Lecture Notes 3

Prof. Marvin Nakayama

Properties of Probability Distributions

- Properties of Distributions:
  Moments, Support, Tail, Shape, Failure Rate
- Specific Distributions
- Be sure to check course webpage for reading assignment and HW
- Download and install Arena.

Properties of Distributions

- Most of the important aspects of distribution are reflected in a few simple properties.
- A step in modeling is matching salient characteristics of real system to properties of theoretical distributions.

Moments

- kth moment and central moment of random variable X are $\mathcal{E}[X^k]$ and $\mathcal{E}[(X-\mu)^k]$.
- Most important *moments* of distribution are its mean and variance.
- For reasonably shaped distributions,
  mean measures central tendency.
  - Mean is often close to the mode—the single most likely value—and equals the median in the case of symmetric distribution.
- Variance measures spread of distribution
  - How far from mean typical values of distribution tend to fall.

- Consider random variable X with
  - mean μ
  - standard deviation σ
- **Chebyshev's inequality**: for any $k > 1$,

$$\mathcal{P}(|X-\mu| > k\sigma) \leq 1/k^2.$$

- Thus, at least 89% of all values fall within three standard deviation of mean.
- For most reasonable distributions, the number is closer to 95% or 99%.

- Generally speaking, matching the true mean and variance is most important step in building credible simulation model.
- Importance of mean is fairly obvious.
- Importance of variance should not be underestimated.
- Often convenient to replace variance with an equivalent measure, namely the *coefficient of variation*

$$c = \frac{\text{std. dev.}}{\text{mean}} = \frac{\sigma}{\mu}.$$

  (The *squared* coefficient of variation, c^2, is also used.)
- This measures variability *relative to the mean.*

- To illustrate, suppose
  - time to build an aircraft is 1 year on average with standard deviation of 3 weeks
  - time to build piano is 1 month on average with standard deviation of 2 weeks.
  - Which is more variable, aircraft or piano manufacturing?
  - Former has higher standard deviation, but latter has much higher coefficient of variation.
- Generally speaking,
  - Coefficient of variation ≥ 1 indicates high degree of variability
  - Coefficient < 1 indicates low to moderate variability.
  - But validity of this rule-of-thumb depends heavily on context.

Support

- *Support* of distribution is its set of possible values.
- For discrete-event simulations,
  - Primarily deal with non-negative random variables; i.e., distns whose support is contained within interval $[0, \infty)$.
  - For example, processing times or interarrival times are clearly non-negative quantities.
- For Monte Carlo simulations (e.g., in finance),
  - Often deal with distributions that allow for negative values, e.g., returns of risky asset.

- Among non-negative distributions, we can distinguish three important classes:

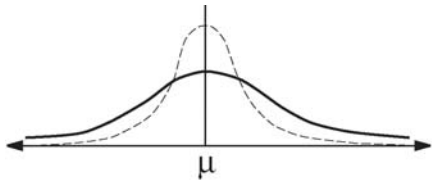
1. Support is all of $[0, \infty)$, so every positive value is possible (e.g., exponential distribution).
2. Support is finite interval $[a, b]$ (e.g., uniform distribution).
3. Support is finite number of points or all integer points (e.g., Poisson distribution).

- Do not be deterred from using distributions with support $[0, \infty)$ simply because very large values are unrealistic.
- For example, in an exponential distribution with mean 1, probability of observing value greater than 7 is less than 0.001.
- Thus, while very large values are possible, they are sufficiently unlikely to be ignored.
- Similar comment applies to normal distribution.
  - Mathematically, normal distribution has support $(-\infty, \infty)$.
  - In practice, values beyond three standard deviations are rarely observed.
  - So normal distribution with mean of 10 and standard deviation of 2 could be reasonable distribution for processing time.

- It is important not to use distribution whose support is discrete set when continuous-support distribution is appropriate.
- For example, should processing time ever follow Poisson distribution?
  - Only if all processing times are integer multiples of some basic unit of time; e.g., on computer.

Tails

- **Tail distribution**: $\bar{F}(x) = 1 - F(x)$, where F is CDF.
- Much of the story that a distribution tells is in the tails.
- Tail of distribution can be studied by examining how fast the following go to 0:
  - $f(x)$ as $|x| \to \infty$
  - $\bar{F}(x)$ as $x \to +\infty$ for right tail
  - $F(x)$ as $x \to -\infty$ for left tail.



