HANDOUT #3 - Summaries of Population Distributions

TOPICS

1. Definition of Population/Process
2. Definition of Random Variable
3. Types of Random Variables
4. Functions which Characterize Random Variables/Populations
5. Families of Distributions Indexed by Parameters
6. Examples of Distributions: Discrete, Continuous, Mixtures
7. Interrelationships between Various Distributions
8. Simulation of Data from Specified Distributions
9. Functions Associated with Reliability/Survival Analysis

Supplemental Reading:

- Chapter 2 in Tamhane-Dunlop book
- *Statistical Distributions* by Forbes, Evans, Hastings, and Peacock

Definition of Population/Process/Random Variable

1. **Statistical Population -** Collection of all possible items or units possessing one or more common characteristics under specified experimental or observational conditions

2. **Process -** Repeatable series of actions that results in an observable characteristic or measurement

   Industrial and Laboratory experiments often are characterized as hypothetical populations. Why?

   Because the population of values does not exist at the beginning of the experiment or study and in fact, may never exist.

   Example 1: Study of the effect of dehydration on ticks by placing 100 ticks to a vessel having a very dry climate. What is the hypothetical population?

   • All ticks in a very dry climate

   Example 2: Study the gain in productivity of a random sample of 25 assembly workers who complete a new training program. What is the hypothetical population?

   • All workers that will sometime in the future attend the training program - a hypothetical population

3. **Random Experiment -** Procedure or operation whose outcome is uncertain and cannot be predicted in advance

   • Expose 3 rats to a potentially toxic chemical and observe the number of survivors 24 hours later.

4. **Sample Space -** Collection of all possible outcomes of a random experiment

   Let E_i be the event rat i was alive at the end of 24 hours

   $\overline{E_i}$ is the event rat i is dead at the end of 24 hours

   $2^3 = 8$ elements in the sample sample, $\mathcal{S}$

   $\mathcal{S} = \{E_1\, E_2\, E_3;\ \overline{E_1}\, E_2\, E_3;\ E_1\, \overline{E_2}\, E_3;\ E_1\, E_2\, \overline{E_3};\ \overline{E_1}\, \overline{E_2}\, E_3;\ \overline{E_1}\, E_2\, \overline{E_3};\ E_1\, \overline{E_2}\, \overline{E_3};\ \overline{E_1}\, \overline{E_2}\, \overline{E_3}\}$

5. **Random Variable (RV) -** A function, Y, which maps sample space to the real line

   $Y : \mathcal{S} \to (-\infty, \infty)$,

   Y assigns a unique numerical value to each element of the sample space

   For each s in $\mathcal{S}$, $Y(s)$ is a real number.

   Let N = number of rats surviving then $N : \mathcal{S} \to \{0, 1, 2, 3\}$

Example 1: Randomly select a sample of water (1 liter) from a river and record the amount, A, of PCB in the container in ppb

The sample space, S, is all possible 1 liter bottles of water from the river

A: $S \Rightarrow [0, \infty)$

Example 2: Randomly select light bulb from distribution center and measure one of the following characteristics of the bulb:

The sample S is all light bulbs in the distribution center

(a) Time, T, to failure of light bulb

$T : S \Rightarrow [0, 10000]$

where 10,000 is the maximum possible life length of the bulb

(b) Amount of protective coating, C, on bulb

$C : S \Rightarrow [0, .28]$

where .28 cm is the maximum possible coating thickness

(c) Determine if bulb is defective or not, with $D = 1$ if defective and $D = 0$ if not defective

$D : S \Rightarrow \{0, 1\}$

(d) Determine Quality, Q, of bulb, with $Q = 0$ if not defective, $Q = 1$ if defective but repairable, and $Q = 2$ if defective and non-repairable

$Q : S \Rightarrow \{0, 1, 2\}$

6. **Types of Random Variables -** Three major classifications:

   (a) **Discrete RV -** Collection of possible values of RV is at most a finite or countably infinite set

   - Random variables D and Q on previous page

   (b) **Continuous RV -** Collection of possible values of RV is one or more intervals on the real line (probability that it assumes any specific value is 0)

   - Random variables T and S on previous page

   (c) **Discrete-Continuous Mixture -** Collection of possible values of RV is one or more intervals on the real line and a set of distinct values

   Example 1: Let Y be the number of fish captured in a randomly placed net in the Gulf of Mexico divided by the length of time the net is in the water, Catch Per Unit Effort (CPUE)

   - 40% of the nets have no fish, Y=0, and the remaining 60% have a value of Y in the interval (0, 500).

   Example 2: Let X be the amount spent on health insurance per member of the household by a randomly selected employee at Texas A&M University.

   - 20% of the employees have no insurance, X=0, and the remaining employees have a value of X in the interval ($1200, $5000).

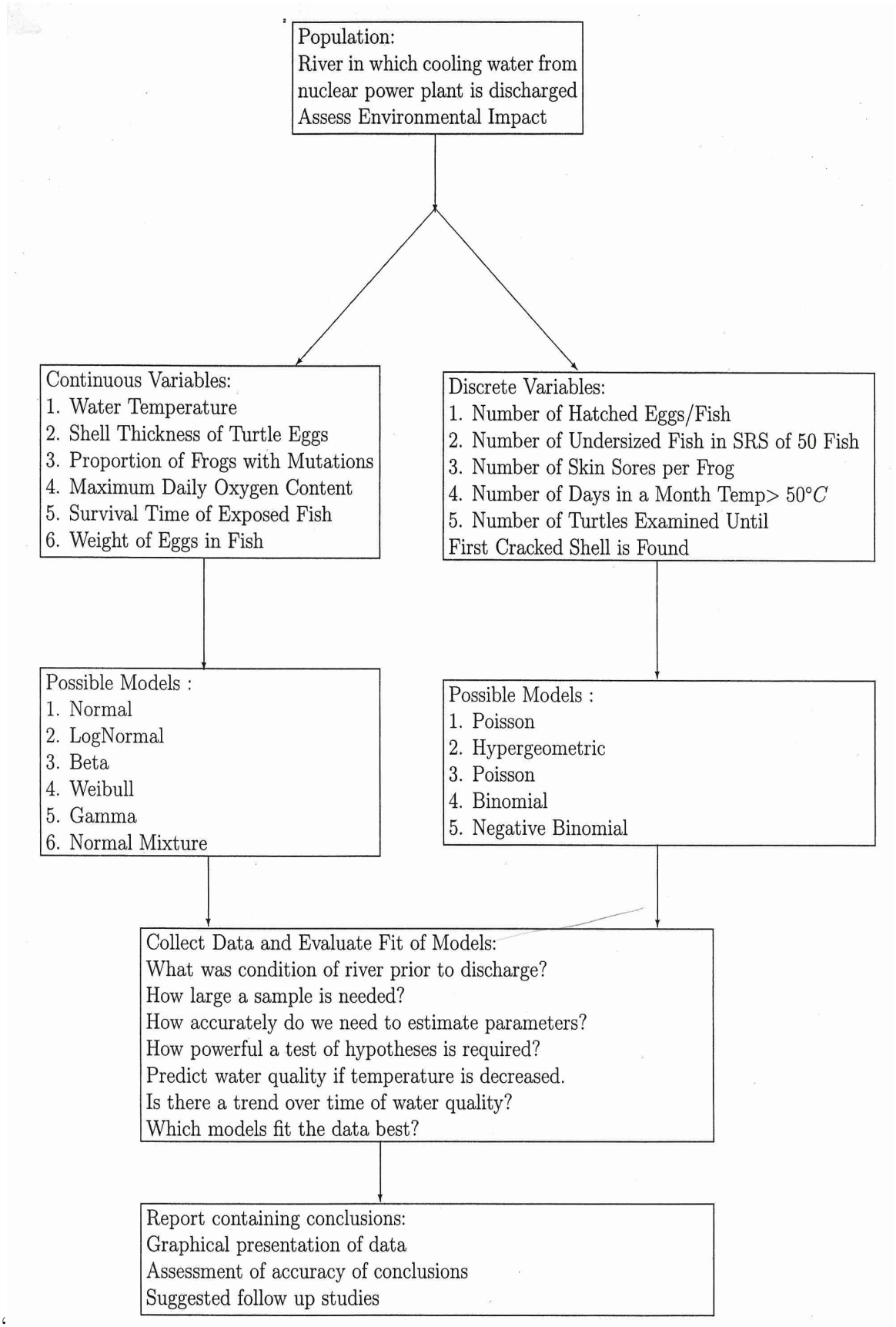
   Example 3: Let G be the growth rate of a randomly selected plant after receiving a prescribed amount of growth stimulant

   - 43% of the plants have no growth, G=0, and the remaining 57% have a value of G in (0, 28) cm.

   Example 4: Yearly payout on a health insurance policy

   - 46% of the policies have no payout, P=0, and the remaining 64% have a value of P in (0, $1,000,000)

The following diagram depicts a variety of RV's that may be defined on a single random experiment


Population:
River in which cooling water from
nuclear power plant is discharged
Assess Environmental Impact
Continuous Variables:
1. Water Temperature
2. Shell Thickness of Turtle Eggs
3. Proportion of Frogs with Mutations
4. Maximum Daily Oxygen Content
5. Survival Time of Exposed Fish
6. Weight of Eggs in Fish
Discrete Variables:
1. Number of Hatched Eggs/Fish
2. Number of Undersized Fish in SRS of 50 Fish
3. Number of Skin Sores per Frog
4. Number of Days in a Month Temp> 50°C
5. Number of Turtles Examined Until
First Cracked Shell is Found
Possible Models :
1. Normal
2. LogNormal
3. Beta
4. Weibull
5. Gamma
6. Normal Mixture
Possible Models :
1. Poisson
2. Hypergeometric
3. Poisson
4. Binomial
5. Negative Binomial
Collect Data and Evaluate Fit of Models:
What was condition of river prior to discharge?
How large a sample is needed?
How accurately do we need to estimate parameters?
How powerful a test of hypotheses is required?
Predict water quality if temperature is decreased.
Is there a trend over time of water quality?
Which models fit the data best?
Report containing conclusions:
Graphical presentation of data
Assessment of accuracy of conclusions
Suggested follow up studies

Characterization/Descriptions of Populations/Processes

Let **Y** be a R.V. associated with a Population/Process

Let **R(Y)** be the possible values of Y

Three Functions Which Completely Describe Y:

1. The Cumulative Distribution Function **(cdf)** of Y, $F(y)$:

$$F(y) = P[Y \leq y] \text{ for } -\infty < y < \infty$$

That is, $F : (-\infty, \infty) \Rightarrow [0, 1]$ F maps $(-\infty, \infty)$ into [0,1]

$F(y)$ is the probability that the next observed value of the r.v. Y is less than or equal to y

$F(y)$ is the proportion of the population having values less than or equal to y

$F(y)$ is the proportion of the output of a process having values less than or equal to y

2. The Probability Mass Function **(pmf)** for discrete r.v.'s or Probability Density Function **(pdf)** for continuous r.v.'s

   (a) For Discrete R.V.'s:

$$f(y) = P[Y = y] = \text{ proportion of population values equal to y}$$

$$\text{cdf, F is related to pmf, f, by } F(y) = P[Y \leq y] = \sum_{t \leq y} f(t)$$

   (b) For Continuous R.V.'s, the pdf is defined as that function, f, such that

$$f(y) \geq 0; \quad \text{with } F(y) = \int_{-\infty}^{y} f(t)dt \quad \Rightarrow f(y) = \frac{dF(y)}{dy}$$

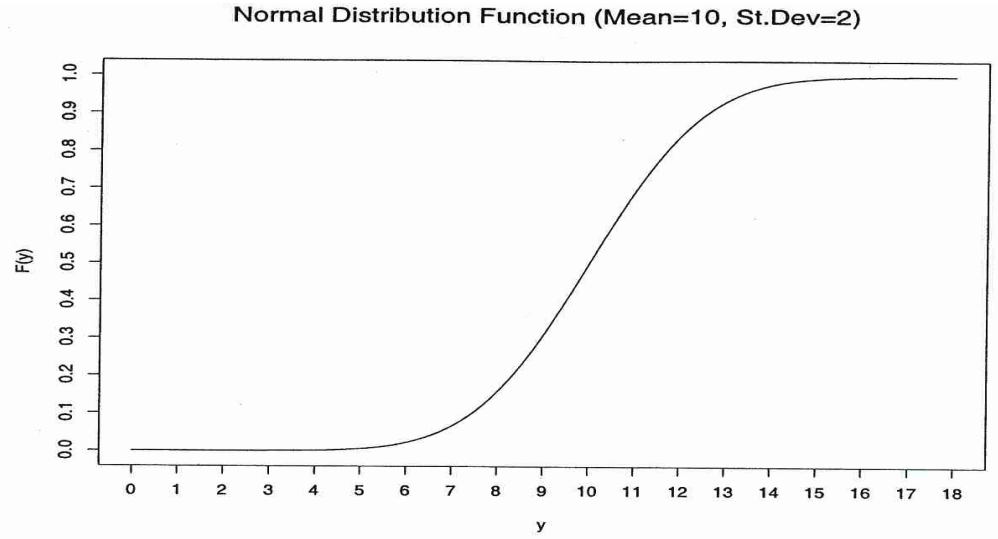
$$P[a \leq Y \leq b] = \int_{a}^{b} f(t)dt = \text{ area under} f() \text{ between a and b}$$

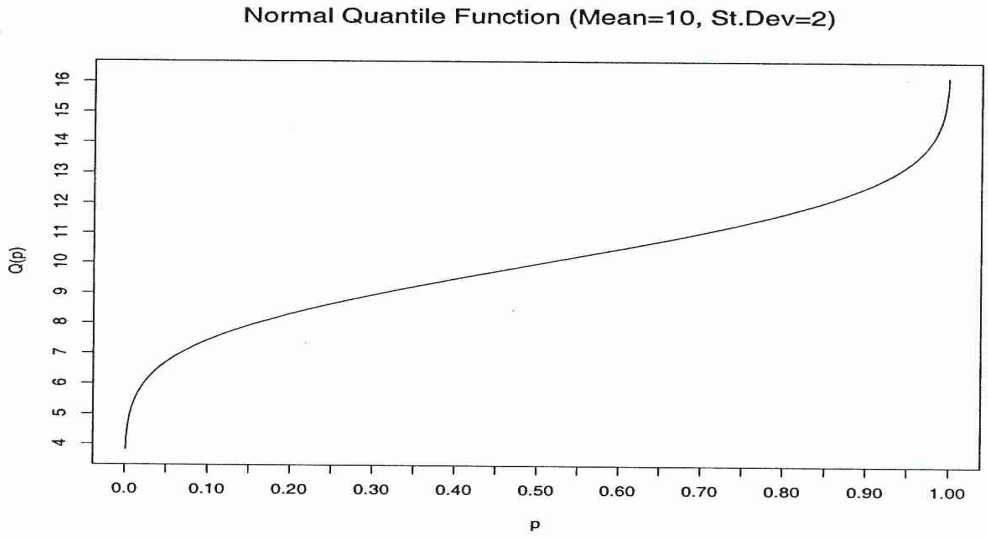
Note: $f(y)$ is the rate of increase in F at y. $f(y)$ **is not a probability**. In fact, it can have values greater than 1.0. For example, the exponential pdf with parameter $\lambda = 5$:

