Pol Sci 630: Problem Set 2 Solutions - Properties of Random Variables

Prepared by: Anh Le (anh.le@duke.edu)

Due Date for Grading: Friday, September 11, 2015, 10 AM
(Beginning of Class)

1. Expected Value and Its Properties

a.

(1/4 point) (DeGroot, p. 216) Suppose that one word is to be selected at random from the sentence 'the girl put on her beautiful red hat'. If X denotes the number of letters in the word that is selected, what is the value of E(X)?

Solution

As the number of letters in a word, X can take on following values: $x \in \{2, 3, 4, 9\}$, with probability as follows:

$$P(X = 2) = \frac{1}{8} \qquad \text{(1 word (“on”) out of 8 words in the sentence)} \tag{1}$$

$$P(X = 3) = \frac{5}{8} \tag{2}$$

$$P(X = 4) = \frac{1}{8} \tag{3}$$

$$P(X = 9) = \frac{1}{8} \tag{4}$$

Therefore,

$$E(X) = \sum_{\text{all } x_i} x_i P(X = x_i) = 3.75$$

b.

(2/4 point) (Degroot p. 216) Suppose that one letter is to be selected at random from the 30 letters in the sentence given in Exercise 4. If Y denotes the number of letters in the word in which the selected letter appears, what is the value of E(Y)?

Solution

Y can take on values $y \in \{2, 3, 4, 9\}$ with probability as follows:

$$P(Y = 2) = \frac{2}{30} \qquad \text{O,N} \tag{5}$$
$$P(Y = 3) = \frac{15}{30} \qquad \text{T,H,E, P,U,T, H,E,R, R,E,D, H,A,T} \tag{6}$$
$$P(Y = 4) = \frac{4}{30} \qquad \text{G,I,R,L} \tag{7}$$
$$P(Y = 9) = \frac{9}{30} \qquad \text{B,E,A,U,T,I,F,U,L} \tag{8}$$

Therefore,

$$E(Y) = \sum_{\text{all } y_i} y_i P(Y = y_i) = \frac{73}{15} = 4.867$$

c.

(1/4 point) (Degroot, p. 224) Suppose that three random variables X_1, X_2, X_3 are uniformly distributed on the interval [0, 1]. They are also independent. Determine the value of $E[(X_1 - 2X_2 + X_3)^2]$.

Solution

$$E[(X_1 - 2X_2 + X_3)^2] = \tag{9}$$
$$= E(X_1^2) + 4E(X_2^2) + E(X_3^2) - 4E(X_1X_2) + 2E(X_1X_3) - 4E(X_2X_3) \tag{10}$$
$$= E(X_1^2) + 4E(X_2^2) + E(X_3^2) - 4E(X_1)E(X_2) + 2E(X_1)E(X_3) - 4E(X_2)E(X_3) \tag{11}$$

Since each X_i is uniformly distributed on $[0, 1]$,

$$E(X_i) = \frac{1}{2} \tag{12}$$
$$E(X_i^2) = \int_0^1 x^2 dx = \frac{1}{3} \qquad \text{law of unconscious statistician} \tag{13}$$

Note: Law of unconscious statistician $E[g(x)] = \int g(x)f(x)dx$. This is an important theorem because it allows us to work with any function of a variable, as long as we know the distribution of that variable.

Alternatively, a common trick to find $E(X^2)$ is:

$$E(X^2) = Var(X) + [E(X)]^2 \tag{14}$$
$$= \frac{1}{12} - \frac{1}{4} = \frac{1}{3} \qquad \text{look up variance of uniform variable} \tag{15}$$

Plug everything back in, we have $E[(X_1 - 2X_2 + X_3)^2] = \frac{1}{2}$

2. Variance and its properties

For this problem, you can use the properties of expected value.

a.

(1/4 point) Prove that $Var(aX + b) = a^2 Var(X)$.

Solution

$$
\begin{aligned}
Var(aX + b) &= E[(aX + b)^2] - (E[(aX + b)])^2 && (16)\\
&= E[a^2X^2 + 2abX + b^2] - a^2[E(X)]^2 - 2abE(X) - b^2 && (17)\\
&= a^2(E(X^2) - [E(X)]^2) && (18)\\
&= a^2 Var(X) \quad \square && (19)
\end{aligned}
$$

b.

(2/4 point) Implement in `R` two functions that calculates the variance of the sum of two variables in two ways. The first calculates `Var(X + Y)`. The second calculates `Var(X) + Var(Y) + 2Cov(X, Y)`.

