StatR 501 HW 2 Key

Scott Rinnan and Elie Gurarie

Thursday, October 15, 2015

Contents

2) Ping-pong 1
3) Analysis of global patterns 4

2) Ping-pong

(a) Formulate a prediction regarding the appeal of table tennis between male and female students.

In general, boys seem to enjoy sports more than girls, so there might be some preference among the boys. Ping pong seems like a sport with universal appeal, so the differences might not be so stark.

(b) Present a table summarizing the total number of responses for male, female and total number of students in each of the five categories.

There are several ways to do this. Here's some compact code that generates a summary table:

```
Students <- read.csv("StudentSurvey.csv") #read the file from current working directory
p <- Students$Pingpong
s <- Students$Sex
SummaryTable <- data.frame(cbind(table(p,s),Total = table(p)))
SummaryTable
```

```
##   Female Male Total
## 1     13    6    19
## 2     14   13    27
## 3     13   15    28
## 4      4    9    13
## 5      2    4     6
```

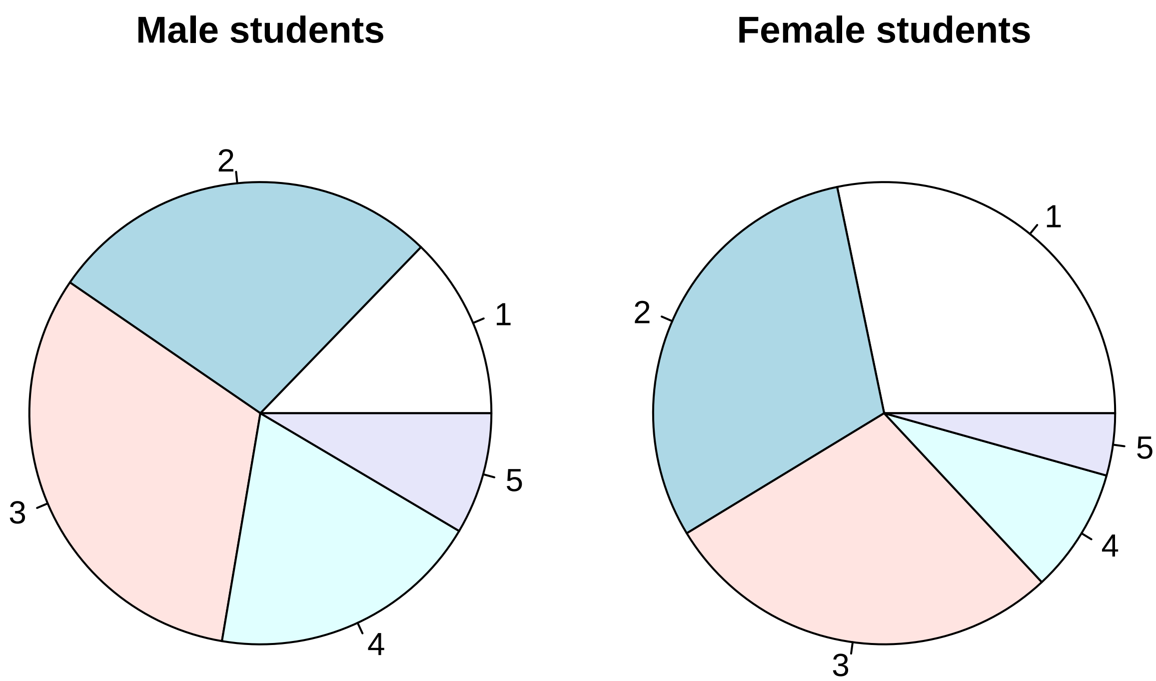
Here is an optional last row with the sex totals:

```
SummaryTable <- rbind(SummaryTable, Total = colSums(SummaryTable))
SummaryTable
```

```
##       Female Male Total
## 1         13    6    19
## 2         14   13    27
## 3         13   15    28
## 4          4    9    13
## 5          2    4     6
## Total     46   47    93
```

