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:

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

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:

```
#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))
library(ggplot2)
ggplot(aes(x=y, y = height), data = mydata) + geom_line(size=2) +
    ylab('F(y)') + ggtitle('Cumulative Distribution Function')
```

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

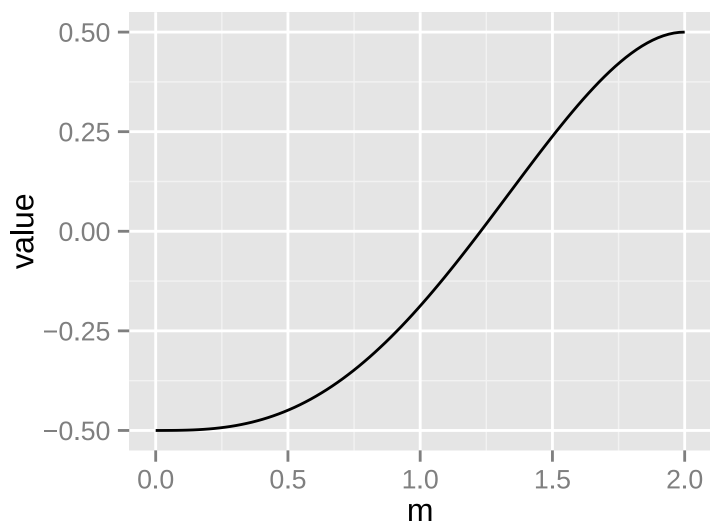
d) Find $E(Y)$.

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
```