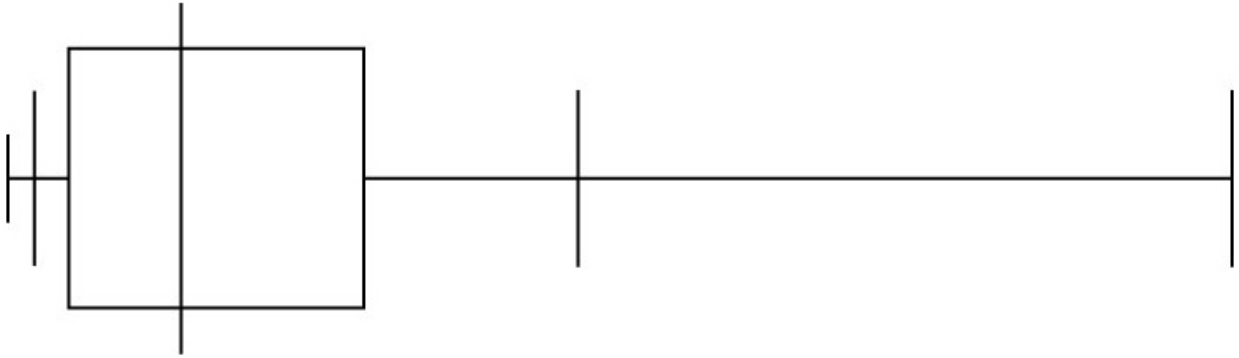
- To substantiate the hypothesis
  - $\bar{X}(219) = 0.399 > 0.270 = \hat{x}_{0.5}$
  - $\hat{v}(219) = 1.478$
    - Underlying distribution is probably skewed to the right
  - $\widehat{cv}(219) = 0.953$
    - Theoretical value for expo. dist. is 1

Summary statistic	Value
Minimum	0.010
Maximum	1.960
Mean	0.399
Median	0.270
Variance	0.144
Coefficient of variation	0.953
Skewness	1.478

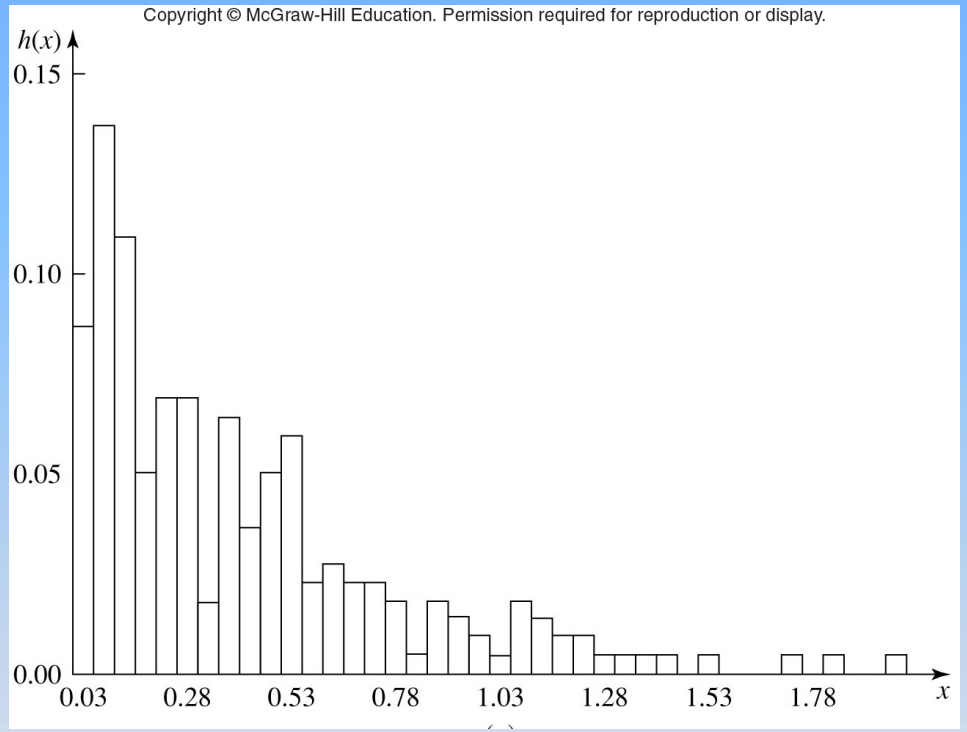
Drive-up Banking Example (3)

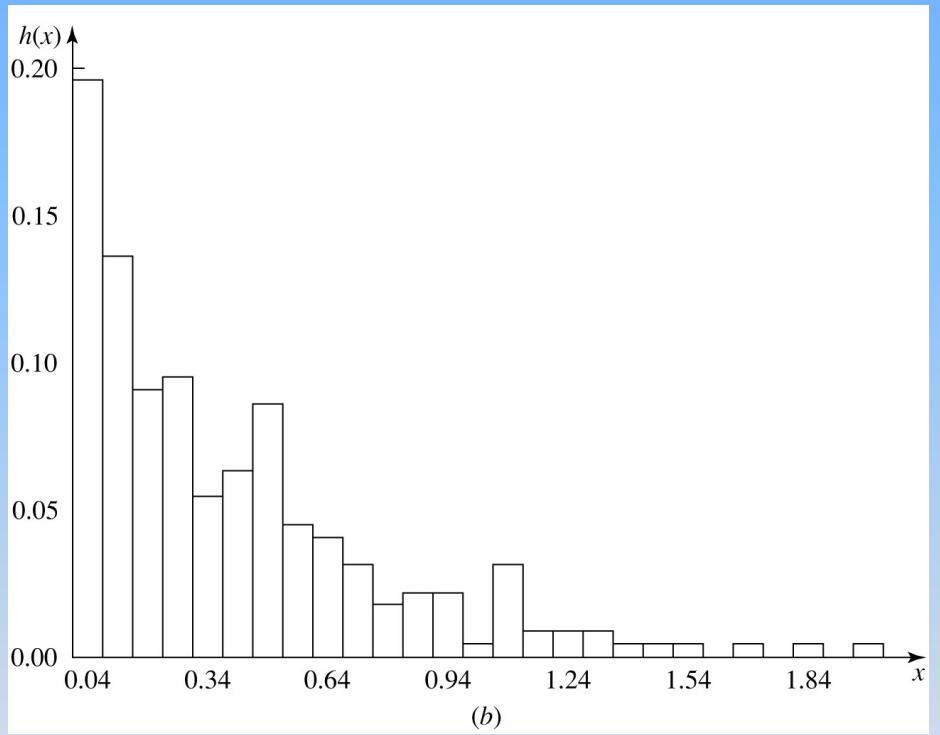
- Quantile summary and box plot

Quantile	Depth	Sample Value(s)		Midpoint
Median	110	0.27		0.270
Quartiles	55.5	0.100	0.545	0.323
Octiles	28	0.050	0.870	0.460
Extremes	1	0.010	1.960	0.985



Drive-up Banking Example (4)

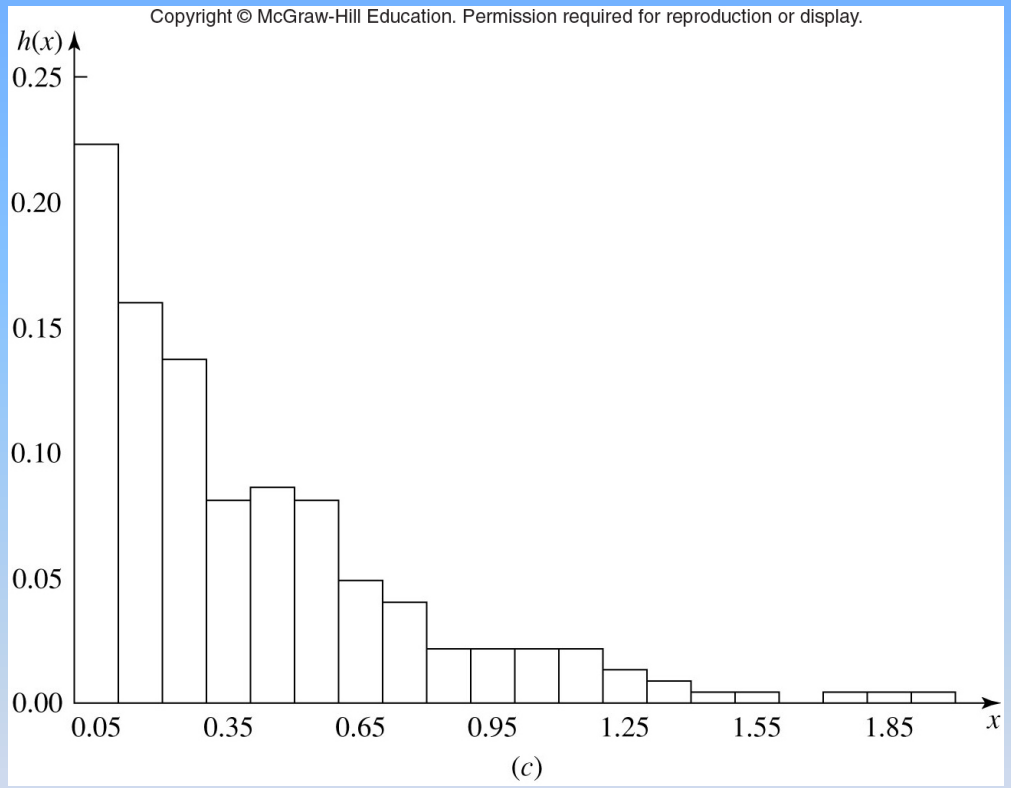




Δb=0.050

Δb=0.075

Drive-up Banking Example (5)



Δb=0.100

- Sturge's rule gives Δb=0.250

Inventory Demand Example (1)

- 156 observations on number of items demanded in a week from an inventory over 3 years

Summary statistic	Value
Minimum	0.000
Maximum	11.000
Mean	1.891
Median	1.000
Variance	5.285
Lexis ratio	2.795
Skewness	1.687

0(59),	1(26),	2(24),	3(18),	4(12),
5(5),	6(4),	7(3),	9(3),	11(2)

Inventory Demand Example (2)

- Lexis ratio $\hat{\tau}(156) = 2.795$
  - Binomial and Poisson not likely
- Skewness $\hat{v}(156) = 1.687$
  - Discrete uniform not likely
- Geometric or negative binomial?
  - Histogram matches the former better

6.5 Activity II: Estimating Parameters

- Need to specify parameter values to completely specify the distribution
- Estimator
  - Numerical function of the data
  - Maximum-likelihood estimators (MLEs) considered here
- Desirable statistical properties of MLEs
  - Unique
  - $E(\hat{\theta}) = \theta$ as $n \to \infty$
  - Invariant
    - MLE of $h(\theta)$ is $h(\hat{\theta})$
  - Asymptotically normally distributed
  - Strongly consistent $\lim_{n\to\infty} \hat{\theta} = \theta$ (w.p. 1)

Maximum-likelihood Estimators

- Given data X_i (i = 1, 2, …, n), likelihood function
  - Discrete (PMF: $p_\theta(x)$)

$$L(\theta) = p_\theta(X_1)p_\theta(X_2) \dots p_\theta(X_n)$$

  - Continuous (PDF: $fp_\theta(x)$)

$$L(\theta) = f_\theta(X_1)f_\theta(X_2) \dots f_\theta(X_n)$$

- Find $\hat{\theta}$ that maximizes $L(\theta)$

MLE Example – Expo. Dist. (1)

- Expo. dist. $f_\beta(x) = \frac{1}{\beta} e^{-\frac{x}{\beta}}$ for x >= 0
- Likelihood function

$$L(\beta) = \frac{1}{\beta} e^{-\frac{X_1}{\beta}} \frac{1}{\beta} e^{-\frac{X_2}{\beta}} \ldots \frac{1}{\beta} e^{-\frac{X_n}{\beta}}$$

$$= \beta^{-n} \exp(-\frac{1}{\beta} \sum_{i=1}^{n} X_i)$$

- Log-likelihood function

$$l(\beta) = lnL(\beta) = -nln(\beta) - \frac{1}{\beta} \sum_{i=1}^{n} X_i$$

MLE Example – Expo. Dist. (2)

$$\frac{dl}{d\beta} = -\frac{n}{\beta} + \frac{1}{\beta^2}\sum_{i=1}^{n} X_i = 0$$

$$\beta = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}(n)$$

$$\frac{dl^2}{d\beta^2} = \frac{n}{\beta^2} - \frac{2}{\beta^3}\sum_{i=1}^{n} X_i < 0 \text{ when } \beta = \bar{X}(n)$$

MLE Example – Geometric Dist.

