MET3220C & MET6480 Computational Statistics

Lecture 8
Parametric Probability Distributions

Continuous Distributions

Key Point: ALWAYS LOOK AT THE DATA!!!!
DOES THE DATA REALY FIT THE DISTRIBUTION?

Continuous Distributions

- Continuous distributions have probabilities for any value(s) within a parameter space.
  - For example, a univariate distribution has probabilities for upper and lower bounds, as well as all values between these bounds.
  - This limits could be $\pm\infty$.
- The probability distribution function $f(x)$ is such that $\int_x f(x)dx = 1$.
  - Probability distribution (or density) function is abbreviated as PDF.
- Note that the probability of an event occurring is the area under the PDF, bounded by the limiting conditions on the event.
- These last two points should make it clear that $f(x) = \partial \Pr\{x\}/\partial x$.
  - This equation is easily written in terms of cumulative probability CDF, $C\{X \le x\}$, because $\partial \Pr\{x\}/\partial x = \partial C\{X \le x\}/\partial x$
  - If we can calculate a a CDF, then we can easily randomly generate a distribution that matches the CDF and corresponding PDF.
    - Particularly so if we can determine X(C) from C(X).

Fitting Parameters for Continuous Distributions

Distribution	E[x]	Var[x]
Gaussian	μ	σ^2
Log-normal	$\exp[\mu + \sigma^2/2]$	$(\exp[\sigma^2] - 1)\exp[2\mu + \sigma^2]$
Gamma	$\alpha\beta$	$\alpha\beta^2$
Exponential	β	β^2
Chi-squared	ν	2ν
Pearson III	$\zeta + \alpha\beta$	$\alpha\beta^2$
Beta	$p / (p + q)$	$(pq)/[(p + q)^2(p + q + 1)]$
GEV	$\zeta - \beta[1 - \Gamma(1 - \kappa)] / \kappa$	$\beta^2[\Gamma(1 - 2\kappa) - \Gamma^2(1 - \kappa)] / \kappa^2$
Gumbel	$\zeta + \gamma\beta$	$\beta\pi / \sqrt{6}$
Weibull	$\beta\,\Gamma(1 + 1 / \alpha)$	$\beta^2[\Gamma(1 + 2 / \alpha) - \Gamma^2(1 - \kappa)] / \kappa^2$
Mixed Exponential	$w\beta_1 + (1 - w)\beta_2$	$w\beta_1^2 + (1 - w)\beta_2^2 + w(1 - w)(\beta_1 - \beta_2)^2$

μ = mean, σ = standard deviation

Gaussian Distribution

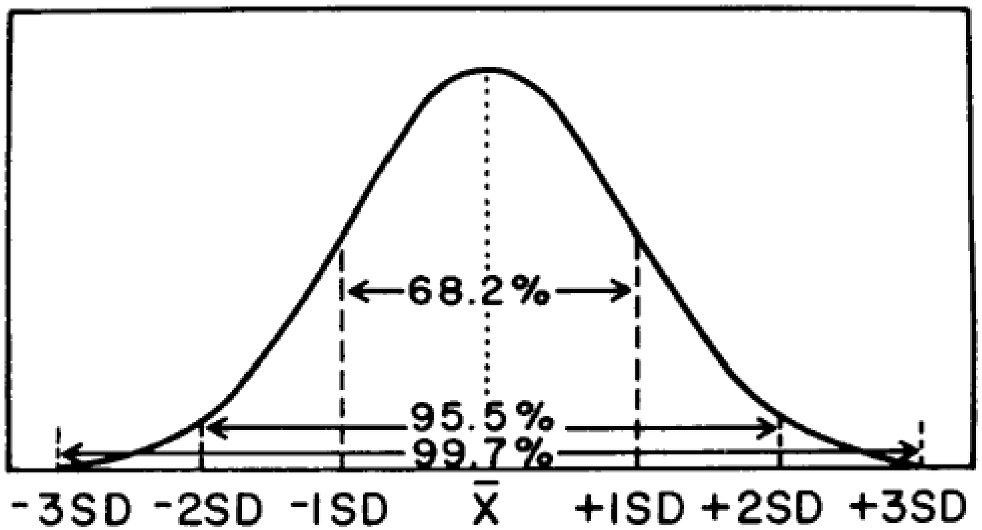
- A Gaussian distribution (bell curve) is relatively common, particularly when describing differences.
  - If a Gaussian distribution is normalized, meaning the area under the curve is equal to unity (one), then this special case of the Gaussian distribution is sometimes called a normal distribution.
  - Definitions do vary: Wilks defines the Gaussian distribution as I have defined a normal distribution.
- Estimates of a sum (or mean) will have a Gaussian distribution if the samples are (1) independent, and (2) of sufficient number.
  - The above statement is the **central limit theorem**.
  - The sufficient number is small if the population from which the samples are taken (and the sum calculated) has a near Gaussian distribution. It is larger (>100) for highly non-Gaussian PDFs.

Gaussian Distribution: The Formula

- A normal distribution is described by two parameters: a mean (μ) and a standard deviation (σ).
- A Gaussian distribution (not a pdf) would also have an amplitude.

$$pdf = f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \quad -\infty < x < \infty$$

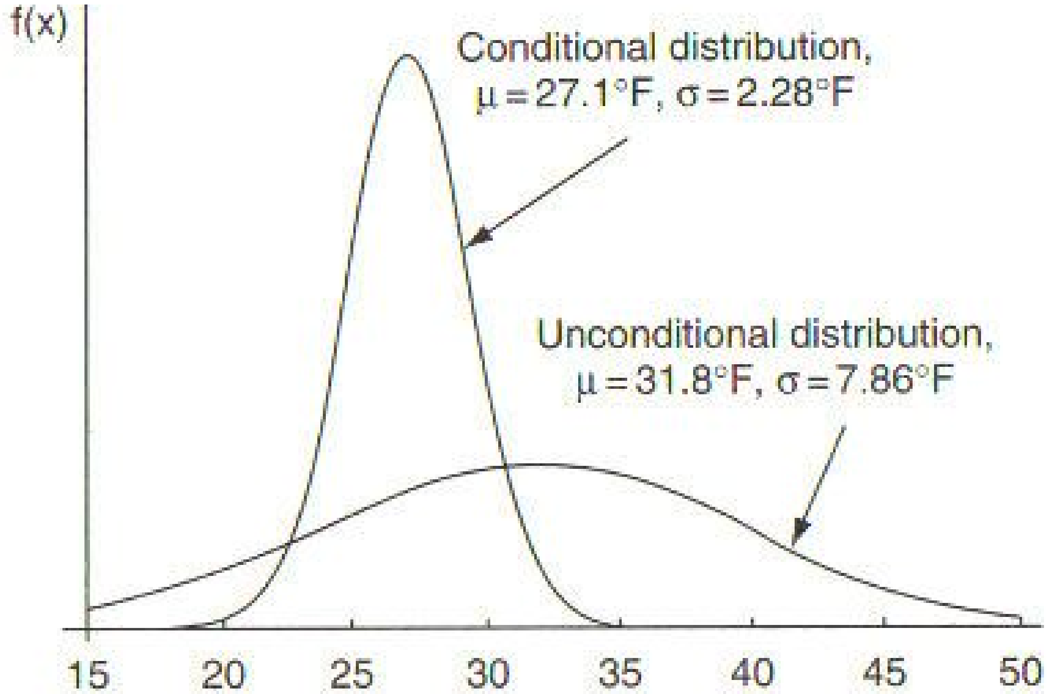
- Think about how the the standard deviation influences the shape of $f(x)$.
  - Larger σ implies a wider peak, and a smaller amplitude.