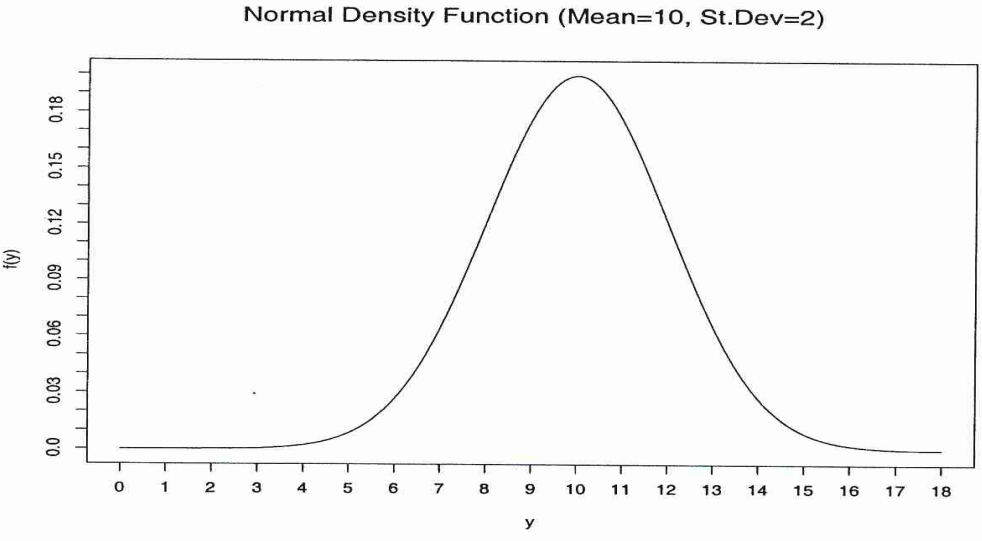
$$f(y) = 5e^{-5y} \text{ for } y \geq 0 \Rightarrow f(.04) = 5e^{-5(.04)} = 4.09 > 1$$

Using the definition, we have $f(y) = \frac{dF(y)}{dy} = \lim_{\Delta \to 0} \frac{F(y+\Delta) - F(y)}{\Delta} = \lim_{\Delta \to 0} \frac{P(Y \epsilon (y, y+\Delta))}{\Delta}$

Therefore, for very small Δ, $\Delta \cdot f(y) \approx P[Y \epsilon (y, y + \Delta)]$


Normal Distribution Function (Mean=10, St.Dev=2)
F(y)
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
y


Normal Quantile Function (Mean=10, St.Dev=2)
Q(p)
4 5 6 7 8 9 10 11 12 13 14 15 16
0.0 0.10 0.20 0.30 0.40 0.50 0.60 0.70 0.80 0.90 1.00
p


Normal Density Function (Mean=10, St.Dev=2)
f(y)
0.0 0.03 0.06 0.09 0.12 0.15 0.18
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
y

Suppose D is a discrete random variable with possible values 0, 1, 2, 3, 4, 5 and probabilities:

$$f(d) = P(D = d) = \begin{cases} .03 & if \quad d = 0 \\ .16 & if \quad d = 1 \\ .35 & if \quad d = 2 \\ .25 & if \quad d = 3 \\ .15 & if \quad d = 4 \\ .06 & if \quad d = 5 \end{cases}$$

The pmf for D is $f(d) = P(D = d)$ with values given above.

The cdf for D is obtained from the expression: $F(d) = P(D \leq d) = \sum_{i=0}^{d} f(d)$, that is,

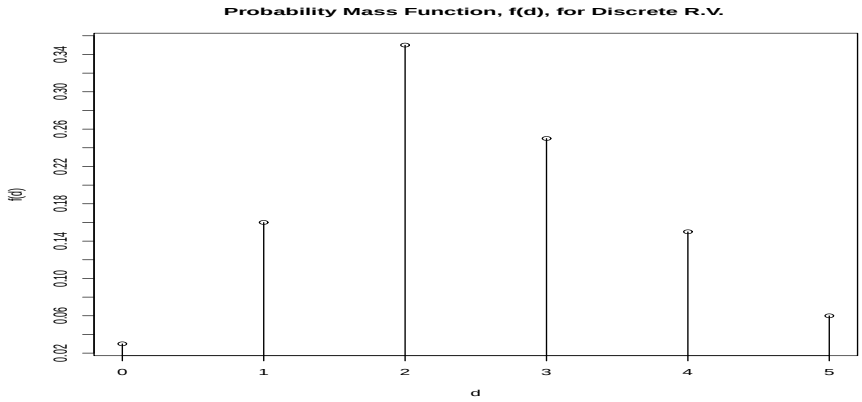
$$F(d) = P(D \leq d) = \begin{cases} 0 & if \quad d < 0 \\ .03 & if \quad 0 \leq d < 1 \\ .19 & if \quad 1 \leq d < 2 \\ .54 & if \quad 2 \leq d < 3 \\ .79 & if \quad 3 \leq d < 4 \\ .94 & if \quad 4 \leq d < 5 \\ 1 & if \quad 5 \leq d \end{cases}$$

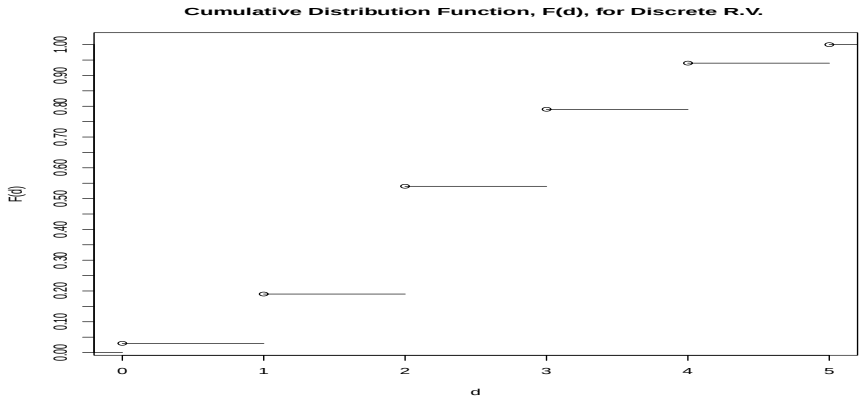
A graph of the cdf and pmf for the discrete r.v. D are given on the next page along with the quantile function, $Q(u)$ for $0 \leq u \leq 1$, which is the inverse of the cdf:

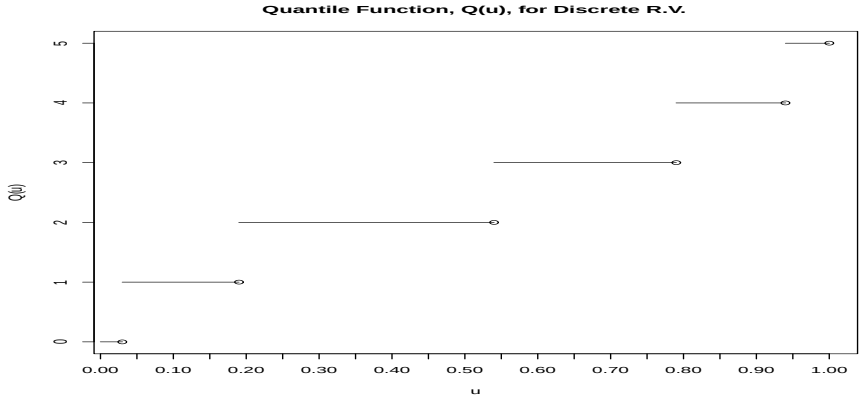
$$Q(u) = \begin{cases} 0 & if \quad 0.03 \leq u < 0.19 \\ 1 & if \quad 0.19 \leq u < 0.54 \\ 2 & if \quad 0.54 \leq u < 0.79 \\ 3 & if \quad 0.79 \leq u < 0,94 \\ 4 & if \quad 0.94 \leq u < 1 \\ 5 & if \quad 1 \leq u \end{cases}$$

Note: The pmf of D is obtained from the cdf by $f(d_j) = F(d_j) - F(d_{j-1})$.

That is, $f(3) = F(3) - F(2) = .79 - .54 = .25$


Probability Mass Function, f(d), for Discrete R.V.
f(d)
0.02
0.06
0.10
0.14
0.18
0.22
0.26
0.30
0.34
0
1
2
3
4
5
d

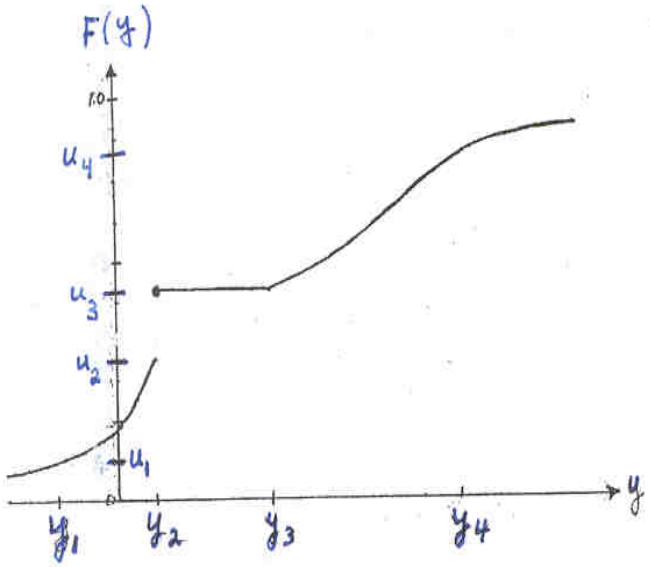

Cumulative Distribution Function, F(d), for Discrete R.V.
F(d)
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
0
1
2
3
4
5
d


Quantile Function, Q(u), for Discrete R.V.
Q(u)
0
1
2
3
4
5
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
u

Discrete - Continuous Mixture (point mass at $Y = y_2$)

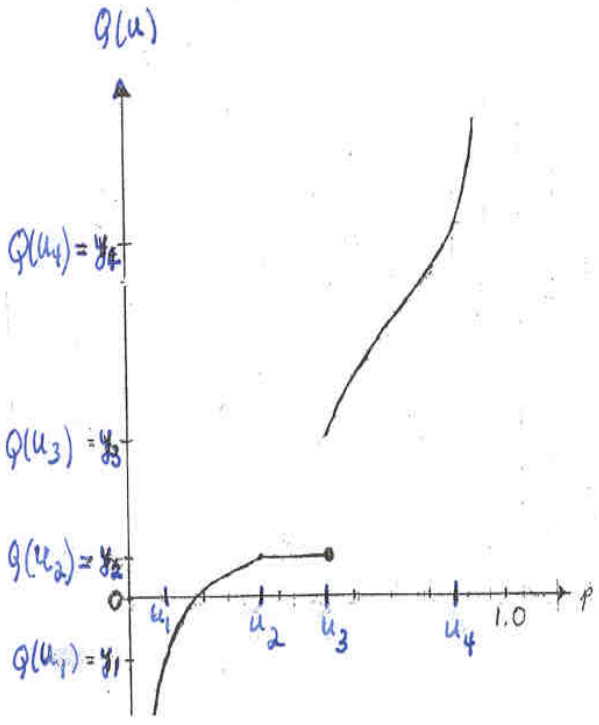
cdf

$F(y)$



Quantile

$Q(u)$



3. The **Quantile Function** of Y, $Q(u)$:

   Definition 1: Inverse of the cdf: $Q(u) = F^{-1}(u)$

   $Q : [0, 1] \to (-\infty, \infty)$

   - Case 1: For a continuous, strictly increasing cdf, $F(\cdot)$

     $$Q(u) = y_u \quad \text{if and only if} \quad F(y_u) = u$$

   - Case 2: For a discrete or discrete-continuous mixture r.v. the inverse of the cdf may not be defined for one of the following reasons:
     a. For specified $u\epsilon[0, 1]$ there may exists many real numbers y_u for which $F(y_u) = u$

     b. For specified $u\epsilon[0, 1]$ there is no real number y_u for which $F(y_u) = u$

   In either case, the inverse of F would not be a valid function because it violates the definition of a function which states every value in the domain is assigned to one and only one value in the range of the function.

   For example, see discrete-continuous mixture graphs on previous page:

   1. For any value y satisfying $y_2 \leq y \leq y_3$ $F(y) = u_3$. Therefore, by the definition in Case 1,

      $$F(y_2) = u_3 \quad \text{and} \quad F(y_3) = u_3, \quad \text{in fact, for all } \; y \in [y_2,\ y_3], \quad F(y) = u_3$$

      Thus, by the definition of an inverse function,

      $$Q(u_3) = y \quad \text{for all } \; y \in [y_2,\ y_3]$$

      But this violates the definition of a function (a given value in the domain is mapped to multiple distinct values.)

   2. For all u satisfying $u_2 < u < u_3$, there is no real number y_u for which $F(y_u) = u$.

      That is, there exists values in the domain which are not mapped by the function. Once again we have violated the definition of a function.

Thus, we have the following **Alternative Definitions:**

Definition 2: For $u\epsilon(0,1)$, the $100u$-quantile of the r.v. Y (or cdf F) is the real number $y_u = Q(u)$ such that

$$Q(u) = y_u = inf\{y : F(y) \geq u\}$$

That is, $Q(u)$ is the smallest value of y for which $F(y) \geq u$.

Special Case: If the r.v. Y is bounded below, that is, $Y \geq a$, then we define $Q(0) = a$.
Note: If we did not make the above specification, then Q(0) would be undefined because
$F(y) = 0$ if $y < a$ thus $inf\{y : F(y) \geq 0\} = -\infty$ but $Y \geq a > -\infty$
Thus, we have the following: if $-\infty < a \leq Y \leq b < \infty$ then $Q(0) = a$ and $Q(1) = b$

Note that in our example of a discrete-continuous mixture distribution,

1. $Q(u_3) = y_2$ (y_2 is the smallest value of y for which $F(y) \geq u_3$)
2. For all u satisfying $u_2 < u < u_3$, $\quad Q(u) = y_2$

   (For u satisfying $u_2 < u < u_3$, $F(y_2) = u_3 > u$ and $F(y) < u$ for all $y < y_2$).

Note, the graph of $(u, Q(u))$ is a rotation of the mirror image of the graph of $(y, F(y))$:

- Jumps in the cdf F become flat regions in Q
- Flat regions in F become jumps in Q.

Remark: For $u\epsilon(0,1)$, if y_u is the $100u$ quantile of the r.v. Y or cdf F then

1. $P[Y \leq y_u] \geq u$ AND $P[Y \geq y_u] \geq 1 - u$

   or in terms of the cdf F

2. $F(y_u) \geq u$ AND $F(y_u^-) \leq u$

That is, $y_u = Q(u)$ is that value of Y such that at least $100u\%$ of the population values are less than or equal to y_u and that value of Y such that at least $100(1-u)\%$ of the population values are greater than or equal to y_u.

