Week 4 Live Session

w203 Instructional Team

January 28, 2016

Exploratory Data Analysis Debrief

Homework 3 Presentation

Random Variables Group Discussion

In weeks 1 and 2, we discussed the different types of variables that can be found in a dataset. Is it correct to call these random variables? If not, how are they related to random variables?

Modeling A Fair Die

The range of a discrete random variable is a discrete set of real numbers, O. Much of the time, the range will simply be a finite set of numbers. For any value k in the range, $0 \leq P(X = k) \leq 1$. The sum over all values k in the range is 1, i.e. $\sum_{k \in O} P(X = k) = 1$.

The expected value (or population mean) of a discrete random variable X is the weighted average of the values in the range of X.

$$E(X) = \sum_{k \in O} k \cdot P(X = k)$$

One very useful feature of R is its ability to generate random numbers from a wide variety of distributions.

R Exercise: How can we represent a fair (six-sided) die in R?

```
(die <- 1:6)
```

```
## [1] 1 2 3 4 5 6
```

```
(p.die <- rep(1/6,6))
```

```
## [1] 0.1666667 0.1666667 0.1666667 0.1666667 0.1666667 0.1666667
```

Find the expected value.

```
sum(die*p.die)
```

```
## [1] 3.5
```

Roll the die 10 times, plot the result, and compute the sample mean.

```
s1 <- sample(die, size=10, prob=c(1/6,1/6,1/6,1/6,1/6,1/6), replace=T)
s1 <- sample(die, size=10, prob=p.die, replace=T)
```

Roll the die 1000 times, plot the result, and compute the sample mean.

```
s2 <- sample(die, size=1000, prob=p.die, replace=T)
s2 <- table(sample(die, size=1000, prob=p.die, replace=T))
```

Modeling a (Possibly) Fair Coin

Suppose you have a fair coin. With each flip, the coin can land on heads or tails (we will assume that the coin will not land on its side for the purposes of this example).

How can we represent these possible outcomes of the coin?

```
coin <- c("head","tail")
```

How can we represent the outcome of a random experiment, in which you flip the coin 10 times? Run such a simulation and report the number of heads.

```
s <- sample(coin, 10,replace=T)
count <- function(x, n){ length((which(x == n))) }
count(s,"head")
```

```
## [1] 5
```

What if the coin is not fair? Repeat your experiment assuming the probability the coin will land Heads is 30% and Tails 70%. How many heads do you get?

```
S <-  sample(c("head","tail"), 10, prob=c(0.3,0.7),replace=T)
    as.data.frame(table(S))
```

```
##      S Freq
## 1 head    4
## 2 tail    6
```

What would happen in the limit as the number of tosses approaches infinity?

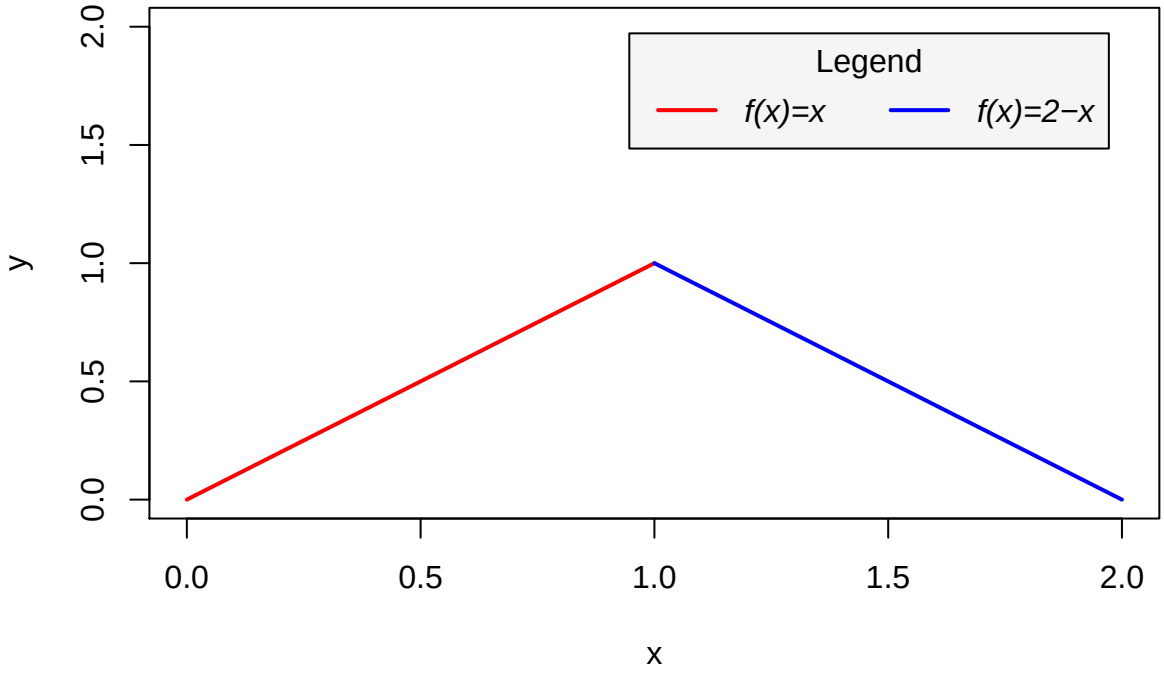
The 'Pyramid' Distribution

A continuous random variable has the following PDF.