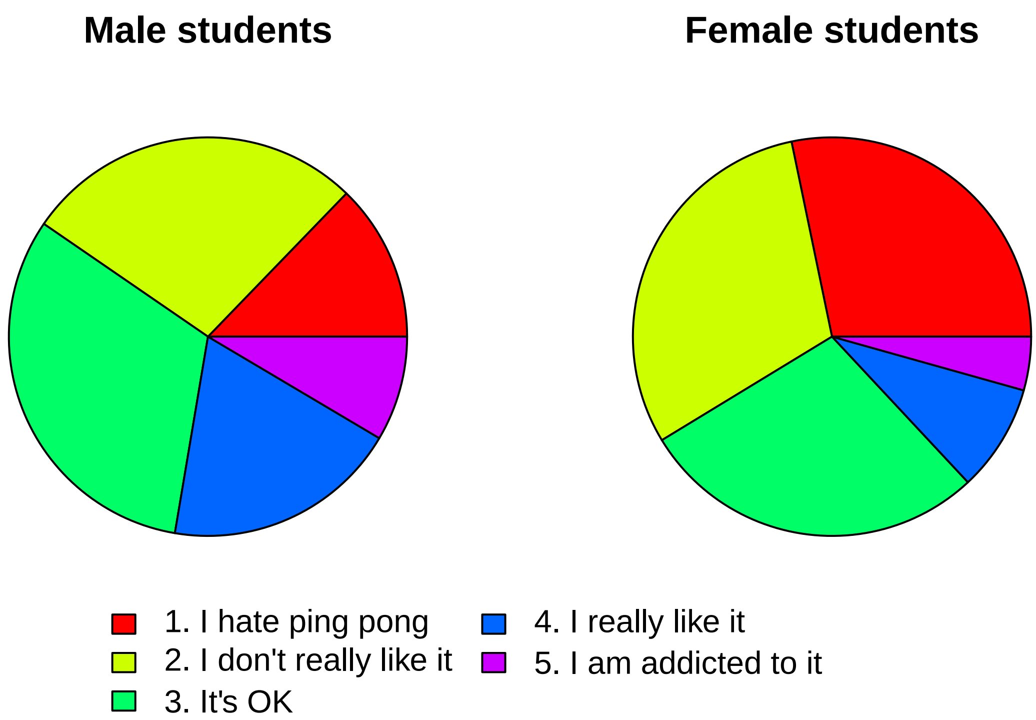
(c) Make side by side pie charts of pingpong enjoyment, one for males and one for females. Label each pie.

```
par(mfrow=c(1,2), mar=c(0,0,2,0))
pie(table(p[s=="Male"]), main="Male students")
pie(table(p[s=="Female"]), main="Female students")
```



Here is a somewhat more customized plot that looks better on the page.

```
cols <- rainbow(5)
labels <- c("1. I hate ping pong", "2. I don't really like it",
            "3. It's OK", "4. I really like it", "5. I am addicted to it")
par(mfrow=c(1,2), mar=c(4,0,2,0))
pie(table(p[s=="Male"]), main="Male students", lab="", col=cols)
legend(-.5,-1,legend=labels, fill=cols, ncol=2, adj=0, xpd=NA, bty="n")
pie(table(p[s=="Female"]), main="Female students", lab="", col=cols)
```



There's quite a bit going on here, e.g. the customized color palette using `rainbow()`, or the creation of a legend, which can only straddle both pies because of the subtle `xpd = NA` argument.

(d) Produce a 2x5 matrix (call it `M1`) summarizing the proportional distribution of male and female students in each category such that $\sum_{i=1}^{5} P_{male,i} = 1$ and $\sum_{i=1}^{5} P_{female,i} = 1$.

```
M1 <- as.matrix(table(s,p)/rowSums(table(s,p)))
M1
```

```
##         p
## s                 1          2          3          4          5
##   Female 0.28260870 0.30434783 0.28260870 0.08695652 0.04347826
##   Male   0.12765957 0.27659574 0.31914894 0.19148936 0.08510638
```

