STAT 450/460

Handout 2: Continuous Random Variables

Fall 2016

Chapter 4: Continuous random variables

To define a continuous random variable, we must first define the *cumulative distribution function*, often notated as $F(y)$. A CDF can be defined for **any** random variable Y.

Definition: A function, $F(y) = P(Y \leq y)$, $y \in \mathcal{R}$, is a CDF if and only if:

1. $\lim_{y \to -\infty} F(y) = 0$, and $\lim_{y \to \infty} F(y) = 1$
2. $F(y)$ is nondecreasing: $F(y_1) \leq F(y_2)$ if $y_1 \leq y_2$
3. $F(y)$ is right-continuous: $\lim_{y \to y_0^+} F(y) = F(y_0)$

Recall from handout 1; $Y \equiv$ number of heads out of 3 flips. The pmf was:

y	p(y)
0	1/8 = 0.125
1	3/8 = 0.375
2	3/8 = 0.375
3	1/8 = 0.125

The CDF would look like the following:

> ***Definition***: A random variable Y with distribution function $F(y)$ is said to be *continuous* if $F(y)$ is continuous, for $-\infty < y < \infty$.

A typical CDF for a continuous random variable:

The derivative of $F(y)$ (**if it exists**) is also extremely important for theoretical statistics. The derivative (if it exists) is notated by $f(y)$ and is called the **probability density function** (pdf) of Y.

> ***Definition*** Let $F(y)$ be the distribution function for a continuous random variable Y. Then $f(y)$, given by
>
> $$f(y) = \frac{dF(y)}{dy} = F'(y)$$
>
> is called the **probability density function** (pdf) for the random variable Y wherever $F'(y)$ exists.

It follows from this definition, and from the Fundamental Theorem of Calculus, that:

$$F(y) = \int_{-\infty}^{y} f(t)dt = P(Y \leq y).$$

Properties of a pdf: If $f(y)$ is a probability density function for a continuous random variable, then:

1. $f(y) \geq 0$ for all y; $-\infty < y < \infty$
2. $\int_{-\infty}^{\infty} f(y)dy = 1$
3. $P(Y = y) = 0$: $\int_{y}^{y} f(t)dt = 0$
4. $P(a \leq Y \leq b) = \int_{a}^{b} f(y)dy$; note that the inclusion of endpoints ($\leq$ vs $<$) doesn't matter for continuous random variables.

Expectation:

- $E(Y) = \int_{-\infty}^{\infty} yf(y)dy$
- $E(g(Y)) = \int_{-\infty}^{\infty} g(y)f(y)dy$
- $Var(Y) = \int_{-\infty}^{\infty}(y-\mu)^2 f(y)dy = E(Y^2) - E(Y)^2$
- $E(aY + b) = aE(Y) + b$
- $Var(aY + b) = a^2 Var(Y)$

MGFs:

- $M_Y(t) = E(e^{tY}) = \int_{-\infty}^{\infty} e^{ty} f(y)dy$ for $t \in \{-h, h\}$

Common Continuous Random Variables

- Uniform
- Exponential
- Gamma (survival times)
- Weibull (survival analysis)
- Rayleigh (physics)
- Maxwell (physics)
- Normal (!!!)
- Cauchy - the straw man of pdfs
- Beta - used to model probabilities; $y \in [0, 1]$

Example

$$f(y) = \begin{cases} ky^2(2-y) & 0 \le y \le 2 \\ 0 & otherwise \end{cases}$$

For this pdf, the *support* is $[0, 2]$: the support is defined to be the region where $f(y) > 0$.