You should use vectorized operation and check that two functions return the same result. You may not use R's built-in `var()` and `cov()` functions.

Solution

```r
sumVar1 <- function(X, Y) {
  Z <- X + Y
  return(sum((Z - mean(Z))**2) / (length(Z) - 1))
}

sumVar2 <- function(X, Y) {
  varX <- sum((X - mean(X))**2) / (length(X) - 1)
  varY <- sum((Y - mean(Y))**2) / (length(Y) - 1)
  covXY <- sum((X - mean(X)) * (Y - mean(Y))) / (length(X) - 1)
  return(varX + varY + 2 * covXY)
}

set.seed(1)
X <- rnorm(100) ; Y <- rnorm(100)
sumVar1(X, Y)

## [1] 1.722583

sumVar2(X, Y)

## [1] 1.722583
```

c.

(1/4 point) (Degroot, p. 232) Suppose that one word is selected at random from the sentence 'the girl put on her beautiful red hat'. If X denotes the number of letters in the word that is selected, what is the value of $Var(X)$?

Solution

Notice that the distribution of X is the same as in Question 1a), therefore $E(X) = 3.75$ and

$$E(X^2) = \sum_{\text{all } x_i} x_i^2 P(X = x_i) = \frac{73}{4}$$

Thus,

$$Var(X) = E(X^2) - [E(X)]^2 = \frac{67}{16} \tag{20}$$

3. Binomial distribution

(Credit to Jan) This problem is taken from Pitman (1993) Probability

Suppose a fair coin is tossed n times. Find a simple formula in terms of n and k for the following probability: $Pr(k\ heads | k-1\ heads\ or\ k\ heads)$. Please pay close attention to the formula, particularly what event is conditioned on what events. (Ch. 2.1, Problem 10 b) (p. 91)

Hint 1: Use the binomial distribution to model this.

Hint 2: Use $Pr(A|B) = \dfrac{Pr(A \cap B)}{Pr(B)}$ with $A = k\ heads$ and $B = k - 1\ heads\ or\ k\ heads)$

Solution (Credit to Jan)

$Pr(k\ heads | k-1\ heads\ or\ k\ heads)$

$$= \frac{Pr(k\ heads \cap (k-1\ heads\ or\ k\ heads))}{Pr(k\ heads) + Pr(k-1\ heads)}$$

$$= \frac{Pr(k\ heads)}{Pr(k\ heads) + Pr(k-1\ heads)}$$

$$= \frac{\binom{n}{k} 0.5^k 0.5^{n-k}}{\binom{n}{k} 0.5^k 0.5^{n-k} + \binom{n}{k-1} 0.5^{k-1} 0.5^{n-(k-1)}}$$

$$= \frac{\binom{n}{k} 0.5^n}{\binom{n}{k} 0.5^n + \binom{n}{k-1} 0.5^n}$$

$$= \frac{\binom{n}{k}}{\binom{n}{k} + \binom{n}{k-1}}$$

$$= \frac{\dfrac{n!}{(n-k)!k!}}{\dfrac{n!}{(n-k)!k!} + \dfrac{n!}{(n-(k-1))!(k-1)!}}$$

$$= \frac{\frac{n!}{(n-k)!k!} * \frac{n-k+1}{n-k+1}}{\frac{n!}{(n-k)!k!} * \frac{n-k+1}{n-k+1} + \frac{n!}{(n-k+1)!(k-1)!} * \frac{k}{k}}$$
$$= \frac{\frac{n!(n-k+1)}{(n-k+1)!k!}}{\frac{n!(n-k+1)}{(n-k+1)!k!} + \frac{n!}{(n-k+1)!k!}}$$
$$= \frac{n!(n-k+1)}{n!(n-k+1) + n!k}$$
$$= \frac{n-k+1}{n-k+1+k}$$
$$= \frac{n-k+1}{n+1}$$

4. Plotting distribution

For this problem, you'll need to Google some R techniques (e.g. side-by-side / overlapping plot). Also, label the axes and the plots accordingly.

a.

(1/4 point) Download a variable you are interested in, using `WDI`. Plot the histogram, density plot, boxplot, and normal quantile plot.

