Statistics with R – Beginner Level

Section 4

Building Charts

Lesson 21 – Histograms

```
demo <- read.csv("demographics.csv")

View(demo)

##########
### how to build a histogram with ggplot2
##########

### we will build the histogram for the variable income

### load the package

require(ggplot2)

### build the histogram
### on the y axis we will represent the counts (absolute
frequencies)

ggplot()+geom_histogram(data=demo, aes(x=income))

### change the bins color and border

ggplot()+
       geom_histogram(data=demo, aes(x=income), fill="red",
color="black")

### represent the density on the y axis (relative
frequencies)

ggplot()+
  geom_histogram(data=demo, aes(x=income, y=..density..),
fill="red", color="black")

##########

### create a facet grid (multiple histograms)
### we will create a histogram for each combination of the
variables
### gender and marital status

ggplot()+
  geom_histogram(data=demo, aes(x=income, y=..density..),
fill="red", color="black")+
  facet_grid(gender~marital)

### create multiple histograms on the same chart
### we will build a histogram for each gender category

ggplot()+
  geom_histogram(data=demo, aes(x=income, y=..density..,
fill=gender), color="black")

### N. B. the "fill" variable must be a factor
```

Lesson 22 - Cumulative Frequency Line Charts

```
demo <- read.csv("demographics.csv")

View(demo)

##########
### how to build cumulative frequency line charts with
ggplot2
##########
```

```
### we will create a cumulative frequencies line for the
variable income

### load the packages

require(ggplot2)

require(plyr)   ### we need the count function

### create a data frame with the unique income values

mydata <- count(demo, 'income')

View(mydata)

### compute the cumulative counts and percentages

cumul <- cumsum(mydata$freq)

cumperc <- cumul/nrow(demo)

### add the cumulative frequencies column to the iniatial
mydata matrix

mydata <- cbind(mydata, cumperc)

View(mydata)

### plot the cumulative frequencies line (smooth)

ggplot()+geom_line(data=mydata, aes(x=income, y=cumperc))

### OR plot the stepped line

ggplot()+geom_step(data=mydata, aes(x=income, y=cumperc))

#################

### create grouped cumulative frequencies lines
### we will build a cumulative frequencies line chart for
the variable income
### for each gender group
```

```
## first we create two databases, by gender, using the
brackets

male <- demo[demo$gender=="Male",]

female <- demo[demo$gender=="Female",]

View(male)

View(female)

### for the male data frame, we get the unique income
values
### then compute the cumulative relative frequencies

mydata_male <- count(male, "income")

cumulm <- cumsum(mydata_male$freq)

cumpercm <- cumulm/nrow(male)

### add the cumulative relative frequencies column

mydata_male <- cbind(mydata_male, cumpercm)

View(mydata_male)

### the same for the female data frame

mydata_female <- count(female, "income")

cumulf <- cumsum(mydata_female$freq)

cumpercf <- cumulf/nrow(female)

mydata_female <- cbind(mydata_female, cumpercf)

View(mydata_female)

### now we can build the chart

ggplot()+geom_line(data=mydata_male, aes(x=income,
y=cumpercm), color="red")+
  geom_line(data=mydata_female, aes(x=income, y=cumpercf),
color="blue")

### for a stepped line we must replace geom_line with
geom_step

### add a legent to the chart

lgd <- scale_color_manual("Legend", values=c(Male="red",
Female="blue"))

ggplot()+
  geom_line(data=mydata_male, aes(x=income, y=cumpercm,
color="Male"), size=1.3)+
  geom_line(data=mydata_female, aes(x=income, y=cumpercf,
color="Female"), size=1.3)+
  lgd
```

Lesson 23 - Column Charts

```
demo <- read.csv("demographics.csv")

View(demo)

##########
### how to build column charts with ggplot2
##########

### we will create a column chart representing the average
income
### for each education level

### load the package

require(ggplot2)

### build the chart

ggplot(demo, aes(x=educ, y=income, fill=educ))+
  stat_summary(fun.y=mean, geom="bar")

### if you want the same color for the bins
```

```
ggplot(demo, aes(x=educ, y=income))+
  stat_summary(fun.y=mean, geom="bar", fill="red")

### N.B. if the grouping variable is not a factor,
### we must convert it into a factor

### to create a clustered bar chart (by the variable
gender)
### position_dodge will put the columns side by side


ggplot(demo,aes(x=educ, y=income, fill=gender)) +
  stat_summary(fun.y=mean, geom="bar",
position=position_dodge())

### to stack the columns we use position_stack

ggplot(demo,aes(x=educ, y=income, fill=gender)) +
  stat_summary(fun.y=mean, geom="bar",
position=position_stack())
```