$$f_X(x) = \begin{cases} x, & 0 \leq x < 1 \\ 2 - x, & 1 \leq x < 2 \\ 0, & \text{otherwise} \end{cases}$$

```
x1 <- seq(0,1,0.1)
x2 <- seq(1,2,0.1)
y1 <- x1
y2 <- 2-x2
plot(c(0,2), c(0,2), type="n", xlab="x", ylab="y")
lines(x1, y1, col="red", lwd=2)
lines(x2, y2, col="blue", lwd=2)
legend("topright",
```

```
       inset=.05,
       cex = 1,
       title="Legend",
       c("f(x)=x","f(x)=2-x"),
       horiz=TRUE,
       lty=c(1,1),
       lwd=c(2,2),
       col=c("red","blue"),
       bg="grey96",
       text.font=3)
```



a. Find the corresponding CDF and plot it.

$$\text{For } x \in (0,1), F_X(x) = \int_0^x f_x(u)du = x^2/2$$

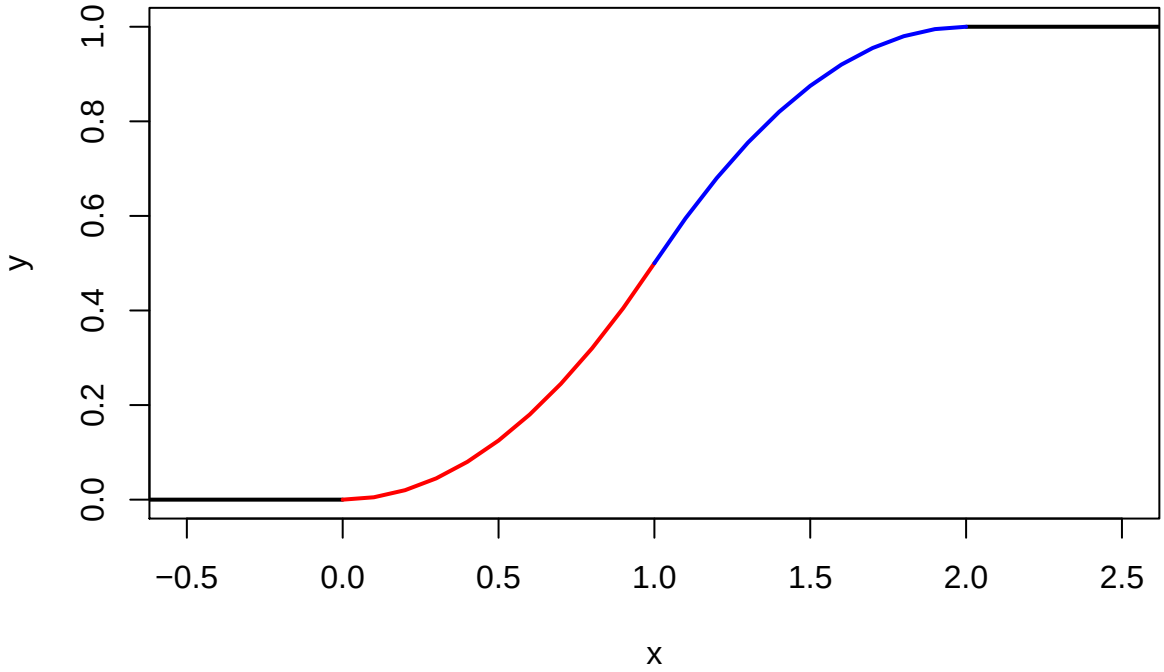
$$\text{For } x \in (1,2), F_X(x) = \int_0^1 f_x(x)dx + \int_1^x f_x(u)du = \int_0^1 udu + \int_1^x (2-u)du = \left[\frac{u^2}{2}\right]_0^1 + \left[2u - \frac{u^2}{2}\right]_0^x = \frac{1}{2} + 2x - \frac{1}{2}x^2$$

$$F_X(x) = \begin{cases} 0, & x < 0 \\ \frac{x^2}{2}, & 0 \le x < 1 \\ \frac{1}{2} + 2x - \frac{1}{2}x^2, & 1 \le x < 2 \\ 1, & 2 < x \end{cases}$$

```r
x0 <- seq(-1,0,0.1)
x1 <- seq(0,1,0.1)
x2 <- seq(1,2,0.1)
x3 <- seq(2,3,0.1)
y0 <- 0*x0
y1 <- (x1*x1)/2
y2 <- 2*x2 - ((x2^2)/2)-1
y3 <- (0*x3)+1

plot(c(-0.5,2.5), c(0,1), type="n", xlab="x", ylab="y")

lines(x0, y0, lwd=2)
lines(x3, y3, lwd=2)
lines(x1, y1, col="red", lwd=2)
lines(x2, y2, col="blue", lwd=2)
```



b. Compute the expectation of the variable.