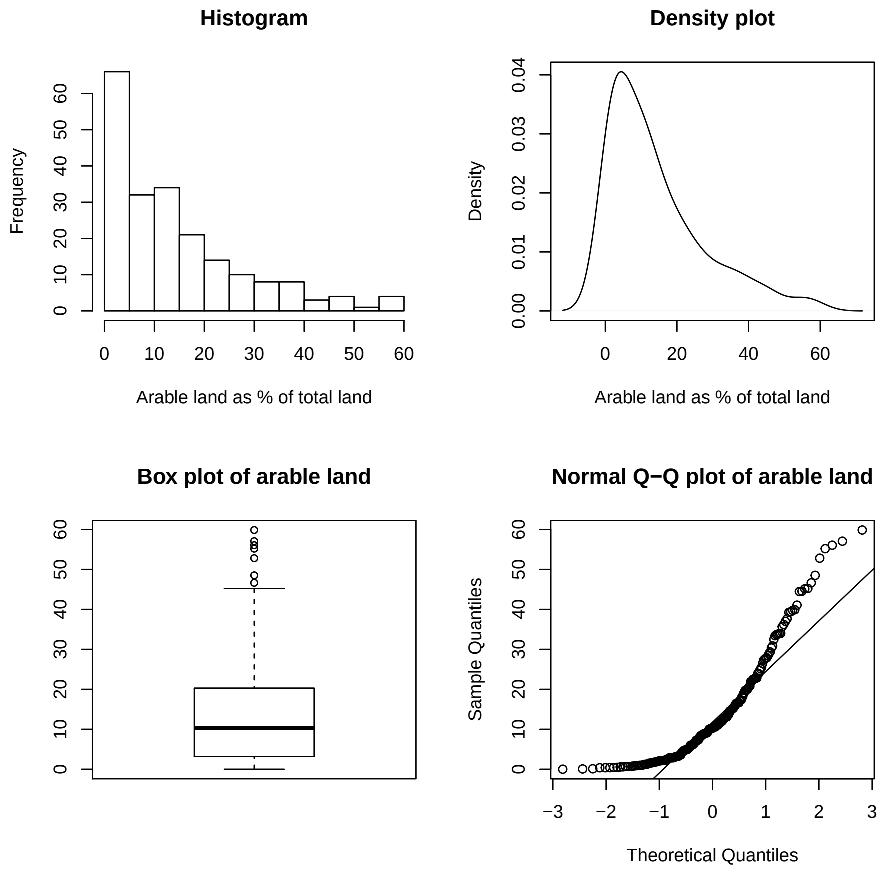
```
# install.packages("WDI")
library(WDI)

## Loading required package:  RJSONIO

d_land <- WDI(indicator = c("AG.LND.ARBL.ZS", "NY.GDP.PCAP.KD"),
              start=2010, end=2010, extra=TRUE)
d_land <- d_land[d_land$region != "Aggregates", ]

# Rename column
colnames(d_land)[colnames(d_land) == "AG.LND.ARBL.ZS"] <- "arable_land_pct"
colnames(d_land)[colnames(d_land) == "NY.GDP.PCAP.KD"] <- "gdp_percapita"

xlabel <- "Arable land as % of total land"
par(mfrow=c(2, 2))
hist(d_land$arable_land_pct, main = "Histogram", xlab = xlabel)
plot(density(d_land$arable_land_pct, na.rm = TRUE), main = "Density plot", xlab = xlabel)
boxplot(d_land$arable_land_pct, main = "Box plot of arable land")
qqnorm(d_land$arable_land_pct, main = "Normal Q-Q plot of arable land")
qqline(d_land$arable_land_pct)
```



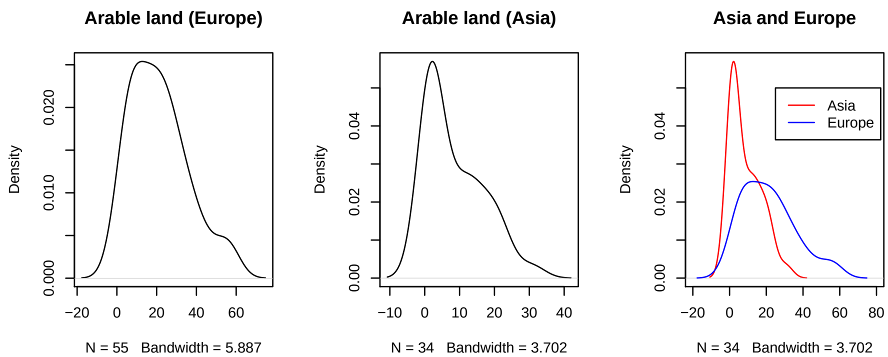
b.