Lesson 24 - Mean Plot Charts

```
demo <- read.csv("demographics.csv")

View(demo)

##########
### how to build mean plot charts with ggplot2
##########

### we will create a mean plot representing the average
income
### for each gender category

### load the package

require(ggplot2)

### create the dataframe with the means of the gender
groups
```

```
aggdata <- aggregate(demo$income, by=list(demo$gender),
FUN=mean)

View(aggdata)

### draw the plot
### the x axis is defined as discrete, with convenient
labels

ggplot()+geom_line(data=aggdata, aes(x=(1:2),
y=aggdata$x))+
  scale_x_discrete(name="Gender", labels=c("Female",
"Male"))+
  scale_y_continuous(name="Income", limits=c(72, 85))

### change line color and thickness

ggplot()+geom_line(data=aggdata, aes(x=(1:2), y=aggdata$x),
color="red", size=1.3)+
  scale_x_discrete(name="Gender", labels=c("Female",
"Male"))+
  scale_y_continuous(name="Income", limits=c(72, 85))

##########

### build the chart with a polytomous factor
### (we'll get a broken line)

### the factor will be education level (educ)

aggdata <- aggregate(demo$income, by=list(demo$educ),
FUN=mean)

View(aggdata)


ggplot()+geom_line(data=aggdata, aes(x=(1:5),
y=aggdata$x))+
  scale_x_discrete(name="Education Level",
labels=c("College degree", "Did not complete high school",
"High school degree", "Post-undergraduate degree", "Some
college"))+
  scale_y_continuous(name="Income", limits=c(64, 116))
```

```
##############

### build a grouped mean plot
### the grouping variable will be the car category (carcat)

## create three data frames for the economy, standard and
luxury cars

demo_ec <- demo[demo$carcat=="Economy",]

demo_st <- demo[demo$carcat=="Standard",]

demo_lu <- demo[demo$carcat=="Luxury",]

# compute the mean income for each education level and for
each data frame (car category)

agg_ec <- aggregate(demo_ec$income, by=list(demo_ec$educ),
FUN=mean)

agg_st <- aggregate(demo_st$income, by=list(demo_st$educ),
FUN=mean)

agg_lu <- aggregate(demo_lu$income, by=list(demo_lu$educ),
FUN=mean)

View(agg_ec)

View(agg_st)

View(agg_lu)

## plot the three lines on the same graph

ggplot()+
  geom_line(data=agg_ec, aes(x=(1:5), y=agg_ec$x),
color="green")+
  geom_line(data=agg_st, aes(x=(1:5), y=agg_st$x),
color="red")+
  geom_line(data=agg_lu, aes(x=(1:5), y=agg_lu$x),
color="blue")+
  scale_x_discrete(name="Education Level",
labels=c("College degree", "Did not complete high school",
"High school degree", "Post-undergraduate degree", "Some
college"))+
  scale_y_continuous(name="Income", limits=c(15, 220))


## add legend

lgd <- scale_color_manual(name="Legend",
values=c(Economy="green", Standard="red", Luxury="blue"))

ggplot()+
  geom_line(data=agg_ec, aes(x=(1:5), y=agg_ec$x,
color="Economy"))+
  geom_line(data=agg_st, aes(x=(1:5), y=agg_st$x,
color="Standard"))+
  geom_line(data=agg_lu, aes(x=(1:5), y=agg_lu$x,
color="Luxury"))+
  scale_x_discrete(name="Education Level",
labels=c("College degree", "Did not complete high school",
"High school degree", "Post-undergraduate degree", "Some
college"))+
  scale_y_continuous(name="Income", limits=c(15, 220))+lgd

### N.B. in the code above, the color argument is found in
the aesthetics section
```

Lesson 25 - Scatterplot Charts

```
hw <- read.csv("hw.csv")

View(hw)

##########
### how to build scatterplot charts with ggplot2
##########

### we will create a scatterplot with the variables
### height and weight

### load the package

require(ggplot2)
```

```
### create the plot

ggplot()+geom_point(data=hw, aes(x=height, y=weight))+
  scale_x_continuous(limits=c(150,193))

#########

### build a clustered scatterplot
### by gender

lgd <- hw$gender

### get points of different colors

ggplot()+geom_point(data=hw, aes(x=height, y=weight,
color=lgd))+
  scale_x_continuous(limits=c(150,193))

### get points of different shapes

ggplot()+geom_point(data=hw, aes(x=height, y=weight,
shape=lgd))+
  scale_x_continuous(limits=c(150,193))

### get points of both different shapes and colors

ggplot()+geom_point(data=hw, aes(x=height, y=weight,
shape=lgd, color=lgd))+
  scale_x_continuous(limits=c(150,193))

###########

### add a trendline to the scatterplot

### create a linear model
### with weight as the dependent variable and height as the
explainer

model <- lm(weight~height, data=hw)

print(model)

### get the minimum and the maximum height
```

```
minh <- min(hw$height)

maxh <- max(hw$height)

### create a new vector height

height <- c(minh, maxh)

print(height)

### predict the weight based on the height, with the model
above

fit <- predict(model, data.frame(height))

print(fit)

### create a data frame with the line end points

endpoints <- data.frame(height, fit)

View(endpoints)

### build the scatter plot with trend line

ggplot()+
  geom_point(data=hw, aes(x=height, y=weight))+
  geom_line(data=endpoints, aes(x=height, y=fit),
color="red", size=1)
```

Lesson 26 - Boxplot Charts

```
demo <- read.csv("demographics.csv")

View(demo)

##########
### how to build boxplot charts with ggplot2
##########

### we will create a boxplot for the variable income
```

```r
### for each gender category

### load the package

require(ggplot2)

### create the plot

ggplot()+geom_boxplot(data=demo, aes(x=gender, y=income))+
  scale_x_discrete(labels=c("Female", "Male"))

### N.B. if the grouping variable is not a factor
### make sure you convert it into a factor first

### set the color of the outliers

ggplot()+geom_boxplot(data=demo, aes(x=gender, y=income),
outlier.colour="red")+
  scale_x_discrete(labels=c("Female", "Male"))

### set the shape of the outliers

ggplot()+geom_boxplot(data=demo, aes(x=gender, y=income),
outlier.colour="red", outlier.shape=4)+
  scale_x_discrete(labels=c("Female", "Male"))

### add a legend

lgd <- demo$gender

ggplot()+geom_boxplot(data=demo, aes(x=gender, y=income,
fill=lgd), outlier.colour="red")+
  scale_x_discrete(labels=c("Female", "Male"))

###########

### build a clustered boxplot
### we will group the boxplots by gender and marital status
### the legend will represent the two marital statuses

lgd <- demo$marital

ggplot()+geom_boxplot(data=demo, aes(x=gender, y=income,
fill=lgd))+
  scale_x_discrete(labels=c("Female", "Male"))
```