$$E(X) = \int_0^1 x f_x(x)dx + \int_1^2 x f_x(x)dx = \int_0^1 x^2 dx + \int_1^2 (2x - x^2)dx = \left[\frac{1}{3}x^3\right]_0^1 + \left[x^2 - \frac{1}{3}x^3\right]_1^2 = 1$$

c. Prove that the expectation of any random variable with a PDF that's symmetric around some number k is k.

Suppose Y is any random variable with PDF f_Y that's symmetric around k. This means that for any v, $f_Y(k - v) = f_Y(k + v)$.

The expectation we want is

$$E(Y) = \int_{-\infty}^{\infty} y f_Y(y) dy$$

.

To compute this, we start with a (very simple) variable substitution. Let $y = k + z$. We can then write our integral as,

$$E(Y) = \int_{-\infty}^{\infty} (k+z) f_Y(k+z) dz$$
$$= k \int_{-\infty}^{\infty} f_Y(k+z) dz + \int_{-\infty}^{\infty} z f_Y(k+z) dz$$

Notice that the left integral is just the integral of a probability density function, so its value is 1. Therefore,

$$E(Y) = k + \int_{-\infty}^{\infty} z f_Y(k+z) dz$$

We can break the right integral into a term for $z < 0$ and a term for $z > 0$:

$$E(Y) = k + \int_{-\infty}^{0} z f_Y(k+z) dz + \int_{0}^{\infty} z f_Y(k+z) dz$$

To show that the two integrals cancel each other out, we need to change variables again. Let z = -v. Notice that when we do this, we get a -1 from the substitution, another -1 from $dz/dv = -1$, and a final -1 from switching the bounds of integration:

$$\int_{-\infty}^{0} z f_Y(k+z) dz = \int_{\infty}^{0} (-v) f_Y(k-v)(-1) dv = -\int_{0}^{\infty} v f_Y(k-v) dv$$

We know, however, that $f_Y(k-v) = f_Y(k+v)$. Substituting in and switching our integration variable back to z, this integral therefore equals,

$$-\int_{0}^{\infty} v f_Y(k+v) dv = -\int_{0}^{\infty} z f_Y(k+z) dz$$

Subsituting back in above, we have,

$$E(Y) = k - \int_{0}^{\infty} z f_Y(k+z) dz + \int_{0}^{\infty} z f_Y(k+z) dz = k$$

d. Compute the variance of the random variable.

$$Var(X) = E[X^2] - (E[X]^2) = \int_{-\infty}^{\infty} x^2 f_x(x) dx - 1$$

$$Var(X) = \int_{0}^{1} x^2 f_x(x) dx + \int_{1}^{2} x^2 f_x(x) dx = \int_{0}^{1} x^3 dx + \int_{1}^{2} (2x^2 - x^3) dx - 1$$

$$Var(X) = \left[\frac{1}{4}x^4\right]_0^1 + \left[\frac{2}{3}x^3 - \frac{1}{4}x^4\right]_1^2 - 1 = \frac{1}{4} + \frac{16}{3} - \frac{16}{4} - \frac{2}{3} + \frac{1}{4} - 1 = \frac{14}{3} - \frac{14}{4} - 1 = \frac{56-42}{12} - 1 = \frac{1}{6}$$

Exponential Decay

The lifespan of a radioactive element is a random variable, X, with the following probability density function.

$$f_X(x) = \begin{cases} ce^{-2x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

where c is a constant.

a. compute the constant c.

$$F_X(x) = \int_0^\infty ce^{-2x}dx = 1$$

$$F_X(x) = c\int_0^\infty e^{-2x}dx = c\left[-\frac{1}{2}e^{-2x}dx\right]_0^\infty = c\frac{1}{2} = 1$$

$$c = 2$$

b. compute $P(X > 2)$.

$$P(X > 2) = \int_2^\infty 2e^{-2x}dx = \left[-e^{-2x}dx\right]_2^\infty = e^{-4}$$

Modeling an American Delicacy

Suppose that the number of calories in a Twinkie is a normally-distributed random variable. Given than 50% of Twinkies have over 100 calories, and 68.2% of Twinkies are in the interval from 85 calories to 115 calories, what fraction of Twinkies have more than 140 calories?

Next use R to simulate the production of 1000 Twinkies. Plot the result and compute the fraction of Twinkies in your sample that have more than 140 calories.