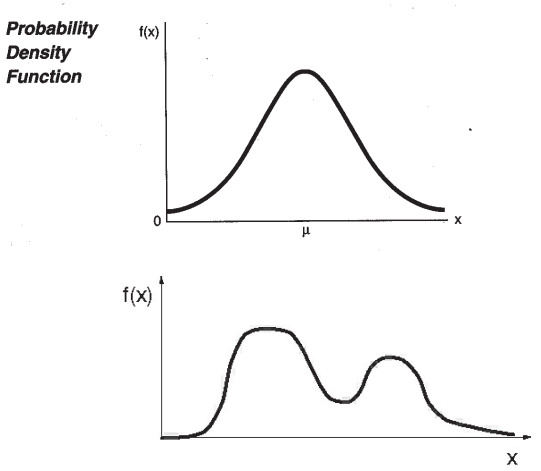
- For non-negative distributions, primarily interested in right tails
  - Probabilities of exceptionally high values.
- Phenomena with high degree of randomness are candidates for moderately thick tails.
  - Gamma distributions (including exponential) are suitable for such phenomena.
- Common criticism of normal distribution (in many settings) is that its tails are insufficiently thick

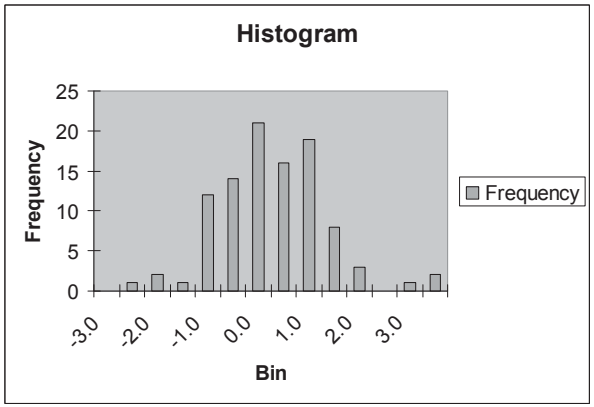
Shape

- Shape is rather nebulous but important set of considerations in choosing distribution.
- Indeed, if histogram of data is available, then shape should be primary criterion in choosing of distribution.
- Two specific aspects of shape to keep in mind are the following:

1. *Unimodality vs. multimodality:*
   - Unimodal distribution has single local maximum (hump)
   - Multimodal distribution can have several.
2. *Skewness vs. symmetry.*

Shape: Unimodal vs. Multimodal





- Histograms of real data are almost always multimodal.
- Whether or not to represent the data with multimodal theoretical distribution is modeling decision.

- *Skewness vs. symmetry.*
  - Histograms of observations with occasional large values tend to be skewed to the right.
    - ▷ Might model data with distribution having skewed density or mass function.
    - ▷ Exponential, for example, is extremely skewed.
  - Quantities subject to small random fluctuations are often symmetric about their mean.
    - ▷ Some processing times are of this type.
- Other aspects of shape may be difficult to describe precisely.
- But do not hesitate to consider them if they are evident to the eye and have meaningful interpretation.

Failure (or Hazard) Rate

- Probably less familiar way of looking at distributions than issues above.
- But it is of primary importance in many settings.
- It is a bit more technical, so we motivate it with a couple of examples.

- Consider front two tires on a car.
- Suppose one of the front two tires has a blowout and must be replaced.
- Now, of the two front tires, one is new and other is partly used.
- Which one will need to be replaced first?
- We cannot predict exactly when either will need to be replaced next, but we suspect that new one will last longer.
- Generally speaking, *the longer a tire has been in use, the greater the probability that it will need to be replaced shortly.*
- This is an example of an **increasing failure rate** (IFR) distn.

- Consider now time elapsed from when you leave a telephone message for a prospective employer (or, if you prefer, a prospective date) to when you obtain a reply.
- *The longer it's been since you called, the lower the probability that you are about to be called back.*
- This is an example of **decreasing failure rate** (DFR) distn.

- Let X be random variable representing some time period (lifetime of component, interarrival time, etc.).
- Let F be CDF of X and let f be its density.
- Then probability that X will run out in the interval $[t, t+\Delta]$, *given* that X has lasted t time units so far, is

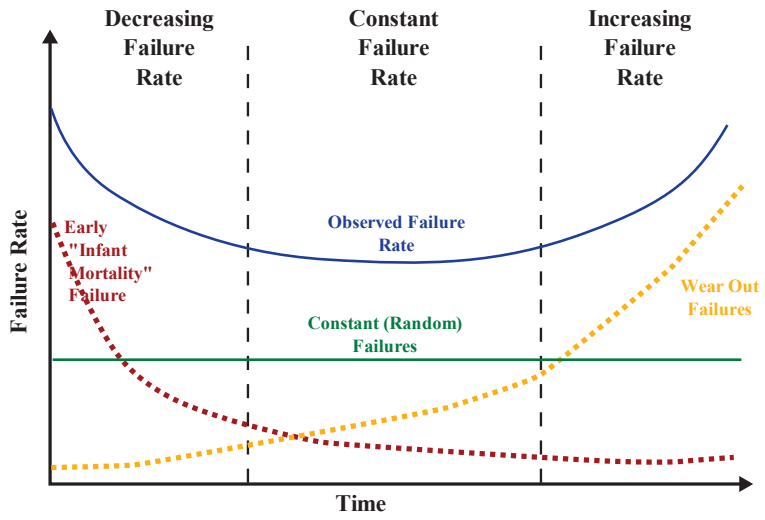
$$\mathcal{P}(X \leq t+\Delta | X \geq t) = \frac{\mathcal{P}(t \leq X \leq t+\Delta)}{\mathcal{P}(X \geq t)} \approx \frac{f(t)\ \Delta}{1-F(t)}.$$

- Approximation becomes exact as Δ becomes small.
- *Failure rate* (also called the *hazard rate*) is function

$$h(t) = \frac{f(t)}{1-F(t)}.$$

- Interpretation: $h(t)$ is the instantaneous probability that X runs out just after time t, given that X has survived until time t.