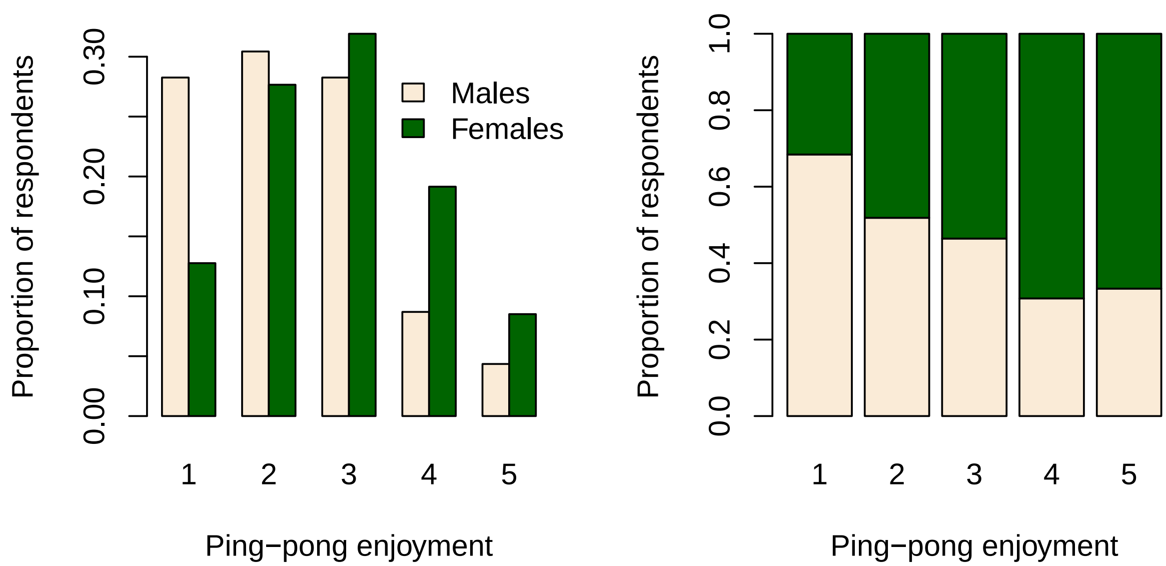
(e) Produce a 5x2 matrix (call it `M2`) summarizing the proportion for each response of male and female respondents: $P_{male,i} + P_{female,i} = 1$ for each category i.

```
M2 <- as.matrix(table(p,s))/rowSums(table(p,s))
M2
```

```
##    s
## p      Female      Male
##   1 0.6842105 0.3157895
##   2 0.5185185 0.4814815
##   3 0.4642857 0.5357143
##   4 0.3076923 0.6923077
##   5 0.3333333 0.6666667
```

(f) Produce two barplots using the following commands: `barplot(M1, beside = TRUE)` and `barplot(t(M2))`. Add a label the x-axis and customize the colors of the columns so they are not the (boring) grey default. Use the `legend()` command to add a legend identifying your unique colors with different sexes.

```r
par(mfrow=c(1,2))
cols <- c("antiquewhite", "darkgreen")
barplot(M1, beside=TRUE, col=cols,
        xlab="Ping-pong enjoyment", ylab="Proportion of respondents")
legend(9,.3, fill=cols, legend=c("Males", "Females"), bty="n",xpd=NA)
barplot(t(M2), col=cols, xlab="Ping-pong enjoyment", ylab="Proportion of respondents")
```



(g) What conclusions do you draw from these tables and plots with respect to your initial prediction? Which of the four output plots do you feel is most informative? Why?

Both male and female students showed a wide range of interest in table tennis, though there were far fewer females in the top categories, with only 10% in the 4 and 5 categories, compared to 27.5% of male students in the top two categories. It is not so easy to compare these two samples directly in the pie charts - though you can see clearly that the palest wedges are larger for the females than the males. The first barplot contains the most information, showing the relative distribution of male and female table tennis enjoyment across all categories. The second barplot is a good illustration of the trend of fewer and fewer females at high levels of enjoyment; however, it makes a somewhat artificial visual equivalence between the responses in category 5 (of which there were only 6) to responses in other categories (e.g.~category 3 has 28 responses).