Graphic from http://homepage.univie.ac.at/Franz.Vesely/cp0102/dx/img579.png

Distributions For Conditional Probabilities

- The pdf for a conditional probability can have a very different shape than the unconditional probability.
- For example, consider the pdf for January daily maximum temperatures at Canandaigua: mean = 31.8°F, σ = 7.86°F.
- If the data set is restricted to those days when the temperature at Ithica was 25°F, then the mean is 27.1°F, and σ = 2.28°F



CDF For a Gaussian Distribution

- The technique for determining a CDF is often the integration of the corresponding pdf.

$$CDF(x) = \int_0^x pdf(x')dx'$$

- However, the Gaussian function is non-integratable.
- One approach to solving this problem is a lookup table.
  - Table B.1 in Wilks' book shows the probabilities in terms of z values: $z = (x - \mu) / \sigma$.
  - z scores are numbers of standard deviations above (positive values) or below (negative values) the mean.
- The lookup table shows $\Pr\{Z \leq z\}$
- Note that the Gaussian function is symmetric.
  - Therefore $\Pr\{Z \leq z\} = 1 - \Pr\{Z \geq -z\}$

Approximating the Gaussian CDF

- When a good approximation is sufficient, there is a relatively simple function that can be used as an approximate CDF, $\Phi(z)$.

$$\Phi(z) = \frac{1}{2}\left[1 \pm \sqrt{1 - \exp\left(\frac{-2z^2}{\pi}\right)}\right]$$

  - Where the positive root is used for $z > 0$, and the negative root for $z < 0$
  - Where z is the number of standard deviations from the mean.

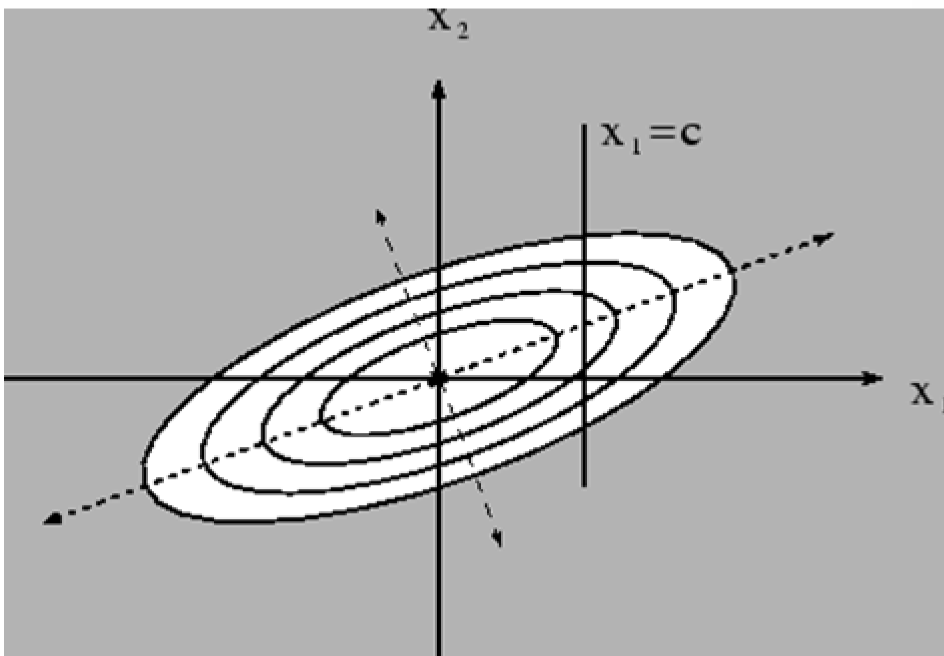
- The maximum errors (in probability) using this approximation are about 0.003 when z = ±1.65.
- This can be inverted to solve for z as a function of the value of the CDF.

$$z = \left[-\frac{\pi}{2}\ln\left[1 - [2\Phi(z) - 1]^2\right]\right]^{1/2}$$

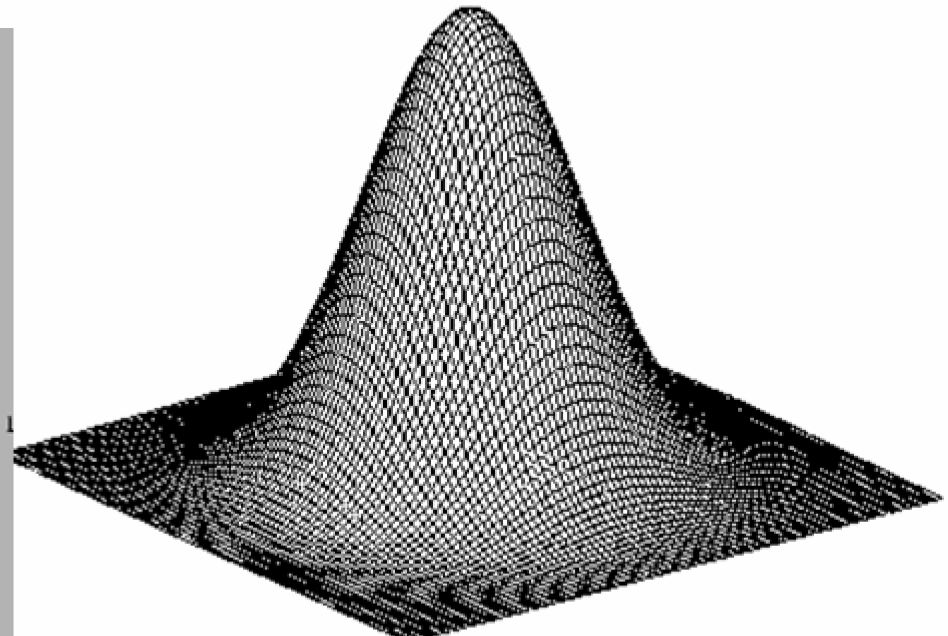
Two Dimensional Gaussian Distributions

- Two dimensional Gaussian PDFs are also common, particularly when showing differences in two spatial dimensions.