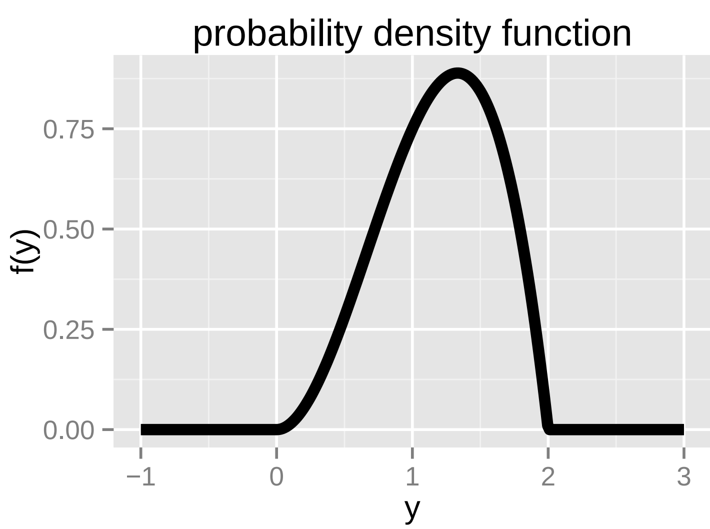
Tasks

a) Find k such that $f(y)$ is a pdf, and graph the pdf.
b) Find the CDF, $F(y)$, and graph it.
c) Find $p(1 < Y < 2)$.
d) Find $E(Y)$.
e) Find $Var(Y)$.
f) Find the median, m.

a) Find k such that $f(y)$ is a pdf, and graph the pdf.

R code to plot pdf:

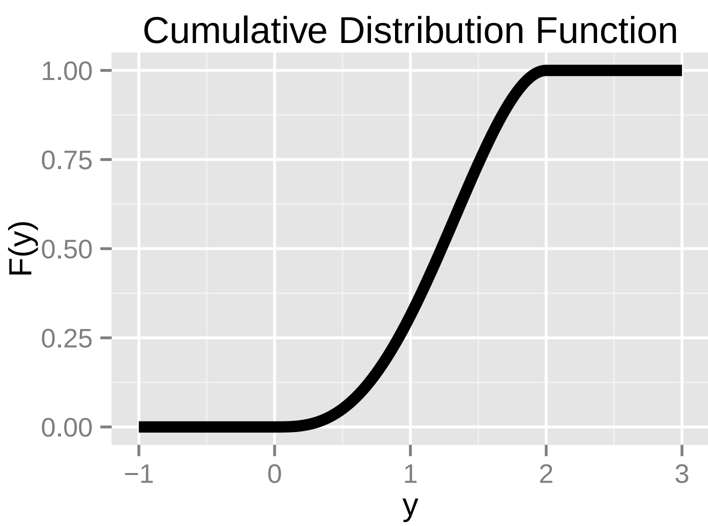
```
f.y <- function(y) {
  pdf <-  ifelse( y < 0 | y > 2,0, 0.75*y^2*(2-y))
  return(pdf)
}
yvals <- seq(-1,3,length=300)
mydata <- data.frame(y =  yvals, height= f.y(yvals))
library(ggplot2)
ggplot(aes(x=y, y = height), data = mydata) + geom_line(size=2) +
    ylab('f(y)') + ggtitle('probability density function')
```



b) Find the CDF, $F(y)$, and graph it.

R code to plot CDF:

```r
#Have to modify since we have three regions to define instead of just 2:
F.y <- function(y) {
  CDF <- rep(NA,length(y))
  region1 <- which(y < 0)
  region2 <- which(0<=y & y <=2)
  region3 <- which(y>2)
  CDF[region1] <- 0
  CDF[region2] <- 0.5*y[region2]^3-3*y[region2]^4/16
  CDF[region3] <- 1
  return(CDF)
}
yvals <- seq(-1,3,length=300)
mydata <- data.frame(y =  yvals, height= F.y(yvals))
ggplot(aes(x=y, y = height), data = mydata) + geom_line(size=2) +
    ylab('F(y)') + ggtitle('Cumulative Distribution Function')
```



c) Find $p(1 < Y < 2)$.

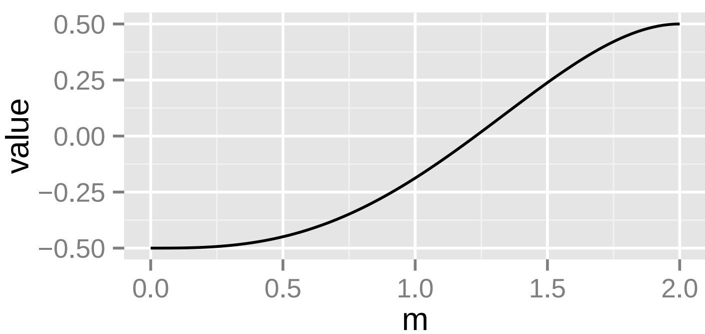
d) Find $E(Y)$.

e) Find $Var(Y)$.

f) Find m, the median of Y.