- If $h(t)$ increases with time, then F is an IFR distribution.
- If $h(t)$ decreases with time, then F is DFR distribution.
- Most distributions are neither IFR nor DFR.
- In practice, lifetimes of many electrical or mechanical components have what are called *bathtub-shaped* failure rates:
  - Failure rate starts off high due to defects or "lemons" (this is called infant mortality).
  - Then levels off for a period.
  - Finally increases again as component ages.
- Unfortunately, none of the commonly used distributions has bathtub-shaped failure rate
  - So we must decide whether the increasing part or the decreasing part is more important.



[Source: Wikipedia]

- When would it be appropriate to model following quantities with IFR or DFR distributions?
  - Time between bus arrivals.
  - Time between machine breakdowns.
  - Time between earthquakes.
  - Time on hold waiting for technical support at a call center.

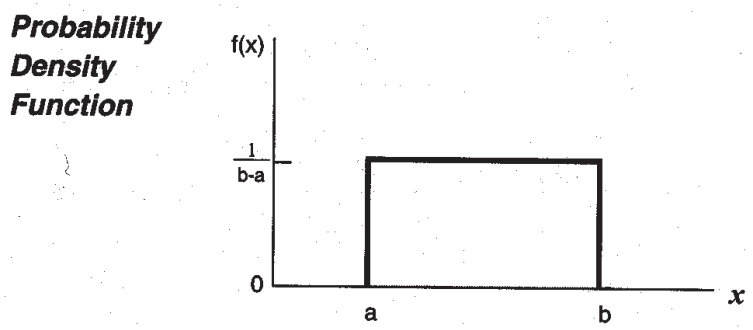
Some Special Distributions

- We now examine some specific distributions.
- Give suggestions when appropriate to use each.

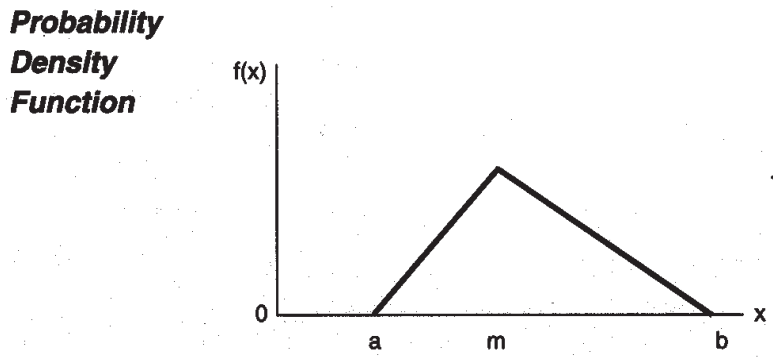
Uniform and Triangular

- These densities are sometimes used when little data is available but something is known about range of likely values.
- Uniform density on an interval $[a, b]$
  - horizontal line at height $1/(b-a)$ (so that area under density is 1).
- Triangular density on $[a, b]$ with peak at m
  - increases linearly from a to m
  - decreases linearly from m to a.

Density of Uniform Distribution



Density of Triangular Distribution



- Uniform density is sometimes used to model ignorance, since it makes all values in a range equally likely.
- In practice, the abrupt drop off in uniform density outside $[a, b]$ makes uniform somewhat artificial.
- In contrast, triangular density predicts a gradual drop off away from mode.
- Both uniform and triangular distributions are useful in eliciting information from experts in absence of data.
- For example, a factory worker would be more comfortable specifying a minimum time and a maximum time for an operation than specifying mean and variance.
- For same reason, these distributions are useful in communicating model in non-technical terms.

Exponential

- Exponential distribution with mean $\beta > 0$ has density

$$f(x) = (1/\beta)\exp(-x/\beta), \quad x \geq 0,$$

and $f(x) = 0$ for $x < 0$.

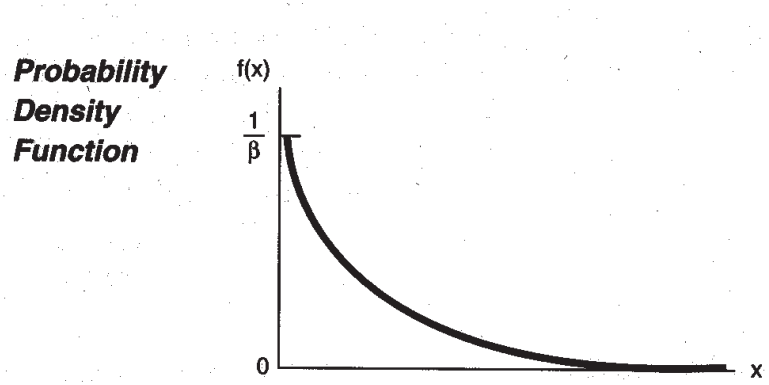
- Cumulative distribution function

$$F(x) = 1 - \exp(-x/\beta), \quad x \geq 0,$$

and $F(x) = 0$ for $x < 0$.