$$pdf = f(x) = \frac{1}{\sigma_x \sigma_y \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} \right) \right], \quad -\infty < x < \infty$$



Graphic from www.westgard.com/ lesson34.htm



Graphic from http://www.sia.uq.edu.au/physics/light/fred.gif

Log-Normal Distributions

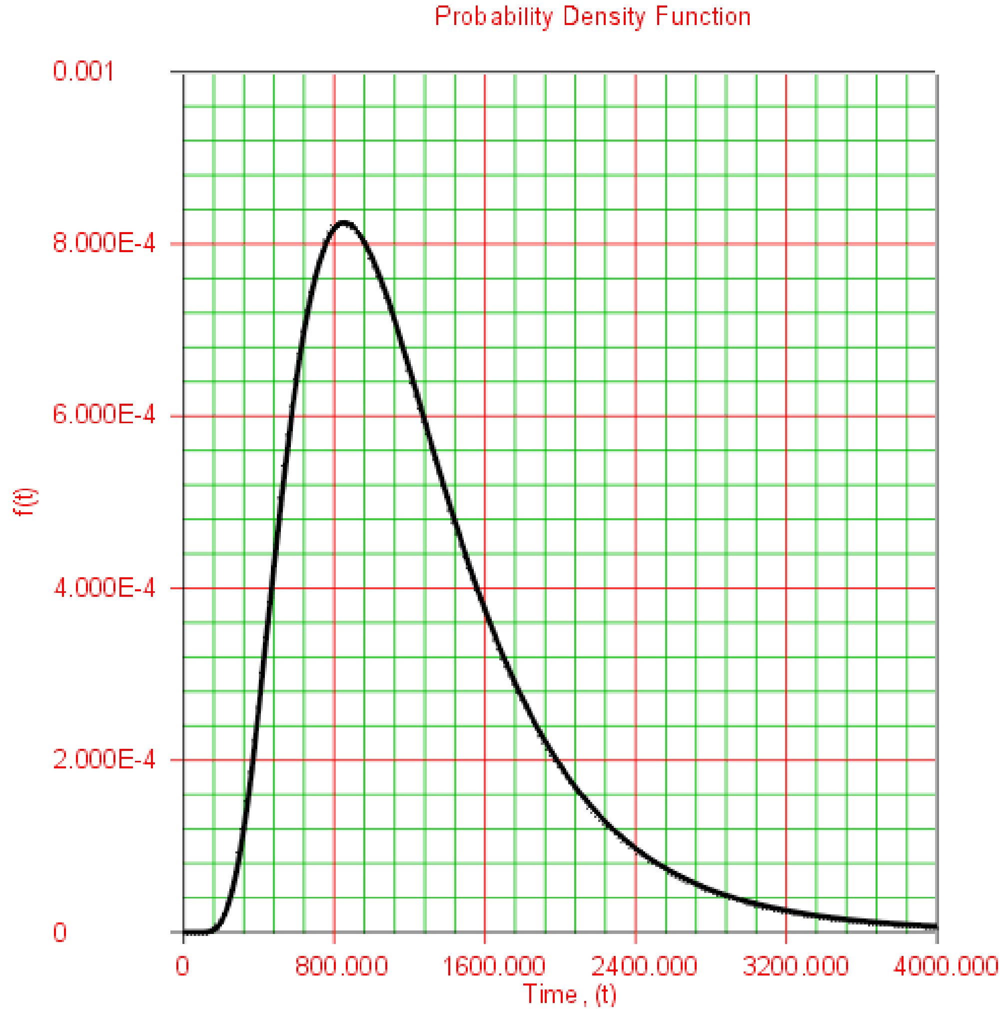
- There are many occurrences of distributions that have
  - (1) only positive values, and
  - (2) peak is displaced to the left.
- Some of these distributions are log-normal distributions.
  - A transformation of variables is used: $y = \ln(x)$

$$pdf = f(x) = \frac{1}{x\sigma_y\sqrt{2\pi}} \exp\left[-\frac{\left(\ln(x) - \mu_y\right)^2}{2\sigma_y^2}\right], \quad -\infty < y < \infty, \; y = \ln(x)$$

  - Where μ_y and σ_y are the mean and standard deviation of the transformed variable y.

- The mean of x is $\exp[\mu + \sigma^2/2]$, and
  The standard deviation of x is $(\exp[\sigma^2] - 1) \exp[2\mu + \sigma^2]$,
  - Where μ and σ are the mean and standard deviation of the transformed variable y.