- PMF: $p_p(x) = p(1-p)^x$ for x = 0,1, …
- Likelihood function

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

- Log-likelihood function

$$l(p) = lnL(p) = nlnp + \sum_{i=1}^{n} X_i \ln(1-p)$$

$$\frac{dl}{dp} = \frac{n}{p} - \frac{\sum_{i=1}^n X_i}{1-p} = 0$$

$$p = \frac{1}{\bar{X}(n)+1}$$

$$\frac{dl^2}{dp^2} = -\frac{n}{p^2} - \frac{\sum_{i=1}^n X_i}{(1-p)^2} < 0$$

Confidence Interval of Parameters

- $\sqrt{n}(\hat{\theta}-\theta) \rightarrow N(0, \delta(\theta))$

  $$\delta(\theta) = -\frac{n}{E\left(\frac{d^2 l}{d\theta^2}\right)}$$

  $$\Rightarrow \frac{(\hat{\theta}-\theta)}{\sqrt{\frac{\delta(\theta)}{n}}} \rightarrow N(0,1)$$

- 100(1-α) percent confidence interval

$$\hat{\theta} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\delta(\theta)}{n}}$$

Parameter C.I. Example

- Inventory demand example
- 90% confidence interval of p of Geometric distribution

$$E\left(\frac{dl^2}{dp^2}\right) = -\frac{n}{p^2} - \frac{\frac{n(1-p)}{p}}{(1-p)^2} = -\frac{n}{p^2(1-p)}$$

$$\delta(p) = p^2(1-p)$$

$$\hat{p} \pm 1.645\sqrt{\frac{\hat{p}^2(1-\hat{p})}{n}} = 0.346 \pm 0.037$$

Sensitivity to Parameters

- Run simulations for the parameters set at the lower endpoint/center/upper endpoint of the confidence interval
- Check if the performance measure varies much
  - If yes, sensitive to parameters
    - Need better parameter estimate
    - Usually entail collecting more data

MLE of Multiple Parameters

- E.g. for gamma distribution

$$L(\alpha,\beta) = \frac{\beta^{-n\alpha}\left(\prod_{i=1}^{n} X_i\right)^{\alpha-1} exp\left[-\frac{1}{\beta}\sum_{i=1}^{n} X_i\right]}{\left(\Gamma(\alpha)\right)^{n}}$$

$$l(\alpha,\beta) = \ln L(\alpha,\beta)$$

Solve $\frac{\partial l}{\partial \alpha} = 0$ and $\frac{\partial l}{\partial \beta} = 0$ simultaneously

Or,

$$T = \left[\ln \bar{X}(n) - \sum_{i=1}^{n} \frac{\ln X_i}{n}\right]^{-1}$$

Look up Table 6.21

T	$\hat{\alpha}$	T	$\hat{\alpha}$	T	$\hat{\alpha}$	T	$\hat{\alpha}$
0.08	0.068	2.10	1.189	6.40	3.357	16.50	8.413
0.09	0.076	2.20	1.240	6.60	3.458	17.00	8.663
0.10	0.083	2.30	1.291	6.80	3.558	17.50	8.913
0.11	0.090	2.40	1.342	7.00	3.658	18.00	9.163
0.12	0.098	2.50	1.393	7.20	3.759	18.50	9.414
0.13	0.105	2.60	1.444	7.40	3.859	19.00	9.664
0.14	0.112	2.70	1.495	7.60	3.959	19.50	9.914
0.15	0.119	2.80	1.546	7.80	4.059	20.00	10.164
0.16	0.126	2.90	1.596	8.00	4.159	20.50	10.414
0.17	0.133	3.00	1.647	8.20	4.260	21.00	10.664
0.18	0.140	3.10	1.698	8.40	4.360	21.50	10.914
0.19	0.147	3.20	1.748	8.60	4.460	22.00	11.164
0.20	0.153	3.30	1.799	8.80	4.560	22.50	11.414
0.30	0.218	3.40	1.849	9.00	4.660	23.00	11.664
0.40	0.279	3.50	1.900	9.20	4.760	23.50	11.914
0.50	0.338	3.60	1.950	9.40	4.860	24.00	12.164
0.60	0.396	3.70	2.001	9.60	4.961	24.50	12.414
0.70	0.452	3.80	2.051	9.80	5.061	25.00	12.664
0.80	0.507	3.90	2.101	10.00	5.161	30.00	15.165
0.90	0.562	4.00	2.152	10.50	5.411	35.00	17.665
1.00	0.616	4.20	2.253	11.00	5.661	40.00	20.165
1.10	0.669	4.40	2.353	11.50	5.912	45.00	22.665
1.20	0.722	4.60	2.454	12.00	6.162	50.00	25.166
1.30	0.775	4.80	2.554	12.50	6.412		

Finding MLE

- Generally not as simple as the examples
- Numerical methods must be used in many cases

Input-model Uncertainty

- Model uncertainty
  - Not sure about distribution family for input
- Parameter uncertainty
  - Unsure about distribution parameters
- A confidence interval for a simulation performance measure would be ideal
  - Take into account both sampling variability of the simulation model (Ch. 9) and input model uncertainty

6.6 Activity III: Determining How Representative the Fitted Distributions Are

- None of the fitted distributions will be exactly correct
  - Goal: accurate enough for intended purposes of the model
- Heuristic procedures
- Goodness-of-fit tests

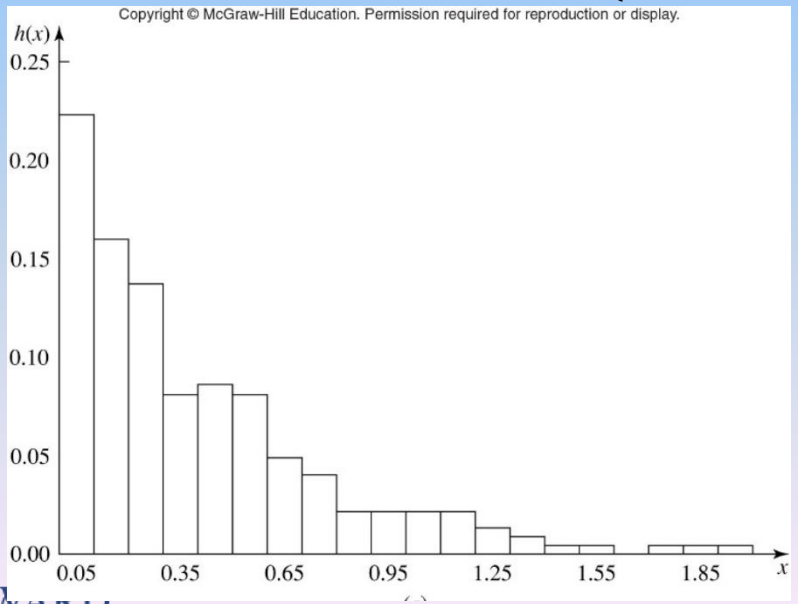
Heuristic Procedures (1)

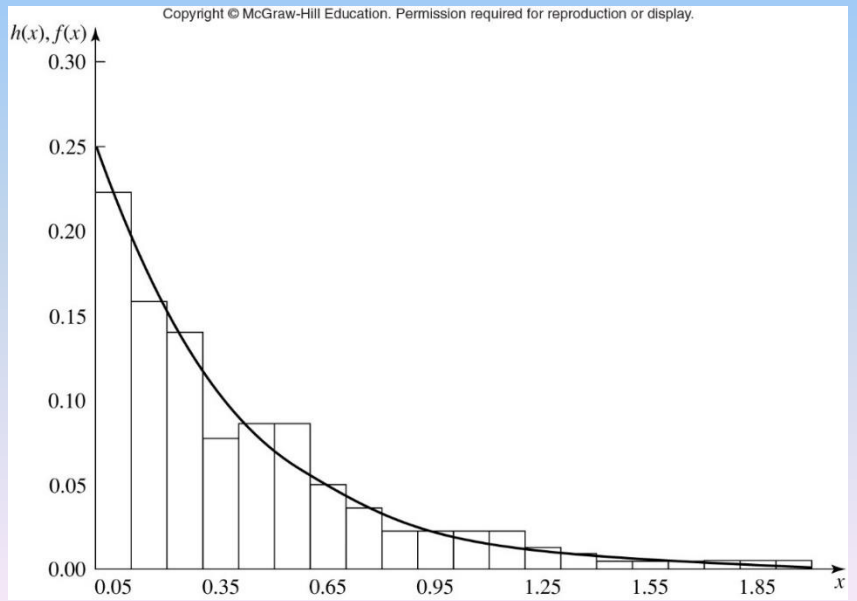
- Density-histogram plots and frequency comparisons
- For continuous data
  - Histogram intervals $[b_0,b_1)$, $[b_1,b_2)$, … , $[b_{k-1},b_k)$
  - Calculate $r_j = \int_{b_{j-1}}^{b_j} \hat{f}(x)dx$
  - Plot both h_j and r_j in the *j*th histogram interval for j=1,2,…,k
- For discrete data
  - Calculate $r_j = \hat{p}(x_j)$
  - Plot both h_j and r_j versus x_j for all relevant values of x_j

Density-Histogram Plot Example

- Drive-up Banking Example
- Hypothesized an expo. distribution, MLE $\hat{\beta} = 0.399$

$$\hat{f}(x) = \begin{cases} 2.506e^{-\frac{x}{0.399}} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$