- Alternatively specify exponential with *rate* $\lambda = 1/\beta$.
  - density becomes $f(x) = \lambda\exp(-\lambda x)$
  - CDF becomes $F(x) = 1 - \exp(-\lambda x)$

Density of Exponential Distribution



- Failure rate of exponential is then

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{\lambda e^{-\lambda t}}{1 - (1 - e^{-\lambda t})} = \lambda, \quad t \geq 0.$$

- Thus, failure rate is constant for all $t \geq 0$, which is known as the *memoryless* property
  - Most distinguishing feature of exponential distribution.
  - *Only* exponential has this property.
- If, for instance, bus interarrival times follow an exponential distribution, then
  - No matter how long you've been waiting, probability of an arrival remains constant.
  - No matter how long you've been waiting, expected time until next arrival is $1/\lambda$, the original mean time between arrivals.
  - Fortunately, most real bus lines do better than this.

- Exponential is generally only suitable when time until next occurrence of an event is truly random.
- Exponential distribution is often used to model lifetimes of electronic components (e.g., light bulbs, CPUs in computers).
- However, it is probably not reasonable to model lifetimes of mechanical components with exponential distributions. Why?
- Variance of an exponential is square of its mean.
  - Consequently, exponential has coefficient of variation equal to one.
  - This indicates high variability, consistent with exponential as model of random phenomena.

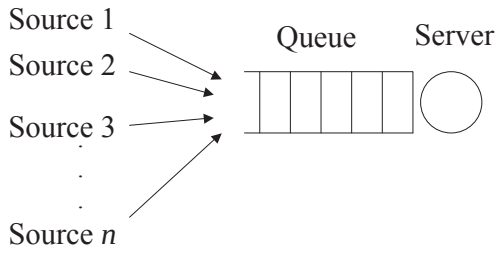
Poisson Process

Poisson process $[N(t) : t \geq 0]$ with rate $\lambda > 0$.

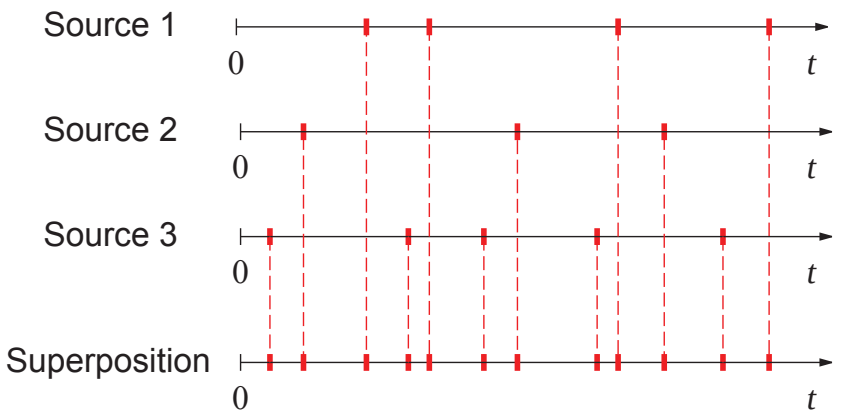
- Renewal process with exponentially distributed inter-event times:
  - Events occur repeatedly over time; e.g.,
    - ▷ Arrivals of customers to bank.
    - ▷ Generation of messages at node in network
    - ▷ Jumps in price of a stock
  - Time between events IID exponential with rate λ (i.e., mean $1/\lambda$), with λ constant over time.
- $N(t) =$ number of events that occurred by time t.
- For each time t, $N(t)$ has Poisson distn with parameter λt; i.e.,

$$\mathcal{P}\{N(t) = k\} = \frac{(\lambda t)^k e^{-\lambda t}}{k!}, \quad \text{for } k = 0, 1, 2, \ldots.$$

- Consider arrival process to queue.
- Can frequently view arrivals as aggregate *superposition* of arrivals from variety of statistically independent sources.



Arrival Times



- Example: Fast-food restaurant
  - Sources: n different neighborhoods
  - Arrival epochs corresponding to source i are times at which vehicle from ith neighborhood visits restaurant.
- Example: Telephone switch
  - Sources: n different telephones in area
  - Arrival epochs are times at which calls are initiated from phone in question.
- Reasonable to assume that number n of sources is large.

- Aggregate arrival process to queue (also known as "superposition process") often looks approximately like a Poisson process.
- **Palm-Khintchine Theorem:**
  - Under mild conditions on individual sources (each is negligible compared to superposition), superposition process of n sources is approximately Poisson process as $n \to \infty$.
- Can generalize Poisson process to have non-constant rate function $\lambda(t)$.
  - $\lambda(t)$ is arrival rate at time t.
  - Known as **non-homogeneous** (or **non-stationary**) **Poisson process**.
  - Useful to model systems in which many events occur during certain time periods and fewer at other time periods.