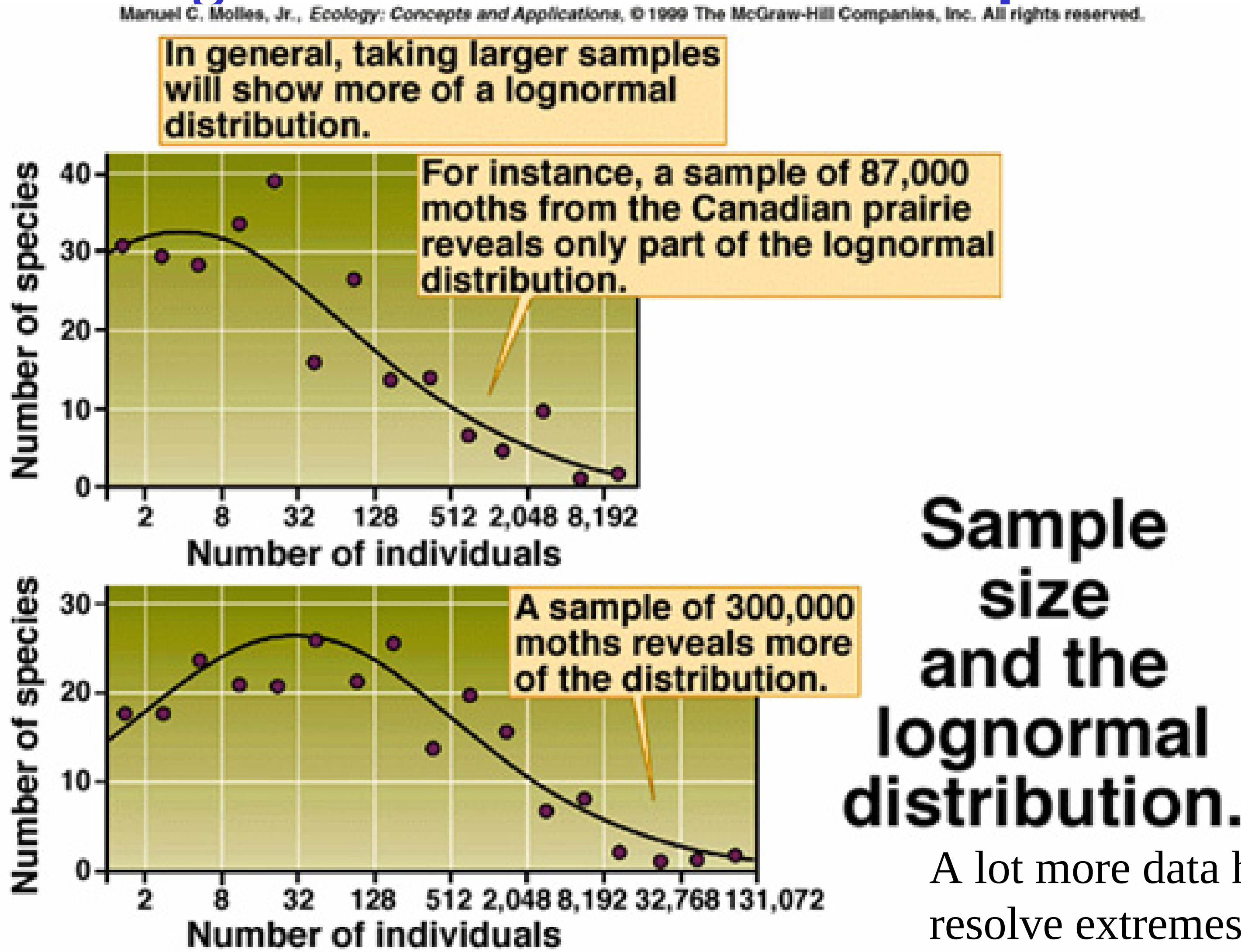
Log-Normal Distribution Example



- Features:
  - (1) only positive values, and
  - (2) peak displaced to the left.
- If the x-axis was plotted in log coordinates, then the distribution would appear to be Gaussian.

Graphic from http://www.weibull.com/AccelTestWeb/characteristics_of_the_lognormal_distribution.htm

Log-Normal Distribution Example



A lot more data helps resolve extremes

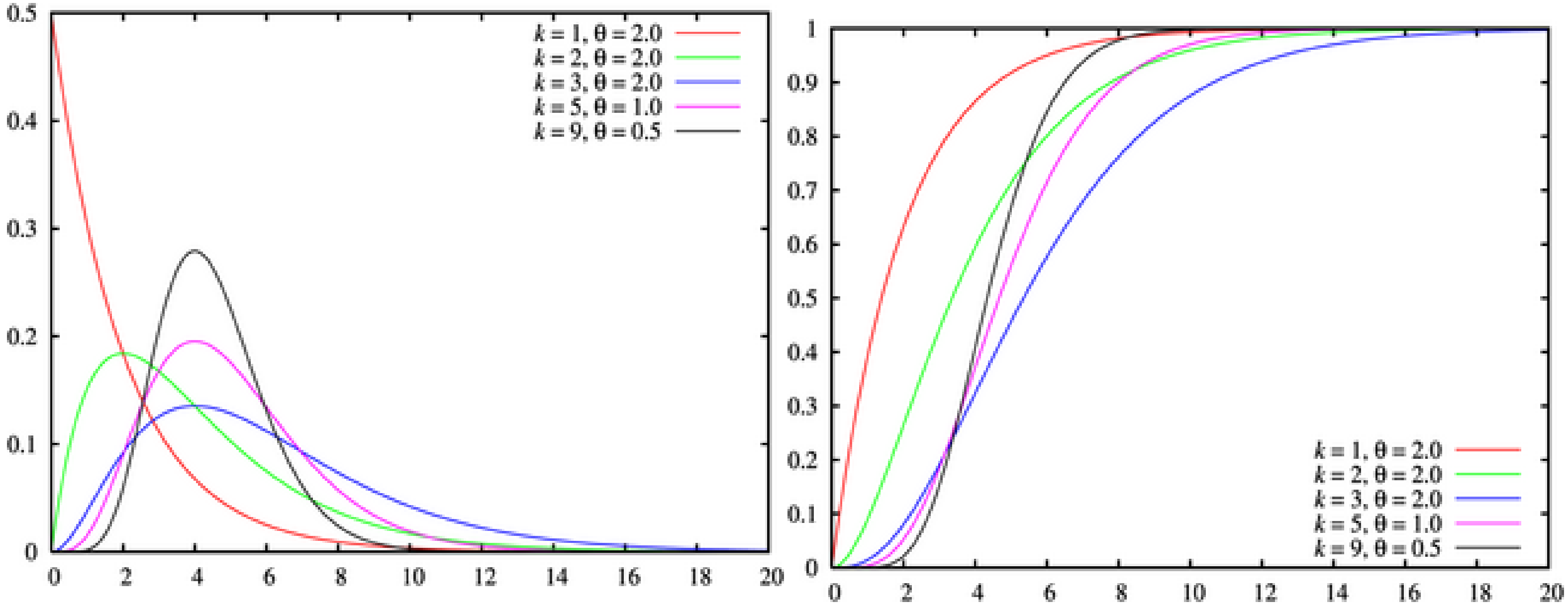
Graphic from http://www.biology.lsu.edu/heydrjay/1202/Chapter53/lognormal%20distribution.jpg

Gamma Distributions

- Gamma distributions are asymmetric, and skewed to the right (meaning the peak is to the left of the mean).
- They are well suited to describe variables that have a peak close to a limit.
  - For example, wind speed or precipitation.
- There are several different (but equivalent) forms of the gamma distribution. Each has two fitting parameters
- The fitting parameters are a shape parameter α , and a scaling parameter β.
  - Alternatively, it can be written with an inverse scale factor.

$$f(x) = \frac{(x/\beta)^{\alpha-1} \exp(-x/\beta)}{\beta\Gamma(\alpha)}, \quad for\ x,\ \alpha,\ \beta > 0$$