(1/4 point) Plot the density plots of that variable for Europe and Asia, 1) side by side (Hint: `par(mfrow=c(?, ?))`), and 2) overlapping in the same plot.

```
par(mfrow=c(1, 3))
europe_density <- density(
  d_land[d_land$region == "Europe & Central Asia (all income levels)", "arable_land_pct"],
  na.rm=TRUE)
asia_density <- density(
  d_land[d_land$region == "East Asia & Pacific (all income levels)", "arable_land_pct"],
  na.rm=TRUE)
plot(europe_density, main = "Arable land (Europe)")
plot(asia_density, main = "Arable land (Asia)")

# Overlaying
```

```r
plot(asia_density, xlim = c(-20, 80), col='red', main = "Asia and Europe")
lines(europe_density, col='blue')
legend(25, .05, c("Asia", "Europe"),
  lty=c(1,1), # gives the legend appropriate symbols (lines)
  lwd=c(1,1),col=c("red","blue"))
```

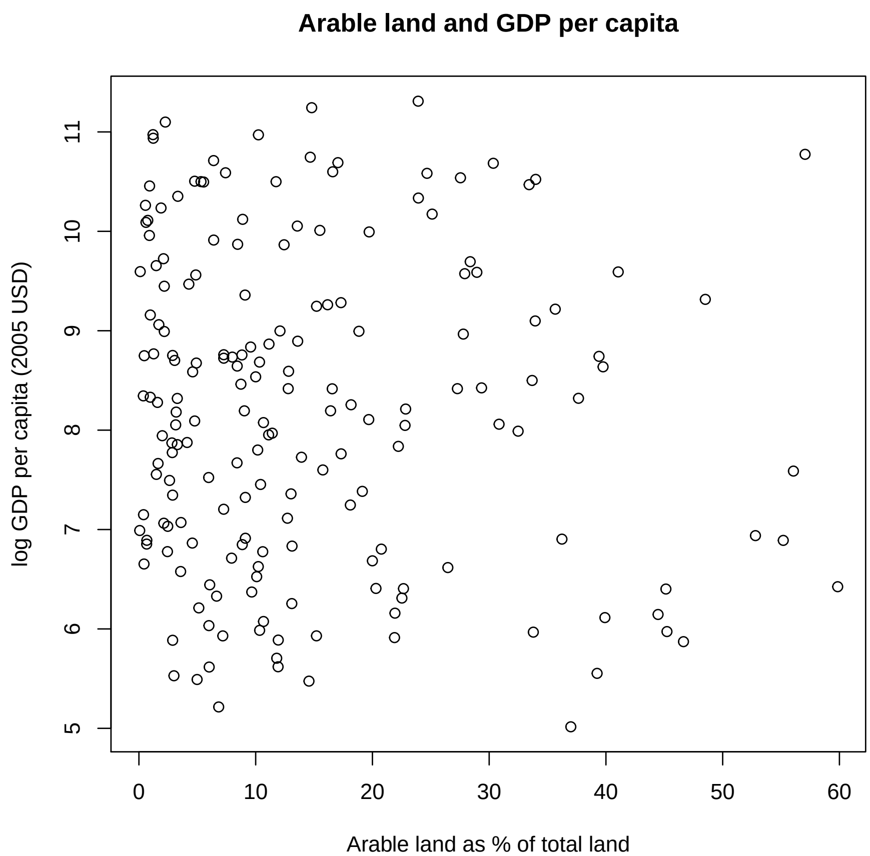



```r
# Tutorial for legend: http://www.r-bloggers.com/adding-a-legend-to-a-plot/
```

c.

(1/4 point) Draw the scatterplot of that variable against another variable.

```r
plot(d_land$arable_land_pct, log(d_land$gdp_percapita),
     xlab = "Arable land as % of total land",
     ylab = "log GDP per capita (2005 USD)",
     main = "Arable land and GDP per capita")
```



d.

(1/4 point) Label the point that represents your country (Hint: Tutorial) and color it red (Some Googling involved)

```
par(mfrow=c(1, 1))
plot(log(gdp_percapita) ~ arable_land_pct,
     data = d_land,
     xlab = "Arable land as % of total land",
     ylab = "log GDP per capita (2005 USD)",
     main = "Arable land and GDP per capita")
d_land_Vietnam <- d_land[d_land$country == "Vietnam", ]
with(d_land_Vietnam,
     text(log(gdp_percapita) ~ arable_land_pct, labels = "Vietnam",
          pos = 3, col = 'red'))
points(d_land_Vietnam$arable_land_pct, log(d_land_Vietnam$gdp_percapita),
```

```
        pch = 16, col = 'red')
```