For distributions having cdf F strictly increasing on the support of the corresponding pdf and continuous, we can determine $Q(u)$ from $F(y)$ using the relationship:

$$y_u = Q(u) \text{ if and only if } F(y_u) = u.$$

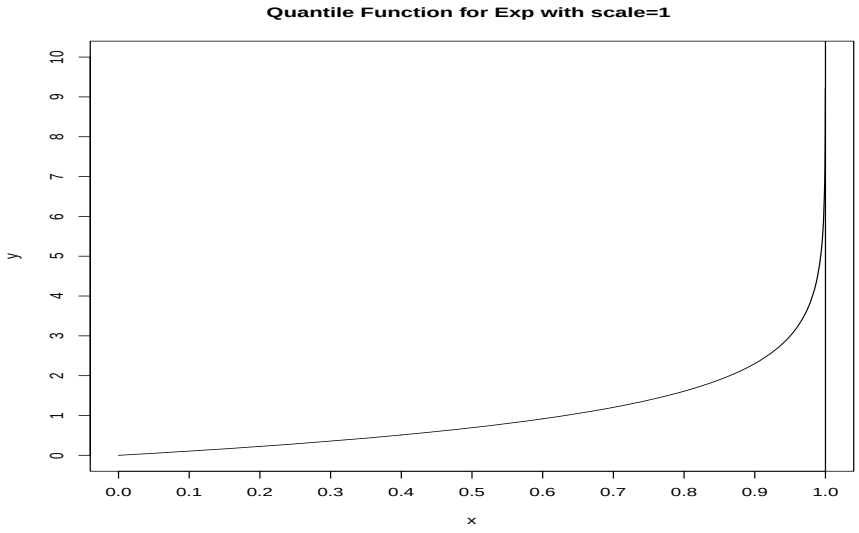
For example, suppose the random variable Y has cdf given by the following:

$$F(y) = \begin{cases} 1 - e^{-y} & \text{if } y \geq 0 \\ 0 & \text{if } y < 0 \end{cases}$$

For $u > 0$, $u = F(y_u) = 1 - e^{-y_u} \Rightarrow log(1-u) = -y_u \Rightarrow$

$Q(u) = y_u = -log(1-u)$

Because $Y \geq 0$ we set $Q(u) = 0$ for $u = 0$



For further reading on the topic of quantile functions see Casella-Berger, *Statistical Inference* or John Rice, *Mathematical Statistics and Data Analysis*.

Location/Scale Families of CDF's

In many cases, the cdf or pdf of a r.v. Y is specified as a member of a family of distributions which are indexed by parameters:

$$Y \quad \text{has a pdf in the family} \quad \{f(y;\theta) : \theta \epsilon \Theta\}$$

The following examples will illustrate this notation:

Example 1 Y has pdf given by, for $-\infty < y < \infty$

$$f(y,\theta_1,\theta_2) = \frac{1}{\sqrt{2\pi}\sqrt{\theta_2}} e^{-\frac{1}{2\theta_2}(y-\theta_1)^2} \quad \text{for} \quad \theta \epsilon \Theta = \{(\theta_1,\theta_2) : \theta_1 \epsilon (-\infty,\infty), \theta_2 \epsilon (0,\infty)\}$$

Example 2 Y has pmf given by, for $y = 0, 1, \ldots, n$

$$f(y;\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y} \quad \text{for} \quad \theta \epsilon \Theta = [0,1]$$

Example 3 Y has pdf given by, for $0 < y < \infty$

$$f(y,\theta_1,\theta_2) = \frac{1}{\Gamma(\theta_1)\theta_2^{\theta_1}} y^{\theta_1 - 1} e^{-y/\theta_2} \quad \text{for} \quad \theta \epsilon \Theta = \{(\theta_1,\theta_2) : \theta_1, \theta_2 > 0\}$$

Note: $\Gamma(c) = \int_0^\infty y^{c-1} e^{-y} dy$ is referred to as the gamma function

The following are some special cases of family of distributions:

1. The parameter θ in a family of pdf's for the r.v. Y, $\{f_Y(y;\theta) : \theta \epsilon \Theta\}$ is said to be a **Location Parameter** if the distribution of $W = Y - \theta$ does not depend on θ, that is, if the pdf of W, $\quad f_W(w) = f_Y(w + \theta)$ does not depend on θ.

   - W is referred to as the **Standard** member of a location family if $\theta = 0$.

2. The parameter θ in a family of pdf's for the r.v. Y, $\{f_Y(y;\theta) : \theta \epsilon \Theta\}$ is said to be a **Scale Parameter** if the distribution of $W = Y/\theta$ does not depend on θ, that is, if the pdf of W $f_W(w) = \theta f_Y(\theta w)$ does not depend on θ.

   - W is referred to as the **Standard** member of a scale family if $\theta = 1$.

3. The parameters θ_1 and θ_2 in a family of pdf's for the r.v. Y, $\{f_Y(y;\theta_1,\theta_2) : \theta \epsilon \Theta\}$ are said to be a **Location-Scale Parameters** if the distribution of $W = (Y - \theta_1)/\theta_2$ does not depend on θ_1 nor θ_2, that is, if the pdf of $W \quad f_W(w) = \theta_2 f_Y(\theta_2 w + \theta_1)$ does not depend on θ_1 nor θ_2.

   - W is referred to as the **Standard** member of a location-scale family if $\theta_1 = 0$ and $\theta_2 = 1$.

The following examples will illustrate these types of families:

Example 1. Let Y have a $N(\theta, 1)$ distribution.

Then θ is a location parameter as demonstrated by

$$f(y, \theta, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\theta)^2} \quad \text{for} \quad \theta \epsilon (-\infty, \infty)$$

Let $W = Y - \theta \quad \Rightarrow \quad f_W(w) = f_Y(w + \theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[(w+\theta)-\theta]^2} = \frac{1}{\sqrt{2\pi}} e^{-.5[w]^2}$

The pdf of W does not depend on θ.

W is the standard member of the normal family of distributions, that is, W has a normal distribution with parameter values 0 and 1, that is, $N(0,\ 1)$.

Example 2. Let Y have an Exponential Distribution with parameter λ, that is,

$$f(y; \lambda) = \begin{cases} \frac{1}{\lambda} e^{-y/\lambda} & if \quad y \geq 0 \\ 0 & if \quad y < 0 \end{cases}$$

Let $W = Y/\lambda$ then the pdf of W is given by

$$f_W(w) = \lambda f(\lambda w) = \lambda \left(\frac{1}{\lambda} e^{-(\lambda w)/\lambda} \right) = e^{-w}$$

The pdf of W does not depend on λ.

Thus, λ is a scale parameter and W has an exponential distribution with $\lambda = 1$.

W is the standard member of the exponential family.

Example 3. Let Y have a $N(6, \theta^2)$ distribution. Is θ a scale parameter in this distribution?

$$f(y, 6, \theta) = \frac{1}{\sqrt{2\pi}\theta} e^{-\frac{1}{2\theta^2}(y-6)^2} \quad \text{for} \quad \theta \epsilon (0, \infty)\}$$

Let W $= Y/\theta$ then the pdf of W is given by

$$f_W(w) = \theta f(\theta w) = \theta \left(\frac{1}{\sqrt{2\pi}\theta} e^{-\frac{1}{2\theta^2}[(\theta w)-6]^2} \right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2\theta^2}[\theta w - 6]^2} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2\theta^2}[\theta^2 w^2 - 12\theta w + 36]}$$

Thus, the pdf of W depends on θ and therefore θ is not a scale parameter. It would be a shape parameter.

If Y had a $N(0,\ \theta^2)$ distribution, would θ be a scale parameter in this distribution?

Example 4. Let Y have a $N(\theta_1, \theta_2^2)$ distribution. Are (θ_1, θ_2) location- scale parameters in this distribution?

$$f(y, \theta_1, \theta_2) = \frac{1}{\sqrt{2\pi}\theta_2} e^{-\frac{1}{2\theta_2^2}(y-\theta_1)^2} \quad \text{for} \quad \theta \epsilon \Theta = \{(\theta_1, \theta_2) : \theta_1 \epsilon (-\infty, \infty), \theta_2 \epsilon (0, \infty)\}$$

Let W $= \frac{Y-\theta_1}{\theta_2}$ then the pdf of W is given by

$$f_W(w) = \theta_2 f(\theta_2 w + \theta_1) = \theta_2 \frac{1}{\sqrt{2\pi}\theta_2} e^{-\frac{1}{2\theta_2^2}[(\theta_2 w + \theta_1) - \theta_1]^2} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2\theta^2}[\theta_2^2 w^2]} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[w^2]}$$

The pdf of W does not depend on θ_1 nor θ_2. Therefore, θ_1 and θ_2 are location-scale parameters for the family of distributions for Y.

Is the location parameter of a distribution equal to the mean of the distribution?

Is the scale parameter of a distribution equal to the standard deviation of the distribution?

The answer is both yes and no:

1. For the normal distribution the location parameter is the mean and scale parameter is the standard deviation.
2. For the Cauchy distribution, the mean and standard deviation do not exist but the Cauchy has both a location and scale paramenter.
3. For the double exponential distribution, the location parameter is the mean but the standard deviation is $\sqrt{2}$ times the scale parameter
4. For the 3-parameter Weibull distribution, both the mean and standard deviation are functions of all three paramenters.

Why is it important to be able to designate a family of distributions as being a location/scale family?

A few important reasons are given next:

1. If $F(y, \theta_1, \theta_2)$ is a cdf with location/scale parameters (θ_1, θ_2), then to tabulate the probability distribution we only need a table for the standard member of the family, F^*, where F^* has $\theta_1 = 0, \theta_2 = 1$. For any other member of the family we can find its probabilities by using the following expression:

   Let Y be a random variable with cdf F a member of the family and Y^* be a random variable with cdf F^*, then

   $$Y^* = \frac{Y - \theta_1}{\theta_2} \quad \text{"equals in distribution"}$$

   $$F(y) = P[Y \leq y] = P\left[Y^* \leq \frac{y - \theta_1}{\theta_2}\right] = F^*\left(\frac{y - \theta_1}{\theta_2}\right)$$

   That is to find a probability associated with Y just look up the standardized value in the Y^* table.

   This is what is done with the normal distribution.

   Example: Suppose we have a normal distribution with $\mu = 5, \sigma = 2.3$ and we want to know what proportion of the population has values less than 1.7:

   $P[Y \leq 1.7] = P\left[Y^* \leq \frac{1.7-5}{2.3}\right] = F^*(-1.435) = pnorm(-1.435) = 0.0756$ using the R function, **pnorm**.

2. Similarly, we can determine a percentile for the general member of the family using the percentiles from the standard member of the family:

   Let $Q_Y(u)$ be the quantile function for Y which has location-scale parameters (θ_1, θ_2) and let $Q_{Y^*}(u)$ be the quantile function for the standard member of the family. Then,

   $$Q_Y(u) = \theta_1 + \theta_2 Q_{Y^*}(u)$$

   Example: To find the 95th percentile of a normal distribution with $\mu = 5, \sigma = 2.3$. Look up the 95th percentile in the standard normal table, $Q_{Y^*}(.95) = 1.645 = qnorm(.95)$ and then compute

   $Q_Y(.95) = \theta_1 + \theta_2 Q_{Y^*}(.95) = 5 + (2.3)(1.645) = 8.7835$

Alternatively, use the R function, **qnorm**, to obtain qnorm(.95,5,2.3) = 8.783163

There are many examples of distributions having parameters which are neither location nor scale parameters. The following tables (Cassela & Berger) and figures will illustrate such distributions:

Table of Common Distributions

Discrete Distributions

Bernoulli(p)

pmf	$P(X=x\|p) = p^x(1-p)^{1-x}; \quad x=0,1; \quad 0 \le p \le 1$
mean and variance	$\mathrm{E}X = p, \quad \mathrm{Var}\,X = p(1-p)$
mgf	$M_X(t) = (1-p) + pe^t$

Binomial(n,p)

pmf	$P(X=x\|n,p) = \binom{n}{x}p^x(1-p)^{n-x}; \quad x=0,1,2,\ldots,n; \quad 0 \le p \le 1$
mean and variance	$\mathrm{E}X = np, \quad \mathrm{Var}\,X = np(1-p)$
mgf	$M_X(t) = [pe^t + (1-p)]^n$
notes	Related to Binomial Theorem (Theorem 3.2.2). The *multinomial* distribution (Definition 4.6.2) is a multivariate version of the binomial distribution.

Discrete uniform

pmf	$P(X=x\|N) = \frac{1}{N}; \quad x=1,2,\ldots,N; \quad N=1,2,\ldots$
mean and variance	$\mathrm{E}X = \frac{N+1}{2}, \quad \mathrm{Var}\,X = \frac{(N+1)(N-1)}{12}$
mgf	$M_X(t) = \frac{1}{N}\sum_{i=1}^{N} e^{it}$

Geometric(p)

pmf	$P(X=x\|p) = p(1-p)^{x-1}; \quad x=1,2,\ldots; \quad 0 \le p \le 1$
mean and variance	$\mathrm{E}X = \frac{1}{p}, \quad \mathrm{Var}\,X = \frac{1-p}{p^2}$

mgf $M_X(t) = \frac{pe^t}{1-(1-p)e^t}, \quad t < -\log(1-p)$

notes $Y = X - 1$ is negative binomial$(1, p)$. The distribution is *memoryless*: $P(X > s|X > t) = P(X > s - t)$.

Hypergeometric

pmf $P(X = x|N, M, K) = \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}; \quad x = 0, 1, 2, \ldots, K;$
$M - (N - K) \leq x \leq M; \quad N, M, K \geq 0$

mean and variance $\mathrm{E}X = \frac{KM}{N}, \quad \mathrm{Var}\, X = \frac{KM}{N}\frac{(N-M)(N-K)}{N(N-1)}$

notes If $K \ll M$ and N, the range $x = 0, 1, 2, \ldots, K$ will be appropriate.

Negative binomial(r, p)

pmf $P(X = x|r, p) = \binom{r+x-1}{x}p^r(1-p)^x; \quad x = 0, 1, \ldots; \quad 0 \leq p \leq 1$

mean and variance $\mathrm{E}X = \frac{r(1-p)}{p}, \quad \mathrm{Var}\, X = \frac{r(1-p)}{p^2}$

mgf $M_X(t) = \left(\frac{p}{1-(1-p)e^t}\right)^r, \quad t < -\log(1-p)$

notes An alternate form of the pmf is given by $P(Y = y|r, p) = \binom{y-1}{r-1}p^r(1-p)^{y-r}$, $y = r, r+1, \ldots$. The random variable $Y = X + r$. The negative binomial can be derived as a gamma mixture of Poissons. (See Exercise 4.32.)

Poisson(λ)

pmf $P(X = x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}; \quad x = 0, 1, \ldots; \quad 0 \leq \lambda < \infty$

mean and variance $\mathrm{E}X = \lambda, \quad \mathrm{Var}\, X = \lambda$

mgf $M_X(t) = e^{\lambda(e^t-1)}$

Discrete Distributions - Examples

1. **Discrete Uniform R.V.** - Possible values: $1, 2, \ldots, N$ with equal probability, $\frac{1}{N}$

- Used in computing the odds in sporting events and card games
- During WWII, the Allied forces wanted to estimate the number of tanks placed in combat by the Germans. Because the tanks were consecutively numbered, the tank numbers in any given area formed a discrete uniform distribution. The Allied forces had a sample of such numbers from which they could estimate the total number of German tanks in a given region.