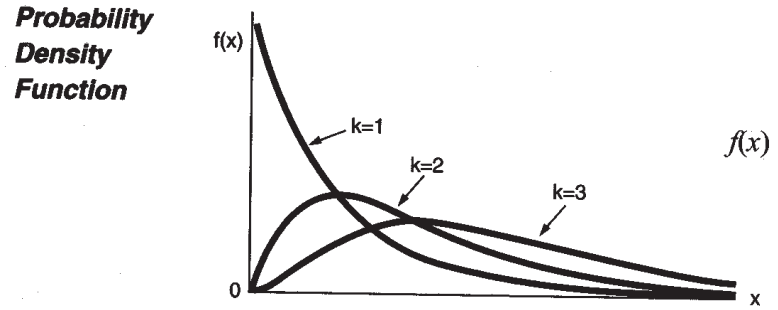
Erlang

- Erlang distribution is close relative of exponential with lower variability (relative to its mean).
- Imagine process that takes k (sequential) stages
  - Each stage requires an indep exponential time with mean β.
- Then total time required, adding over k stages, has an Erlang distribution with
  - *shape parameter* k
  - *scale parameter* β
- Density function:

$$f(x) = \frac{\beta^{-k} x^{k-1} e^{-x/\beta}}{(k-1)!}, \quad x \geq 0$$

- When $k = 1$, we get exponential again.

Density of Erlang Distribution



- Mean of an Erlang is $k\beta$
- Variance is $k\beta^2$
- Consequently, squared coefficient of variation is

$$c^2 = \frac{1}{k}$$

- By increasing k, we obtain c^2's of 1/2, 1/3, 1/4, etc.

Gamma

- Gamma distribution generalizes Erlang
  - No longer require shape parameter to be an integer
  - There is no longer interpretation of stages.
- With shape parameter now denoted by α, gamma density becomes

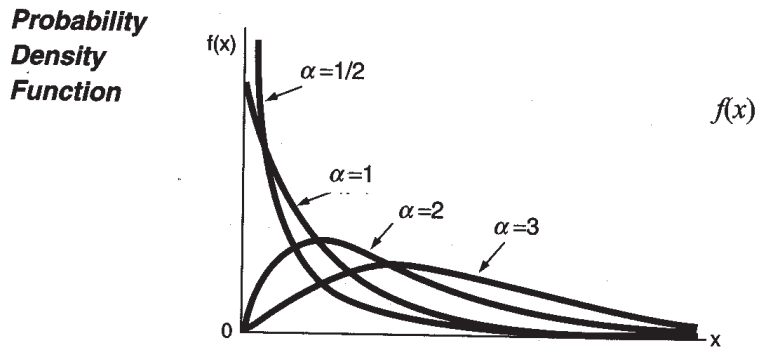
$$f(x) = \frac{\beta^{-\alpha}\, x^{\alpha-1}\, e^{-x/\beta}}{\Gamma(\alpha)}, \quad x \geq 0,$$

and $f(x) = 0$ for $x < 0$, where

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}\, e^{-t}\, dt$$

is so-called gamma function (generalization of factorial, i.e., $\Gamma(k+1) = k!$ for integers $k \geq 0$).

Density of Gamma Distribution



- Mean: $\alpha\beta$
- Variance: $\alpha\beta^2$
- As α increases, hump of gamma density moves to right.
- Gamma family is versatile collection of distributions with support $[0, \infty)$.
- Gammas are useful in modeling task times.

Beta Distribution

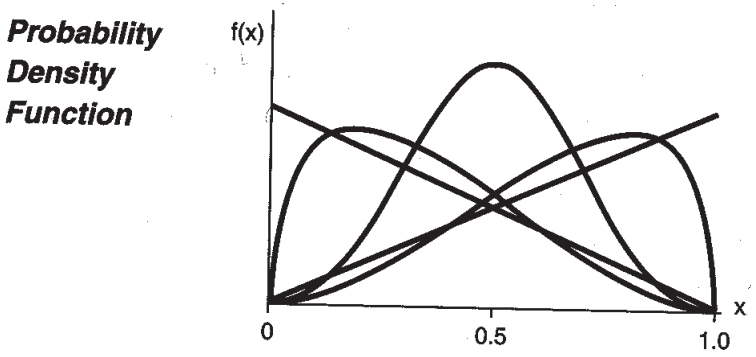
- Standard beta family of distributions has support $[0, 1]$
- Beta family has two parameters: $\beta > 0$ and $\alpha > 0$.
- Varying parameters leads to very different shapes for density.
- Density function:

$$f(x) = \frac{x^{\beta-1}\,(1-x)^{\alpha-1}}{B(\beta,\alpha)}, \quad 0 \leq x \leq 1,$$

where $B(\beta,\alpha) = \int_0^1 t^{\beta-1}\,(1-t)^{\alpha-1}\, dt$ is the beta fcn.