3) Analysis of global patterns

(a) Load the data in CountryData.csv. Create a separate vector for each of the columns in the data.

```r
#this will read the file from the current working diretory
CountryData <- read.csv("CountryData.csv")
```

(b) Create a data frame of the 10 countries with the lowest and highest GDP per capita, the highest and lowest birth rates, and the lowest literacy. Present this as a table in your document. Comment on any patterns that you identify in these columns.

```r
Poorest <- CountryData$Country[order(CountryData$GDP)][1:10]
Richest <- CountryData$Country[order(CountryData$GDP, decreasing = TRUE)][1:10]
LeastBabies <- CountryData$Country[order(CountryData$Birthrate)][1:10]
MostBabies <- CountryData$Country[order(CountryData$Birthrate, decreasing = TRUE)][1:10]
LowestLiteracy <- CountryData$Country[order(CountryData$Literacy)][1:10]
Development <- data.frame(Poorest, Richest, LeastBabies, MostBabies, LowestLiteracy)
Development
```

```
##                           Poorest              Richest
## 1  Congo, Democratic Republic of                Qatar
## 2                         Liberia           Luxembourg
## 3                         Burundi            Singapore
## 4                        Zimbabwe               Norway
## 5                         Eritrea               Brunei
## 6        Central African Republic United Arab Emirates
## 7                           Niger        United States
## 8                    Sierra Leone            Hong Kong
## 9                          Malawi          Switzerland
## 10                           Togo          Netherlands
##                LeastBabies                    MostBabies LowestLiteracy
## 1                Hong Kong                         Niger           Mali
## 2                    Japan                          Mali    South Sudan
## 3                  Germany                        Uganda          Niger
## 4                  Andorra                   Afghanistan   Burkina Faso
## 5                    Italy                  Sierra Leone         Guinea
## 6                    Macau                  Burkina Faso           Chad
## 7                 Guernsey                       Somalia       Ethiopia
## 8                  Austria                        Angola   Sierra Leone
## 9  Bosnia and Herzegovina                       Liberia          Benin
## 10               Lithuania Congo, Democratic Republic of        Senegal
```

Note that there is considerable overlap between the countries with lowest GDP, highest birthrate, and lowest literacy, suggesting connections between these indices of development. In particular, there are 4 countries in common in the list of poorest and highest birthrate, 4 countries in common between the most babies and lowest literacy, and 2 in common for the highest birthrate and lowest literacy (Niger and Burkina Faso make all three lists). The countries on all three lists are predominantly sub-Saharan African, with only one non-African country (Afghanistan) making the list. Note that you can ask R to report these results using a combination of `match()` or `%in%` and subsampling, for example:

```r
sum(Poorest %in% MostBabies)
```

```
## [1] 4
```

```r
CountryData$Continent[match(MostBabies, CountryData$Country)]
```

```
##  [1] Africa Africa Africa Asia   Africa Africa Africa Africa Africa Africa
## Levels: Africa Asia Europe North America Oceania South America
```

Somewhat interestingly, the overlap between wealthiest countries and the those with lowest birth rate is somewhat weaker, with only Hong Kong making both lists. The wealthiest countries are distributed throughout the globe (Asia, Europe, North America, Oceania), while the fewest babies are primarily European countries at various levels of development and well-developed Asian economies.

(c) Identify the 10 countries with the highest and lowest densities, respectively, and present two tables that include their population, area and percentage of water coverage.

The easiest way to start this problem is to create a new column of CountryData called density.

```
CountryData$Density <- CountryData$Population/CountryData$Area
```

We can then use a subset of both rows (to order by density) and columns (ie those specified in the question), and subset again to the first 10 rows.

```
CountryData[order(CountryData$Density), c("Density","Population","Area","Water")][1:10,]
```

```
##        Density Population    Area Water
## 144 0.02606175      56452 2166086    NA
## 5   0.24644705       3000   12173  0.00
## 12  1.06382979         50      47  0.00
## 109 1.80480788    2822900 1564100  0.68
## 3   1.99624060     531000  266000  0.00
## 89  2.53396832    2088669  824268  0.12
## 175 2.95407750   22722835 7692024  0.76
## 162 3.09176699     318452  103000  2.67
## 90  3.09295189    1800098  582000  2.58
## 132 3.20473691     525000  163820  4.77
```