2. **Bernoulli R.V.** - Possible values: 0 (failure) and 1 (success), with probabilities $1 - p$ and p

- Inspector determines if part is defective or good
- Patient has the disease or not
- Gene is mutated or not

3. **Binomial R.V.** - Possible values: $0, 1, 2, \ldots, n$ with probabilities:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \text{ for } k = 0, 1, 2, \ldots, n$$

- The number of successes, S, in a fixed number, n, of independent Bernoulli Trials with identical probabilities, p for each trial. S is distributed $B(n, p)$
- A circuit board has 25 individual components with a .01% probability of an individual component being defective and the events that components are defective being independent events. Let C be the number of defective components on a randomly selected circuit board.
- A veterinarian inspects 50 deer for ticks. Let T be the number of deer having ticks.

4. **Binomial R.V.** - The number of Type A items, X, in a random sample of n units selected **with replacement** from a population containing N units consisting of two types of Units - Type A and Type B (not Type A).

- Let M be the number of Type A units in the population. Then $p = M/N$ in the Binomial pmf
- Sampling with replacement does not occur very often in practice but statisticians use sampling with replacement in approximating sampling distributions.

5. **Negative Binomial R.V.** - The number of trials, B, until the rth success in series of independent identically distributed (i.i.d.) Bernoulli trials

   Possible values: $r, r+1, r+2, \cdots$ with probabilities: $P(B=k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}$, for $k = r, r+1, \ldots$

- Let D be the number of patients examined until the 5th patient with a rare eye disease is found

- In a series of iid Bernoulli trials, the Binomial distribution has a fixed number of trials and a random number of successes

- In a series of iid Bernoulli trials, the Negative Binomial distribution has a fixed number of successes and a random number of trials

6. **Geometric R.V.** - The number of trials until the first success in series of i.i.d. Bernoulli trials

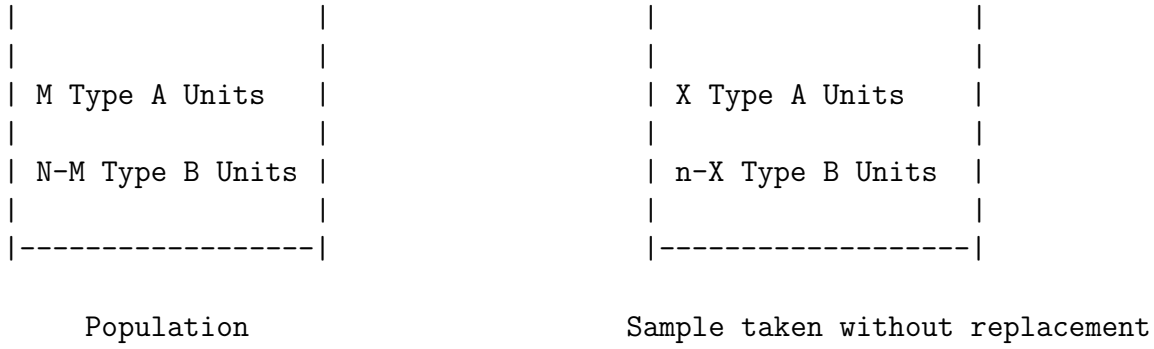
- Special case of Negative Binomial with $r = 1$

7. **Hypergeometric R.V.** - The number of Type A items, X in a random sample of n units selected **without replacement** from a population containing N units consisting of M Type A Units and $N - M$ Type B Units.

- Possible values: $k = 0, 1, 2, \ldots, n$ provided $M - (N - n) \leq k \leq M$ with probabilities:

  $P(X = k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$ for $k = 0, 1, 2, \ldots, n$ and $M - (N - n) \leq k \leq M$

- Let X be the number of Type A units in a random sample of n units from the population of N units, sampling without replacement. The possible values of X are the integers between $max(0, M - (N - n))$ and $min(n, M)$



8. **Poisson R.V.** - The number of event occurrences, Y during a specified period of time or space of a Poisson process

   Poisson Postulates: The events occur according the following set of conditions:

   P1. The probability that exactly one event occurs in a very short period of time of length Δt is $\lambda \cdot \Delta t + o(\Delta t)$, where $\lambda > 0$ is a fixed constant

   P2. The probability of more than one event occurring during Δt is $o(\Delta t)$

   P3. The number of events occurring during the time interval Δt is independent of the number occurring in any other time interval.

- A more mathematical formulation of the Poisson postulates is given in the Casela-Berger book.
- λ is the average number of event occurrences in a unit period of time
- The Poisson postulates can be formulated in space as well as time where Δt is a unit of space in place of a unit of time.
- - Possible values: $0, 1, 2, \ldots$ with probabilities:

  $P(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!}$, for $y = 0, 1, 2, \ldots$

- Suppose the number of requests for assistance arrive at a computer service desk according to the Poisson postulates with $\lambda = 5$ per second. Let S be the number of requests in a given hour.
- Let F be the number of persons who are killed in plane accidents during a given year
- Let P be the number of lead particles in a square inch of painted wall in an old building
- The Poisson distribution is the limit of a Binomial distribution as n approaches infinity and p_n approaches 0 in such a manner that np_n approaches λ in the limit.
- Both the negative binomial and Poisson distributions are used to model R.V.'s in studies which do not exactly follow the about definitions: For example, we may randomly select plants in a field treated with an insecticide and count the number of insects on each of the plants. In a future handout, we will discuss how to measure how well a given distribution fits a given data set. *No model is correct but many are useful.* Quote from Dr. George Box.

The interrelationships between the various discrete distributions are given in the following display.

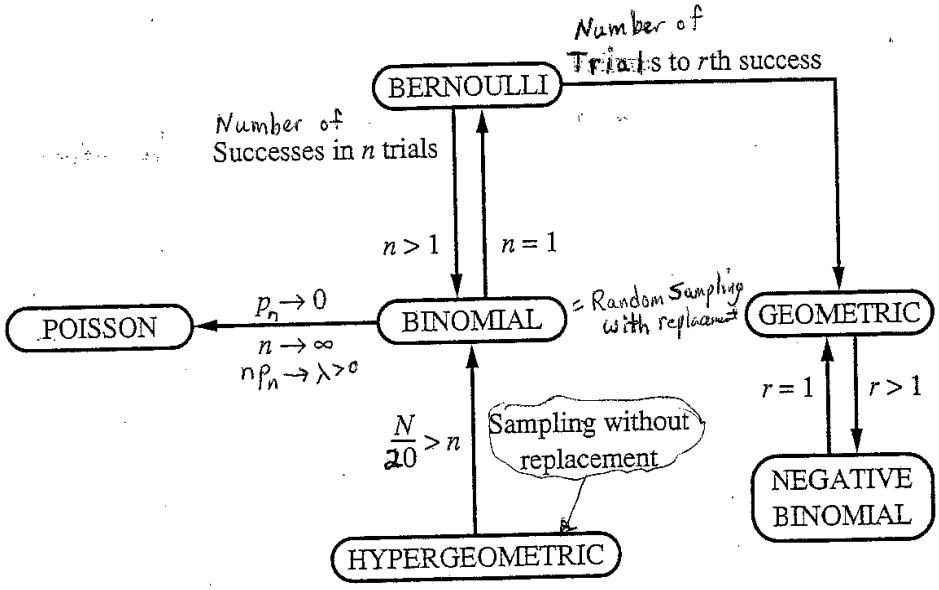


FIGURE
Relationships between the distributions
Statistical Analysis for Engineers/Scientists
J. Wesley Barnes

Examples of Continuous Distributions

Beta(α, β)

pdf	$f(x\|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 \le x \le 1, \quad \alpha > 0, \quad \beta > 0$
mean and variance	$\mathrm{E}X = \frac{\alpha}{\alpha+\beta}, \quad \mathrm{Var}\, X = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
mgf	$M_X(t) = 1 + \sum_{k=1}^{\infty} \left(\prod_{r=0}^{k-1} \frac{\alpha+r}{\alpha+\beta+r} \right) \frac{t^k}{k!}$
notes	The constant in the beta pdf can be defined in terms of gamma functions, $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$. Equation (3.2.18) gives a general expression for the moments.

Cauchy(θ, σ)

pdf	$f(x\|\theta,\sigma) = \frac{1}{\pi\sigma} \frac{1}{1+\left(\frac{x-\theta}{\sigma}\right)^2}, \quad -\infty < x < \infty; \quad -\infty < \theta < \infty, \quad \sigma > 0$
mean and variance	do not exist
mgf	does not exist
notes	Special case of Student's t, when degrees of freedom $= 1$. Also, if X and Y are independent n$(0,1)$, X/Y is Cauchy.

Chi squared(p)

pdf	$f(x\|p) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{(p/2)-1} e^{-x/2}; \quad 0 \le x < \infty; \quad p = 1, 2, \ldots$
mean and variance	$\mathrm{E}X = p, \quad \mathrm{Var}\, X = 2p$
mgf	$M_X(t) = \left(\frac{1}{1-2t}\right)^{p/2}, \quad t < \frac{1}{2}$
notes	Special case of the gamma distribution.

Double exponential(μ, σ)

pdf	$f(x\|\mu,\sigma) = \frac{1}{2\sigma} e^{-\|x-\mu\|/\sigma}, \quad -\infty < x < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0$
mean and variance	$\mathrm{E}X = \mu, \quad \mathrm{Var}\, X = 2\sigma^2$
mgf	$M_X(t) = \frac{e^{\mu t}}{1-(\sigma t)^2}, \quad \|t\| < \frac{1}{\sigma}$
notes	Also known as the *Laplace* distribution.

Exponential(β)

pdf $f(x|\beta) = \frac{1}{\beta}e^{-x/\beta}, \quad 0 \le x < \infty, \quad \beta > 0$

mean and variance $\mathrm{E}X = \beta, \quad \mathrm{Var}\, X = \beta^2$

mgf $M_X(t) = \frac{1}{1-\beta t}, \quad t < \frac{1}{\beta}$

notes Special case of the gamma distribution. Has the *memoryless* property. Has many special cases: $Y = X^{1/\gamma}$ is *Weibull*, $Y = \sqrt{2X/\beta}$ is *Rayleigh*, $Y = \alpha - \gamma \log(X/\beta)$ is *Gumbel*.

F

pdf $f(x|\nu_1, \nu_2) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} \frac{x^{(\nu_1-2)/2}}{\left(1+\left(\frac{\nu_1}{\nu_2}\right)x\right)^{(\nu_1+\nu_2)/2}};$
$0 \le x < \infty; \quad \nu_1, \nu_2 = 1, \ldots$

mean and variance $\mathrm{E}X = \frac{\nu_2}{\nu_2-2}, \quad \nu_2 > 2,$

$\mathrm{Var}\, X = 2\left(\frac{\nu_2}{\nu_2-2}\right)^2 \frac{(\nu_1+\nu_2-2)}{\nu_1(\nu_2-4)}, \quad \nu_2 > 4$

moments (mgf does not exist) $\mathrm{E}X^n = \frac{\Gamma\left(\frac{\nu_1+2n}{2}\right)\Gamma\left(\frac{\nu_2-2n}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \left(\frac{\nu_2}{\nu_1}\right)^n, \quad n < \frac{\nu_2}{2}$

notes Related to chi squared ($F_{\nu_1,\nu_2} = \left(\frac{\chi^2_{\nu_1}}{\nu_1}\right) / \left(\frac{\chi^2_{\nu_2}}{\nu_2}\right)$, where the χ^2s are independent) and t ($F_{1,\nu} = t^2_\nu$).

Gamma(α, β)

pdf $f(x|\alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1}e^{-x/\beta}, \quad 0 \le x < \infty, \quad \alpha, \beta > 0$

mean and variance $\mathrm{E}X = \alpha\beta, \quad \mathrm{Var}\, X = \alpha\beta^2$

mgf $M_X(t) = \left(\frac{1}{1-\beta t}\right)^\alpha, \quad t < \frac{1}{\beta}$

notes Some special cases are exponential ($\alpha = 1$) and chi squared ($\alpha = p/2$, $\beta = 2$). If $\alpha = \frac{3}{2}$, $Y = \sqrt{X/\beta}$ is *Maxwell*. $Y = 1/X$ has the *inverted gamma distribution*. Can also be related to the Poisson (Example 3.2.1).

Logistic(μ, β)

pdf $f(x|\mu, \beta) = \frac{1}{\beta} \frac{e^{-(x-\mu)/\beta}}{[1+e^{-(x-\mu)/\beta}]^2}, \quad -\infty < x < \infty, \quad -\infty < \mu < \infty, \quad \beta > 0$

mean and variance $\mathrm{E}X = \mu, \quad \mathrm{Var}\, X = \frac{\pi^2\beta^2}{3}$

mgf $M_X(t) = e^{\mu t}\Gamma(1-\beta t)\Gamma(1+\beta t), \quad |t| < \frac{1}{\beta}$

notes The cdf is given by $F(x|\mu,\beta) = \frac{1}{1+e^{-(x-\mu)/\beta}}$.

Lognormal(μ, σ^2)

pdf $f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\frac{e^{-(\log x-\mu)^2/(2\sigma^2)}}{x}, \quad 0 \le x < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0$

mean and variance $\mathrm{E}X = e^{\mu+(\sigma^2/2)}, \quad \mathrm{Var}\, X = e^{2(\mu+\sigma^2)} - e^{2\mu+\sigma^2}$

moments (mgf does not exist) $\mathrm{E}X^n = e^{n\mu+n^2\sigma^2/2}$

notes Example 2.3.5 gives another distribution with the same moments.

Normal(μ, σ^2)

pdf $f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/(2\sigma^2)}, \quad -\infty < x < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0$

mean and variance $\mathrm{E}X = \mu, \quad \mathrm{Var}\, X = \sigma^2$

mgf $M_X(t) = e^{\mu t+\sigma^2 t^2/2}$

notes Sometimes called the *Gaussian* distribution.

Pareto(α, β)

pdf $f(x|\alpha,\beta) = \frac{\beta\alpha^\beta}{x^{\beta+1}}, \quad a < x < \infty, \quad \alpha > 0, \quad \beta > 0$

mean and variance $\mathrm{E}X = \frac{\beta\alpha}{\beta-1}, \quad \beta > 1, \quad \mathrm{Var}\, X = \frac{\beta\alpha^2}{(\beta-1)^2(\beta-2)}, \quad \beta > 2$

mgf does not exist

---

t

pdf $f(x|\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\frac{1}{\sqrt{\nu\pi}}\frac{1}{\left(1+\left(\frac{x^2}{\nu}\right)\right)^{(\nu+1)/2}}, \quad -\infty < x < \infty, \quad \nu = 1, \ldots$

mean and variance $\mathrm{E}X = 0, \quad \nu > 1, \quad \mathrm{Var}\, X = \frac{\nu}{\nu-2}, \quad \nu > 2$

moments (mgf does not exist) $\mathrm{E}X^n = \frac{\Gamma\left(\frac{n+1}{2}\right)\Gamma\left(\frac{\nu-n}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{\nu}{2}\right)}\nu^{n/2}$ if $n < \nu$ and even,
$\mathrm{E}X^n = 0$ if $n < \nu$ and odd.

notes Related to F ($F_{1,\nu} = t_\nu^2$).

Uniform(a, b)

pdf	$f(x\|a,b) = \frac{1}{b-a}, \quad a \le x \le b$
mean and variance	$\mathrm{E}X = \frac{b+a}{2}, \quad \mathrm{Var}\,X = \frac{(b-a)^2}{12}$
mgf	$M_X(t) = \frac{e^{bt}-e^{at}}{(b-a)t}$
notes	If $a = 0$ and $b = 1$, this is a special case of the beta ($\alpha = \beta = 1$).

Weibull(γ, β)

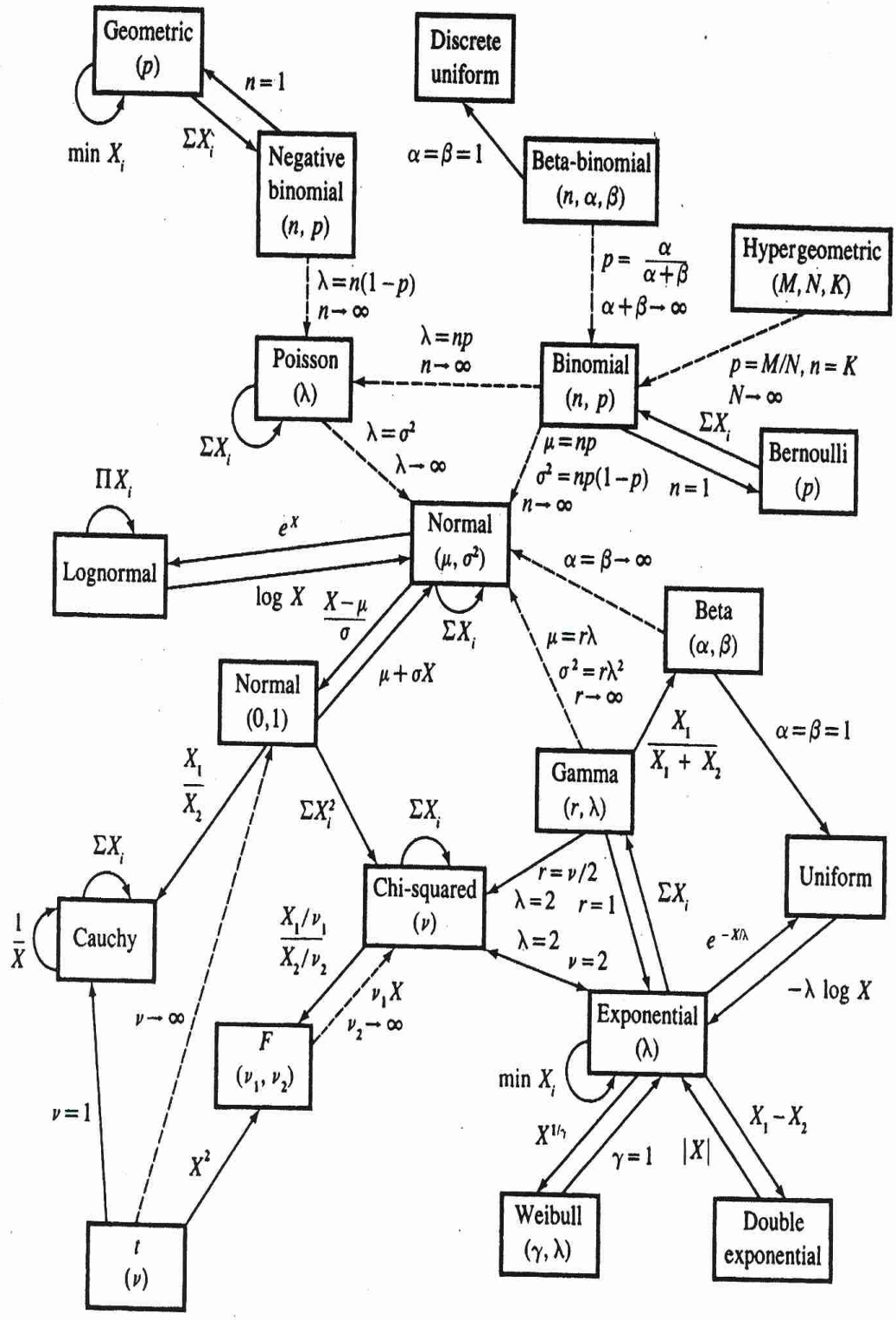
pdf	$f(x\|\gamma,\beta) = \frac{\gamma}{\beta}x^{\gamma-1}e^{-x^\gamma/\beta}, \quad 0 \le x < \infty, \quad \gamma > 0, \quad \beta > 0$
mean and variance	$\mathrm{E}X = \beta^{1/\gamma}\Gamma\left(1+\frac{1}{\gamma}\right), \quad \mathrm{Var}\,X = \beta^{2/\gamma}\left[\Gamma\left(1+\frac{2}{\gamma}\right) - \Gamma^2\left(1+\frac{1}{\gamma}\right)\right]$
moments	$\mathrm{E}X^n = \beta^{n/\gamma}\Gamma\left(1+\frac{n}{\gamma}\right)$
notes	The mgf exists only for $\gamma \ge 1$. Its form is not very useful. A special case is exponential ($\gamma = 1$).

Note that the Weibull distribution is often expressed with alternative parameters as

$$f(y|\gamma,\alpha) = \frac{\gamma}{\alpha}\left(\frac{y}{\alpha}\right)^{\gamma-1} e^{-(y/\alpha)^\gamma} \quad \text{for} \quad y \ge 0$$

That is, $\beta = \alpha^\gamma$

This is the form used in R and SAS.



Relationships among common distributions. Solid lines represent transformations and special cases, dashed lines represent limits. Adapted from Leemis (1986).

From Casella-Berger, "Statistical Inference"

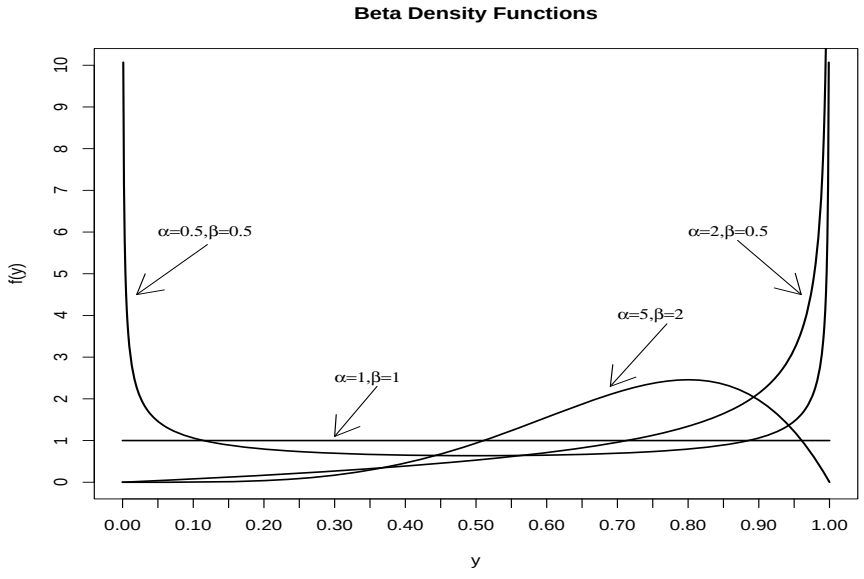
Continuous Distributions

1. **Uniform**(a, b) **R.V.** - probability outcome of r.v. Y is in interval (c, d) is proportional to the length of the interval.

- Symmetric distribution on the interval (a, b)
- Used in simulations and modelling events which are equally likely

2. **Beta**(α, β) **R.V.** - Generalization of the Uniform on $[0, 1]$ R.V.

- Can be symmetric, right-skewed, or left-skewed depending on the values of α and β
- Distribution is Uniform on $[0, 1]$ if $\alpha = \beta = 1$
- Let p be the proportion of defectives in the plant producing a given product, where a company has many plants. A model for p could be the Beta
- Model for R.V.s with bounded realizations. Let Z have a Beta distribution on $[0, 1]$ and let $Y = (b - a)Z + a$. Then Y has a Beta distribution on the interval $[a, b]$



3. **Normal**(μ, σ^2) **R.V.** - Also referred to as a Gaussian R.V.

- Used to model r.v.s with symmetric distributions having tails of moderate weight, not too many extreme values
- Used in many statistical applications due to the Central Limit Theorem

4. **Chi-squared**(ν) **R.V.**

- Right skewed distribution which becomes more symmetric (normal like) as ν becomes large
- Special case of the Gamma and Weibull R.V.s
- parameter ν is referred as degrees of freedom
- Related to the standard normal$(\mu = 0, \sigma = 1)$ distribution through the relationship:

  If $Z_1, Z_2, \ldots, Z_k$ are i.i.d. $N(0, 1)$ r.v.s, then $Y = Z_1^2 + Z_2^2 + \cdots + Z_k^2$ has a Chi-squared distribution with $\nu = k$.

5. **Student t**(ν) **R.V.** - Generally just referred to as the t-distribution

- Symmetric distribution with tails heavier than standard normal
- Converges to a standard normal as ν converges to ∞
- parameter ν is referred as degrees of freedom
- Used to model populations in which extreme values occur more frequently than would be expected in a normal distribution, for example, financial data
- Related to the Chi-squared and standard normal distribution:

  Let Z have a $N(0, 1)$ distribution, W have a Chi-squared distribution with parameter ν with Z and W independent. Then the r.v. $Y = \frac{Z}{\sqrt{W/\nu}}$ has a t-distribution with parameter ν

6. **Fisher**(ν_1, ν_2) **R.V.** - Generally referred to as an F with degrees of freedom, df $= (\nu_1, \nu_2)$

- Right skewed distribution
- Most often used as a test statistic in ANOVA and Regression
- Related to the Chi-squared distribution:

  Let W_1 have a Chi-squared distribution with parameter ν_1, W_2 have a Chi-squared distribution with parameter ν_2, with W_1 and W_2 independent. Then the r.v. $Y = \frac{W_1/\nu_1}{W_2/\nu_2}$ has an F-distribution with parameters ν_1, ν_2

- If T has a t-distribution with parameter ν then T^2 has an F-distribution with parameters $\nu_1 = 1, \nu_2 = \nu$

β_1 = skewness
β_2 = kurtosis

df	μ	σ^2	β_1	β_2
1	—	—	—	—
2	0	—	—	—
5	0	1.7	0	9
9	0	1.3	0	4.2
12	0	1.2	0	3.75
20	0	1.1	0	3.38
25	0	1.09	0	3.29
30	0	1.07	0	3.23
∞	0	1.00	0	3

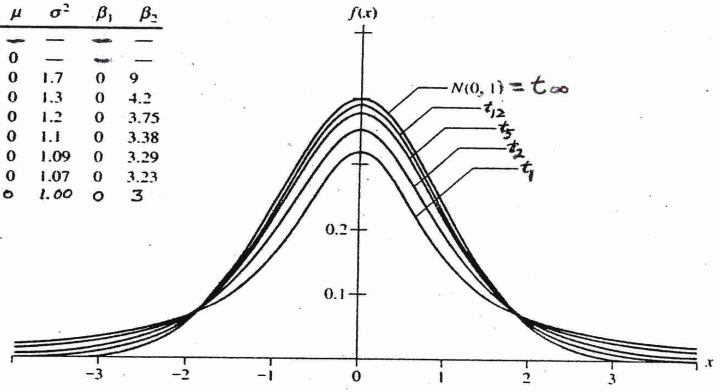
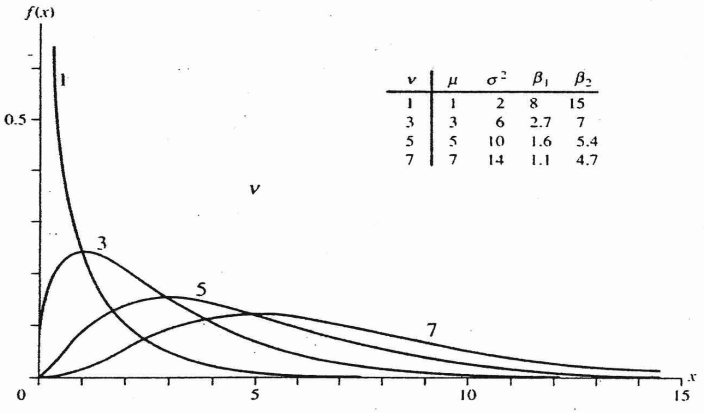


FIGURE 4-6
The relationship between the standard normal distribution and the t distribution.



ν	μ	σ^2	β_1	β_2
1	1	2	8	15
3	3	6	2.7	7
5	5	10	1.6	5.4
7	7	14	1.1	4.7

FIGURE 4-5
Representative χ^2 distribution plots with low degrees of freedom.

n	m	μ	σ^2	β_1	β_2
1	10	1.25	4.69	33.3	109
3	6	1.50	5.25	—	—
5	10	1.25	1.35	14.9	53.9
10	20	1.11	0.43	3.37	9.89

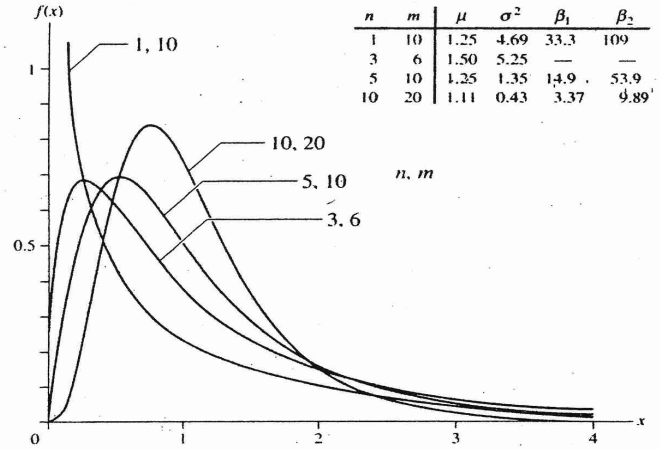
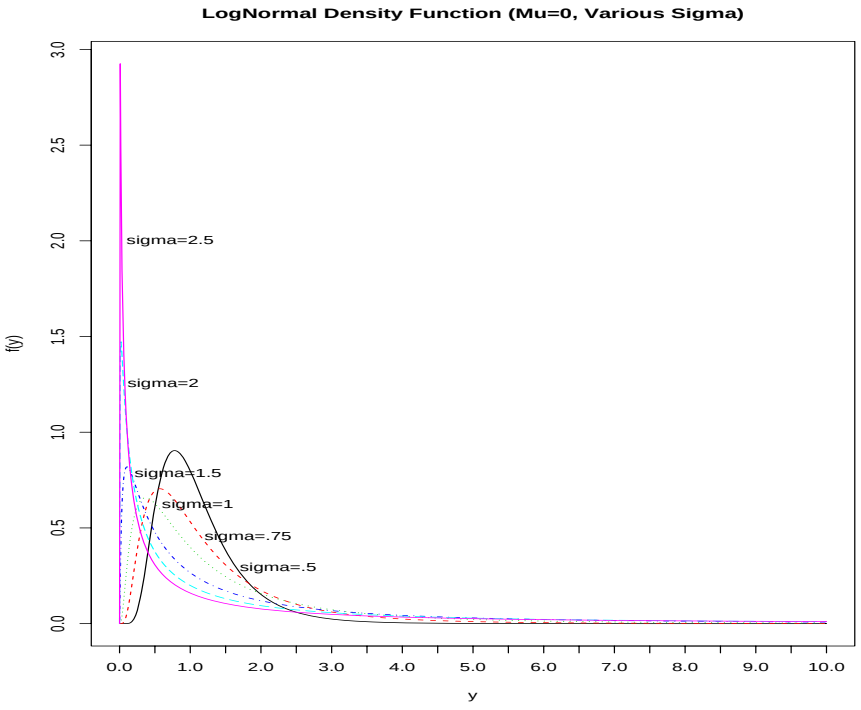


FIGURE 4-7
Plots of selected F distributions.

7. **Lognormal**(μ, σ^2) **R.V.**

- Right skewed distribution
- Used to model the growth of plants, tumors, and in reliability, the time to failure of a device or the time until a tumor is no longer detectable
- Related to the normal distribution:

  Let Y have a $N(\mu, \sigma^2)$ distribution then $X = e^Y$ has a lognormal distribution, that is, $log(X)$ has a normal distribution



8. **Cauchy**(θ_1, θ_2) **R.V.**

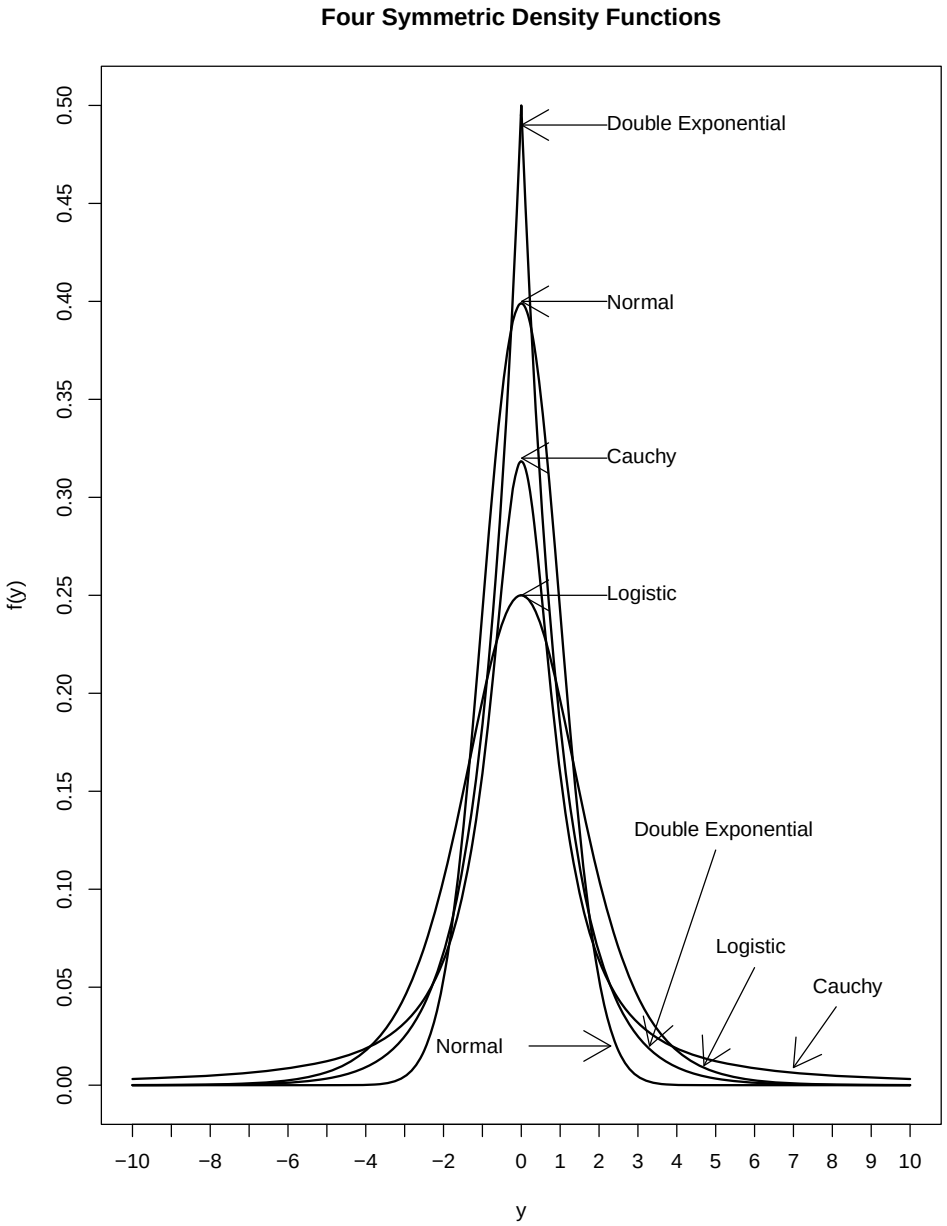
- Symmetric distribution with very heavy tails, so heavy that its mean and variance do not exist
- Models populations in which very large, relative to the point of symmetry, θ_1, occur much more frequently than would be expected in a normal distribution - financial models
- If T has a Student t-distribution with $\nu = 1$ then T has a Standard Cauchy distribution, $(\theta_1 = 0, \theta_2 = 1)$
- If Z_1 and Z_2 are independent $N(0, 1)$ r.v.s then $Y = Z_1/Z_2$ has a standard Cauchy distribution

9. **Double Exponential**(θ_1, θ_2) **R.V.** - Also, referred to as Laplace R.V.

- Symmetric distribution with a sharp peak and tails heavier than the normal distribution but lighter than Cauchy

10. **Logistic**(θ_1, θ_2) **R.V.**

- Symmetric distribution with tail weights in between the double exponential and Cauchy



11. **Exponential**(β) **R.V.**

- In a Poisson process with λ being the average number of occurrences in a unit of time, let T be the time between occurrence of 2 events. T has an exponential distribution with $\beta = 1/\lambda$
- Used in reliability or survival analysis to model the time until failure or death when the time to failure/death has a **memoryless property**, that is, given that the device has survived until at least time y, the probability that the device will function an additional t time units is equal to the probability that the device will survive until time t,

  $P[T \geq t + y | T \geq y] = P[T \geq t]$

  Also, referred as a constant failure rate

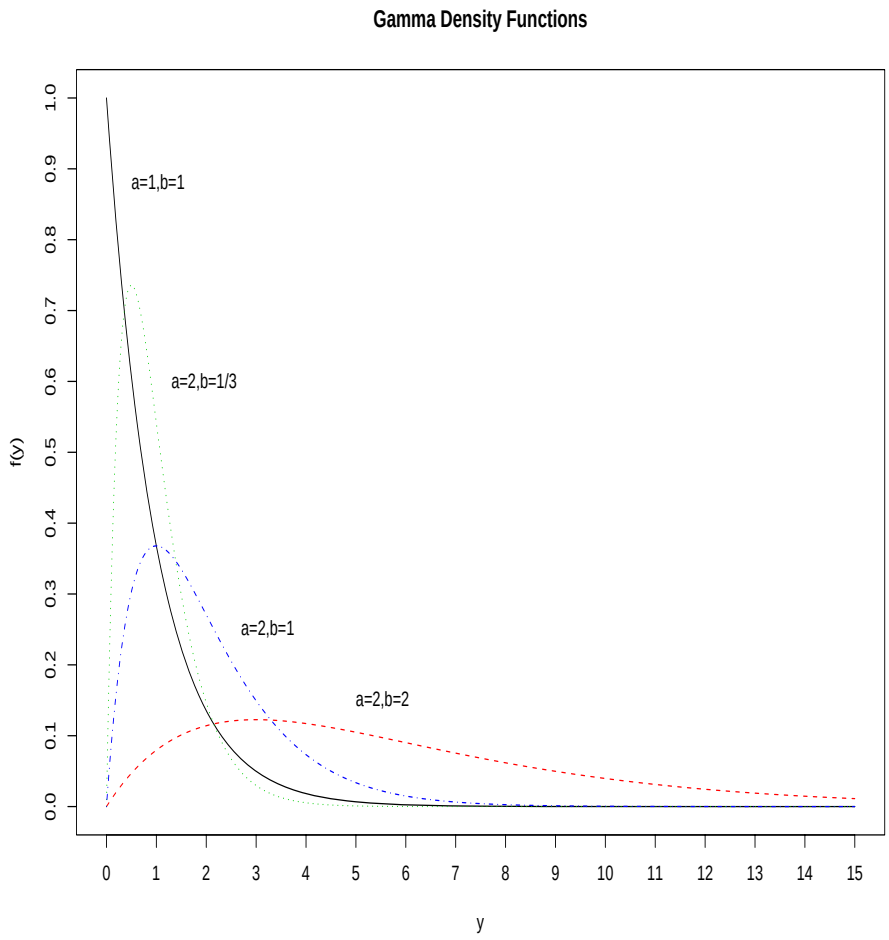
  The failure times of most devices/human subjects do not have the memoryless property over their total lifetime but may have over a limited range

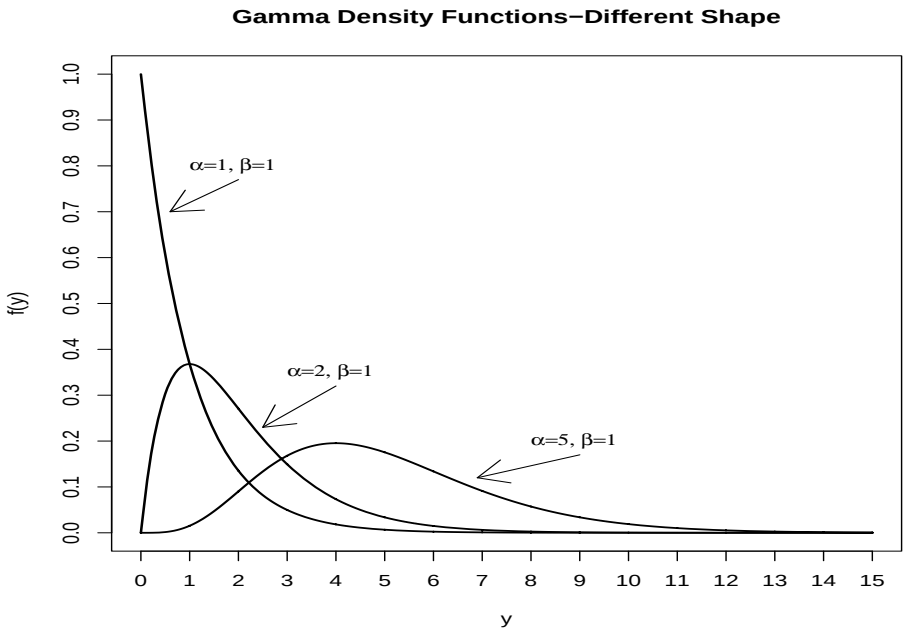
12. **Gamma**(α, β) **R.V.**

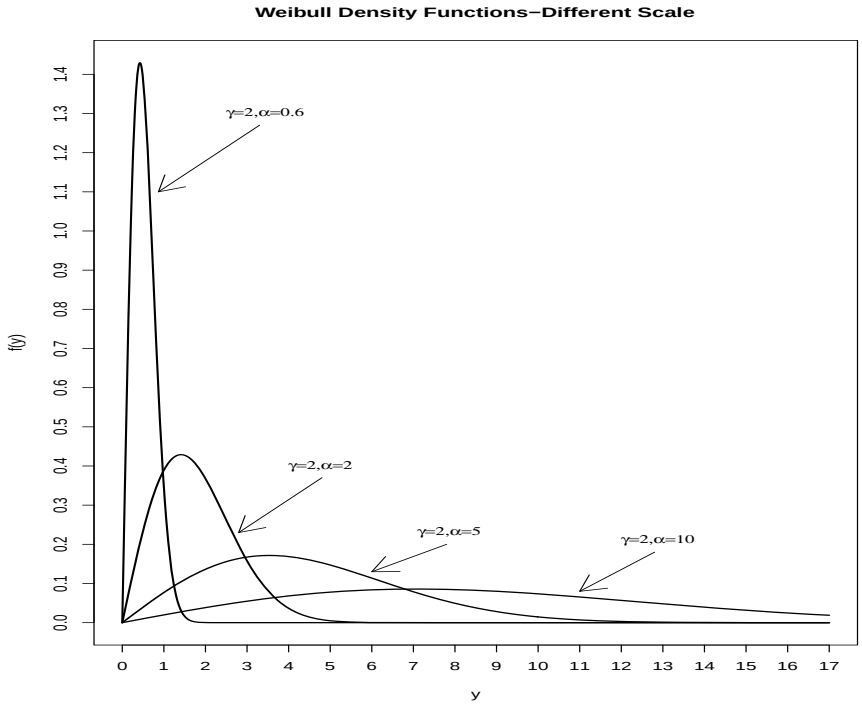
- Let T be the time between k events in a Poisson process with parameter λ then T has a Gamma distribution with $\alpha = k, \beta = 1/\lambda$
- It follows that if $E_1, E_2, \ldots, E_k$ are i.i.d. Exponential r.v.s with parameter β then $T = E_1 + E_2 + \ldots + E_k$ has a Gamma distribution with parameters $\alpha = k, \beta$. This form of the Gamma distribution is referred to as the Erlang distribution
- Exponential(β) distribution is a Gamma($\alpha = 1, \beta$) distribution
- Chi-squared(ν) distribution is a Gamma($\alpha = \nu/2, \beta = 2$) distribution
- If Y has a Gamma(α, β) distribution then $W = \sqrt{Y/\beta}$ has a Maxwell distribution which is used to model particle speeds in gases under special conditions.

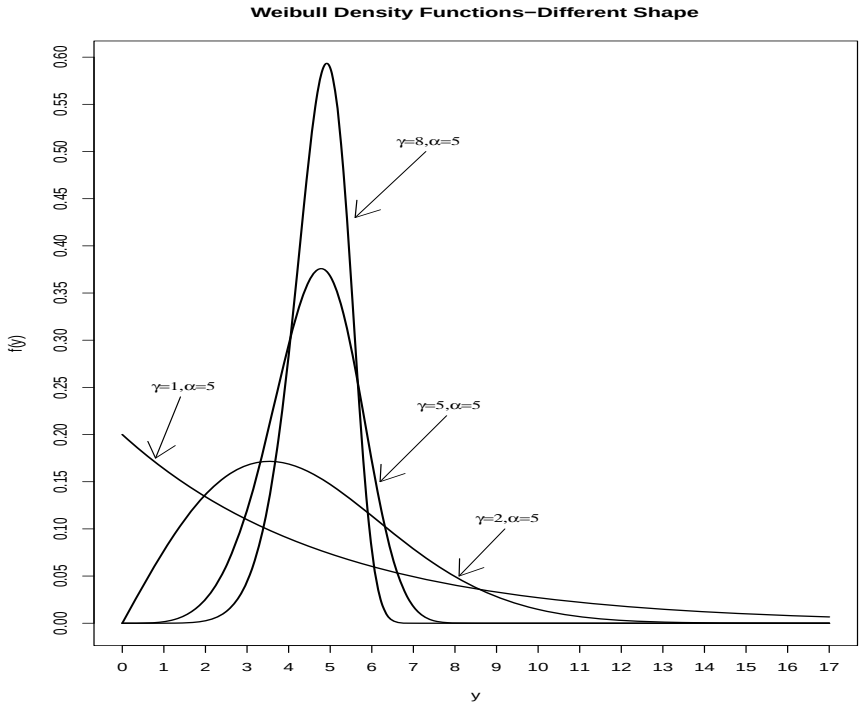
13. **Weibull**(γ, β) **R.V.**

- Generalization of the exponential distribution so that the failure rate is time raised to a power
- Let Y have an exponential(β) distribution and let $X = Y^{1/\gamma}$ then X has a Weibull(γ, β) distribution
- Exponential(β) is a special case of Weibull($\gamma = 1, \beta$)
- The Weibull distribution is a special case of a class of distributions called the Extreme Value Distributions used to model the extreme observations in a populations, either minimums or maximums. For example, largest crack in a 40 feet long pipe, shortest time to death of 100 ticks exposed to an environment of high temperature/low humidity
- Alternative parametrization of the Weibull has parameters (α, γ) with $\alpha = (\beta)^{1/\gamma}$ Need to check software to determine which parametrization they are using.


Gamma Density Functions
a=1,b=1
a=2,b=1/3
a=2,b=1
a=2,b=2
f(y)
y


Gamma Density Functions−Different Shape
α=1, β=1
α=2, β=1
α=5, β=1
f(y)
y


Weibull Density Functions−Different Scale
γ=2,α=0.6
γ=2,α=2
γ=2,α=5
γ=2,α=10
f(y)
y


Weibull Density Functions−Different Shape
γ=8,α=5
γ=1,α=5
γ=5,α=5
γ=2,α=5
f(y)
y

Mixture Distribution

In a number of situations neither a discrete nor a continuous distribution will adequately model the population/process. The population/process is a combination of the realizations from several discrete and/or continuous distributions. We model such situations using Mixture Distributions. In other situations, we may have several populations/processes producing the observed data.

The following examples will illustrate these situations.

Example 1: A central warehouse receives the output from 5 production facilities, e.g., Firestone Tires. The tires are inspected and D_{ij} is the deviation from the specified adhesive strength of Tire j from Facility i. Suppose D_{ij} has a $N(\mu_i, \sigma_i^2)$ distribution, that is, the distribution of D may be differ from facility to facility. Let p_i be the proportion tires in the warehouse from Facility i with $\sum_{i=1}^{5} p_i = 1$. Let X be the measurement obtained from a randomly selected tire in the central warehouse. What is the distribution of D?

Example 2: The summer ozone level data in Houston where the data is the combined data from 10 detection devices placed within 1000 meters of 10 different chemical plants.

In general, if the population of interest is a combination of the values from k distinct populations, where Population i has pdf f_i and cdf F_i, then the Mixture Population has cdf and pdf given by

$$F(x) = \sum_{i=1}^{k} p_i F_i(x) \qquad f(x) = \sum_{i=1}^{k} p_i f_i(x) \quad \text{with} \quad \sum_{i=1}^{k} p_i = 1$$

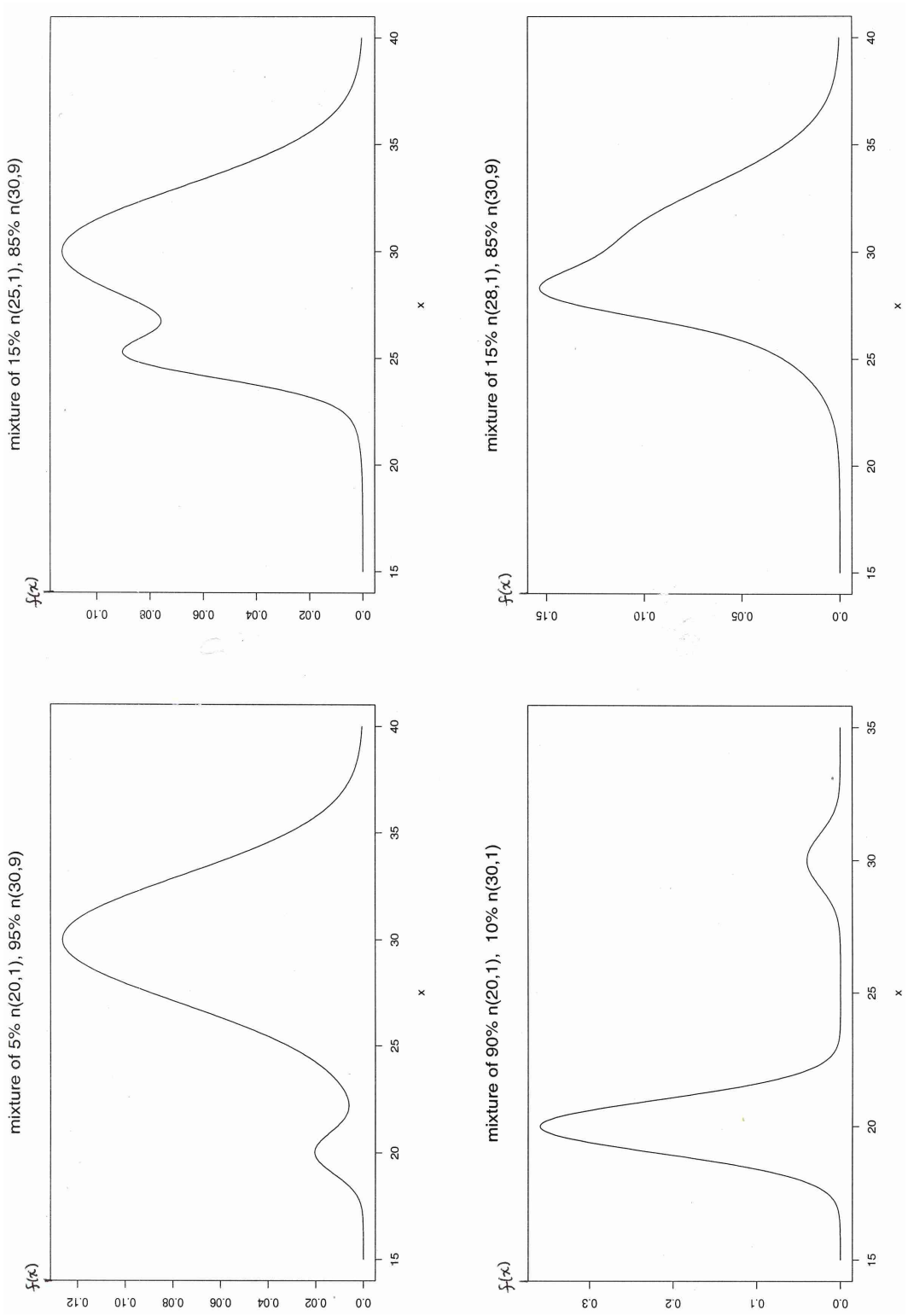
The graphs on the following pages illustrate mixtures of two normal populations.

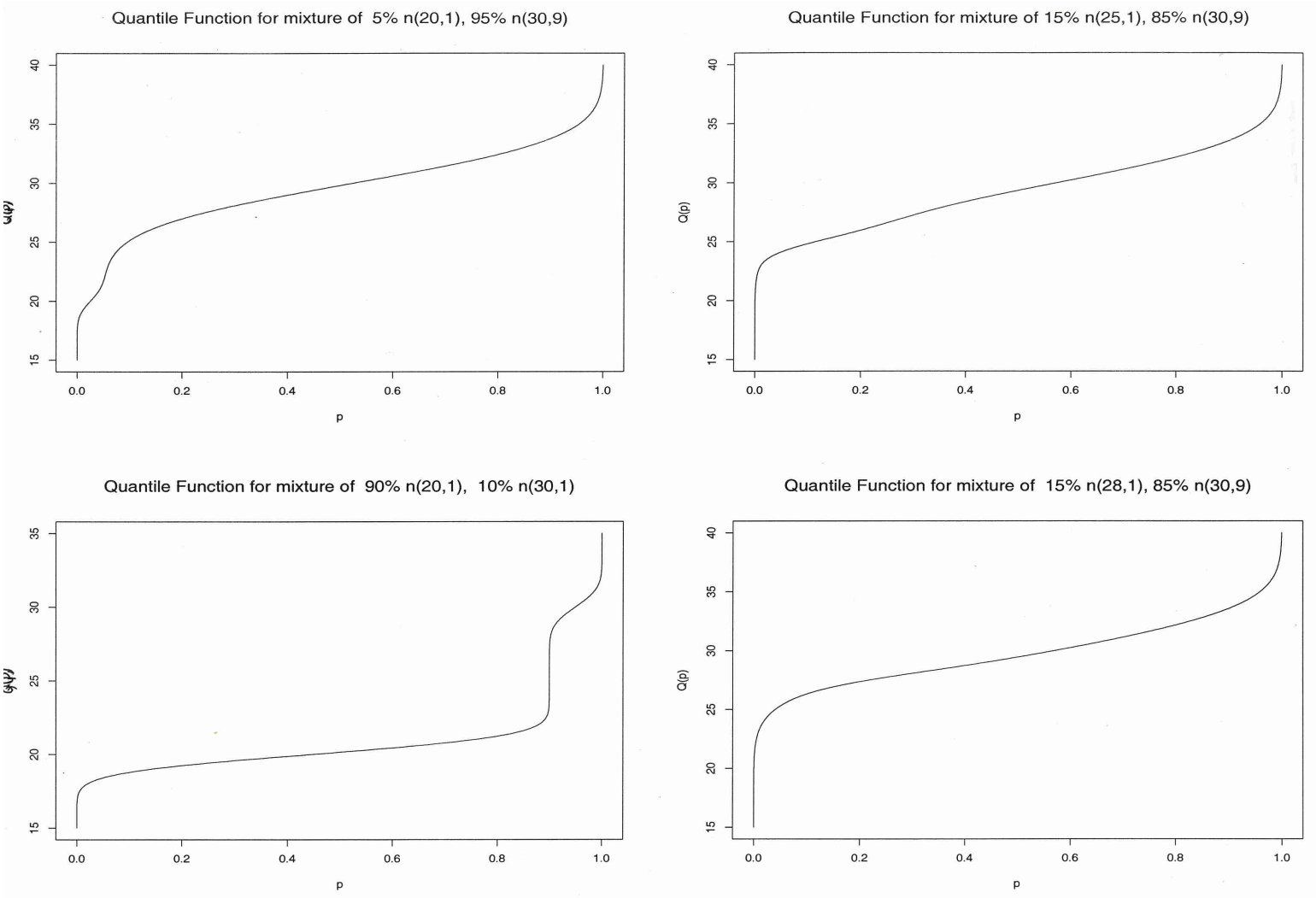
For the situation where we have a random variable, Y, which has probability p of equalling 0 and a continuous cdf, F^* on $(0, \infty)$, then the cdf of Y is

$$F(y) = pI(y \geq 0) + (1 - p)F^*(y)$$

Example: The National Oceanic and Atmospheric Administration (NOAA) wants to determine the amount of game fish caught in shrimp nets. Let Y be the Catch per Unit Effort (CPUE) for a shrimp boat whose nets have been in the water H hours. That is, $Y = N/H$, where N is the total number of game fish caught in a net which was H hours in the water. A proportion of nets, say 20%, will have $N = 0$ and hence $Y = 0$ whereas the values with $Y > 0$ will have a right skewed continuous distribution, such as the log-normal distribution. The cdf of Y is given by

$$F(y) = .2I(y \geq 0) + .8F^*(y), \quad \text{where } F^* \text{ is the cdf of log-normal distribution}$$


mixture of 5% n(20,1), 95% n(30,9)
mixture of 15% n(25,1), 85% n(30,9)
mixture of 90% n(20,1), 10% n(30,1)
mixture of 15% n(28,1), 85% n(30,9)
f(x)
x


Quantile Function for mixture of 5% n(20,1), 95% n(30,9)
Quantile Function for mixture of 15% n(25,1), 85% n(30,9)
Quantile Function for mixture of 90% n(20,1), 10% n(30,1)
Quantile Function for mixture of 15% n(28,1), 85% n(30,9)
Q(p)
p
15
20
25
30
35
40
0.0
0.2
0.4
0.6
0.8
1.0

Simulation of Observations from Specified Distributions

Simulate Observation from Strictly Increasing Continuous cdf F

Let Y have a strictly increasing continuous cdf F_Y. Then,

the quantile function of Y, $Q_Y(u) = F^{-1}(u)$ is a well defined function.

Let U have a Uniform on (0,1) distribution and define the r.v. $W = Q_Y(U)$.

The cdf for U is

$F_U(u) = u$ for $0 \leq u \leq 1$ and $F_U(u) = 0$ for $u < 0$; $F_U(u) = 1$ for $u > 1$

Claim: W has cdf $F_W(w) = F_Y(w)$ for all w.

$$F_W(w) = P[W \leq w] = P[Q_Y(U) \leq w] = P[F_Y(Q_Y(U)) \leq F_Y(w)] = P[U \leq F_Y(w)] = F_Y(w)$$

That is, W is a realization from the distribution of Y.

Thus, we only need a method for generating observations from the Uniform on (0,1) distribution in order to obtain realizations from any continuous distribution that has $Q(u) = F^{-1}(u)$ expressed in a closed form.

Example:

Generate 1000 observations from an Exponential distribution with $\beta = 4$.

$$F_Y(y) = 1 - e^{-y/4} \text{ for } y \geq 0$$

1. Find $Q_Y \quad \Rightarrow \; u = F(y_u) = 1 - e^{-y_u/4}$
2. Solve for $y_u = -4log(1-u) = Q_Y(u)$
3. Generate observation from Uniform on (0, 1) distributions
   Using R: U = runif(1) = .27
4. The observation from an Exp(4) distribution would be
   $Y = Q_Y(.27) = -4log(1-.27) = 1.259$
5. Repeat steps 3. and 4. 1000 times

The above method of generating random observations only works when the cdf and hence quantile function can be expressed in a closed form. For example, the normal distribution and gamma distribution do not have cdf with a close form. They can only be expressed as indefinite integrals:

$$F(t) = \int_0^t \frac{1}{\Gamma(\alpha)\beta^\alpha} y^t e^{-y/\beta} \mathrm{d}y$$

For a particular value of t, the integral must be solved numerically. A similar problem occurs with the normal cdf.

Special algorithms exist for generating observations from these distributions. For example, Cassella-Berger have an algorithm for generating 2 independent observations from a $N(\mu, \sigma^2)$ distribution.

1. Generate 2 observations from a $U_{(0,1)}$ distribution: U_1, U_2
2. Let $R = \sqrt{-2log(U_1)}$
3. Let $Z_1 = R\cos(2\pi U_2)$ and $Z_1 = R\sin(2\pi U_2)$
   Z_1 and Z_2 have independent $N(0, 1)$ distributions
4. Let $Y_1 = \mu + \sigma Z_1$ and $Y_2 = \mu + \sigma Z_2$
   Y_1 and Y_2 have independent $N(\mu, \sigma^2)$ distributions

Simulation from Discrete cdf F

Let D have a discrete distribution with cdf F and pmf

$$f(d_i) = p_i \quad \text{for} \quad d_1 < d_2 < \cdots < d_k$$

$$F(d) = \sum_{d_i \le d} p_i$$

Generate an observation, U, from a Uniform (0,1) distribution.

To obtain an observation on D, let $D = F^{-1}(U) = inf\{d : F(d) \ge U\}$

$$D = \begin{cases} d_1 & \text{if} \quad 0 \le U \le F(d_1) \\ d_2 & \text{if} \quad F(d_1) < U \le F(d_2) \\ \vdots & \vdots \\ d_i & \text{if} \quad F(d_{i-1}) < U \le F(d_i) \\ \vdots & \vdots \\ d_{k-1} & \text{if} \quad F(d_{k-2}) < U \le F(d_{k-1}) \\ d_k & \text{if} \quad F(d_{k-1}) < U \le 1 = F(d_k) \end{cases}$$

Prove that the above method results in D having cdf F.

For U having a Uniform on (0,1) distribution, the cdf for U is given by

$$G(u) = P[U \le u] = \begin{cases} 0 & if \quad u < 0 \\ u & if \quad 0 \le u \le 1 \\ 1 & if \quad u \ge 1 \end{cases}$$

Let f be the pmf of D. Prove that $f(d_j) = F(d_j) - F(d_{j-1})$

$$\begin{aligned} f(d_j) = P[D = d_j] &= P[F(d_{j-1}) < U \le F(d_j)] \\ &= P[U \le F(d_j)] - P[U \le F(d_{j-1})] \\ &= G(F(d_j)) - G(F(d_{j-1})) \\ &= F(d_j) - F(d_{j-1}) \end{aligned}$$

Example: Generate observations from a geometric distribution with $p = .2$:

$$f(i) = p(1-p)^{i-1} = (.2)(.8)^{i-1} \quad \text{for} \quad i = 1, 2, 3, \ldots$$

$$F(i) = \sum_{k=1}^{i} f(i)$$

Suppose we observe $U = .63$ then the corresponding realization from a Geometric distribution with $p = .2$ would be that integer C such that $F(C-1) < .63 \leq F(C)$. Determine C:

$$F(1) = .2, \quad F(2) = .2 + (.2)(.8) = .36, \quad F(3) = .36 + (.2)(.8)^2 = .488$$

$$F(4) = .488 + (.2)(.8)^3 = .590, \quad F(5) = .590 + (.2)(.8)^4 = .672$$

Therefore, we have that with $U = .63$, $C = inf\{y : F(y) \geq .63\}$ which implies

$$F(4) = .590 < .63 < .672 = F(5) \quad \Rightarrow C = 5$$

The following several pages provide SAS and r code for generating random samples from specified distributions and drawing graphs of pdfs and cdfs.

The following table from *Modern Applied Statistics with S* by W.N. Venables and B.D. Ripley describes the r function names and parameters for a number of standard probability distributions. The first letter of the function name indicates which of the probability functions it describes. For example:

- **dnorm** is the density function(pdf) of the normal distribution
- **pnorm** is the cumulative disribution function(cdf) of the normal distribution
- **qnorm** is the quantile function of the normal distribution

Example Let Y have a $N(\mu = 2,\ \sigma = 5)$ distribution.

The value of the pdf at Y=4 is $f(4) = dnorm(4, 2, 5) = .0735$

The value of the cdf at Y=4 is $F(4) = pnorm(4, 2, 5) = .65542$

The value of the quantile at u=.95 is $Q(.95) = qnorm(.95, 2, 5) = 10.224$

Table 5.1: S function names and parameters for standard probability distributions.

Distribution	S name	Parameters
beta	`beta`	`shape1, shape2`
binomial	`binom`	`size, prob`
Cauchy	`cauchy`	`location, scale`
chi-squared	`chisq`	`df`
exponential	`exp`	`rate`
F	`f`	`df1, df2`
gamma	`gamma`	`shape, rate`
geometric	`geom`	`prob`
hypergeometric	`hyper`	`m, n, k`
log-normal	`lnorm`	`meanlog, sdlog`
logistic	`logis`	`location, scale`
negative binomial	`nbinom`	`size, prob`
normal	`norm`	`mean, sd`
Poisson	`pois`	`lambda`
T	`t`	`df`
uniform	`unif`	`min, max`
Weibull	`weibull`	`shape, scale`
Wilcoxon	`wilcox`	`m, n`

These functions can be used to replace statistical tables. For example, the 5% critical value for a (two-sided) t test on 11 degrees of freedom is given by `qt(0.975, 11)`, and the P value associated with a Poisson(25)-distributed count of 32 is given by (by convention) `1 - ppois(31, 25)`. The functions can be given vector arguments to calculate several P values or quantiles.

SAS Program to Generate Random Values

The following SAS program will generate 10 observations from a N(0,1) distribution and 10 observations from a Uniform on (0, 1) distribution:

```
DATA;
DO I=1 TO 10;
X = RANNOR(0);
U = RANUNI(0);
OUTPUT;
END;
RUN;
PROC PRINT X U;
RUN;
```

The following are the functions for a number of other distributions:

1. RANPOI(0,L) - POISSON DISTRIBUTION WITH PARAMETER λ = L
2. RANTRI(O,H) - TRIANGULAR DISTRIBUTION WITH PARAMETER H
3. RANBIN(0,N,P) - BINOMIAL DISTRIBUTION WITH PARAMETERS N and P
4. RANCAU(0) - CAUCHY DISTRIBUTION WITH LOC = 0, SCALE = 1
5. RANEXP(0) - EXPONENTIAL DISTRIBUTION WITH SCALE = 1
6. RANGAMMA(0,A) - GAMMA DISTRIBUTION WITH SHAPE = A, SCALE = 1

The following R commands will generate 10 observations from a N(0,1) and one observation from a Uniform on (0,1) distribution.

```
y = rnorm(10,0,1)
u = runif(1,0,1)
```

Survival Analysis and Reliability Theory

In survival analysis and reliability theory, we are interested in the time to the occurrence of an event: death, failure of a machine, cancer-free examination. Let T be the time at which the event occurs, with cdf F and pdf f, then three functions related to the r.v. T are

1. **Survival Function** is the probability that the event occurs after time t:

$$S(t) = P[T > t] = 1 - F(t)$$

(probability device works greater than t units of time).

2. **Hazard Function** (Failure Rate or Intensity Function) is the risk of failure of a device at time t given device is working at time t:

$$h(t) = \frac{f(t)}{S(t)} \quad \Rightarrow \quad (\Delta t)h(t) \approx P[T \leq t + \Delta t | T > t] \quad \text{for very small } \Delta t$$

$$\begin{aligned} h(t) = \frac{f(t)}{S(t)} = \frac{1}{S(t)} F'(t) &= \frac{1}{P[T > t]} \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{F(t + \Delta t) - F(t)}{P[T > t]} \\ &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{P[t < T \leq t + \Delta t]}{P[T > t]} \\ &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} P[t < T \leq t + \Delta t | T > t] \end{aligned}$$

$h(t)$ is generally reported as the number of failures per unit of time. It specifies the instantaneous rate of failure at time t given that the device is working at time t.

3. **Cumulative Hazard Function**:

$$H(t) = \int_0^t h(\tau) d\tau$$

Accumulated instantaneous risk at time t.

For a r.v. having a continuous strictly increasing cdf, $F(t)$, the following table displays the interrelationships between the various functions: pdf - $f(t)$; Survival function - $S(t)$; Hazard function - $h(t)$; Cumulative hazard function - $H(t)$

	Lifetime Distribution Relationships			
	$f(t)$	$S(t)$	$h(t)$	$H(t)$
$f(t)$	$*$	$S(t) = \int_t^\infty f(\tau)d\tau$	$h(t) = \frac{f(t)}{\int_t^\infty f(\tau)d\tau}$	$H(t) = -ln\left(\int_t^\infty f(\tau)d\tau\right)$
$S(t)$	$f(t) = -S^{'}(t)$	$*$	$h(t) = \frac{-S^{'}(t)}{S(t)}$	$H(t) = -ln(S(t))$
$h(t)$	$f(t) = h(t)e^{-\int_0^t h(\tau)d\tau}$	$S(t) = e^{-\int_0^t h(\tau)d\tau}$	$*$	$H(t) = \int_0^t h(\tau)d\tau$
$H(t)$	$f(t) = H^{'}(t)e^{-H(t)}$	$S(t) = e^{-H(t)}$	$h(t) = H^{'}(t)$	$*$

Note: For a Discrete distributions we have the following relationships:

1. pmf: $f(t) = P[T = t]$ with $\sum_t f(t) = 1$
2. cdf: $F(t) = P[T \leq t] = \sum_{\tau \leq t} f(\tau)$
3. Survival: $S(t) = P[T > t] = \sum_{\tau > t} f(\tau) = 1 - \sum_{\tau \leq t} f(\tau) = 1 - F(t)$

Exponential Dist.

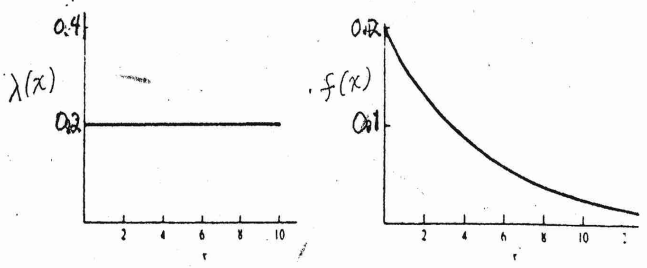
$\lambda(x) = \frac{1}{\beta}$ for $x > 0$

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

EX $\beta = 5$

$\lambda(x) = .2$

Failure Rate pdf



(a)

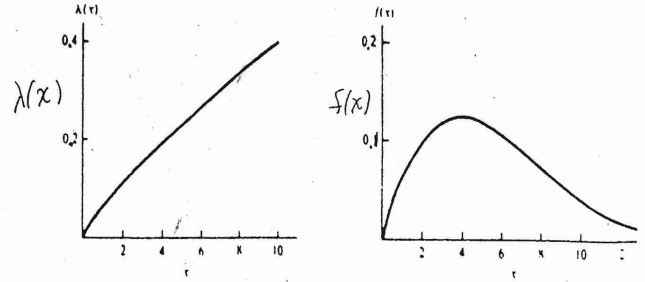
Weibull Dist.

$\lambda(x) = \gamma x^{\gamma-1}/\beta$ for $x > 0$

$f(x) = \frac{\gamma}{\beta} x^{\gamma-1} e^{-\frac{1}{\beta}x^{\gamma}}$ for $x > 0$

EX $\gamma = 1.8$, $\beta = 28.26$

$\lambda(x) = 1.8\, x^{1.8-1}/28.26 = .0637 x^{.8}$



(b)

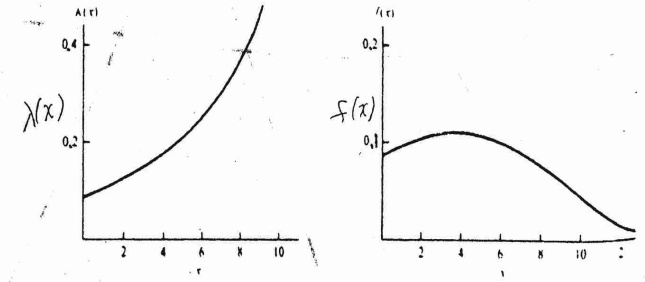
Gompertz Dist.

$\lambda(x) = c e^{bx}$ for $x > 0$

$f(x) = c \exp\left\{bx - \frac{c}{b} e^{bx} + \frac{c}{b}\right\}$ for $x > 0$

EX $b = \ln(1.2)$ $c = .087$

$\lambda(x) = .087\, e^{.1823x}$ for $x > 0$



(c)

FIGURE 3.3-2 Plot of the failure rate $\lambda(x)$ and the p.d.f. $f(x)$ of the following three distributions: (a) exponential with $\beta = 5$; (b) Weibull with $\beta = 6.4$ and $\gamma = 1.8$; (c) Gompertz with $b = \ln(1.2)$ and $c = 0.087$.

Hogg - Ledolter
Applied Statistics for
Engrg's & Physical Scientists

Leemis (1995) "Reliability, Probability Models an STATISTICAL Methods"

TABLE 4.2 TWO-PARAMETER UNIVARIATE LIFETIME DISTRIBUTIONS

Distribution	$f(t)$	$S(t)$	$h(t)$	$H(t)$	Parameters
Weibull	$\kappa\lambda^{\kappa}t^{\kappa-1}e^{-(\lambda t)^{\kappa}}$	$e^{-(\lambda t)^{\kappa}}$	$\kappa\lambda^{\kappa}t^{\kappa-1}$	$(\lambda t)^{\kappa}$	$\lambda>0;\kappa>0$
Gamma	$\frac{\lambda(\lambda t)^{\kappa-1}e^{-\lambda t}}{\Gamma(\kappa)}$	$1-I(\kappa,\lambda t)$	$\frac{\lambda(\lambda t)^{\kappa-1}e^{-\lambda t}}{\Gamma(\kappa)[1-I(\kappa,\lambda t)]}$	$-\log(1-I(\kappa,\lambda t))$	$\lambda>0;\kappa>0$
Uniform	$\frac{1}{b-a}$	$\frac{b-t}{b-a}$	$\frac{1}{b-t}$	$-\log\left(\frac{b-t}{b-a}\right)$	$0\le a<b$
Log normal	$\frac{1}{\sigma t\sqrt{2\pi}}e^{\frac{-(\log t-\mu)^2}{2\sigma^2}}$	$\int_t^{\infty} f(\tau)d\tau$	$\frac{f(t)}{S(t)}$	$-\log S(t)$	$-\infty<\mu<\infty;\sigma>0$
Log logistic	$\frac{\lambda\kappa(\lambda t)^{\kappa-1}}{[1+(\lambda t)^{\kappa}]^2}$	$\frac{1}{1+(\lambda t)^{\kappa}}$	$\frac{\lambda\kappa(\lambda t)^{\kappa-1}}{1+(\lambda t)^{\kappa}}$	$\log[1+(\lambda t)^{\kappa}]$	$\lambda>0;\kappa>0$
Inverse Gaussian	$\sqrt{\frac{\lambda}{2\pi t^3}}e^{-\frac{\lambda}{2\mu^2 t}(t-\mu)^2}$	$\int_t^{\infty} f(\tau)d\tau$	$\frac{f(t)}{S(t)}$	$-\log S(t)$	$\lambda>0;\mu>0$
Exponential Power	$(e^{1-e^{\lambda t^{\kappa}}})e^{\lambda t^{\kappa}}\lambda\kappa t^{\kappa-1}$	$e^{(1-e^{\lambda t^{\kappa}})}$	$e^{\lambda t^{\kappa}}\lambda\kappa t^{\kappa-1}$	$e^{\lambda t^{\kappa}}-1$	$\lambda>0;\kappa>0$
Pareto	$\frac{\kappa\lambda^{\kappa}}{t^{\kappa+1}}$	$\left(\frac{\lambda}{t}\right)^{\kappa}$	$\frac{\kappa}{t}$	$\kappa\log\left(\frac{t}{\lambda}\right)$	$\lambda>0;\kappa>0$
Gompertz	$\delta\kappa^{t}e^{[-\delta(\kappa^{t}-1)/\log\kappa]}$	$e^{[-\delta(\kappa^{t}-1)/\log\kappa]}$	$\delta\kappa^{t}$	$\frac{\delta(\kappa^{t}-1)}{\log\kappa}$	$\kappa>1;\delta>0$