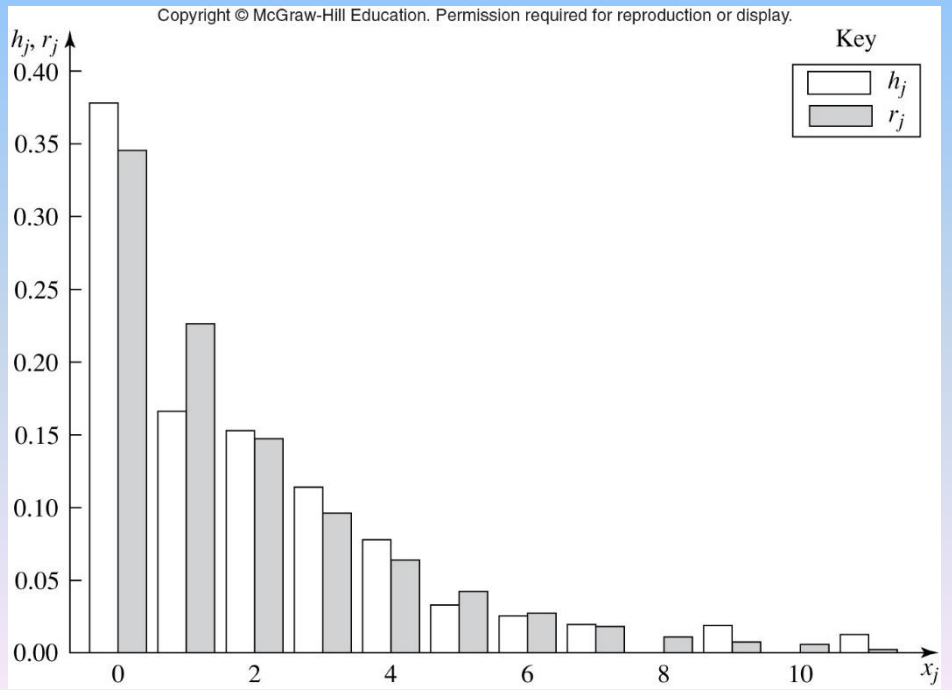
Frequency Comparison Example

- Inventory demand example
- Hypothesized a geometric distribution, MLE $\hat{p} = 0.346$

$$\hat{p}(x) = \begin{cases} 0.346(0.654)^x & \text{if } x = 0,1,2,\ldots \\ 0 & \text{otherwise} \end{cases}$$

$$r_j = \hat{p}(x_j) = \hat{p}(j-1)$$

for j = 1,2, …, 12

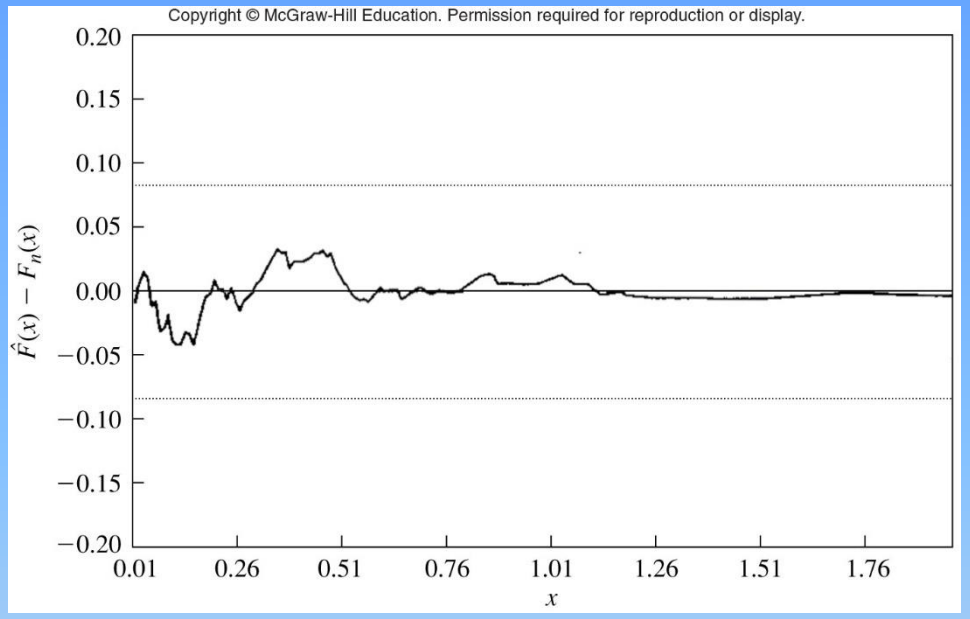


Heuristic Procedures (2)

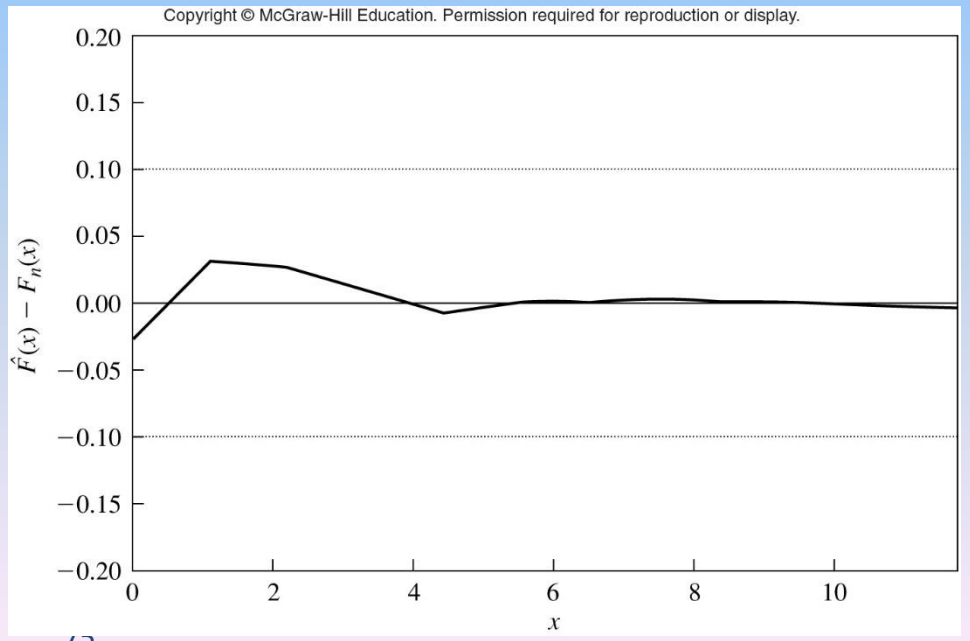
- Distribution-function-differences plots
  - A comparison of the individual probabilities of the fitted distribution and of that of the underlying distribution (approx. by empirical distribution)
  - $F_n(x) = \frac{\text{number of } x_i's \leq x}{n}$
  - Not easy to eyeball for differences or similarities in the S-shaped curves of $\hat{F}(x)$ and $F_n(x)$
  - Instead, plot the differences between $\hat{F}(x)$ and $F_n(x)$ over the range of the data
    - If perfect fit, should be on the x axis

Distribution-function-differences Plot Examples

- Drive-up Banking Example

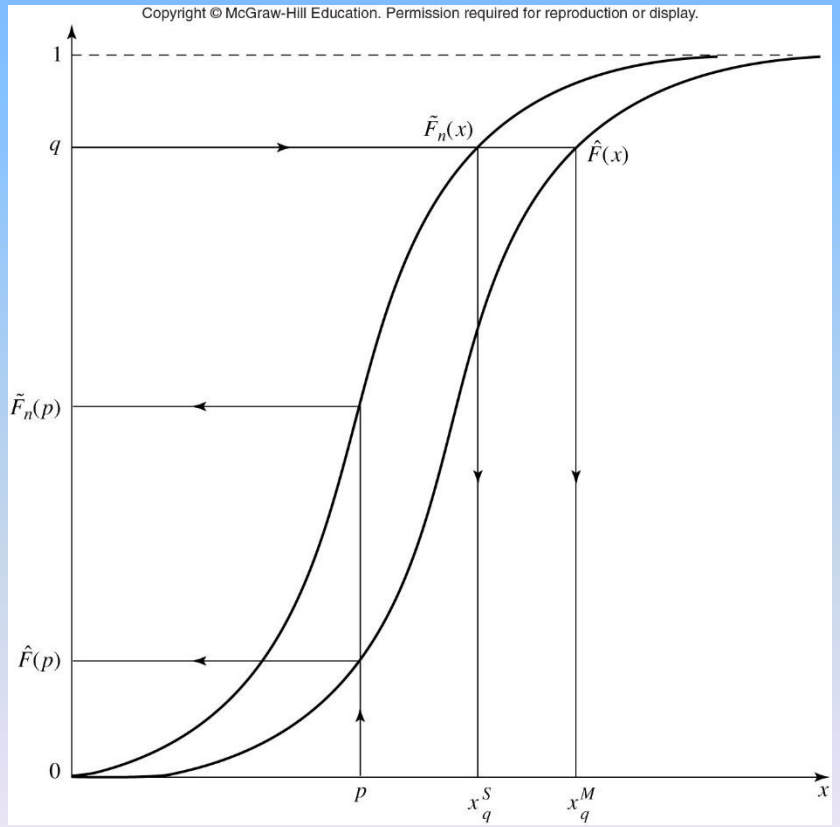


- Inventory demand example



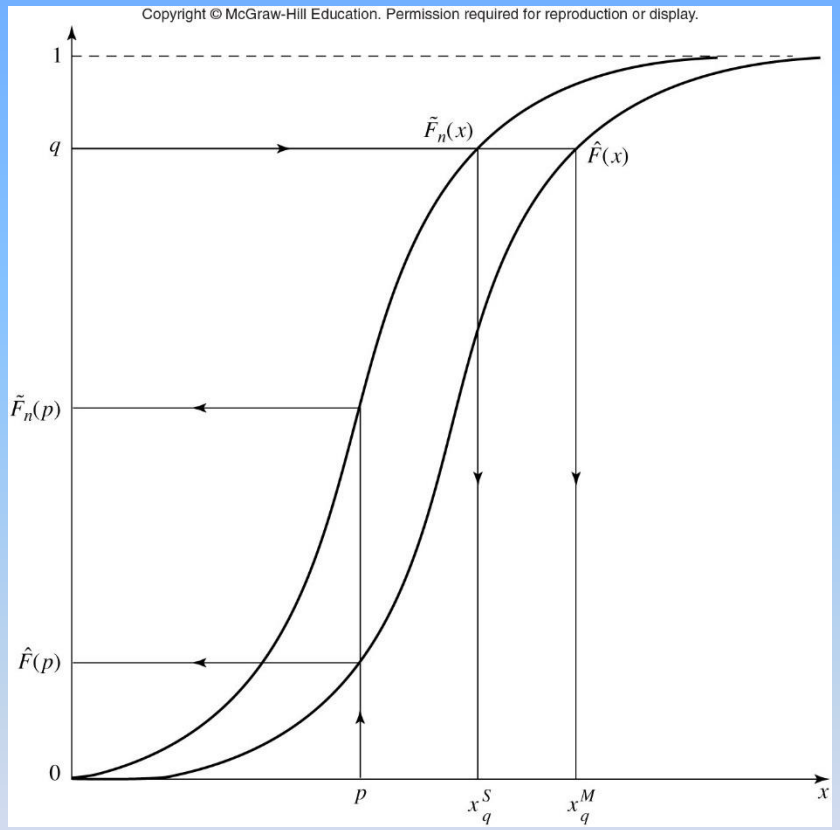
Heuristic Procedures (3)

- Probability plots
- New empirical distribution $\tilde{F}_n(X_{(i)}) = \frac{i-0.5}{n}$
- Quantile-quantile (Q-Q) plot
  - For continuous data sets
  - The q_i-quantile of the fitted distribution function $\hat{F}(x)$ vs. the q_i-quantile of $\tilde{F}_n(x)$
  - $q_i = \frac{i-0.5}{n}$
  - If the two distributions are the same and the sample size is large, the plot will be approx. the 45° line.
  - For small to moderate sample sizes, it may deviate from the 45° line.
  - Requires $\hat{F}^{-1}(x)$, may need numerical approximation



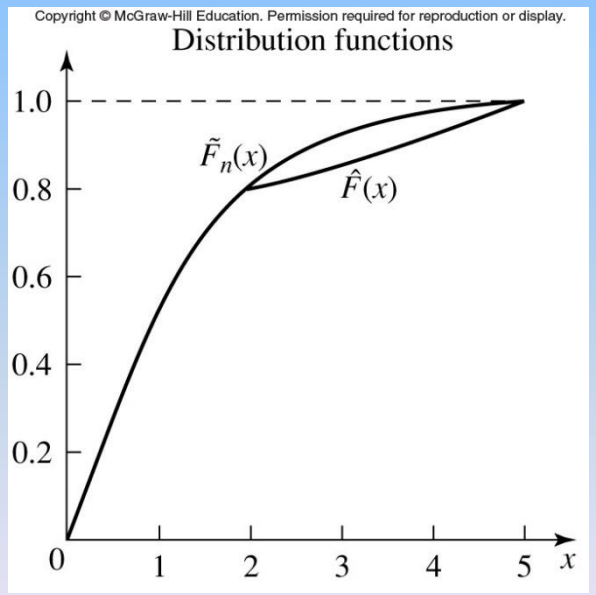
Heuristic Procedures (3) (cont'd)

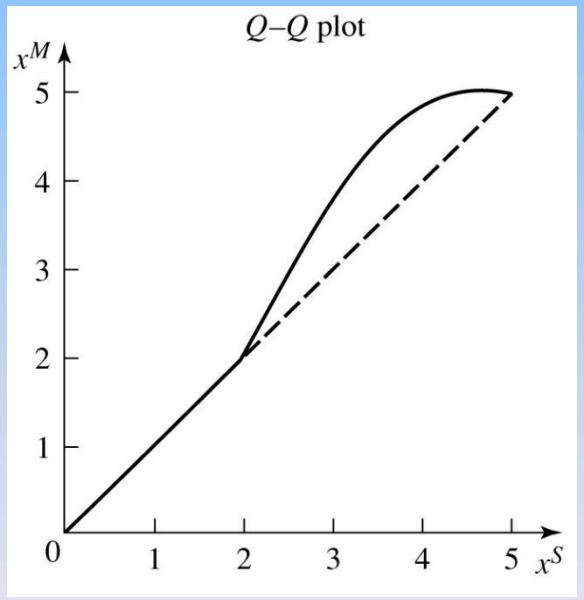
- Probability-probability (P-P) plot
  - For continuous and discrete data sets
  - Plot $\hat{F}(p)$ vs. $\widetilde{F_n}(p)$ for p values on the abscissa
  - If the two distributions are the same and the sample size is large, the plot will be approx. the 45° line.
  - For small to moderate sample sizes, it may deviate from the 45° line.

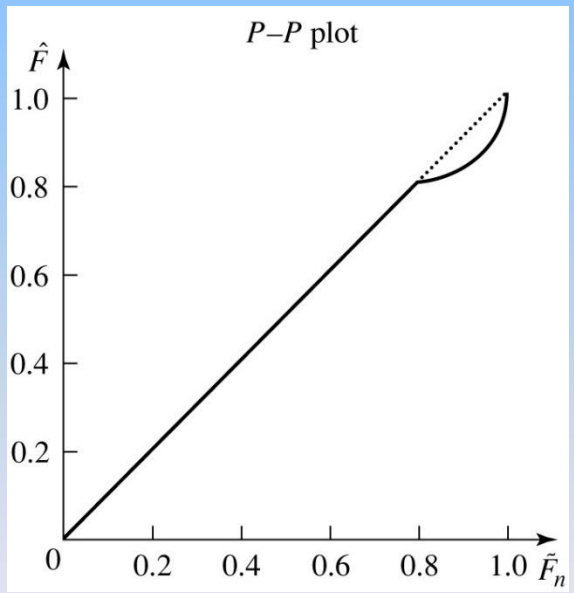


Heuristic Procedures (3) (cont'd)

- Q-Q plots amplify differences between the tails of the two distributions

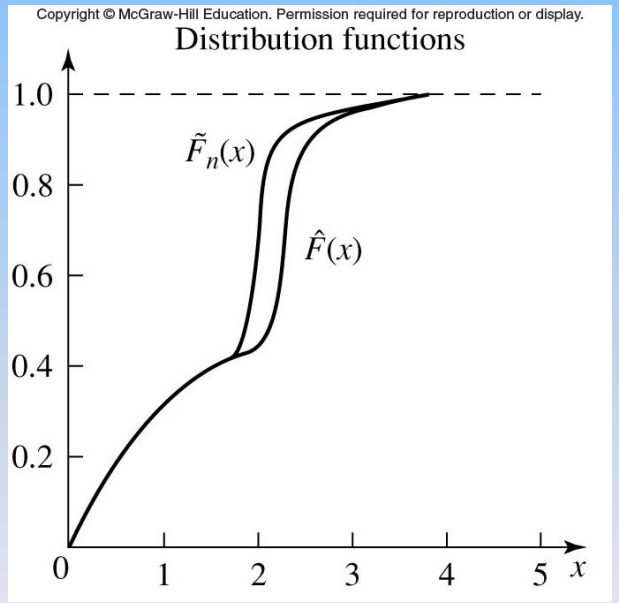


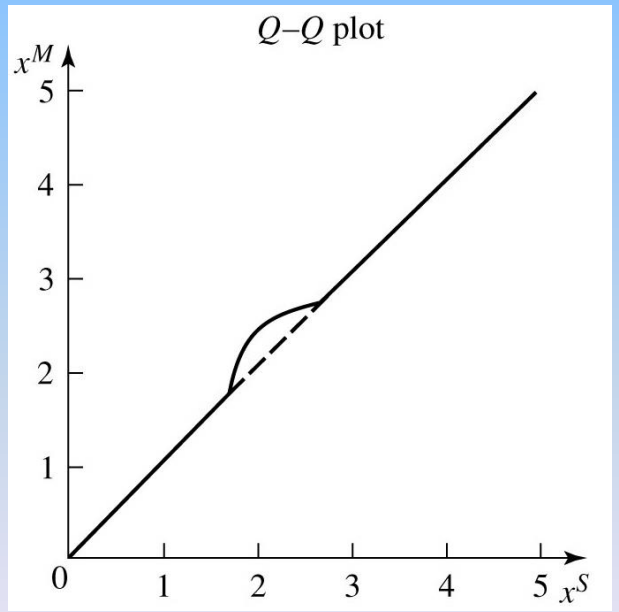


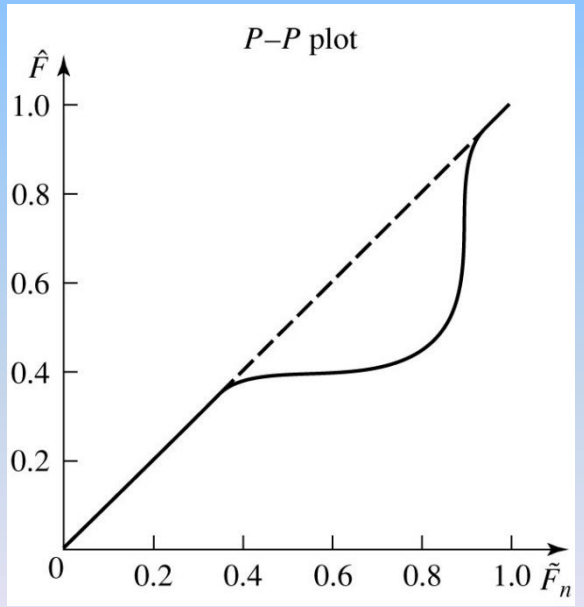


Heuristic Procedures (3) (cont'd)

- P-P plots amplify differences between the middle of the two distributions







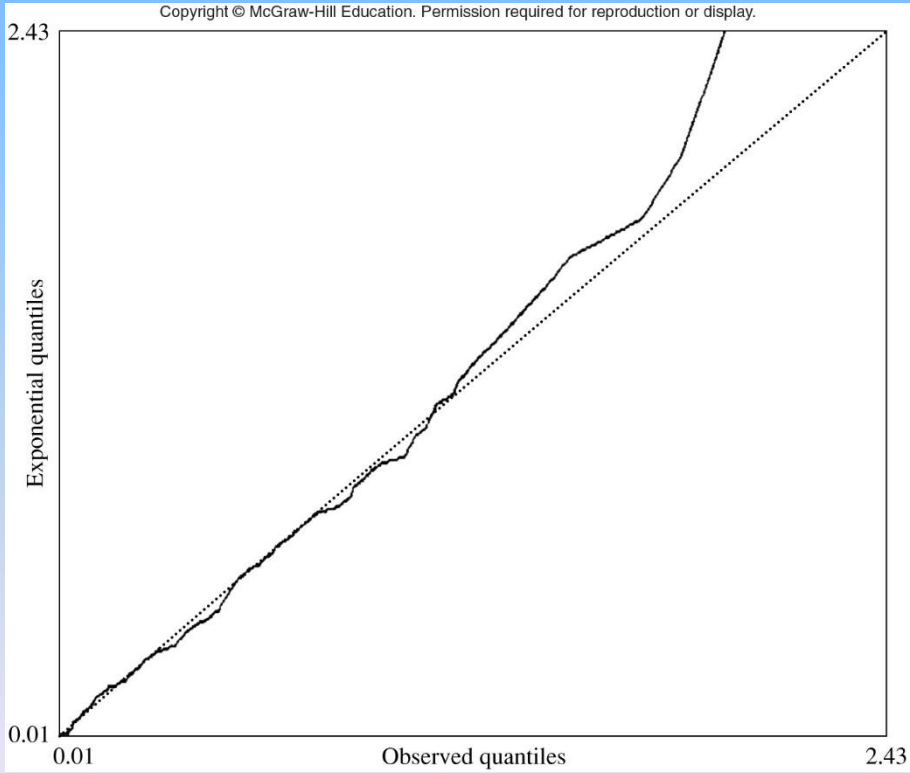
Heuristic Procedures (3) (cont'd)

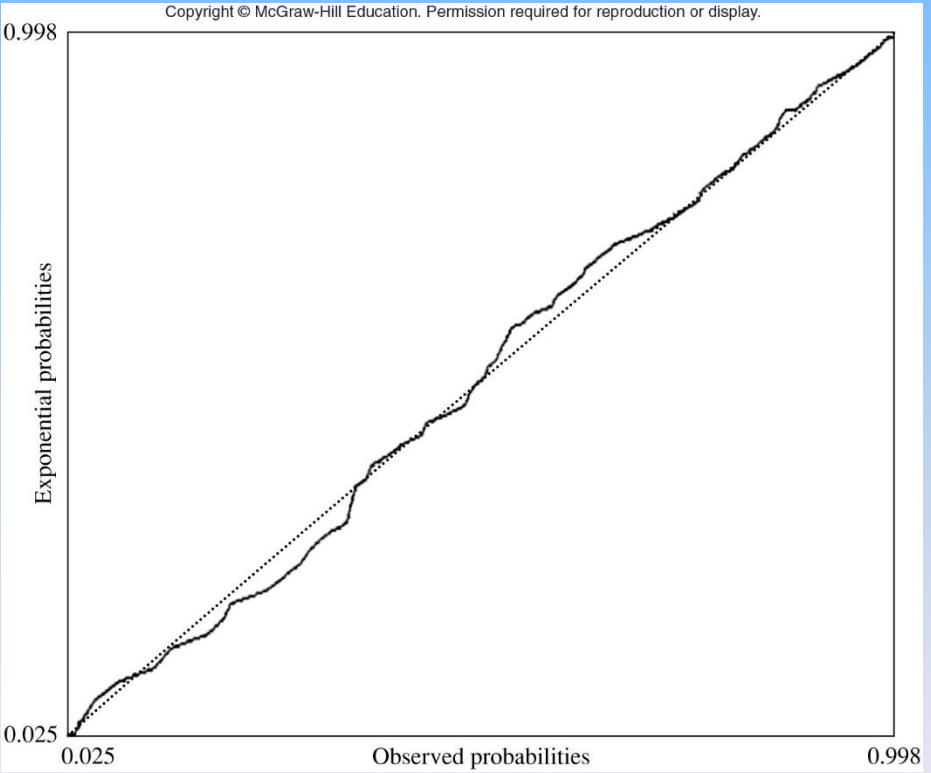
- Ties in sample values
- Let $Y_1, Y_2, \ldots, Y_l$ be the distinct values in $X_1, X_2, \ldots, X_n$

$$\tilde{F}_n(Y_i) = q_i = \left(\text{proportion of } X_j's \leq Y_i\right) - \frac{0.5}{n}$$

Q-Q Plot and P-P Plot Example

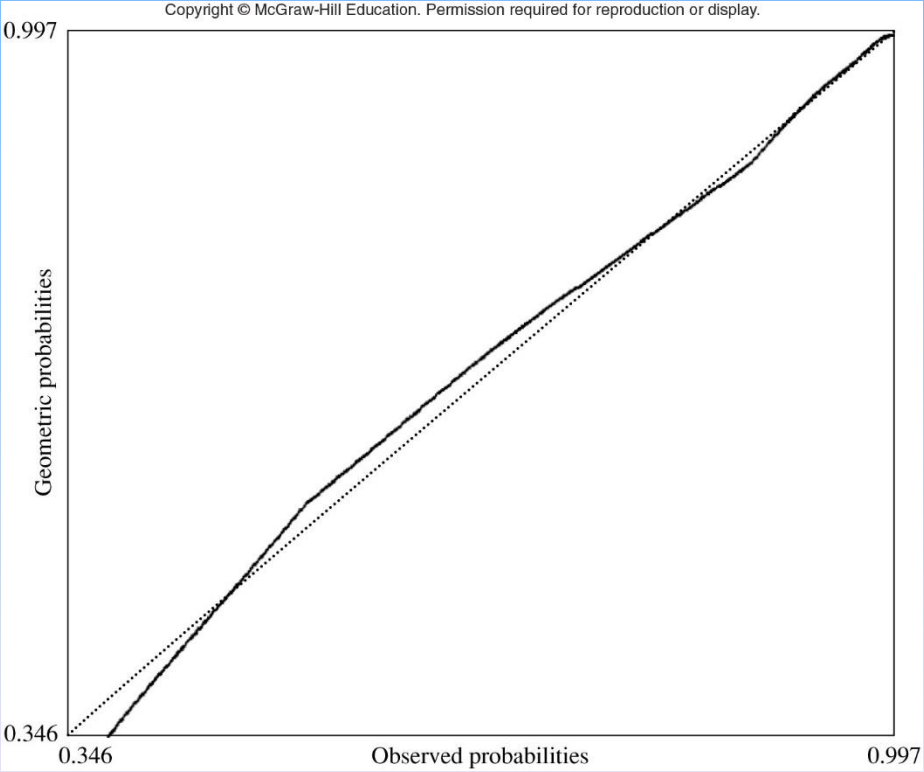
- Drive-up Banking Example





P-P Plot Example

- Inventory demand example



Activity III: Determining How Representative the Fitted Distributions Are

- Goodness-of-fit tests
  - Statistical hypothesis test
  - Used to assess whether the observations are an independent sample from a particular distribution with distribution function $\hat{F}$
  - Test null hypothesis H_0: The Xi's are IID random variables with distribution function $\hat{F}$
    - For small to moderate sample sizes
      - Not sensitive to subtle disagreements, rather for detecting gross differences
    - For large sample sizes
      - H_0 virtually never exactly true
      - We just need a "good enough" distribution

Examples of Goodness-of-Fit Tests

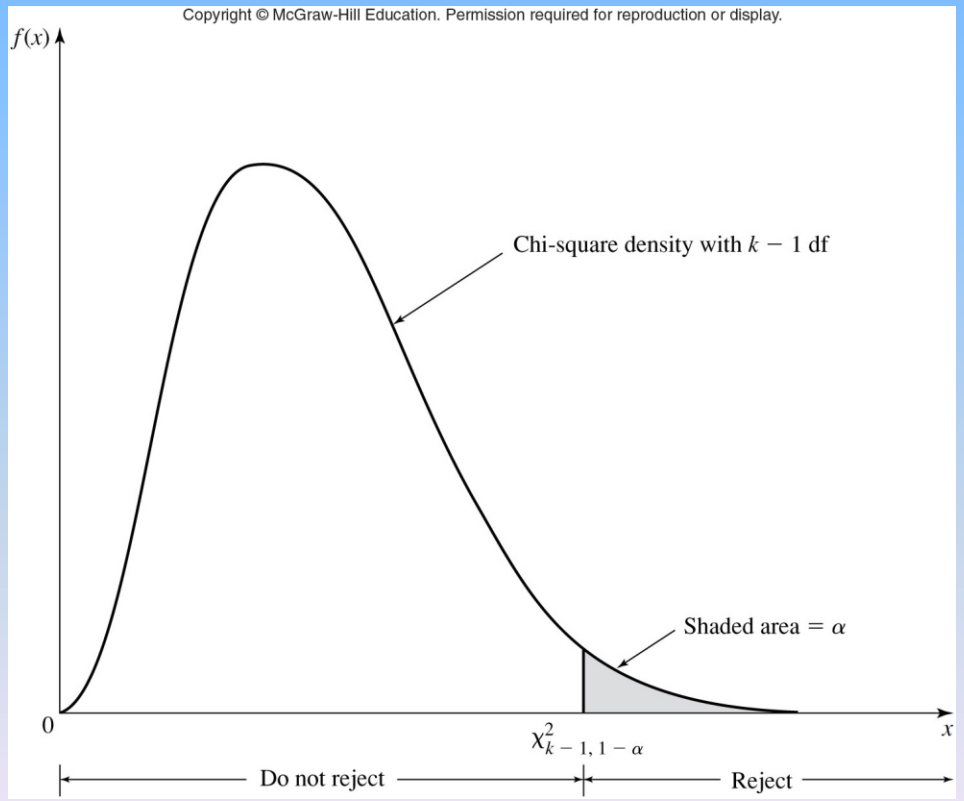
- Chi-square tests
- Kolmogorov-Smirnov tests
- Anderson-Darling tests
- Poisson-Process tests

Chi-Square Tests (1)

- A more formal comparison between data histogram and the fitted distribution
- Divide the entire range of the fitted distribution into k adjacent intervals $[a_0,a_1)$, $[a_1,a_2)$, … , $[a_{k-1},a_k)$ (a_k/a_0 may be +/- ∞)
  - N_j = number of X_i's in $[a_{j-1},a_j]$ (j = 1,2,…,k)
  - Continuous case: $p_j = \int_{a_{j-1}}^{a_j} \hat{f}(x)dx$
  - Discrete case: $p_j = \sum_{a_{j-1} \le x_i \le a_j} \hat{p}(x_i)$
  - Test statistics: $\chi^2 = \sum_{j=1}^{k} \frac{(N_j - np_j)^2}{np_j}$
    - np_j: the expected number of samples in $[a_{j-1},a_j]$
    - Reject H_0 if χ^2 is too large.

Chi-Square Tests (2)

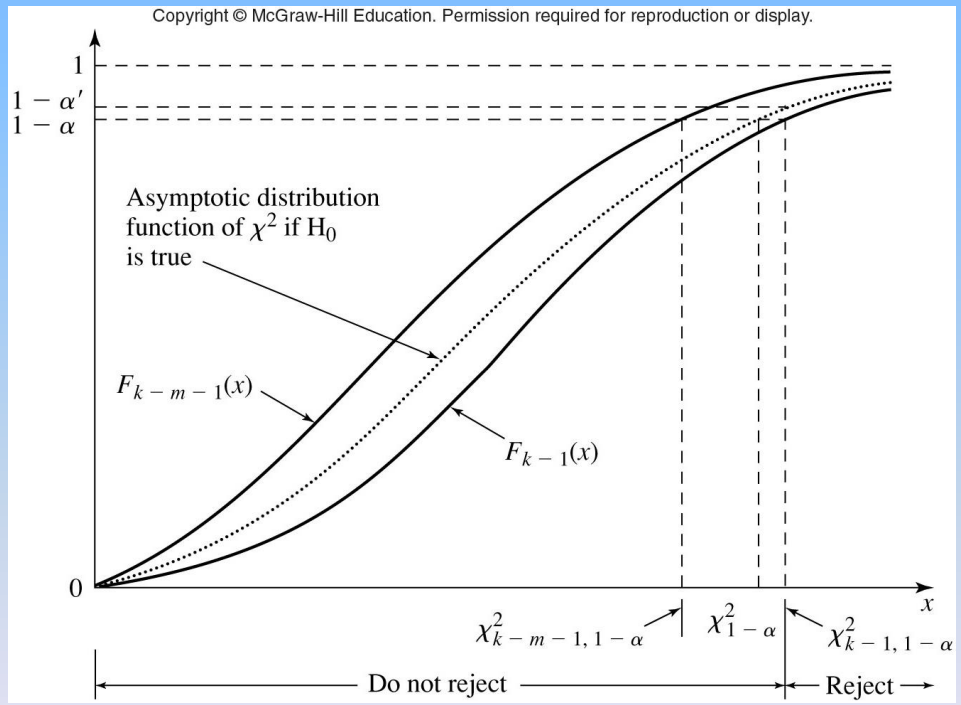
- Case 1: All parameters of the fitted distribution are known (e.g. in empirical testing of random-number generations)
- If H_0 is true, χ^2 converges (as n $\rightarrow \infty$) to a chi-square distribution with k-1 degree of freedom (df), the same as gamma[(k-1)/2,2]
- A test with approx. level α rejects H_0 if $\chi^2 > \chi^2_{k-1,1-\alpha}$ (look up Table T.2)
  - Valid (i.e. of level α) asymptotically as n $\rightarrow \infty$



Chi-Square Tests (3)

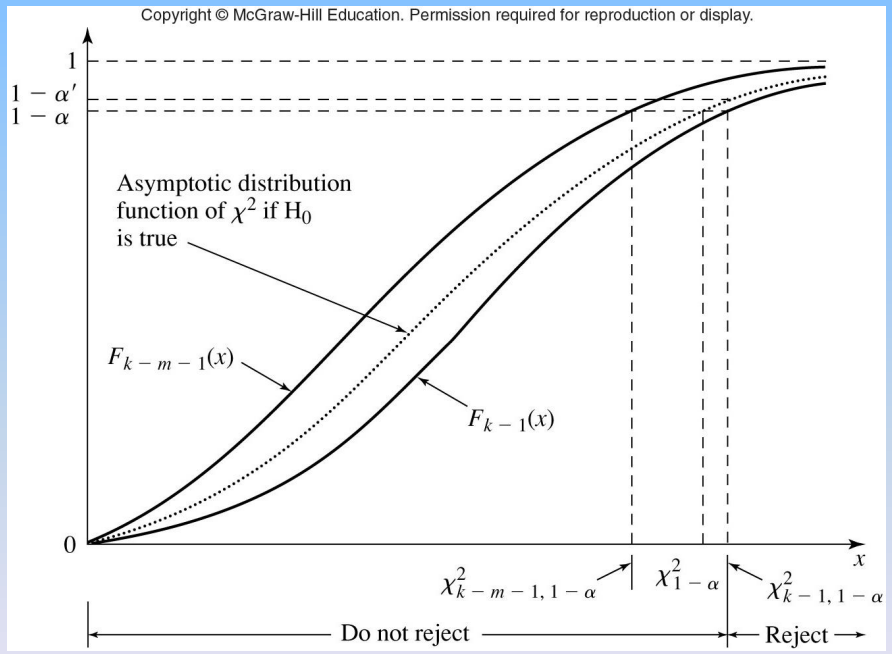
- Case 2: Estimate m (>= 1) parameters of the fitted distribution by MLE
- If H_0 is true, χ^2 converges (as $n \to \infty$) to a distribution lying between chi-square distribution with k – 1 and k – m – 1df
- Critical point $\chi^2_{1-\alpha}$
- Reject H_0 if $\chi^2 > \chi^2_{k-1,1-\alpha}$
- Do not reject H_0 if

$$\chi^2 < \chi^2_{k-m-1,1-\alpha}$$



Chi-Square Tests (4)

- What if $\chi^2_{k-m-1,1-\alpha} \leq \chi^2 \leq \chi^2_{k-1,1-\alpha}$?
- Reject H_0 only if $\chi^2 > \chi^2_{k-1,1-\alpha}$
  - Type I error (rejecting a true H_0) probability is no larger than α
  - At the cost of loss of power (probability of rejecting a false H_0)
  - Usually m <= 2, k large, difference between $\chi^2_{k-m-1,1-\alpha}$ and $\chi^2_{k-1,1-\alpha}$ won't be too large



Chi-Square Tests (5)

- Choosing the number and size of the intervals
  - No definitive guideline
  - Recommendation: guarantee a valid and unbiases test
    - k >= 3
    - Equiprobable approach: $p_1=p_2=\ldots=p_k$ (approx. for discrete data)
    - $np_j >= 5$ (j = 1,2,…k)
  - For the same data set, different ways of having intervals may lead to different conclusions

Chi-Square Test Example 1

- Drive-up Banking Example
- n = 219
- $\hat{F}(x) = 1 - e^{-\frac{x}{0.399}}$ for x >= 0
- k = 20 intervals, p_j = 1/k = 0.05, np_j = 219×0.05=10.950
  - Satisfies the guidelines

j	Interval	N_j	np_j	$\frac{(N_j - np_j)^2}{np_j}$
1	[0, 0.020)	8	10.950	0.795
2	[0.020, 0.042)	11	10.950	0.000
3	[0.042, 0.065)	14	10.950	0.850
4	[0.065, 0.089)	14	10.950	0.850
5	[0.089, 0.115)	16	10.950	2.329
6	[0.115, 0.142)	10	10.950	0.082
7	[0.142, 0.172)	7	10.950	1.425
8	[0.172, 0.204)	5	10.950	3.233
9	[0.204, 0.239)	13	10.950	0.384
10	[0.239, 0.277)	12	10.950	0.101
11	[0.277, 0.319)	7	10.950	1.425
12	[0.319, 0.366)	7	10.950	1.425
13	[0.366, 0.419)	12	10.950	0.101
14	[0.419, 0.480)	10	10.950	0.082
15	[0.480, 0.553)	20	10.950	7.480
16	[0.553, 0.642)	9	10.950	0.347
17	[0.642, 0.757)	11	10.950	0.000
18	[0.757, 0.919)	9	10.950	0.347
19	[0.919, 1.195)	14	10.950	0.850
20	[1.195, ∞)	10	10.950	0.082
				$\chi^2 = 22.188$

- $\hat{F}(a_j) = \frac{j}{20} \Rightarrow a_j = -0.399 \ln\left(1 - \frac{j}{20}\right), a_0 = 0, a_{20} = \infty$
  - For other continuous distributions, F^{-1} can be evaluated by numerical methods
- χ^2=22.188
- $\chi^2_{19,0.90} =? \chi^2_{19,0.75} =?$

Chi-Square Test Example 2

- Inventory demand example
- Can only make the p_j's roughly equal
- Mode = 0, $\hat{p}(0) = 0.346$ is the highest value of the mass function
  - Choice of intervals are limited

j	Interval	N_j	np_j	$\frac{(N_j - np_j)^2}{np_j}$
1	{0}	59	53.960	0.471
2	{1, 2}	50	58.382	1.203
3	{3, 4, . . .}	47	43.658	0.256
				$\chi^2 = 1.930$

- $\chi^2_{2,0.90} = ?$

Chi-Square Test Example 3

- 856 ship-loading times
- Fitted distribution: log-logistic
- Test at level $\alpha = 0.1$

k	Statistic	Critical value	Result of test
10	11.383	14.684	Do not reject
20	27.645	27.204	Reject
40	50.542	50.660	Do not reject

Kolmogorov-Smirnov Tests (1)

- Chi-square tests compare a histogram of the data with the fitted distribution
  - Difficult to specify the intervals
  - Valid only in an asymptotic sense
- K-S tests compare an empirical distribution function with the hypothesized one
  - No need to group data
  - Valid for any sample size
  - Tend to be more powerful against many alternative distributions

Kolmogorov-Smirnov Tests (2)

- For discrete data, required critical values must be computed using complicated formulas
- The original form is valid only if all the parameters of the hypothesized distribution are known and the distribution is continuous
  - Has been extended to allow for estimation of the parameters in normal, lognormal, exponential, Weibull and log-logistic
- The original form is often applied for continuous distributions with estimated parameters and discrete distributions
  - Type I error smaller than specified -> loss of power

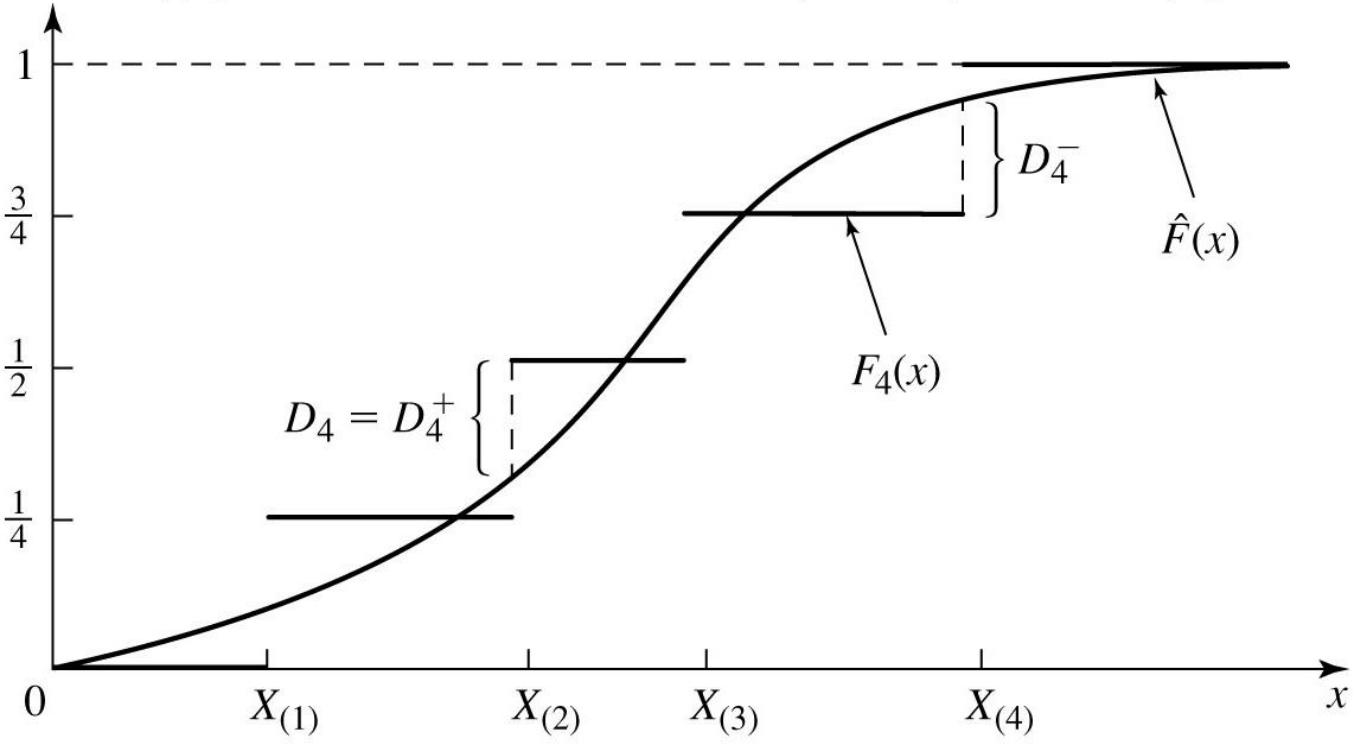
Kolmogorov-Smirnov Tests (3)

- Fitted distribution function $\hat{F}(x)$
- Empirical distribution function

$$F_n(X_{(i)}) = \frac{i}{n} \text{ for i = 1,2,…,n}$$

- Statistic $D_n = \sup_x \{|F_n(x) - \hat{F}(x)|\}$
  - $\sup_x\{A\}$: the smallest value >= all members of A
  - Computation:
    - $D_n^+ = \max_{1 \le i \le n}\{\frac{i}{n} - \hat{F}(X_{(i)})\}, D_n^- = \max_{1 \le i \le n}\{\hat{F}(X_{(i)}) - \frac{i-1}{n}\}$
    - $D_n = \max\{D_n^+, D_n^-\}$
  - Reject H_0 if D_n exceeds $d_{n,1-\alpha}$
  - Critical point $d_{n,1-\alpha}$ depends on $\hat{F}(x)$

Example of K-S Test Statistic Computation



Geometric meaning of the K-S test statistic D_n for n = 4

K-S Tests: Case 1

- Original form
  - All parameters known, continuous data
  - D_n does not depend on the fitted distribution function
  - Adjusted test statistic

$$\left(\sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}}\right) D_n, \qquad \text{reject } H_o \text{ if } > c_{1-\alpha}$$

		$1-\alpha$				
Case	**Adjusted test statistic**	**0.850**	**0.900**	**0.950**	**0.975**	**0.990**
All parameters known	$\left(\sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}}\right) D_n$	1.138	1.224	1.358	1.480	1.628
$N(\bar{X}(n), S^2(n))$	$\left(\sqrt{n} - 0.01 + \frac{0.85}{\sqrt{n}}\right) D_n$	0.775	0.819	0.895	0.955	1.035
$expo(\bar{X}(n))$	$\left(D_n - \frac{0.2}{n}\right)\left(\sqrt{n} + 0.26 + \frac{0.5}{\sqrt{n}}\right)$	0.926	0.990	1.094	1.190	1.308

K-S Tests: Case 2

- Hypothesized distribution is $N(\mu, \sigma^2)$
  - μ, σ² unknown
- $\hat{F}(x) = \Phi\{\frac{[x-\bar{X}(n)]}{\sqrt{S^2(n)}}\}$
  - Adjusted test statistic

$$\left(\sqrt{n} - 0.01 + \frac{0.85}{\sqrt{n}}\right) D_n, \qquad \text{reject } H_o \text{ if } > c'_{1-\alpha}$$

Case	Adjusted test statistic	$1-\alpha$ 0.850	0.900	0.950	0.975	0.990
All parameters known	$\left(\sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}}\right) D_n$	1.138	1.224	1.358	1.480	1.628
$N(\bar{X}(n), S^2(n))$	$\left(\sqrt{n} - 0.01 + \frac{0.85}{\sqrt{n}}\right) D_n$	0.775	0.819	0.895	0.955	1.035
$expo(\bar{X}(n))$	$\left(D_n - \frac{0.2}{n}\right)\left(\sqrt{n} + 0.26 + \frac{0.5}{\sqrt{n}}\right)$	0.926	0.990	1.094	1.190	1.308

K-S Tests: Case 3

- Hypothesized distribution is expo(β)
  - β unknown, MLE =$\bar{X}(n)$
- $\hat{F}(x) = 1 - e^{-\frac{x}{\bar{X}(n)}}$ for x >= 0
  - Adjusted test statistic

$$\left(\sqrt{n} + 0.26 + \frac{0.5}{\sqrt{n}}\right)\left(D_n - \frac{0.2}{n}\right), \qquad \text{reject } H_o \text{ if } > c''_{1-\alpha}$$

Case	Adjusted test statistic	$1-\alpha$ 0.850	0.900	0.950	0.975	0.990
All parameters known	$\left(\sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}}\right)D_n$	1.138	1.224	1.358	1.480	1.628
$N(\bar{X}(n), S^2(n))$	$\left(\sqrt{n} - 0.01 + \frac{0.85}{\sqrt{n}}\right)D_n$	0.775	0.819	0.895	0.955	1.035
$expo(\bar{X}(n))$	$\left(D_n - \frac{0.2}{n}\right)\left(\sqrt{n} + 0.26 + \frac{0.5}{\sqrt{n}}\right)$	0.926	0.990	1.094	1.190	1.308

K-S Tests: Case 4

- Hypothesized distribution is Weibull with α, β unknown
- $\hat{F}(x) = 1 - e^{-\left(\frac{x}{\hat{\beta}}\right)^{\hat{\alpha}}}$ for x >= 0
  - Adjusted test statistic

$$\sqrt{n}D_n \qquad \text{reject } H_o \text{ if } > c^*_{1-\alpha}$$

	$1-\alpha$			
n	**0.900**	**0.950**	**0.975**	**0.990**
10	0.760	0.819	0.880	0.944
20	0.779	0.843	0.907	0.973
50	0.790	0.856	0.922	0.988
∞	0.803	0.874	0.939	1.007

K-S Tests: Case 5

- Hypothesized distribution is log-logistic with α, β unknown
- X_i's are the logarithms of the basic data points
- $\hat{F}(x) = \left(1 + e^{[-(x - \ln \hat{\beta})]\hat{\alpha}}\right)^{-1}$ for $-\infty < x < \infty$
  - Adjusted test statistic

$$\sqrt{n}D_n \qquad \text{reject } H_o \text{ if } > c^{+}_{1-\alpha}$$

	$1-\alpha$			
n	**0.900**	**0.950**	**0.975**	**0.990**
10	0.679	0.730	0.774	0.823
20	0.698	0.755	0.800	0.854
50	0.708	0.770	0.817	0.873
∞	0.715	0.780	0.827	0.886

K-S Test Example

- Drive-up Banking
- $\hat{F}(x) = 1 - e^{-\frac{x}{0.399}}$ for x >= 0
- $D_{219} = 0.047$
- Adjusted test statistic

$$\left(D_{219} - \frac{0.2}{219}\right)\left(\sqrt{219} + 0.26 + \frac{0.5}{\sqrt{219}}\right) = 0.696$$

- Reject H_0 or not?

6.8 Shifted and Truncated Distributions

- Modifying a distribution may provide a better fit in some cases
  - Example: bank teller cannot serve a customer in less than 30 seconds
    - Shift distribution to the right to disallow values less than 30 seconds
  - Shift a distribution to the right by γ units: replace x by x - γ in PDF
  - Exponential distribution original

$$f(x) = \frac{1}{\beta} e^{-\frac{x}{\beta}} \qquad x \geq 0$$

  - Exponential distribution shifted, now has a location param. γ

$$f(x) = \frac{1}{\beta} e^{-\frac{x-\gamma}{\beta}} \qquad x \geq \gamma$$

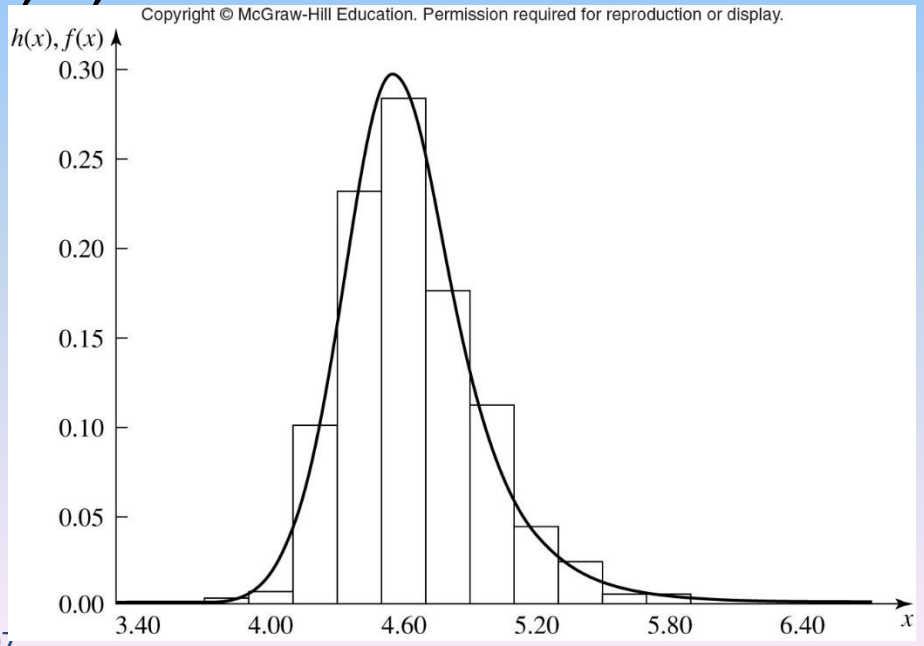
Param. Estimation with the Added γ

- Finding the MLE for γ in addition to the MLEs for the original parameters may not work
  - MLEs for some distribution with γ are not well defined.
- First estimate γ: $\tilde{\gamma} = \frac{X_{(1)}X_{(n)} - X_{(k)}^2}{X_{(1)} + X_{(n)} - 2X_{(k)}}$
  - k is the smallest integer in {2, 3, …, n – 1} such that $X_{(k)} > X_{(1)}$
- Define $X_i' = X_i - \tilde{\gamma} \geq 0$ for i = 1,2,…,n
- Find MLEs of the other parameters using X_i's.

Example

- Unload time of 808 coal trains
- X(1)=3.37, X(2) = 3.68, X(808)=6.32
- $\hat{\gamma} = \frac{X_{(1)}X_{(808)} - X_{(2)}^2}{X_{(1)} + X_{(808)} - 2X_{(2)}} = 3.329$
- $X_i' = \mathrm{X_i} - 3.329$ for i = 1,2,…,808
- MLEs for log-logistic distribution

  $\hat{\alpha} = 7.451, \hat{\beta} = 1.271$



Truncate Distributions

- No random values can be larger than b > 0
- If the range of a PDF f is [0, ∞), define truncated PDF as

$$f^*(x) = f(x)/F(b) \text{ for } 0 \le x \le b$$

$$F(b) = \int_0^b f(x)dx$$

- Example: truncated exponential dist. for [0, 90]

$$F(b) = \int_0^{90} \frac{1}{\beta} e^{-\frac{x}{\beta}} dx = 1 - e^{-\frac{90}{\beta}}$$

$$f^*(x) = \frac{\frac{1}{\beta} e^{-\frac{x}{\beta}}}{1 - e^{-\frac{90}{\beta}}}$$

6.10 Specifying Multivariate Distributions, Correlations, and Stochastic Processes

- Random input variables may be statistically related to each other
  - May form a random vector to be specified by the modeler
  - Could be a correlation between different random input variables
    - In a random vector or stochastic process
    - With their own individual, or marginal, distributions

Example Situations

- A maintenance shop with two service stations
  - First station inspects, second repairs
  - Longer inspection time probably leads to longer reparation time -> positive correlation
- A communication system
  - Large (small) messages tend to come in groups -> positive correlation through several lags
- An inventory system
  - Large orders tends to be followed by small orders

Specifying Multivariate Distributions

- Input random vector $\mathbf{X} = (X_1, X_2, \ldots, X_d)^T$
- R.v.'s within a $\mathbf{X}_k$ are correlated
- R.v.'s across $\mathbf{X}_k$'s are independent
- Maintenance shop example

$$\begin{pmatrix} X_{11} \\ X_{21} \end{pmatrix}, \begin{pmatrix} X_{12} \\ X_{22} \end{pmatrix}, \ldots, \begin{pmatrix} X_{1n} \\ X_{2n} \end{pmatrix}$$

- Multivariate (joint) distribution function

$$F(\boldsymbol{x}) = P(\boldsymbol{X} \leq \boldsymbol{x}) = P(X_1 \leq x_1, \ldots, X_d \leq x_d)$$

  - Implies marginal distribution
  - Embodies relationships between the r.v.'s
- Difficult to estimate the entire multivariate distribution
  - Will look at certain useful cases

Multivariate Normal Distribution

- A springboard to other more useful distributions

$$f(\boldsymbol{x}) = (2\pi)^{-\frac{n}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left[-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^T\sum^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2}\right]$$

Σ: the covariance entry with (i,j)th entry $\sigma_{ij} = \sigma_{ji} = Cov(X_i, X_j)$

$|\Sigma|$: determinant of Σ

- Fit to d-dimensional data $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}} = (\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_d)^T$$

$$\hat{\sigma}_{ij} = \frac{\sum_{k=1}^{n}(X_{ik} - \bar{X}_i)(X_{jk} - \bar{X}_j)}{n}$$

Multivariate Lognormal Distribution

- $\mathbf{X} = (X_1, X_2, \ldots, X_d)^T$ has a multivariate lognormal distribution iff $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_d)^T = (\ln X_1, \ln X_2, \ldots, \ln X_d)^T$ has a multivariate normal distribution

$$X = (e^{Y_1}, e^{Y_2}, \ldots, e^{Y_d})^T$$

$$E(X_i) = e^{\mu_i + \frac{\sigma_{ii}}{2}}$$

$$Var(X_i) = e^{2\mu_i + \sigma_{ii}}(e^{\sigma_{ii}} - 1)$$

$$Cov(X_i, X_j) = \frac{e^{\sigma_{ij}} - 1}{\sqrt{(e^{\sigma_{ii}} - 1)(e^{\sigma_{jj}} - 1)}}$$

- Fit to data
  - Take the natural logarithms of data
  - Estimate $\boldsymbol{\mu}$ and Σ for multivariate normal dist.

Specifying Arbitrary Marginal Distributions and Correlations

- Allow for correlation between various pairs of input random variables, while not imposing an overall multivariate distribution
  - Fit distributions to each of the univariate random variables involved
  - Estimate correlations

$$\hat{\sigma}_{ij} = \frac{\sum_{k=1}^{n}(X_{ik} - \bar{X}_i)(X_{jk} - \bar{X}_j)}{n}$$

$$\hat{\rho}_{ij} = \frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}}$$

  - Generate such random vectors: Sec 8.5.5

Specifying Stochastic Processes (1)

- A sequence of input random variables are modeled as being draws from the same distribution with autocorrelation between themselves
- Autoregressive (AR) processes
  - $X_i = \mu + \phi_1(X_{i-1} - \mu) + \phi_i(X_{i-2} - \mu) + \cdots + \phi_p\left(X_{i-p} - \mu\right) + \varepsilon_i$
  - ϕ_i: constants for X_i's to have a stationary marginal distribution
  - ε_i: IID r.v.'s with mean 0 and particular variance to control X_i
- Autoregressive moving-average (ARMA) processes
  - X_i similar to that of AR, with weighted ε_i
- Use linear regression to estimate the unknown parameters
- X_i's generally restricted to have normal distribution
  - Use AR processes as "base" for ARTA models

Specifying Stochastic Processes (2)

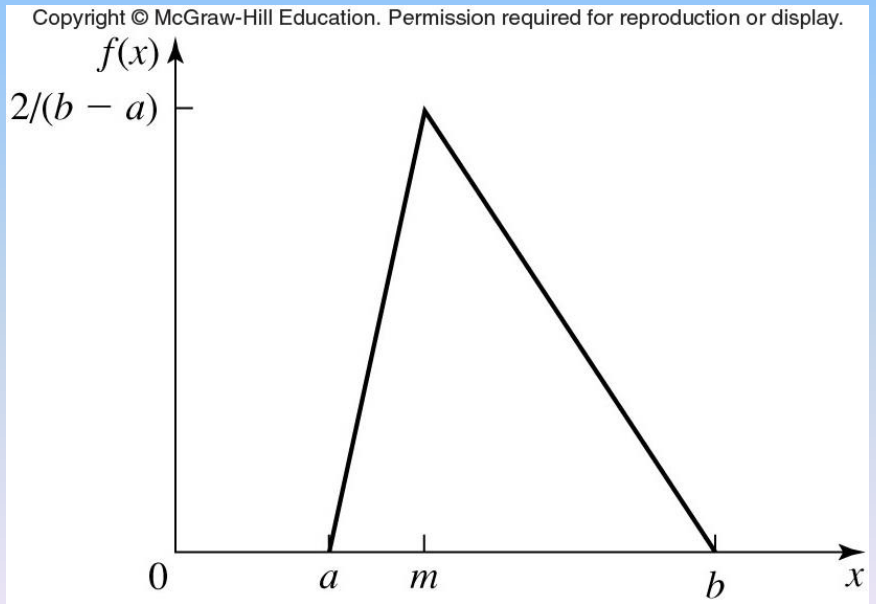
- Autoregressive-to-anything (ARTA) processes
  - Can exactly match the desired autocorrelation structures out to a specified lag p and the desired stationary marginal distribution
  - Specify a AR process $\{Z_i\}$ with N(0,1) marginal distribution
  - $X_i = F^{-1}[\Phi(Z_i)]$
    - $\Phi(Z_i)$ has a U(0,1) distribution
    - F^{-1} is the inverse of the desired stationary marginal distribution F