Perhaps this is less intuitive, but is slightly more computationally efficient: Taking only the first 10 rows of the order vector and using that to subset CountryData directly.

```
CountryData[order(CountryData$Density)[1:10], c("Density","Country", "Population","Area","Water")]
```

```
##        Density          Country Population    Area Water
## 144 0.02606175        Greenland      56452 2166086    NA
## 5   0.24644705 Falkland Islands       3000   12173  0.00
## 12  1.06382979 Pitcairn Islands         50      47  0.00
## 109 1.80480788         Mongolia    2822900 1564100  0.68
## 3   1.99624060   Western Sahara     531000  266000  0.00
## 89  2.53396832          Namibia    2088669  824268  0.12
## 175 2.95407750        Australia   22722835 7692024  0.76
## 162 3.09176699          Iceland     318452  103000  2.67
## 90  3.09295189         Botswana    1800098  582000  2.58
## 132 3.20473691         Suriname     525000  163820  4.77
```

The first approach generates the entire sorted data.frame and then subsets it, whereas this approach only ever generates the 10-row data.frame (but still sorts the entire CountryData$Density data.frame). If you prefer the first approach, that's fine unless you're dealing with a very large dataset.

Here's how we get the densest countries:

```
CountryData[order(CountryData$Density,decreasing=TRUE)[1:10], c("Density","Country",
                                                              "Population","Area","Water")]
```

```
##       Density      Country Population Area Water
## 16        Inf Vatican City        500    0  0.00
## 220 18560.000        Macau     556800   30  0.00
## 212 17500.000       Monaco      35000    2  0.00
## 93  12488.439     Dominica    9378818  751  0.72
## 211  7150.282    Singapore    5076700  710  1.43
## 225  6428.986    Hong Kong    7097600 1104  4.53
## 187  4906.833    Gibraltar      29441    6  0.00
## 129  1628.755      Bahrain    1234596  758  0.00
## 195  1321.544        Malta     417608  316  0.00
## 181  1195.667      Bermuda      64566   54  0.00
```

Here's another way to tackle this problem that might be even more intuitive and makes use of the %in%

operator. Again, start with generating the density column.

```
CountryData$Density <- CountryData$Population/CountryData$Area
```

Now create two vectors of the least and most dense countries.

```
LeastDense <- CountryData$Country[order(CountryData$Density)][1:10]
MostDense <- CountryData$Country[order(CountryData$Density, decreasing = TRUE)][1:10]
```

Now

```
CountryData[CountryData$Country %in% LeastDense, c("Density","Country", "Population","Area","Water")]
```

```
##         Density          Country Population    Area Water
## 3   1.99624060   Western Sahara     531000  266000  0.00
## 5   0.24644705 Falkland Islands       3000   12173  0.00
## 12  1.06382979 Pitcairn Islands         50      47  0.00
## 89  2.53396832          Namibia    2088669  824268  0.12
## 90  3.09295189         Botswana    1800098  582000  2.58
## 109 1.80480788         Mongolia    2822900 1564100  0.68
## 132 3.20473691         Suriname     525000  163820  4.77
## 144 0.02606175        Greenland      56452 2166086    NA
## 162 3.09176699          Iceland     318452  103000  2.67
## 175 2.95407750        Australia   22722835 7692024  0.76
```

```
CountryData[CountryData$Country %in% MostDense, c("Density","Country", "Population","Area","Water")]
```