Gamma Distribution

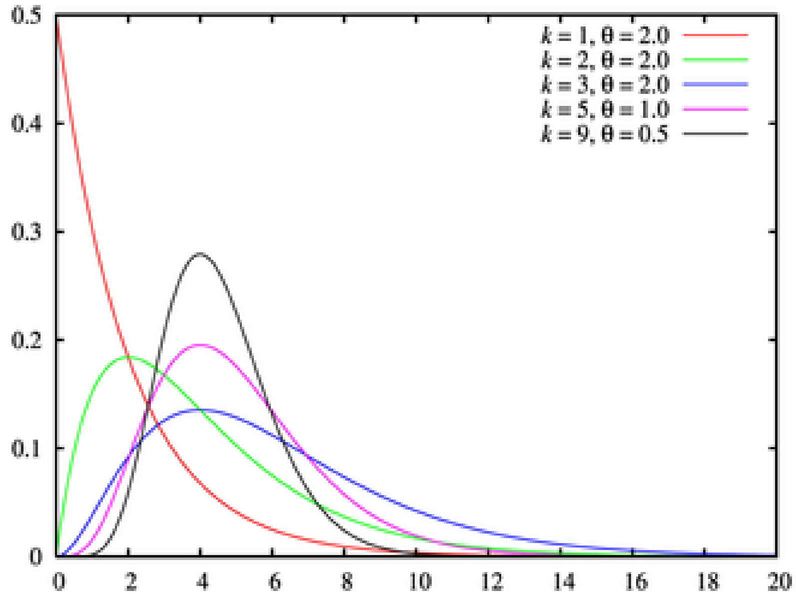


- The above examples use *k* and the shape parameter, and θ as the scale parameter.
- The left plot is the PDF, and the right plot is a CDF
- For a constant scale parameter, a smaller shape parameter will results in the peak being shifted further to the left

$$f(x;k,\theta) = x^{k-1}\frac{e^{-x/\theta}}{\theta^k\,\Gamma(k)} \text{ for } x > 0$$

Graphic from http://en.wikipedia.org/wiki/Gamma_distribution

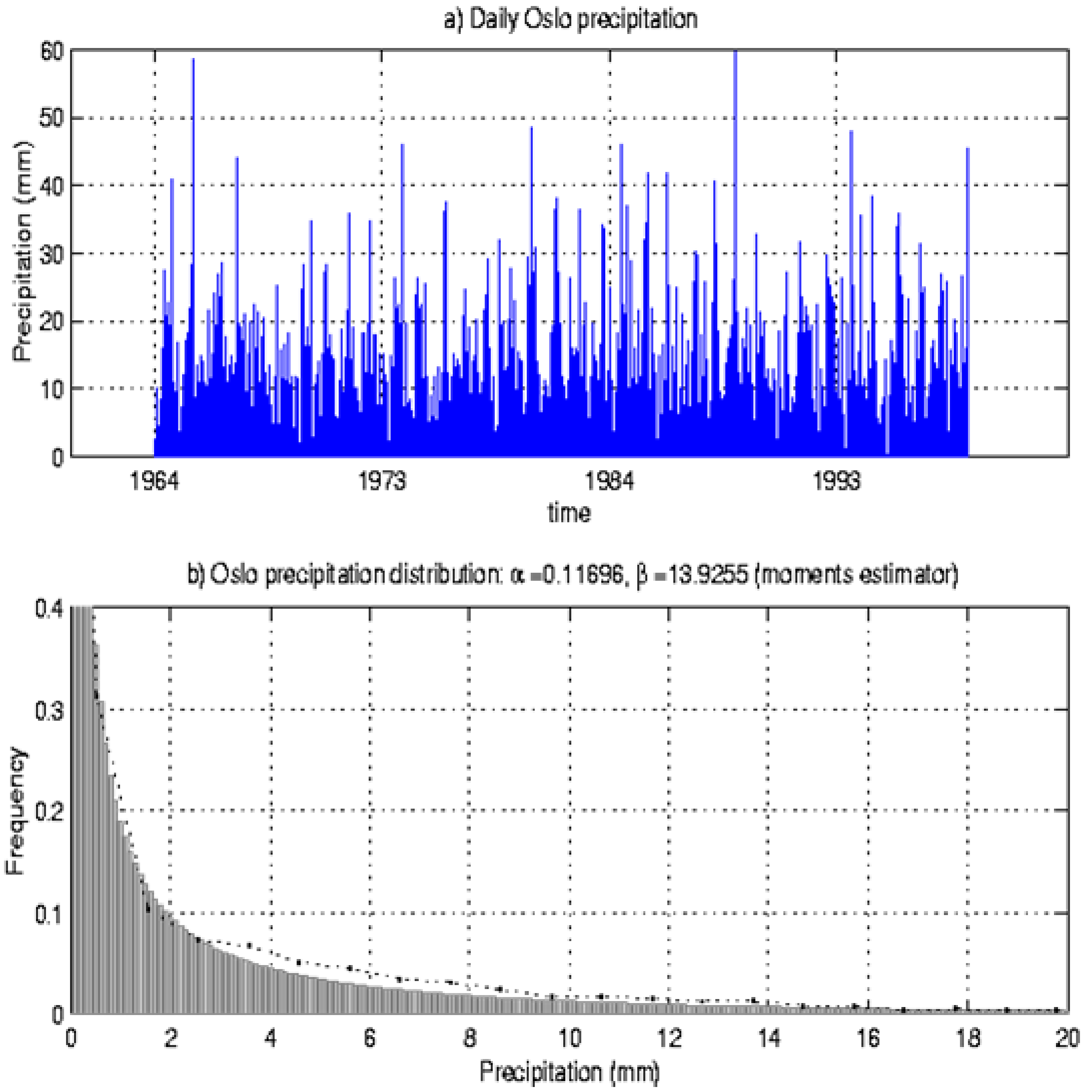
Gamma Distribution Parameters



$$f(x;k,\theta) = x^{k-1}\frac{e^{-x/\theta}}{\theta^k\,\Gamma(k)} \text{ for } x > 0$$

- For a shape parameter *k* =1, the equation simplifies greatly to an exponential distribution.
  - The y-intercept is 1/ *θ*.
- For a shape parameter *k* > 1, the y-intercept is zero.
  - Larger values of *k* result in less skewness, and shift the peak to the right.
  - For *k* > 50 or 100, the distribution is approximately Gaussian.

Gamma Distribution Example



- Time series of daily precipitation at Olso (top)
- The distribution function for daily precipitation in Oslo between 1883 and 1964 (bottom), with the dashed line showing the distribution for the above time period.