- Mean $\beta/(\beta+\alpha)$, and variance $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
- Suppose X is standard beta, i.e., with support $[0, 1]$.
  - Set $Y = a + (b-a)X$.
  - Then Y is general beta with support $[a, b]$.

Density of Beta Distribution



Normal

- Density function

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

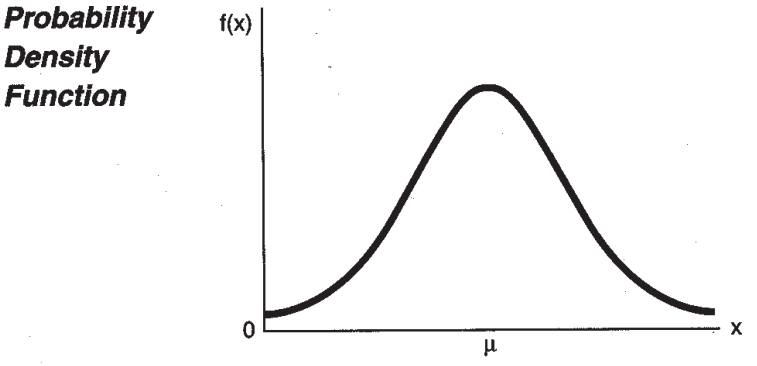
  - μ is mean
  - σ is standard deviation

- One main use of normal distribution is for statistical analysis of simulation results.

- But sometimes used to model processing times or stock returns.

- ARENA truncates normal random variables so that they never lie more than three standard deviations from mean.
  - This makes it possible to avoid negative values.

Density of Normal Distribution



- Suppose that quantity X that we are trying to model can be expressed as sum of other random variables as follows:

$$X = Y_1 + Y_2 + \cdots + Y_n.$$

- In great generality, various versions of central limit theorem assert that, for n large,

$$X \overset{\mathcal{D}}{\approx} \mathcal{N}(\mu_n, \sigma_n^2)$$

  for some constants μ_n and σ_n^2.

- Note that Y_i's need not be identically distributed, nor do they need to be independent of one another.

- **Moral of Story:** If X is sum of large number of other random quantities, then X can be approximately modeled as normal random variable.

Lognormal

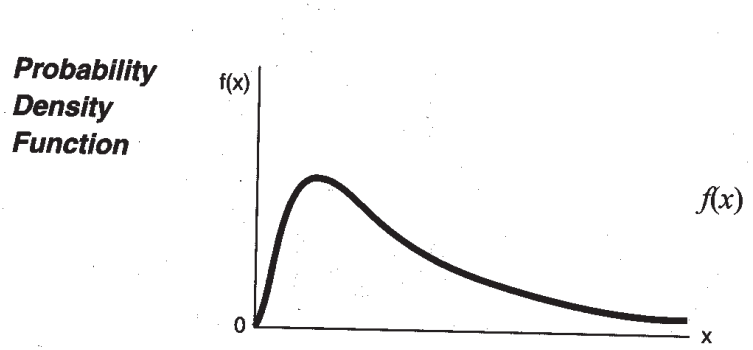
- Random variable Y has lognormal distribution $\mathcal{LN}(\mu, \sigma^2)$ if $Y = e^X$ and X has normal distribution $\mathcal{N}(\mu, \sigma^2)$.
- Support $(0, \infty)$.
- Lognormal dist has fairly heavy tail, so it has lot of variability.
- Lognormal r.v.s can occasionally take on very large values.
- Density function:

$$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(\ln x - \mu)^2}{2\sigma^2}\right\}, \quad x \geq 0,$$

and $f(x) = 0$ for $x < 0$.

- Mean: $\exp\{\mu + \sigma^2/2\}$ (**not** μ)
- Variance: $e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$ (**not** σ^2)

Density of Lognormal Distribution



- Lognormal distribution widely used to model
  - machinery reliability
  - stock prices
- Because of central limit theorem, lognormal distribution may be appropriate when taking product of many IID random variables. Why?
- **Example:**
  - If X_0 is initial value of stock and R_i is return on ith day, then value on nth day is

$$X_n = X_0(1 + R_1)(1 + R_2)\cdots(1 + R_n)$$

  - If R_i are IID, then X_n is approximately lognormal for n large.

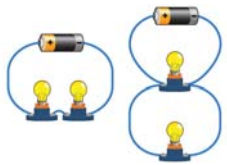
Weibull

- Component lifetimes often modeled with Weibull dist's.
- Consider lifetime of component that is itself made up of n sub-components.
- Let L be lifetime of component.
- Let $Y_1, \ldots, Y_n$ be lifetimes of n sub-components.
- If sub-components are "in series," then