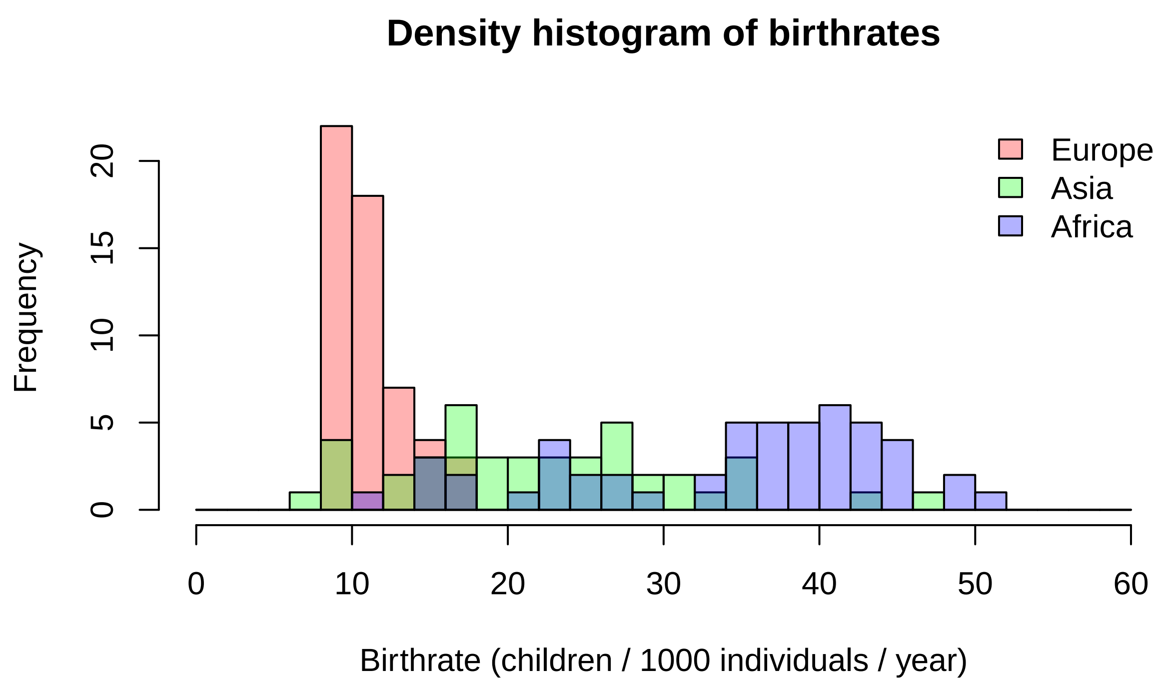
```
##        Density      Country Population Area Water
## 16         Inf Vatican City        500    0  0.00
## 93   12488.439     Dominica    9378818  751  0.72
## 129   1628.755      Bahrain    1234596  758  0.00
## 181   1195.667      Bermuda      64566   54  0.00
## 187   4906.833    Gibraltar      29441    6  0.00
## 195   1321.544        Malta     417608  316  0.00
## 211   7150.282    Singapore    5076700  710  1.43
## 212  17500.000       Monaco      35000    2  0.00
## 220  18560.000        Macau     556800   30  0.00
## 225   6428.986    Hong Kong    7097600 1104  4.53
```

Note that these results are somewhat difficult to interpret, because of the extremely large range (orders of magnitude) between the largest and smallest countries, both by area and population. Anomalous countries (like Vatican City) yield nonsensical results (infinite density). Still, generally we note that countries with the lowest densities tend to be quite arid, or inhospitably tropical, and the highest densities are in city-states and/or islands.

(d) Using the lab as a model, create an overlapping *frequency* histogram of birth rates in Europe, Asia, and Africa in three different, transparent colors. Add a legend to the plot identifying the continents. Make sure that the axes are appropriately labeled and the plot has a meaningful title. Experiment with the bin widths to find one that you feel best illustrates the patterns.

```
hist(CountryData$Birthrate[CountryData$Continent=="Europe"], breaks=seq(0,60,2), col=rgb(1,0,0,.3),
    xlab="Birthrate (children / 1000 individuals / year)",
    main="Density histogram of birthrates")
hist(CountryData$Birthrate[CountryData$Continent=="Asia"], breaks=seq(0,60,2), add=TRUE,
     col=rgb(0,1,0,.3))
hist(CountryData$Birthrate[CountryData$Continent=="Africa"], breaks=seq(0,60,2), add=TRUE,
     col=rgb(0,0,1,.3))
```

```
legend("topright", fill=rgb(c(1,0,0),c(0,1,0),c(0,0,1),.3),
       legend=c("Europe", "Asia", "Africa"), bty="n")
```

