STA 141A
Fundamentals of Statistical Data Science

Fall 2016

Instructor: Debashis Paul

Lecture 9

Plotting categorical data using **vcd** package

- We can use a mosaic plot to display the frequency distribution for a categorical data.
- In a mosaic plot, frequencies in a multidimensional contingency table are represented by nested rectangular regions whose areas are proportional to the cell frequencies.

```
ftable(Titanic)  # plot "flattened" table Titanic
library(vcd)
mosaic(Titanic, shade = TRUE, legend = TRUE)
mosaic(~ Class + Sex + Age + Survived, data = Titanic, shade = TRUE, legend = TRUE)
# order of variables (categorical) in the formula above determine the nesting of rectangles
mosaic(~ Class + Age + Survived, data = Titanic, shade = TRUE, legend = TRUE)
# plot corresponding to a subset of variables
```

Layered graphics using **ggplot2**

- **ggplot2** package uses a "grammar" of graphics to display graphical data summaries.
- Main components of the display:

1. **Aesthetics :** attributes used in defining the visual perception of the graph; defined thorough a set of mappings relating the variables in a data frame to different aesthetic features of the graph such as position, color, shape and size.
2. **Scales :** used to convert the units of the variables to physical units (pixels and color) used in the display.
3. **Geometric objects or geoms :** used to represent the features in a plot in the final rendering of the graph.
4. **Statistical transformations or stats :** various statistical summaries and transformations that can be added to existing plots to provide further information or to enhance interpretability.
5. **Layers :** used to construct the plots by combining different pieces of graphical objects as separate entities.
6. **Facets :** grid of panels used to display of graph of various strata formed by conditioning on the variables

Use of qplot( ) for bivariate scatterplot

```
data(mtcars)  # cars data used earlier
library(ggplot2)
qplot(wt, mpg, data = mtcars, color = I("red"), size = I(2))  # scatterplot of mpg vs wt
qplot(wt, mpg, data = mtcars, color = cyl, size = I(2))  # color values determined by values of
# the discrete variable cyl = # of cylinders
qplot(wt, mpg, data = mtcars, colour = factor(cyl), size = I(2)) # cyl used as factor
# add a scatterplot smoother and corresponding confidence band per level of the factor cyl
qplot(wt, mpg, data = mtcars, colour = factor(cyl), geom = c("point","smooth"))
# same as above, but the smoother is added as an additional layer
qplot(wt, mpg, data = mtcars, colour = factor(cyl)) +  geom_smooth( )
```

Addition of features in several layers

```
# add a scatterplot smoother for the whole data but color the points for different levels of cyl
qplot(wt, mpg, data = mtcars, colour = cyl) +  geom_smooth(span=2)
# add a scatterplot smoother for each level of cyl, and add corresponding confidence band
qplot(wt, mpg, data = mtcars, colour = factor(cyl)) +  geom_smooth(span=2)
# fit regression line and plot confidence bands for the whole data (no subsetting)
qplot(wt, mpg, data = mtcars, colour = cyl) +  geom_smooth(method = "lm")
# fit regression line for each level of cyl and plot corresponding confidence bands
qplot(wt, mpg, data = mtcars, colour = factor(cyl)) +  geom_smooth(method = "lm")
# also adjust the thickness and transparency of the regression line
qplot(wt, mpg, data = mtcars, colour = cyl)  +  geom_smooth(method = "lm",size=2,alpha = I(1/5))
```

Regression (cont.)

```
# in addition, make the size of the dots proportional to the number of gears (levels of variable gear);
# also add labels to the points
mtcars$label=row.names(mtcars)
qplot(wt, mpg, label=label, data = mtcars)  +
        geom_point(colour = cyl, size = gear)  +
        geom_smooth(method = "lm", alpha = I(1/5)) +  geom_text(size=2)
```

Histograms

```
# plot histogram of mpg with specified binwidth
qplot(mpg, data = mtcars, geom="histogram", binwidth=3)
# plot relative frequencies rather than counts by setting ..density.. as an argument
qplot(mpg, ..density.., data = mtcars, geom="histogram",binwidth=3)
# plot relative frequencies rather than counts by setting ..density.. as an argument
qplot(mpg, ..density.., data = mtcars, geom="histogram",binwidth=3)
# histogram subdivided according to the levels of factor(cyl)
qplot(mpg, data = mtcars, geom="histogram", binwidth=3, fill=factor(cyl))
# plot a kernel density estimate for mpg, bandwidth (bw) = 1.5, xlim sets the range of values
qplot(mpg, data = mtcars, geom="density",bw=1.5,xlim=c(5,50))
```

Kernel density estimate

- We can use density geom to plot a kernel density estimate of the data rather than a histogram

```
# plot a kernel density estimate for mpg, bandwidth (bw) = 1.5, xlim sets the range of values
qplot(mpg, data = mtcars, geom="density",bw=1.5,xlim=c(5,50))
# plot the kernel density estimates for mpg for the strata created by levels of factor(cyl)
qplot(mpg, data = mtcars, geom="density",bw=1.5,colour=factor(cyl),xlim=c(5,50))
# plot the density estimates for different strata in separate panels using the argument facets
# the formulate facets = . ~ cyl determines a lattice of plots with
# number of columns corresponding to the number of distinct values of cyl
qplot(mpg, data = mtcars, geom="density",bw=1.5,xlim=c(5,50),facets = . ~ cyl)
```

Which summary is appropriate ?

- When two are more variables are being considered simultaneously, for numerical summary, use *frequency table*. For graphical summary, may use *mosaic plot*.
- When considering two variables that are both numerical, can use *scatter plot*, *bivariate histogram* (when there is a lot of data), scatterplot smoother such as *lowess regression*, or *simple linear regression* (possibly after some transformation of the variables).
- For many numerical variables at the same time (multivariate data), can apply the procedure above for pairs of variables. Also, can use *correlogram* for an overall summary.
- When one variable is categorical, the other numerical, usually *facetting* (conditioning on the values of the categorical variable, while displaying summaries) is helpful. Also, you can use *subdivided boxplot* for a quick comparison of the distribution of the numerical variable against different levels of the categorical variable.