Solving this using R:

```
library(ggplot2)
integral <- function(m) {
  tosolve <- .5*m^3-3*m^4/16-0.5
  return(tosolve)
}
mvals <- seq(0,2,l=100)
newdat <- data.frame( m = seq(0,2,l=100), value = integral(mvals))
ggplot(aes(x=m,y=value),data=newdat) + geom_line()
```




```
#Kind of looks like the median is around 1.25.
#Let's find the exact root using the R function uniroot():
uniroot(integral,interval=c(0,2))
```

```
## $root
## [1] 1.228528
##
## $f.root
## [1] -1.453698e-05
##
## $iter
## [1] 5
##
## $init.it
## [1] NA
##
## $estim.prec
## [1] 6.103516e-05
```

The Uniform Distribution: $Y \sim UNIF(a, b)$

Y is said to have a $Uniform(a, b)$ distribution, $Y \sim UNIF(a, b)$, if and only if for $b > a$, the density function of Y is:

$$f(y) = \begin{cases} \frac{1}{b-a} & a \leq y \leq b \\ 0 & otherwise \end{cases}$$

Graph of $UNIF(a, b)$ pdf:

It follows that the CDF is:

$$F(y) = \begin{cases} 0 & y < a \\ \frac{y-a}{b-a} & a \leq y \leq b \\ 1 & y > b \end{cases}$$

Is this a valid pdf?

1. $f(y) \geq 0$: True, since $b > a$.
2. **Show** $\int_a^b f(y) = 1$:

- $E(Y) = \frac{a+b}{2}$. **Proof**:

- $Var(Y) = \frac{(b-a)^2}{12}$ **Proof**:

- $M_Y(t) = \frac{e^{bt}-e^{at}}{t(b-a)}$

In R, use the functions `dunif()`, `punif()`, and `runif()` for the pdf, CDF, and to generate $UNIF(a,b)$ random variables, respectively.

Important application of Uniform distribution:

If $U \sim UNIF(0,1)$, and Y is a continuous random variable with CDF $F_Y(y)$, then $F_Y^{-1}(U)$ follows the same distribution as Y. Hence, to generate any continuous variable Y, first generate $UNIF(0,1)$ random variables and apply $F^{-1}(\cdot)$ to those realizations. **HW**

Exponential distribution: $Y \sim EXP(\beta)$

A random variable Y follows the exponential distribution with scale parameter β if and only if:

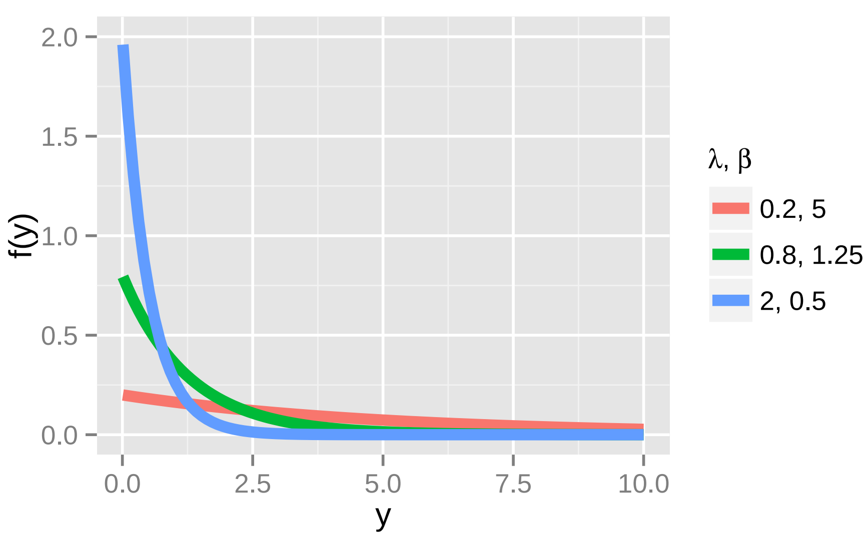
$$f(y) = \begin{cases} \frac{1}{\beta}e^{-y/\beta} & y \geq 0 \\ 0 & otherwise \end{cases}$$

The exponential distribution is often parameterized with a *rate* parameter $\lambda = 1/\beta$, in which case:

$$f(y) = \begin{cases} \lambda e^{-y\lambda} & y \geq 0 \\ 0 & otherwise \end{cases}$$