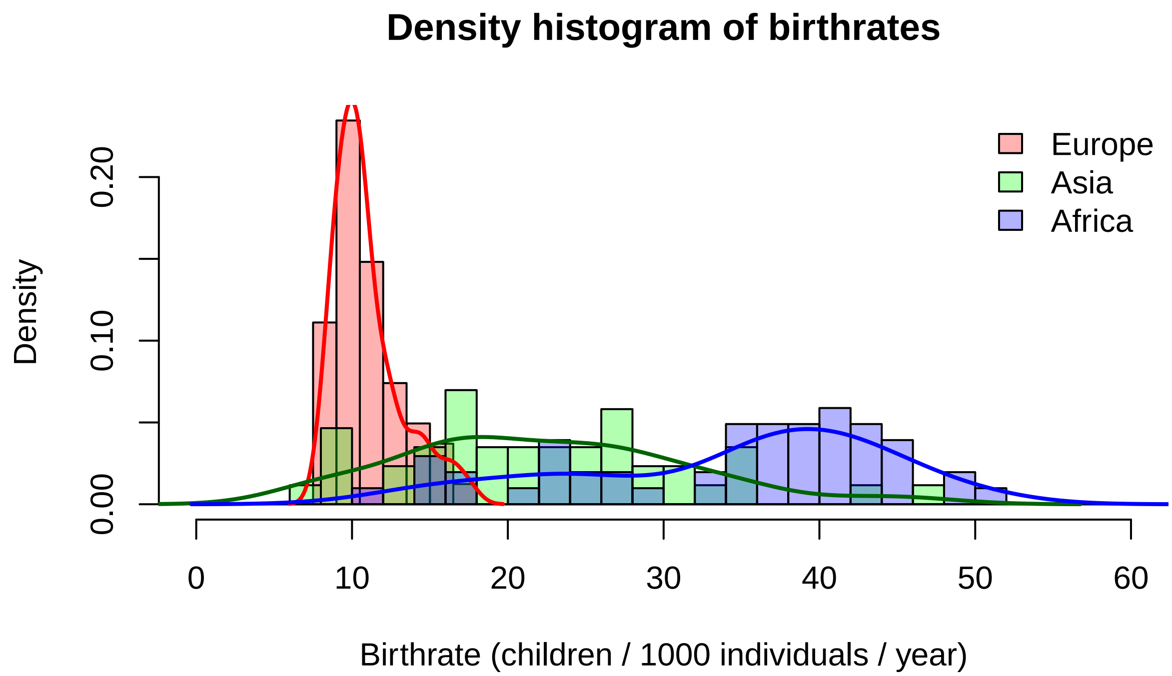


(e) Create a *density* histogram of the same data, and add fitted density lines. Note that unlike a frequency histogram, in a density histogram, the bin widths can be tuned for each individual data set.

```
hist(CountryData$Birthrate[CountryData$Continent=="Europe"], breaks=seq(0,60,1.5),
     col=rgb(1,0,0,.3), freq=FALSE,
     xlab="Birthrate (children / 1000 individuals / year)",
     main="Density histogram of birthrates")
hist(CountryData$Birthrate[CountryData$Continent=="Asia"], breaks=seq(0,60,2), add=TRUE,
     col=rgb(0,1,0,.3), freq=FALSE)
hist(CountryData$Birthrate[CountryData$Continent=="Africa"], breaks=seq(0,60,2), add=TRUE,
     col=rgb(0,0,1,.3), freq=FALSE)

lines(density(CountryData$Birthrate[CountryData$Continent=="Europe"], na.rm=TRUE), col="red", lwd=2)
lines(density(CountryData$Birthrate[CountryData$Continent=="Asia"], na.rm=TRUE), col="darkgreen", lwd=2)
lines(density(CountryData$Birthrate[CountryData$Continent=="Africa"], na.rm=TRUE), col="blue", lwd=2)

legend("topright", fill=rgb(c(1,0,0),c(0,1,0),c(0,0,1),.3),
       legend=c("Europe", "Asia", "Africa"), bty="n")
```



Note that I made the bin widths for Europe somewhat narrower, because the distribution is more concentrated.

(f) Summarize the patterns in these distribution, commenting on the center, the spread, and the modality (i.e. number of humps).

The centers of these distributions are clearly highest in Africa (around 40 children/1000 individuals), with, however, what appear to be several modes at lower birth rates. The center of the distribution is somewhat lower for Asia, with a very broad range from the lowest bin to the highest bin. Birth rates tend to be lowest in Europe (the mode is near 10 children/1000 individuals), with, additionally, a very narrow range. In no country do birth rates appears to be higher than 20.