$$L = \min_{1 \leq i \leq n} Y_i$$

- If sub-components are "in parallel," then

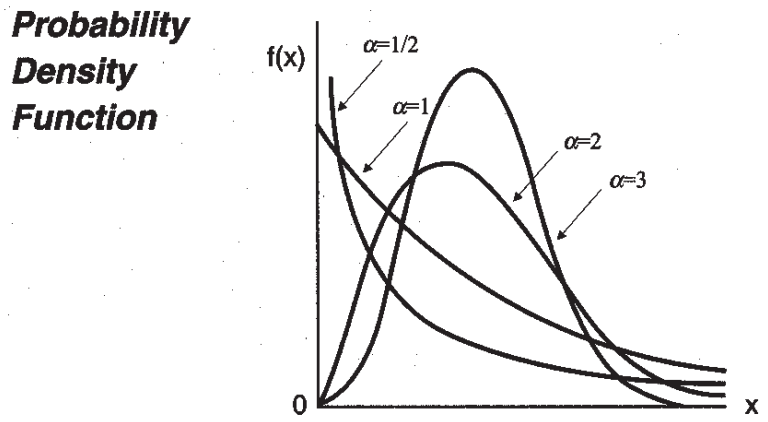
$$L = \max_{1 \leq i \leq n} Y_i$$



[source: guides.machinescience.org]

- *Extreme-value theory* develops approximations to distribution of L when n is large and Y_i's are IID.
  - Under certain conditions, L will have an approximate Weibull distribution when n is large.
- Weibull family has two parameters:
  - shape parameter α
  - scale parameter β
- Density is skewed to right but becomes more symmetric as α increases.

Density of Weibull Distribution



- Density function

$$f(x) = \alpha\,\beta^{-\alpha}\,x^{\alpha-1}\,e^{-(x/\beta)^{\alpha}}, \quad x \geq 0,$$

  and $f(x) = 0$ for $x < 0$.
- Cumulative distribution function

$$F(x) = 1 - e^{-(x/\beta)^{\alpha}}, \quad x \geq 0,$$

  and $F(x) = 0$ for $x < 0$.
- Mean: $\frac{\beta}{\alpha}\Gamma\left(\frac{1}{\alpha}\right)$
- Variance: $\frac{\beta^2}{\alpha}\left(2\Gamma\left(\frac{2}{\alpha}\right) - \frac{1}{\alpha}\left[\Gamma\left(\frac{1}{\alpha}\right)\right]^2\right)$
- Failure rate:

$$h(t) = \alpha\,\beta^{-\alpha}\,t^{\alpha-1}$$

  so Weibull is IFR if $\alpha > 1$ and DFR if $\alpha < 1$.
- What happens if $\alpha = 1$?

Mixtures

- Mixture of distributions is appropriate when
  - Observations can be categorized into n categories, $n > 1$
  - Probability of arbitrary observation being in category i is p_i
  - Each category i has its own CDF F_i
  - If $n = 2$
    - ▷ Some fraction of observations follows one distribution
    - ▷ Rest follow another distribution

- To illustrate, consider length of jobs at photocopier.
- Suppose copy jobs completed without paper jams take exponentially distributed amount of time, with mean of 1 minute.
  - High degree of randomness: jobs easily vary from 1 copy of 1 page to 10 copies of 20 pages.
  - So exponential distribution is reasonable.
- Suppose further that jobs with paper jams are also exponential but with mean of 2 minutes.
- If 5% of jobs are interrupted by paper jams, then overall density of job durations is

$$f(x) = (0.05)(0.5e^{-0.5x}) + (0.95)(e^{-x})$$

- Mixture of exponentials is called *hyperexponential* distribution.

- More generally, if time to complete task follows density f_i with probability p_i, $i = 1, \ldots, n$, with $p_1 + \cdots + p_n = 1$, then
  - Overall task-time density:

$$f(x) = p_1 f_1(x) + \cdots + p_n f_n(x)$$

  - CDF:

$$F(x) = p_1 F_1(x) + \cdots + p_n F_n(x)$$

  - If each f_i has mean μ_i and variance σ_i^2, then mixture has
    - ▷ mean: $\mu = p_1\mu_1 + \cdots + p_n\mu_n$
    - ▷ variance:

$$\sigma^2 = \left(\sum_{i=1}^{n} p_i(\sigma_i^2 + \mu_i^2)\right) - \left(\sum_{i=1}^{n} p_i\mu_i\right)^2$$

Multivariate Distributions

Example 1:

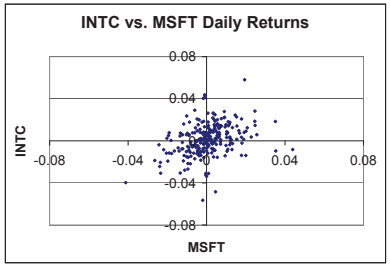
- Suppose a job consists of 2 subtasks to be completed sequentially.
- So total time to complete job is

$$T = X_1 + X_2$$

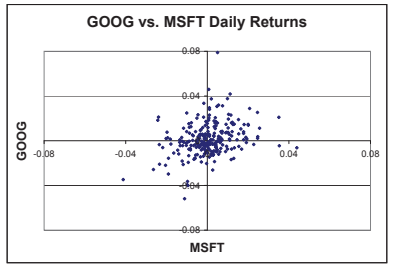
  where X_1 and X_2 are subtask times.
- X_1 and X_2 may be **correlated**.
  - If first subtask takes an unusually long time, then same is likely true for second subtask.
  - Erlang distn assumes task times are IID exponentials.