The exponential distribution serves as a useful model for survival times. Let X represent a survival time. β represents the number of time units per failure, while λ would represent the number of failures per unit time. Graphing several examples:

```
xvals <- seq(0.01,10,l=100)
y1 <- dexp(xvals,rate=0.2) #Note that R's default is to use the rate parameter
y2 <- dexp(xvals,rate=0.8)
y3 <- dexp(xvals,rate=2)
y <- c(y1,y2,y3)
lambdas <- rep(c(0.2,0.8,2),each=100)
mydata <- data.frame(xvals,y,lambda=as.factor(lambdas))
ggplot(aes(x=xvals,y=y),data=mydata) + geom_line(aes(color=lambda),size=2) +
    xlab('y') + ylab('f(y)') +
    scale_color_discrete(name=expression(paste(lambda,', ',beta)),
                         labels=c('0.2, 5','0.8, 1.25','2, 0.5')) + ylim(c(0,2))
```



Note that as λ (the failure rate per unit time) increases, failure times are more distributed toward 0; conversely as β (the amount of time per failure) increases, failure times are more uniformly distributed.

The CDF is an important function to remember:

$$F(y) = \begin{cases} 1 - e^{-\lambda y} & y \geq 0 \\ 0 & otherwise \end{cases}$$

Proof:

Application: Survival function Again let Y denote the survival time; then the survival function is defined to be $S(t) = P(\text{Survive beyond time t}) = P(Y > t) = 1 - F(t) = e^{-\lambda t}$. What happens as the failure rate λ increases?

Important application: Time until first occurrence in a Poisson process

Let $X \sim POI(\lambda)$ represent the number of events per unit time; here λ is the mean arrival rate. The number of occurrences in t time/space units is then $Z \sim POI(t\lambda)$. Let Y denote the time until the first occurrence in the Poisson process. Prove that Y follows an exponential distribution with rate parameter λ, by showing that $P(Y > y) = 1 - F(y) = e^{-\lambda y}$.

Proof:

1. $E(Y) = \beta = 1/\lambda$

2. $M_Y(t) = \frac{1}{1-\beta t}, |t| < 1/\beta$

3. $Var(Y) = \beta^2 = 1/\lambda^2$

Gamma distribution: $Y \sim GAM(\alpha, \beta)$

A random variable Y is said to have a gamma distribution with shape parameter $\alpha > 0$ and scale parameter $\beta > 0$ if and only if the density function of Y is:

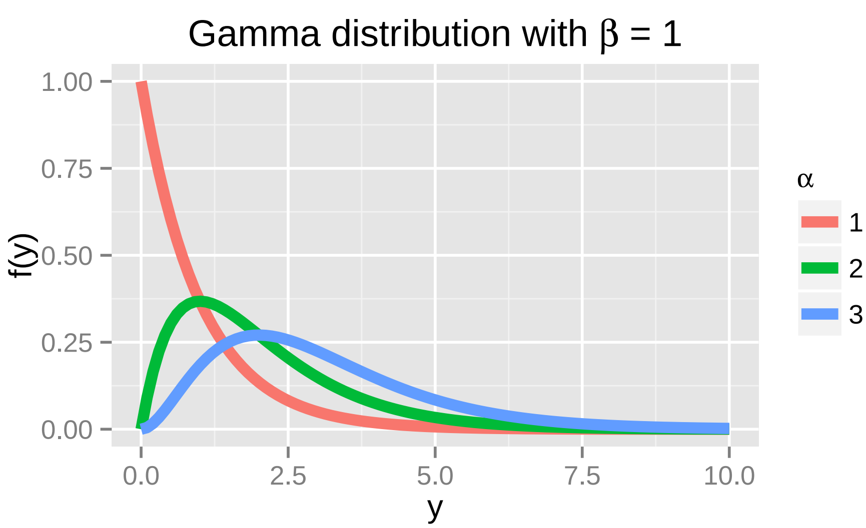
$$f(y) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{-y/\beta} & y \geq 0 \\ 0 & otherwise \end{cases}$$

Like the exponential, the gamma distribution is often parameterized with $\lambda = 1/\beta$ instead. The gamma distribution is often used to model times between failures, or the lengths of time between arrivals in a Poisson process. Special cases of the gamma distribution yield other important well-known distributions:

- $GAM(1, \beta)$ yields the $EXP(\beta)$ distribution
- $GAM(\nu, 2)$ yields the $\chi^2_{2\nu}$ distribution, i.e. the chi-squared distribution with 2ν degrees-of-freedom.