Graphic from www.gfi.uib.no/~nilsg/ kurs/notes/node31.html

Estimating the Gamma Distributions Scale Parameter

- We want to determine the fitting parameters α and β.
- We can solve for these in terms of the mean and the standard deviation of the gamma function.

$$\overline{x} = \alpha\beta$$

$$\sigma = \alpha\beta^2$$

- We can solve these equations for the fitting parameters:

$$\alpha = \overline{x}^2 / \sigma^2$$

$$\beta = \sigma^2 / \overline{x}$$

- What could go wrong with this approach?
  - Good for (shape parameter) $\alpha > 10$
  - Poor estimates of moments lead to problems for smaller α.

More Robust Estimates of Fitting Parameters

- Two better methods are based on *maximum likelihood estimators*.
  - This concept will be explained in later lectures
- Both approach use the same 1st calculation

$$D = \ln\left(\bar{x}\right) - \frac{1}{n}\sum_{i=1}^{n}\ln\left(x_i\right)$$

- The Thom estimators are

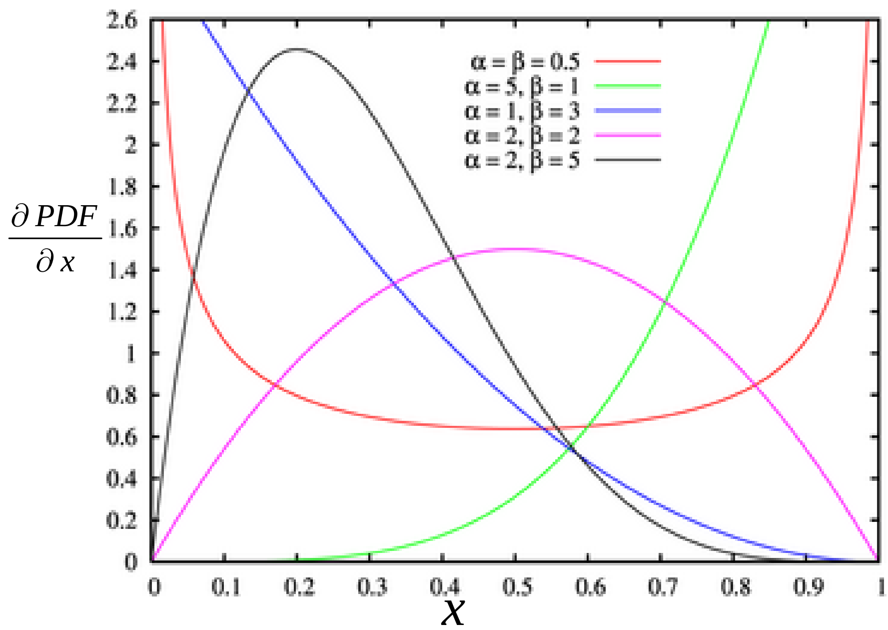
$$\alpha = \frac{1+\sqrt{1+4D/3}}{4D} \quad and \quad \beta = \bar{x}/\alpha$$

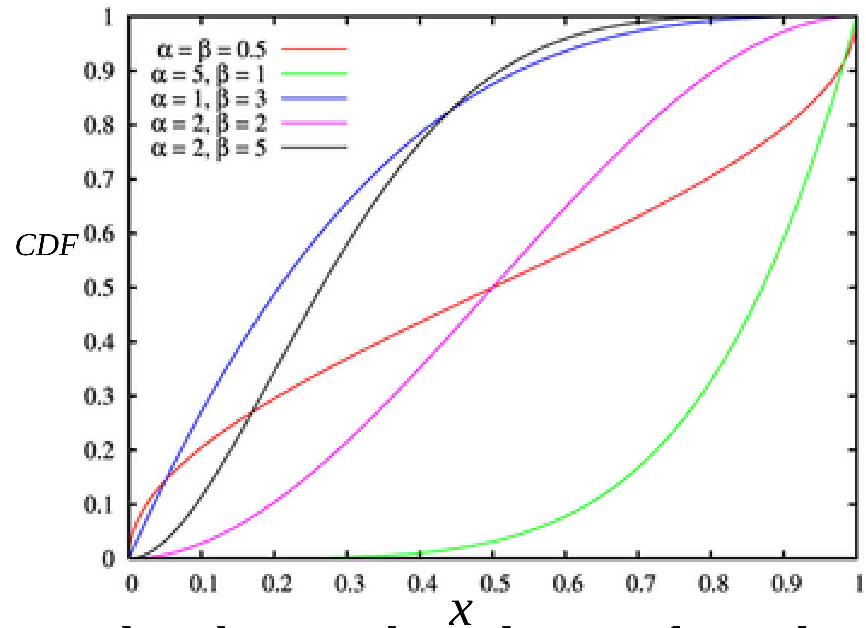
- The other method (Greenwood and Durand, *Technometrics*, 1960)

$$\alpha = \frac{0.5000876 + 0.1648852\,D - 0.0544274\,D^2}{D}, \quad 0 \le D \le 0.5772$$

$$\alpha = \frac{8.898919 + 9.059950\,D + 0.9775373\,D^2}{17.19728\,D + 11.968477\,D^2 + D^3}, \quad 0.5772 \le D \le 17.0$$

Beta Distributions





- Beta distributions have limits of 0 and 1.
  - Applications: RH & cloud cover
- They have two tuning parameters: α, β.
- The B term normalizes the PDF.
- If $\alpha = \beta$, the distribution is symmetric.
- If α and β are exchanged, the $f(x)$ is mirrored around $x = 0.5$.

$$f(x;\alpha,\beta) = \frac{1}{\mathrm{B}(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

$$f(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\,du} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1}(1-x)^{\beta-1}$$

Extreme Value Distributions

- Extreme value distributions usually apply to a small fraction of the events: the extreme events.
  - E.g., floods at a specific location
- The fraction can be artificially increased by using only extreme values in the distribution.
  - E.g., the annual maximum of daily precipitation totals (at a specific location).
- The Generalized Extreme Value (GEV) Distribution is

$$f(x) = \frac{1}{\beta}\left[1 + \frac{\kappa(x-\zeta)}{\beta}\right]^{1-1/\kappa} \exp\left\{-\left[\frac{\kappa(x-\zeta)}{\beta}\right]^{-1/\kappa}\right\}$$

  - Where ζ is a location or shift parameter,
    β is a scale parameter, and
    κ is a shape parameter.