Example 2:

- 3 stocks, e.g., MSFT, INTC, GOOG
- Let X_i denote the return of the ith stock tomorrow.
- Often X_1, X_2, X_3 are **correlated**.
  - If MSFT has an unusually large return tomorrow, then INTC and GOOG are also likely to.



Correlation: 0.366



Correlation: 0.302

- Let $\mathbf{X}$ be a random vector,

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_d \end{pmatrix}$$

where X_i and X_j not necessarily independent for $i \neq j$

- Distribution of $\mathbf{X}$ is specified by the **joint CDF** F of $X_1, X_2, \ldots, X_d$,
$F(\mathbf{x}) = \mathcal{P}(\mathbf{X} \leq \mathbf{x}) = \mathcal{P}(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_d \leq x_d)$
where $\mathbf{x} = (x_1, x_2, \ldots, x_d)^\top$ and superscript $\top$ is transpose.

- Equivalently, when $X_1, \ldots, X_d$ are jointly continuous RVs, distribution of $\mathbf{X}$ is specified by the **joint density** f,

$$f(\mathbf{x}) = \frac{\partial^d}{\partial x_1 \cdots \partial x_d} F(x_1, \ldots, x_d)$$

Multivariate Normal Distribution

- $\mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes d-dimensional normal distribution
  - $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_d)^\top$ is mean vector $\mathcal{E}[\mathbf{X}]$
  - $\boldsymbol{\Sigma}$ is the covariance matrix, with (i, j)th entry

$$\Sigma_{ij} = \mathrm{Cov}(X_i, X_j) = \mathcal{E}[(X_i - \mu_i)(X_j - \mu_j)] = \Sigma_{ji}$$

- Density function of $\mathbf{X} = (X_1, \ldots, X_d)^\top \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$f(\mathbf{x}) = (2\pi)^{-d/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{ -\frac{(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}{2} \right\}$$

  - $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$
  - $\boldsymbol{\Sigma}^{-1}$ is the inverse of $\boldsymbol{\Sigma}$

- If $\mathbf{X} \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then each $X_i \sim \mathcal{N}(\mu_i, \Sigma_{ii})$, i.e., the marginal distribution of each component is normal.

- Correlation coefficient between X_i and X_j is

$$\rho_{ij} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\,\Sigma_{jj}}} = \rho_{ji}$$

- Entire joint distribution of multivariate normal is uniquely determined by its
  - (normal) marginal distributions, and
  - correlation coefficients

- For other non-normal multivariate distributions, this is not true.

- Suppose $\mathbf{X} = (X_1, X_2, \ldots, X_d)^\top \in \mathbf{R}^d$ with

$$\mathbf{X} \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I}_d),$$

i.e., $\mathbf{X}$ is normal random vector with
  - mean vector $\mathbf{0}$
  - covariance matrix $\mathbf{I}_d$ is $d \times d$ identity matrix:

$$\mathbf{I}_d = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

- Then $\mathbf{X}$ is a vector of IID standard normals, i.e., each $X_i \sim \mathcal{N}(0, 1)$, and $X_1, \ldots, X_d$ are independent.

- If $\boldsymbol{\mu} \in \mathbf{R}^d$ is a constant and $\mathbf{A}$ is a $d \times d$ matrix, then

$$\boldsymbol{\mu} + \mathbf{A}\mathbf{X} \sim \mathcal{N}_d(\boldsymbol{\mu}, \mathbf{A}\mathbf{A}^\top).$$

Multivariate Lognormal Distribution

- Recall: if $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Y = e^X \sim \mathcal{LN}(\mu, \sigma^2)$.
- If $\mathbf{X} \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\mathbf{X} = (X_1, X_2, \ldots, X_d)^\top$, then
$$\mathbf{Y} = (Y_1, Y_2, \ldots, Y_d)^\top = (e^{X_1}, e^{X_2}, \ldots, e^{X_d})^\top$$
has d-dimensional lognormal $\mathcal{LN}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution.
- Is the marginal distribution of each Y_i is lognormal?
- Multivariate lognormal sometimes used to model returns of group of stocks.

- Properties of $\mathbf{Y} \sim \mathcal{LN}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
  - Mean $\mathcal{E}[Y_i] = e^{\mu_i + \Sigma_{ii}/2}$
  - Variance $\mathcal{V}[Y_i] = e^{2\mu_i + \Sigma_{ii}}(e^{\Sigma_{ii}} - 1)$
  - Covariance
$$\text{Cov}(Y_i, Y_j) = (e^{\Sigma_{ij}} - 1) \exp\left(\mu_i + \mu_j + \frac{\Sigma_{ii} + \Sigma_{jj}}{2}\right)$$
  - Correlation of Y_i and Y_j is
$$\frac{e^{\Sigma_{ij}} - 1}{\sqrt{(e^{\Sigma_{ii}} - 1)(e^{\Sigma_{jj}} - 1)}}.$$

Other Distributions

- For other probability distributions that are useful in modeling, e.g., see Chapter 6 of Law.