6.11 Selecting a Distribution in the Absence of Data

- Some simulation studies
  - Not possible to gather data on random variables of interest
  - Example: if system does not exist in some form
- Approaches
  - Triangular-distribution approach
  - Beta-distribution approach
  - Lognormal-distribution approach
  - Weibull-distribution approach

Triangular-Distribution Approach

- Identify an interval [a,b] such that $P(a \leq X \leq b) \approx 1$
  - E.g. time to replace a tire
- Decide the mode: the most likely value for X
- Cons
  - [a,b] is subjective
  - No long right tail

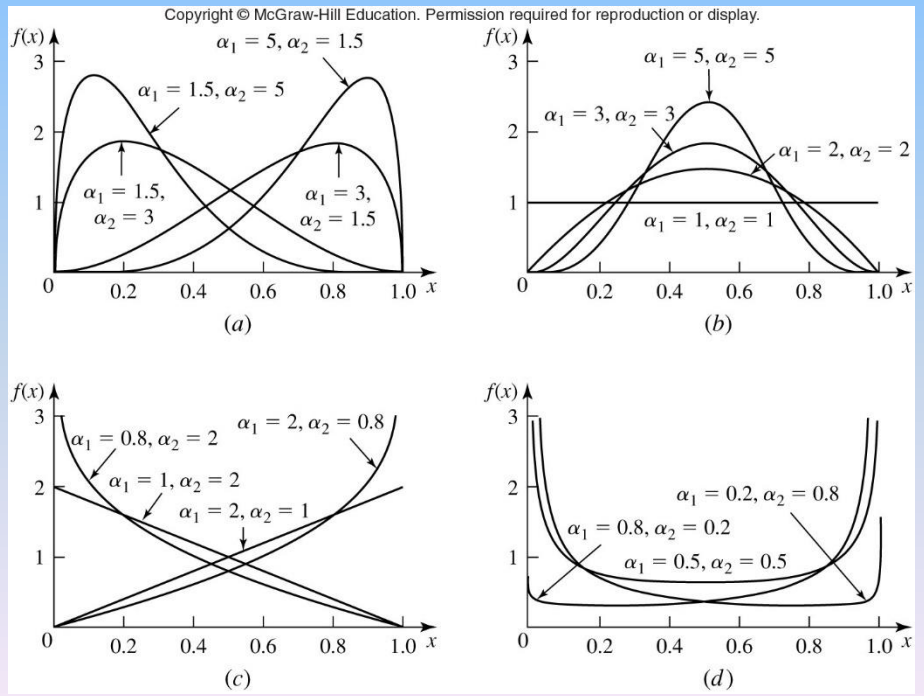


Beta-Distribution Approach (1)

- Identify an interval [a,b] such that $P(a \leq X \leq b) \approx 1$
- Assume X has a beta distribution on [a, b] with shape parameters α_1 and α_2

- $$f(x) = \frac{x^{\alpha_1 - 1}(1-x)^{\alpha_2 - 1}}{B(\alpha_1 - \alpha_2)}$$

  $0 < x < 1$



Beta-Distribution Approach (2)

- If little is known about X, or X is equally likely to take any value, choose $\alpha_1 = \alpha_2 = 1$
- If X models a task time, assume the distribution is skewed to the right $\alpha_2 > \alpha_1 > 1$
- $\mu = a + \frac{\alpha_1(b-a)}{\alpha_1+\alpha_2}$ and $m = a + \frac{(\alpha_1-1)(b-a)}{\alpha_1+\alpha_2-2}$
  - Note $\mu > m$
- $\tilde{\alpha}_1 = \frac{(\mu-a)(2m-a-b)}{(m-\mu)(b-a)}$ $\tilde{\alpha}_2 = \frac{(b-\mu)\tilde{\alpha}_1}{\mu-a}$
- Both approaches may result in large errors

Lognormal-Distribution Approach (1)

- If $\Upsilon \sim N(\mu, \sigma^2), V = e^{\Upsilon} \sim LN(\mu, \sigma^2)$

$$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp \frac{-(\ln x - \mu)^2}{2\sigma^2} \qquad x > 0$$

- $V = e^{\Upsilon} + \gamma \sim LN(\gamma, \mu, \sigma^2)$

$$f(x') = \frac{1}{(x' - \gamma)\sqrt{2\pi\sigma^2}} \exp \frac{-\left(\ln(x' - \gamma) - \mu\right)^2}{2\sigma^2} \qquad x > 0$$

  - Parameters
    - Location: γ
    - Scale: e^{μ}
    - Shape: σ

Lognormal-Distribution Approach (2)

- m: mode, x_q: q-quantile (100qth percentile)
- Assume $0 \leq \gamma < m < x_q < \infty$
- $\tilde{\gamma}$: Lowest value of X
- $\tilde{\sigma} = \frac{-z_q + \sqrt{z_q^2 - 4c}}{2}$
  - z_q: q-quantile of a N(0,1) random variable
  - $c = \ln[\frac{m-\tilde{\gamma}}{x_q-\tilde{\gamma}}]$
- $\tilde{\mu} = \ln(m - \tilde{\gamma}) + (\tilde{\sigma})^2$

Weibull-Distribution Approach (1)

- $Y \sim \text{Weibull}(\alpha, \beta)$

$$f(x) = \alpha\beta^{-\alpha}x^{\alpha-1}e^{-\left(\frac{x}{\beta}\right)^{\alpha}}$$

- $X = Y + \gamma \sim \text{Weibull}(\gamma, \alpha, \beta)$

$$f(x') = \alpha\beta^{-\alpha}(x' - \gamma)^{\alpha-1}e^{-\left(\frac{x'-\gamma}{\beta}\right)^{\alpha}}$$

Weibull-Distribution Approach (2)

- m: mode, x_q: q-quantile (100qth percentile)
- Assume $0 \leq \gamma < m < x_q < \infty$
- Estimate α by solving

$$\frac{m-\gamma}{x_q-\gamma} = \left\{ \frac{\alpha-1}{\alpha \ln\left[\frac{1}{1-q}\right]} \right\}^{\frac{1}{\alpha}}$$

  - Use Newton's method

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

- Estimate β

$$\tilde{\beta} = \frac{m-\gamma}{\left(1-\frac{1}{\tilde{\alpha}}\right)^{\frac{1}{\tilde{\alpha}}}}$$

6.12 Models of Arrival Processes

- Poisson process where the inter-arrival times are IID
- Nonstationary Poisson process where the arrival rate varies with time
- Batch arrivals

Poisson Process (1)

- Event times: $0 = t_0 \leq t_1 \leq t_2 \leq \cdots$
- $N(t) = \max\{i: t_i \leq t\}$: number of events to occur at or before time t
- Poisson process
  - Most commonly used model for arrival process of customers to a queuing system
  - Requirements
    - Customers arrive one at a time
    - Number of arrivals does not depend on the number of arrivals in the earlier time interval
      - $N(t+s) - N(t)$ is independent of $\{N(u), 0 \leq u \leq t)$
    - Number of arrivals does not depend on time of day
      - $N(t+s) - N(t)$ is independent of t for all t,s ≥ 0

Poisson Process (2)

- $N(t+s) - N(t)$ is a Poisson random variable

$$P[N(t+s) - N(t) = k] = P[N(s) = k] = \frac{e^{-\lambda s}(\lambda s)^k}{k!}$$

k = 0,1,2, … and t,s ≥ 0

$$E[N(s)] = \lambda s, E[N(1)] = \lambda \text{ (arrival rate)}$$

- Inter-arrival times are IID exponential random variables with mean $\frac{1}{\lambda}$
  - Vice versa

Nonstationary Poisson Process

- Requirements
  - Customers arrive one at a time
  - Number of arrivals does not depend on the number of arrivals in the earlier time interval
    - $N(t+s)-N(t)$ is independent of $\{N(u), 0 \le u \le t)$
- Arrival rate is allowed to be a function of time: $\lambda(t)$
  - Expectation function: $\Lambda(t) = E[N(t)]$
  - Rate function: $\lambda(t) = \frac{d}{dt}\Lambda(t)$
- $N(t+s)-N(t)$ is a Poisson random variable

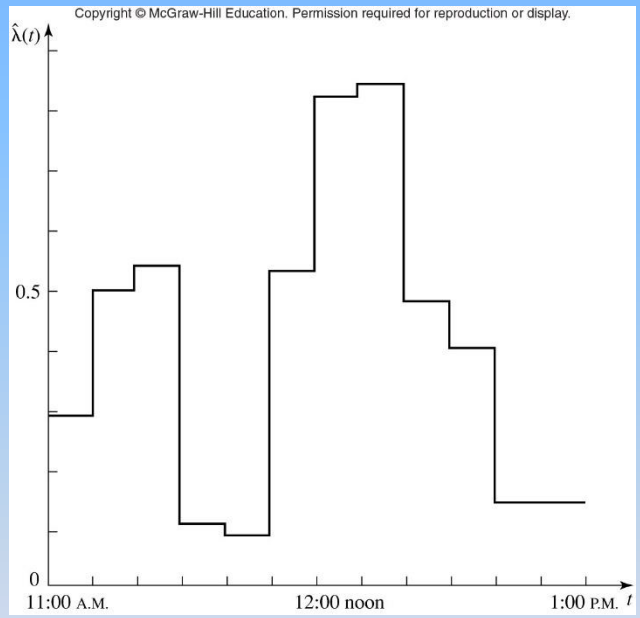
$$P[N(t+s)-N(t)=k] = \frac{e^{-b(t,s)}(b(t,s))^k}{k!}$$

k = 0,1,2, … and t,s ≥ 0

$$b(t,s) = \Lambda(t+s) - \Lambda(t) = \int_t^{t+s} \lambda(y)dy$$

Example of Estimating $\lambda(t)$

- Customer arrival times are collected for a xerographic copy shop between 11AM and 1PM for 8 days
- Divide the 2 hours into 12 10-minute intervals
  - May need to try other widths to have not too ragged and not too smooth plots
- Calculate the average number of arrivals in the intervals over the 8 days
- Divide the average number of arrivals by 10 minutes to obtain the estimate of the arrival rate for a particular interval



Batch Arrivals

- Appropriate when customers arrive in groups
- Compound Poisson process
- $N(t)$ is the number of batches instead of the number of customers
- $X(t)$ is the total number of individual customers to arrive by time t
  - $X(t) = \sum_1^{N(t)} B_i$ for $t \geq 0$
  - B_i's: IID random variables, numbers of customers in the i-th batch,

6.13 Assessing the Homogeneity of Different Data Sets

- Sometimes analysts collect different data sets and would like to know if they are homogeneous
  - If homogeneous, data can be merged
  - Otherwise different distribution is needed for each set
  - E.g. bank service time of different days, message sizes received by different computer
- Kruskal-Wallis hypothesis test for homogeneity
  - Compute the *K-W* test statistic
  - No assumptions made about the distributions
  - No two data values are the same

Kruskal-Wallis Test

- Sample sets sized $n_1, n_2, \ldots, n_k$ $(\sum_{i=1}^{k} n_i = n)$
  - X_{ij}: j-th sample in i-th set
  - $R(X_{ij})$: the rank assigned to among X_{ij} all samples
  - $R_i = \sum_{j=1}^{n_i} R(X_{ij})$
- Null hypothesis
  - H_0: All the population distribution functions are identical
- Alternative hypothesis
  - H_1: At least one of the populations tends to yield larger observations than at least one of the other populations
- K-W test statistic $T = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$
- Reject H_0 at level α if $T > \chi^2_{k-1,1-\alpha}$
  - $\chi^2_{k-1,1-\alpha}$: upper $1-\alpha$ critical value for a chi-square distribution with $k-1$ d.f.