CDF of a GEV Distribution

- The GEV equation can be integrated, resulting in a analytical CDF.

$$CDF(x) = \exp\left\{ -\left[1 + \frac{\kappa(x-\zeta)}{\beta} \right]^{-1/\kappa} \right\}$$

- The CDF can be inverted (solved for *x* as a function of CDF(*x*)).

$$CDF^{-1}(p) = x = \zeta + \frac{\beta}{\kappa}\left\{ \left[-\ln(p) \right]^{-\kappa} - 1 \right\}$$

- Given the fitting parameters, we can determine the extreme value as a function of the probability of that extreme (or greater) occurring.
  - We don't expect the distribution to work for likely occurrences.
  - However, as p becomes smaller, the distribution can be quite realistic.
  - Note that as $p \to 0$, that $\ln(p) \to -\infty$, resulting in rather large x.
- There are three special cases of the GEV Distribution. The two that we will examine are the Gumbel distribution and the Weibull distribution.

Gumbel Distribution

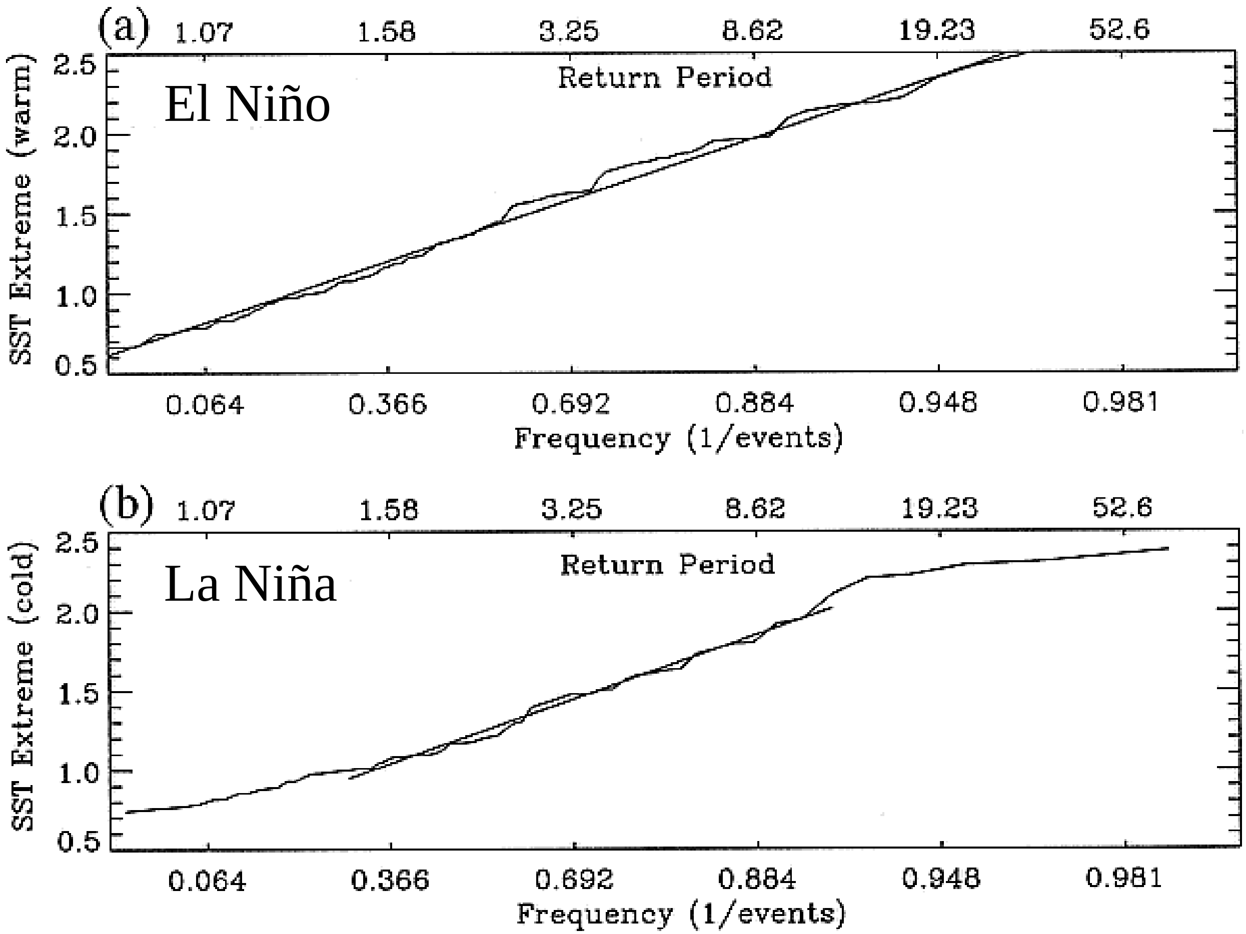
- Typically used to determine the average time between extreme events of the same magnitude or greater.
- The Gumbel distribution is the limit of the GEV distribution, where $\kappa \to 0$.

$$f(x) = \frac{1}{\beta} \exp\left\{ -\exp\left[\frac{(x-\zeta)}{\beta} \right] - \frac{(x-\zeta)}{\beta} \right\}$$

$$CDF(x) = \exp\left\{ -\exp\left[-\frac{(x-\zeta)}{\beta} \right] \right\}$$

- The fitting parameter can be estimated through a method of moments.
  - $\beta = \sigma\sqrt{6} / \pi$
  - $\zeta = \overline{x} - \gamma\beta$
  - Where γ = 0.57721… is Euler's constant.

Gumbel Distribution Example: Simulation of ENSO Extremes



- At the top of each plot is the return period in years.
- At the bottom of each plot is the corresponding frequency in a 40 year period.
- Note the lack of symmetry. This is important in time series analysis.

- Statistical mumbo-jumbo was used to generate 40 years of a sea surface temperature based ENSO index.

Graphics from Caron and O'Brien, *Mon. Wea. Rev.* 126, 2809–2821.

Return Period

- The return period is **average** time between events of a certain magnitude or greater.
- Note that the return period is an average. Three 100-year flood events have been known to happen in within 5 years.
- Suggesting that there might be year-to-year memory of ground water conditions.
- The return period R for an event of magnitude x or greater is
  $R(x) = 1 / \{\omega [1 - CDF(x)]\}$
  - Where ω is the sampling interval.