The gamma distribution is right-skewed:

```
yvals <- seq(0,10,l=100)
fy1 <- dgamma(yvals,shape = 1, scale = 1) #Note that R allows for rate or scale specification
fy2 <- dgamma(yvals,shape = 2, scale = 1)
fy3 <- dgamma(yvals,shape = 3, scale = 1)
mydata <- data.frame(y=rep(yvals,3), f.y = c(fy1,fy2,fy3),alpha=as.factor(rep(c(1,2,3),each=100)))
ggplot(aes(x=y,y=f.y),data=mydata) + geom_line(aes(color=alpha),size=2) +
  ggtitle(expression(paste('Gamma distribution with ' , beta,' = 1'))) +
    xlab('y') + ylab('f(y)') +
    scale_color_discrete(name=expression(alpha))
```



The *gamma function* $\Gamma(\alpha)$ is defined as follows:

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt \text{ where } \alpha > 0$$

It has the following properties:

1. $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$ or $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$
2. $\Gamma(\alpha) = (\alpha-1)!$ if $\alpha \in \mathcal{Z}^+$
3. $\Gamma(1) = 1$
4. $\Gamma(1/2) = \sqrt{\pi}$

- **Proof of 1** (and 2 by inspection):

- **Proof of 3**:

- **Proof of 4**:

- Show that $f(y)$ integrates to 1, and hence that:

$$\int_0^\infty y^{\alpha-1} e^{-y/\beta} dy = \Gamma(\alpha)\beta^\alpha$$

- Show that $E(Y) = \alpha\beta$

- Show that $Var(Y) = \alpha\beta^2$

- Show that $M_Y(t) = \left(\frac{1}{1-\beta t}\right)^{\alpha}$ for $|t| < 1/\beta$

In what follows we will show that if Y is the time until the r^{th} arrival or occurrence in a Poisson process with mean rate λ (λ is the average number of arrivals per time unit), then Y follows a $GAM(r, 1/\lambda)$ distribution.

Specifically, if $X \sim POI(t\lambda)$ is the number of arrivals/occurrences during a time interval t, and Y is the time until the r^{th} arrival or occurrence, then Y follows a $GAM(r, 1/\lambda)$ distribution. We will proceed as follows:

1. Derive the CDF of Y, $F(Y)$
2. Find the pdf $f(y)$ by differentiating $f(y) = \frac{d}{dy}F(y)$

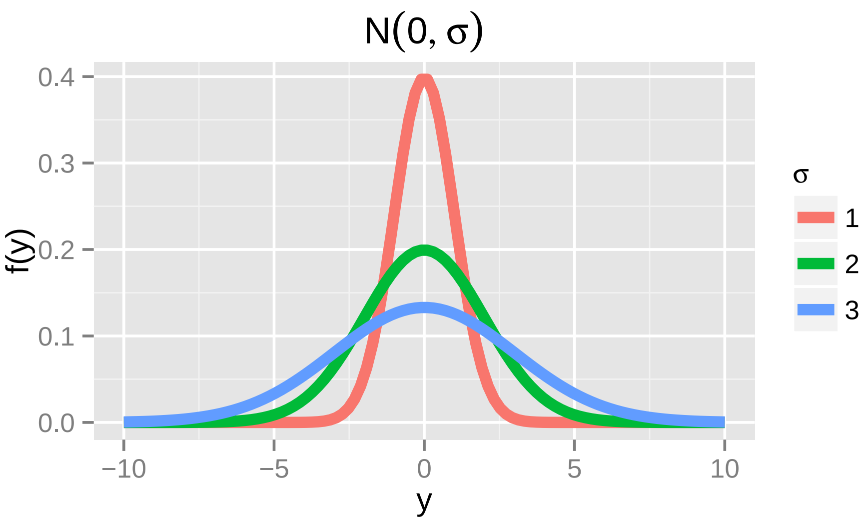
Normal distribution: $Y \sim N(\mu, \sigma^2)$

A random variable Y is said to have a $N(\mu, \sigma^2)$ distribution if, for $\sigma > 0$ and $-\infty < \mu < \infty$, the pdf of Y is:

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \ for \ -\infty < y < \infty$$