Weibull Distribution

- Weibull distributions are the limit of the GEV distribution where $\kappa < 0$.
- They have the distribution

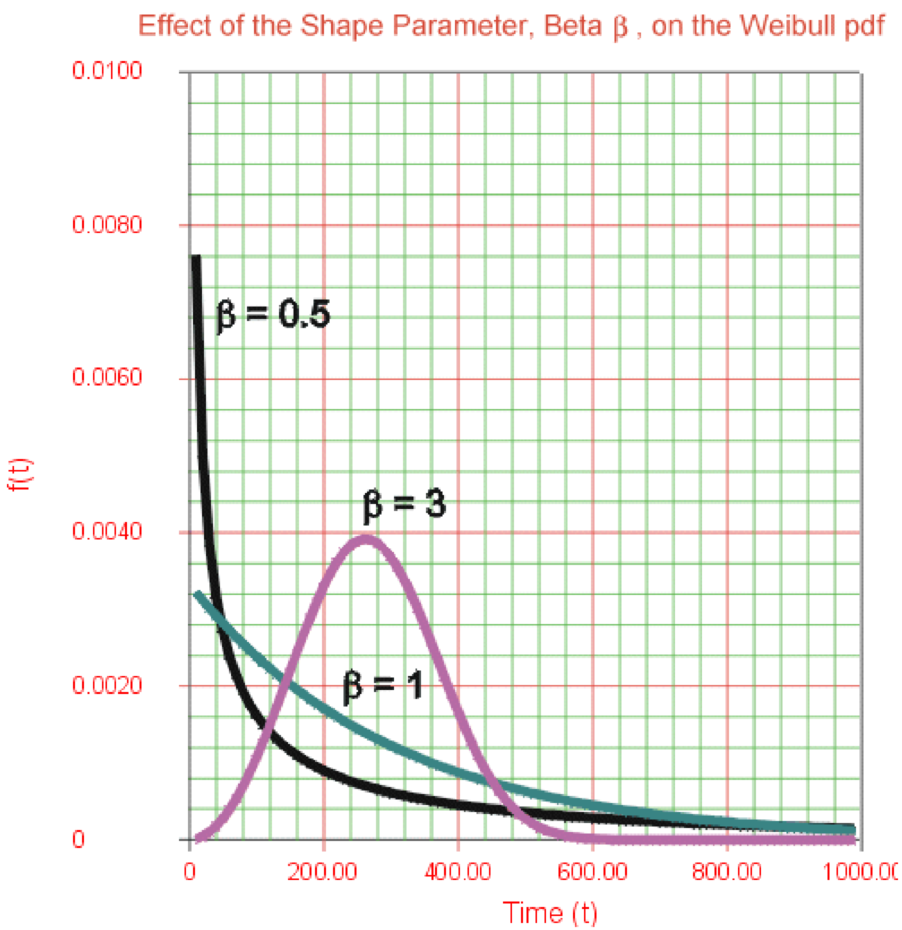
$$f(x) = \left(\frac{\alpha}{\beta}\right)\left(\frac{x}{\beta}\right)^{\alpha-1} \exp\left[-\left(\frac{x}{\beta}\right)^{\alpha}\right]$$

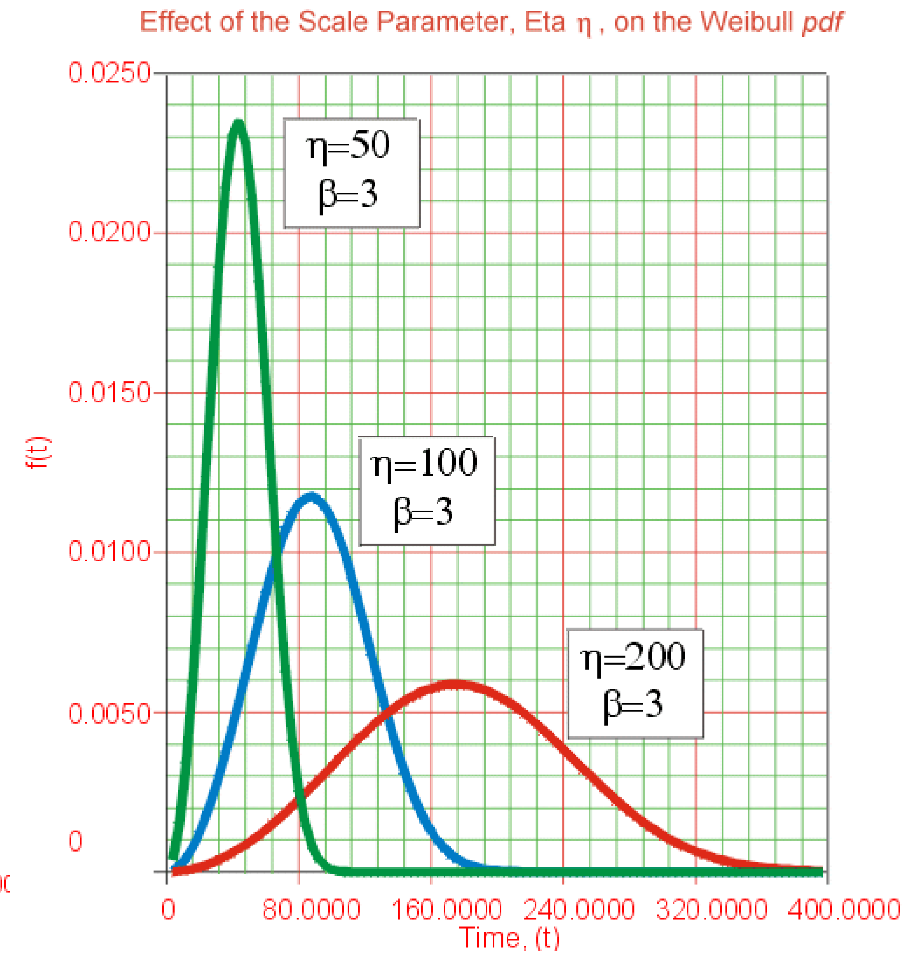
$$CDF(x) = 1 - \exp\left[-\left(\frac{x}{\beta}\right)^{\alpha}\right]$$

- The method of moments does not work for determining the fitting parameters. The gamma functions awkward.

Weibull Distribution Examples

Graphics from [www.weibull.com/ basics/parameters.htm](www.weibull.com/basics/parameters.htm)





- In this example the shape parameter is β (our α), and the scale parameter is η (our β).

Mixtures of Distributions

- For mildly complex physical situations, there is no reason that one type of distribution should fit the data.
- If there are several processes contributing to the physics (e.g., processes for generating rain), then it might be necessary to use a weighted average of several distributions.
- Example: wt1* (Gaussian Distribution 1) +
  wt2 * (Gaussian Distribution 2) +
  (1 – wt1 – wt2) * Weibull Distribution
  - Where 0 < wt1 < 1, 0 < wt2 < 1, and 0 < wt1 + wt2 < 1