The pdf of the normal, of course, is a symmetric bell-shaped curve with a spread that depends on σ^2. Plotting the pdf of $N(0, \sigma^2)$ for various σ^2:

```
yvals <- seq(-10,10,l=100)
fy1 <- dnorm(yvals,mean = 0, sd = 1) #Note that R requires specification of sd, not var!
fy2 <- dnorm(yvals,mean = 0, sd = 2)
fy3 <- dnorm(yvals,mean = 0, sd = 3)
mydata <- data.frame(y=rep(yvals,3), f.y = c(fy1,fy2,fy3),sigma=as.factor(rep(c(1,2,3),each=100)))
ggplot(aes(x=y,y=f.y),data=mydata) + geom_line(aes(color=sigma),size=2) +
  ggtitle(expression(N(0,sigma))) +
    xlab('y') + ylab('f(y)') +
    scale_color_discrete(name=expression(sigma))
```



An important special case of the normal is the $N(0, 1)$, known as the *standard normal distribution*. Letting $Z = (Y - \mu)/\sigma$, which measures the number of standard deviations Y is from the mean, we have:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \ for \ -\infty < z < \infty$$

The standard normal is historically very important; if we want to find a cumulative probability for any $N(\mu, \sigma^2)$ random variable (e.g., $P(Y \leq 3)$), we instead compute the z-score and use the standard normal, tables of which are often found in the backs of most statistics textbooks. Converting to a z-score is of less importance now with the omnipresence of software which can easily calculate cumulative probabilities for any $N(\mu, \sigma^2)$ random variable.

The functions `dnorm()`, `pnorm()`, `qnorm()`, and `rnorm()` are the `R` functions for evaluating the pdf, CDF, finding quantiles, and generating random normal data, respectively.

Showing that the pdf integrates to 1

- $E(Y) = \mu$

Proof:

- $M_y(t) = e^{\mu t + \sigma^2 t^2/2}$

Proof:

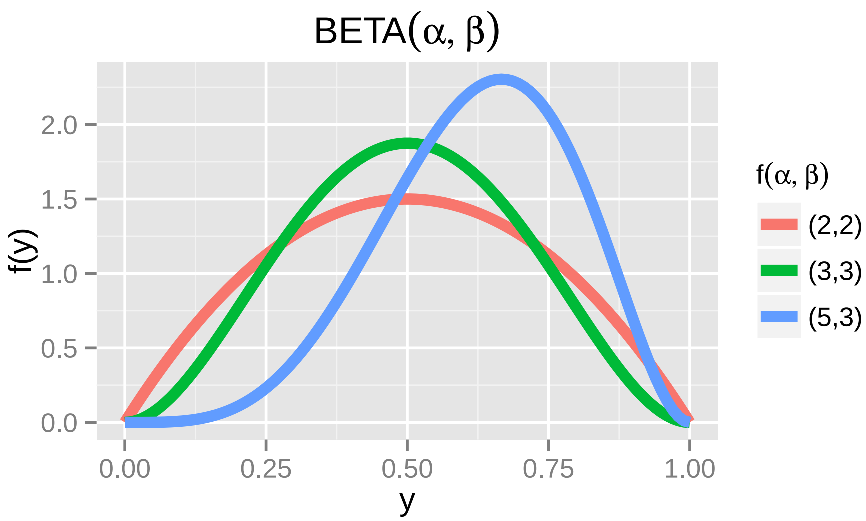
Beta distribution: $Y \sim BETA(\alpha, \beta)$

The Beta distribution is unique in that it is only non-zero for $Y \in [0, 1]$. As such, it is often used to model proportions. A random variable Y is said to have a $BETA(\alpha, \beta)$ distribution fo $\alpha > 0$ and $\beta > 0$ if and only if the pdf of Y is:

$$f(y) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} & 0 \leq y \leq 1 \\ 0 & otherwise \end{cases}$$

Graphs of the pdf:

```
yvals <- seq(0,1,l=100)
fy1 <- dbeta(yvals,shape1 = 2, shape2 = 2)
fy2 <- dbeta(yvals,shape1 = 3, shape2 = 3)
fy3 <- dbeta(yvals,shape1 = 5, shape2 = 3)
mydata <- data.frame(y=rep(yvals,3), f.y = c(fy1,fy2,fy3),pairs=as.factor(rep(c(1,2,3),each=100)))
ggplot(aes(x=y,y=f.y),data=mydata) + geom_line(aes(color=pairs),size=2) +
  ggtitle(expression(BETA(alpha,beta))) +
    xlab('y') + ylab('f(y)') +
    scale_color_discrete(name=expression(f(alpha,beta)),labels=c('(2,2)','(3,3)','(5,3)'))
```



$$E(Y) = \frac{\alpha}{\alpha + \beta}$$

Proof: