Getting to Know Your Data: Graphics, Commercial Software and R

B. Dudek
School of Social Welfare Workshop
4-29-09
“If you haven’t done a graph, then you haven’t done an analysis”

*John Tukey*
“Graphics is for showing the obvious to the ignorant“
- a quote that I’ve inherited from instructors years ago.
- don’t know who to attribute it to.

An old slam that still has vestiges of adherence in some circles........Perhaps growing circles where the "modelers" would seem to prefer an endless litany of tables of coefficients, error components, bootstrapped parameters, etc.
Homer: Here’s good news! According to this eye-catching article, SAT scores are declining at a slower rate!

Lisa: Dad, I think this paper is a flimsy hodgepodge of pie graphs, factoids and Larry King.

Homer: Hey, this is the only paper in America that’s not afraid to tell the truth, that everything is just fine.

*The Simpsons*

*Homer Defined*
Questions

• What?
  – the story, the data, the model

• Why?
  – what is the point you are making?

• Who?
  – the audience

• How?
  – how will the audience view it?

• When?
  – how much time do you have?
Ultimate Question

What is the most efficient set of tools for me?

“The best software is the software I know the best.”

But it is easy to ignore newer options since the array is growing so rapidly.

Locked in to your old software? May limit creativity.
Visualization of Information

- translate symbolic into geometric
- build on natural perception processes
- the translation should utilize visual cognitive processes that enable
  - comparison
  - pattern recognition
  - detection of change

- Based on Bertin, Chambers, Cleveland, Pinker, Wilkinson
- We could formalize this with a treatment of visual perception issues, but I’ll skip that (a few slides at the end, if there is interest)
Tufte Principles of Graphing Excellence

a. the well-designed presentation of interesting data – a matter of *substance*, of *statistics*, and of *design*.

b. consists of complex ideas communicated with clarity, precision, and efficiency

c. that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space
Tufte Axioms

• Show the data.
• Induce the viewer to think about the substance rather and about methodology, technology, and design of the graphic.
• Avoid distorting what the data have to say.
• Present many numbers in a small space.
• Eliminate the Lie Factor.
• Make large data sets coherent.
• Encourage the eye to look at different pieces of the data.
• Reveal the data at several levels of detail – broad overview to fine structure.
• Serve a clear purpose.
• Don’t succumb to ppt (like I have here)
Cleveland Principles

- Clear vision
  - make the data stand out. avoid superfluity
  - visually prominence
  - avoid clutter
  - ........
- Clear understanding
  - major conclusions in graphical form
  - ..... 
- Banking to $45^0$
  - aspect ratios crucial
  - ....
- Care in choice of axes, scales, and tick marks, symbols
- Visual decision making based on knowledge of perceptual processes
  - visual perception from
    - detection
    - assembly
    - estimation
  - pay attention to perceptual hierarchies
Wilkinson Principles

• Deep structure
• “plot types” not a useful construct
• Graphics Grammar Reader Model
  – rooted in linguistics-like cognitive psychology framework from Bertin, Kosslyn, Pinker, Cleveland and others
    • constructing graphs is encoding data
    • but “readers” must decode the graph back into data
  – leads to an affinity with object-oriented programming approaches to graphics
From Wilkinson, 2005

**Figure 19.3** A graphics grammar reader model
Visual Perception Issues

Mackinlay, 1988, based on Cleveland and McGill
The Good

Florence Nightingale on mortality patterns in the Crimean war.

Instantaneous conveyance of a substantial amount of information.
The Good
Linkage Disequilibrium and Haplotype Blocks in a Gene Associated with Autism

Campbell, et al, 2006, probably used “Haploview” software

Fig. 1. MET locus genomic structure, genotyping markers, and definition of haplotype blocks. The MET locus consists of 21 exons spanning 125-kb on chromosome 7q31. Nine SNPs spanning the MET locus were chosen to perform association studies and Taqman Assays-on-Demand were used to determine genotype. The nine genotyping markers defined two distinct linkage disequilibrium blocks. Pairwise linkage disequilibrium (r²) values are indicated. Pairwise r² values are provided in Table 1, which is published as supporting information on the PNAS web site.
The Bad

Yikes!
Did we really need a graph for this?
The Bad (Excel)

Yuk!

Double Yuk!

http://www.softplatz.com/Soft/Network-Internet/Other/3D-Multi-Series-Line-Graph.html
Don’t use Pie Charts

from: http://biostat.mc.vanderbilt.edu/twiki/pub/Main/StatGraphCourse/TEB.pdf
What is worse than a pie chart?

Two, comparison pie charts!
What is worse than two pie charts?

faux 3D exploded pie charts
What is even worse than 3D pie charts?

Doughnut Charts
The Ugly

Tufte’s “worst graph ever.”

From the American Education Magazine (1977?)

The graph presents only five values.
The Uninformed

NY Times Graphic.

What is wrong with this?

From Friendly’s Gallery of Data Visualization
http://www.math.yorku.ca/SCS/Gallery/

*Poll results for these countries were from 1995.
The highlight suggests positive residuals where those countries are “happier” than they should be, based on economics.

But only if you presume a linear model.

Re-expression on a log scale would change the appearance and conclusion dramatically.

From Friendly’s Gallery of Data Visualization
http://www.math.yorku.ca/SCS/Gallery/
Misleading

From “Financial Times

Goal was examination of shrinkage in market values of Banks in late 2008 into 2009

Problem: When data are diameters/radii, this means the areas in the circles don’t scale correctly with the variable
Misleading - fixed

From “Financial Times
Corrected Graph after readers pointed out the error

Alternative graph from R Graph Gallery:
http://addictedtor.free.fr/graphiques/RGraphGallery.php?graph=150
**Misleading**

<table>
<thead>
<tr>
<th>FAMILY INCOME in 2008 dollars</th>
<th>McCAIN</th>
<th>AVERAGE CHANGE IN TAXES</th>
<th>OBAMA</th>
<th>AVERAGE CHANGE IN TAXES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Above $2.87 million (top 0.1%)</td>
<td>-4.4%</td>
<td>-$269,364</td>
<td>+11.5%</td>
<td>+$701,885</td>
</tr>
<tr>
<td>$603,403 to $2.87 million (top 1%)</td>
<td>-3.4%</td>
<td>-$45,361</td>
<td>+8.7</td>
<td>+$115,974</td>
</tr>
<tr>
<td>$226,982 to $603,402</td>
<td>-3.1%</td>
<td>-$7,871</td>
<td>0</td>
<td>-$12</td>
</tr>
<tr>
<td>$160,973 to $226,981</td>
<td>-3.0%</td>
<td>-$4,380</td>
<td>-1.9</td>
<td>-$2,789</td>
</tr>
<tr>
<td>$111,646 to $160,972</td>
<td>-2.5%</td>
<td>-$2,614</td>
<td>-2.1</td>
<td>-$2,204</td>
</tr>
<tr>
<td>$66,355 to $111,645</td>
<td>-1.4%</td>
<td>-$1,009</td>
<td>-1.8</td>
<td>-$1,290</td>
</tr>
<tr>
<td>$37,596 to $66,354</td>
<td>-0.7%</td>
<td>-$319</td>
<td>2.4</td>
<td>-$1,042</td>
</tr>
<tr>
<td>$18,982 to $37,595</td>
<td>-0.5%</td>
<td>-$113</td>
<td>-3.6</td>
<td>-$892</td>
</tr>
<tr>
<td>Up to $18,981</td>
<td>-0.2%</td>
<td>-$19</td>
<td>-5.5</td>
<td>-$567</td>
</tr>
<tr>
<td><strong>Average cut:</strong></td>
<td>-2%</td>
<td>-$1,195</td>
<td>-0.3%</td>
<td>-$160</td>
</tr>
</tbody>
</table>

http://www.washingtonpost.com/wp-dyn/content/story/2008/06/09/ST2008060900950.html
Misleading - Fixed

[Chart showing tax plan comparisons between McCain and Obama]

Tax Plan data from Washington Post reporting of Tax Policy Center analysis. Redrawn to scale with height of bars corresponding to population of each group, as given in original TPC data.

http://chartjunk.karmanaut.com/taxplans/
Don’t forget colorblindness issues - red-green colorblindness illustrated here.

Normal  Protanopia  Deuteranopia
Functions of Graphs

Data Display

Analysis
- Exploration (EDA?)
- Investigation
- Model Display
- Model Comparison
- Diagnostics

Presentation
- Inform
- Persuade
- Stimulate
- “Sell”

Publication
- Efficiency
- Objectivity

Objectives

Principles & Issues
- Efficiency
- Objectivity
- and All of Tufte’s principles
- Ease of Use
- Speed
- Detection
- Compare
- Flexibility
- Perception
- Exposition
- Aesthetics
Did anyone notice anything?

- This permits me to refer to Tufte’s mission to eradicate mindless Powerpoint usage
- Style and colors on previous slide engender individual differences in response
- Simpler would be better
Functions of Graphs

Objectives:
- Exploration (EDA?)
- Investigation
- Model Display
- Model Comparison
- Diagnostics

Presentation:
- Inform
- Persuade
- Stimulate
- “Sell”

Publication:
- Efficiency
- Objectivity

Principles & Issues:
- Ease of Use
- Speed
- Detection
- Compare
- Flexibility

- Perception
- Exposition
- Aesthetics

All of Tufte’s principles
One tool fits all?

Probably not

Factors:
- Economics
- Time
- Skill
- Goals
- Characteristics of the data set
Software Options: Statistics Packages

- **SPSS**
  - Many chart options
  - Graphics can be edited
  - Can export vector graphics
  - Newest versions have a much more powerful graphics engine
  - Expensive

- **SAS**
  - Historically known for poor graphics.
  - Good graphs can be produced with the base package after some training
  - New ODS delivery system is much improved and functional
  - Expensive

- Publication quality of these products??????
Software Options: Statistics Packages

- **Systat**
  - Strong graphics; Wilkinson influence
- **Stata**
  - I’m ignorant. Any comments?
- **BMD?** Anyone still using it?
- **Eviews** – econometrics, but general purpose
- **JMP.** I not yet seen it.
- **DataDesk**
- **Statgraphics**
- **Others** – plenty
Software Options: Statistics Packages

- **R**
  - It is free!
  - Extensive graphics capabilities
  - Unique capabilities in contributed “packages”
  - Export in many formats
  - Infinitely customizable
  - Difficult for the novice statistician / programmer
  - It is free!

- **S/S+**
  - The parent software of R
Software Options:
Mathematics and Engineering Packages

- Mathematica
- MATLAB
- Maple
- DADiSP
Layout/Illustration/Draw Software

- Corel Draw
- The suite of Adobe software (e.g., Illustrator)
- Pagemaker
- etc
Data/Graphics oriented software

- SigmaPlot
- Origin
- Prism
- Epi-info
- Spotfire
- Iris Explorer
- ILog Discovery
- GNUPlot
- GraphPad
- Kaleidascoppe
- Powercalc-GX

- Excel? I have a bad negative bias here

- What have I missed?
  - GIS stuff
  - probably all the Apple/Mac stuff
  - Geoscience, atmospheric science, bioinformatics – specialized software

- Focus here is on social science and some of the life sciences
SigmaPlot (My favorite)

- Can be used with MS products and SPSS
- Opens other data formats
- Menu driven
- Multiple and extensive graphics options
- Curve fitting, functions, smoothing
- Complete control over all graph components
- Easily produces compound graphics
- Exports graphics in multiple formats
- Goal is publication and presentation quality
One Issue

• Presentation and Publication often require care in understanding vector based graphics vs bitmapped graphics.
• The former scales well and is usually required for printing (e.g., ps, pdf, eps, wmf)
• The latter does not scale well thus increasing or decreasing the size often produces undesirable loss of resolution
• What can your software do?
Let’s Explore with R

The R Project for Statistical Computing

Getting Started:

- R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To download R, please choose your preferred CRAN mirror.
- If you have questions about R, like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

Notes:

- R version 2.8.1 has been released on 2008-12-22.
- R News 8.2 has been published on 2008-11-03.
- BIOSC 2009, The 6th workshop on Directions in Statistical Computing, will be held at the Center for Health and Society, University of Copenhagen, Denmark, July 13-14, 2009.
- useR! 2008, has been held at DTU campus, Germany, August 12-14, 2008.

This server is hosted by the Department of Statistics and Mathematics of the Stuttgart University.
# built in to the base R installation
# are several packages that contain
# useful data sets
# one of these is the "mtcars" data set

attach(mtcars)  # load a package that contains a useful graphics function
library(Hmisc)  # examine the structure of the mtcars data frame
str(mtcars)     # list the values of the mtcars variables
mtcars          # plot all data
datadensity(mtcars)
Univariate Description in R

A few graphs from a mouse study, some textbook illustrations, some data sets in R packages, and some data sets available online.

# Opening an SPSS system file:
require(foreign)
mouse1 <- read.spss("mastersmall2_R_nomissing.sav", use.value.labels=TRUE, max.value.labels=Inf, to.data.frame=TRUE)

hist(mouse1$dist15)

# Fill with Color. Histogram with Different Number of Bins
hist(mouse1$dist15, breaks=64, col="red")

# Add a Normal Curve (Based on Dalgaard's book)
x <- mouse1$dist15
h<-hist(x, breaks=64, col="red", xlab="Distance Traveled (cm)",
(main="Histogram with Normal Curve")
xfit<-seq(min(x),max(x),length=64)
yfit<-dnorm(xfit,mean=mean(x),sd=sd(x))
yfit <- yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit, col="blue", lwd=2)
# Simple Kernel Density Plot of Mouse Body Weight

d <- density(mouse1$bwcoc)
plot(d, main="Kernal Density of Mouse Body Weight (g)"

# Filled Kernel Density Plot

d <- density(mouse1$bwcoc)
plot(d, main="Kernal Density of Mouse Body Weight (g)"
polygon(d, col="gold", border="darkgreen")

N = 377   Bandwidth = 0.8363
# More Complicated and More Useful?
# this illustration first draws the histogram,
# then adds the kernel density function,
# and then places a new figure on the first by using a strategy
# for dividing the figures display space and locating the new
# graph with the fig parameter
hist( mouse1$bwcoc,
    col = "light blue",
    probability = "TRUE", breaks=36, xlim = c(14,38),
    ylim = c(0, 1.5*max(density(mouse1$bwcoc)$y)))
lines(density(mouse1$bwcoc),
    col = "red",
    lwd = 3)
op <- par(fig = c(.5,.88,.35,.83),
    new = TRUE)
qqnorm(mouse1$bwcoc,
       xlab = "", ylab = "", main = "",
       axes = FALSE)
qqline(mouse1$bwcoc, col = "red", lwd = 2)
box(lwd=2)
par(op)
# Boxplot of one variable:
boxplot(mouse1$bwcoc)

# Notched Boxplot of Bodyweight against 2 Crossed Factors
# boxes colored for ease of interpretation
boxplot(bwcoc~sex*site, data=mouse1, notch=TRUE, col=(c("gold","darkgreen")), main="Distance Traveled", xlab="Gender and Site")
# A more interesting Boxplot: one DV by an IV with two levels
# Notched and using color as well as other boxplot specs
# Plus rugplots of the individual scores of each group
# shown at right, here

boxplot(mouse1$bwcoc ~ mouse1$sex, boxwex = 0.35, ylab = "Body Weight", col="cornsilk", border="black", notch=TRUE, range=1.5, ylim=c(14.75,33))
fem <- mouse1$sex == "FEMALE"
male <- mouse1$sex == "MALE"
rug(mouse1$bwcoc[fem], side = 2)
rug(mouse1$bwcoc[male], side = 4)
# compare groups with kernel density functions separately done on body weight of males and females
# first, make sure the "sm" library is loaded.
library(sm)
# plot densities
sm.density.compare(mouse1$bwcoc, mouse1$sex, xlab="Mouse Body Weight (grams)")
title(main="Kernel Density Plots of Body Weight of Male and Female Mice")
# Multiple Box Type Plots on the Same Variable
# Illustrate partitioning of the graphic device space
# requires packages: vioplot for violinplot, hdrcde for hdrboxplot and Hmisc for bpplot
library(vioplot)
library(hdrcde)
library(Hmisc)
opar <- par(mfrow=c(1,4), mar=c(2,4,1))
# following three lines attempted to put a fifth plot first
# failed because I can't get hist or kdens to be rotated 90 degrees
#hist(bwcoc, main="Underlying ndensity")
#d <- density(mouse1$bwcoc)
#plot(d, main="Kernal Density")
#polygon(d, col="green", border="red")
boxplot(mouse1$bwcoc, col="gray90", main="standard inboxplot")
hdr.boxplot(mouse1$bwcoc, main="HDR inboxplot")
vioplot(mouse1$bwcoc, col="cornsilk")
title("violin plot")
bpplot(mouse1$bwcoc)
par(opar)
# from R gallery: http://addictedtor.free.fr/graphiques/RGraphGallery.php?graph=151
# a simple mixture of
# - N( 0, 1) with weight .4
# - N( 3, 1) with weight .6
x <- rnorm( 700 , mean = sample( c(0, 3), size = 700, prob = c(.4, .6), replace =TRUE ) )
# axis labels always horizontal
# margin smaller than default
par( la = 1, mar = c(3,3,4,1)+.1 )
# calculate histogram (but do not display)
h <- hist( x, plot = FALSE, breaks = 15 )
# calculate density estimation
d <- density( x )
# pdf("myplot.pdf")
# just set the scene and axes
plot( h, border = NA, freq = FALSE, xlab = "", ylab = "", main= "Use graduated Colors (clipping), Density, and Rugplot" )
# grab the limits of the region
usr <- par( "usr" )
ncolors <- 100
# create the colors
colors <- colorRampPalette( c("yellow","orange","red") )(ncolors)
# horizontal line
abline( h = axTicks(2) , col = "grey", lwd = .5 )
# for each color, clip into a region and redraw the histogram with that color
for( i in 1:ncolors)
    clip( usr[1], usr[2], usr[3] + (i-1) * dy, usr[3] + i*dy )
plot( h, add = TRUE, axes = FALSE, ylab = "", xlab = "",
     col = colors[i], border = NA, freq = FALSE)
# reset the clipping area. See ?clip
do.call( clip, as.list( usr) )
# just to get the boxes right
plot( h, add = TRUE, lwd = .5 , freq = FALSE, xlab = "",
     ylab = "", axes = FALSE )
# adds a tick and translucent line to display the density estimate
lines( d, lwd = 2, col = "#22222288" )
# add a rug (also using translucency)
rug( x, col = "#000088", lwd=.5 )
# box around the plot
box()
#dev.off()
Bivariate and Multivariate Description
# Now do some bivariate and multivariate plots
# Add boxplots to a scatterplot
attach(mouse1)
par(fig=c(0.0,0.8,0.0,0.8), new=TRUE)
plot(bwcoc, brain, xlab="Body Weight", ylab="Brain Weight")
abline(lm(brain~bwcoc), col="red")  # regression line (y~x)
par(fig=c(0.0,0.8,0.55,1), new=TRUE)
boxplot(bwcoc, horizontal=TRUE, axes=FALSE)
par(fig=c(0.65,1,0.0,0.8), new=TRUE)
boxplot(brain, axes=FALSE)
mtext("Scatterplot with Univariate Boxplots", side=3, outer=TRUE, line=-3)
# Scatterplot with univariate frequency histograms of each variable
# requires the `scatterplot.with.hist` script from Verzani
# x variable named first
attach(mouse1)
scatter.with.hist(bwcoc,brain)
# multiple scatterplots
# both linear and lowess functions fit
# a bit garish, but makes the point about capability
library(car)
scatterplot.matrix(~bwcoc+brain+dist15+dist15c|sex,
data=mouse1,main="Relationships by Sex",
col=terrain.colors(6, alpha =1))
# requires use of a method described below to extract only a few variables
# from the data frame
# also requires the MASS package
# modeled after illustration in http://zoonek2.free.fr/UNIX/48_R/04.html
mouse2 <- mouse1[c(7:12)]
library(MASS)
pairs(mouse2,
      gap=0,
      diag.panel = function (x, ...) {
        par(new = TRUE)
        hist(x,
             col = "light blue",
             probability = TRUE,
             axes = FALSE,
             main = "")
        lines(density(x),
              col = "blue",
              lwd = 3)
        rug(x)
      })
# bagplot surrounded by boxplots – lets use the ice cream data set

```r
library(foreign)
ice1 <- read.spss("F:/LearningR/icbase.sav", use.value.labels=TRUE, max.value.labels=Inf, to.data.frame=TRUE)
attach(ice1)

library(aplpack)
# pdf("f:/learningR/icbagplot1.pdf")
par(fig=c(0,0.8,0,0.8), new=TRUE)
bagplot(ice1$temp, ice1$ic, xlab="Temperature", ylab="Per Capita Ice Cream Consumption", col.baghull="tan", col.loophull="cornsilk",)
box()
par(fig=c(0.65,1,0,0.8),new=TRUE)
boxplot(ice1$ic, axes=FALSE, notch=TRUE, col=c("gold","darkgreen"))
par(fig=c(0.65,1,0,0.8),new=TRUE)
boxplot(ice1$ic, axes=FALSE, notch=TRUE, col=c("gold","darkgreen"))
mtext("Ice Cream Consumption and Temperature ", side=3, outer=TRUE, line=-3)
# dev.off()
```
Diagnostics
# Normal Probability Plots
qxnorm(mouse1$bwcoc) qxline(mouse1$bwcoc)

# or use qx.plot and specify comparison dist – options are norm, chisq, t,
qx.plot(mouse1$brain, dist="norm")
qx.plot(mouse1$brain, dist="chisq", df=25)
# Let's do some linear modeling
# some regression illustrations
attach(mtcars)
cor(mtcars, use="complete.obs", method="pearson")
cor.test(mpg,wt, method = "pearson",
   alternative = "two.sided",
   exact = FALSE)
fit1 <- lm(mpg~wt+cyl)
summary(fit1)
anova(fit1)
# pdf("fit1.pdf")
plot(fit1)
# dev.off()
detach(mtcars)
# examples of jittering in scatterplots from Fox text and car
library(car)
data(Vocab)attach(Vocab)

# scatter of quant vars with discrete values
plot(education,vocabulary)
# default jittering
plot(jitter(education),jitter(vocabulary))

#2nd degree jittering
plot(jitter(education,factor=2),jitter(vocabulary,factor=2))
# add linear least squares regression fit and lowess functions
abline(lm(vocabulary~education), lwd=3, lty=2)
lines(lowess(education,vocabulary, f=.2),lwd=3)
# examine scatterplots with very large data sets

detach(Vocab)
attach(SLID)

# a better, more advanced plot would
# 1. make the points smaller
# 2. jitter the points
# 3. scale grayness of points to density
# 4. add a 2d kernel density function

# the sm.density function in the sm library is used:
valid <- complete.cases(wages, education)

library(sm)
sm.density(cbind(education[valid], wages[valid]), nbins=0,
          display='slice', xlab='Education',
          ylab='Wages', col=gray(seq(1,0,length=100)))
points(jitter(education, amount=.25), wages, cex=.45)
box()
lines(lowess(education[valid], wages[valid], f=1/3), lwd=3)
abline(lm(wages~education), lty=2, lwd=3)

0 5 10 15 20
10 20 30 40 50
0 5 10 15 20
25 50 75

library(sm)
sm.density(cbind(education[valid], wages[valid]), nbins=0,
          display='slice', xlab='Education',
          ylab='Wages', col=gray(seq(1,0,length=100)))
points(jitter(education, amount=.25), wages, cex=.45)
box()
lines(lowess(education[valid], wages[valid], f=1/3), lwd=3)
abline(lm(wages~education), lty=2, lwd=3)
remove(valid)
## Two Predictor Multiple Regression Illustration from Lecture

Graphs and analyses produced in lecture by SPSS can also be generated in R

### First, read in the data file - produce an R DataFrame
library(UsingR) # need this for generating violinplots; all other functions are in base package
cohen1 <- read.table("cohench3.csv", header=TRUE, sep=" ", na.strings="NA", dec=".", strip.white=TRUE)
attach(cohen1)

### Section I  Distributional Displays for salary, pubs, and cits

win.graph(8,6) #optional 4 graphs/page
layout(matrix(c(1,2,3,4),2,2))

hist(salary,probability=TRUE,breaks=6,
ylim= c(0, 1.2*max(density(salary)$y)),
main="Histogram/Density Salary")
lines(density(salary),col= "red",lwd = 3)
qqnorm(salary,main="Normal Probability Plot of Salary");qqline(salary)

boxplot(salary,xlab="Salary",ylab="")

violinplot(salary,col="cornsilk",names="Salary")
Some Packages have built in multiple graph functions

# Simple linear least squares regression with # built in multiple graph production # Illustration using Verzani's package library(UsingR) data(home) attach(home) x = old; y = new simple.lm(x,y) lm(formula = y ~ x)

lm.res = simple.lm(x,y) coef(lm.res) simple.lm(x,y,show.residuals=TRUE)
# Bootstrapping capability in R
# illustrate with a capability to bootstrap qq.plot CI
# using residuals from a linear model
# return to the MTCARS illustration
library(car)
attach(mtcars)
fit1 <- lm(mpg~wt+cyl)
qq.plot(fit1,simulate=TRUE)
# Mosaic Plot Example
library(vcd)
mosaic(HairEyeColor, shade=TRUE, legend=TRUE)
One-way ANOVA displaying 3 groups

Contrast coefficients based on group means and sizes

# Natural Graphics - Pruzek
# GRANOVA - graphical ANOVA
# 1-way layout, 3 groups
# granova is a package that does 1 or 2-way
# anovas with some very interesting graphics
# use mouse data set. brain size as a function
# of laboratory
require(foreign)
mouse1 <- read.spss("mastersmall2_R_nomissing.sav", use.value.labels=TRUE, max.value.labels=Inf, to.data.frame=TRUE)
library(granova)
attach(mouse1)
granova.1w(brain, group = site, size.line = -3)
# prob distributions by Bret Larget
# calls code from the prob.R script
source("prob.R")
gbinom(20, 0.3)
gbinom(100, 0.5, a = 45, b = 55, scale = T)
gbinom(200, 0.3, scale = T, quantile = 0.9)
gnorm(145, 22, a = 120, b = 150)

Note that simpler approaches also work for computing and displaying probability distribution values:

```r
.x <- 0:4
plot(.x, dbinom(.x, size=4, prob=0.5), xlab="Number of Successes", ylab="Probability Mass", main="Binomial Distribution: Trials = 4, Probability of success = 0.5", type="h")
points(.x, dbinom(.x, size=4, prob=0.5), pch=16)
abline(h=0, col="gray")
remove(.x)
```

Instruction: binomial and normal

![Binomial Distribution](image)

- **Binomial Distribution**
  - $n = 20$, $p = 0.3$
  - Possible Values
  - Probability

![Binomial Distribution](image)

- **Binomial Distribution**
  - $n = 100$, $p = 0.5$
  - $P(45 \leq Y \leq 55) = 0.728747$

![Binomial Distribution](image)

- **Binomial Distribution**
  - $n = 200$, $p = 0.3$
  - The 0.9 quantile = 68

![Normal Distribution](image)

- **Normal Distribution**
  - $\mu = 145$, $\sigma = 22$
  - Possible Values
  - Probability Density
  - $P(120 < X < 150) = 0.462$
  - $P(X < 120) = 0.1279$
  - $P(X > 150) = 0.4101$
Instruction: the “t” distribution

```r
x <- seq(-5.5, 5.5, by = .05)
jpeg("zandt.jpg", width = 800, height = 600, quality = 100)
par(lwd = 3, font.axis = 2, font.lab = 2, cex.lab = 1.3)
plot(x, dnorm(x), type = "l", col = "black", lwd = 3, ylab = "f(x)",
     main = "Std Normal Distribution and t Distributions (df=2,3,5,10,25)"
     abline(h = -.001, lwd = 3)
points(x, dt(x, df = 2), type = "l", lty = 2, col = "green", lwd = 3)
points(x, dt(x, df = 3), type = "l", lty = 3, col = "orange", lwd = 2)
points(x, dt(x, df = 5), type = "l", lty = 4, col = "purple", lwd = 2)
points(x, dt(x, df = 10), type = "l", lty = 5, col = "blue", lwd = 2)
points(x, dt(x, df = 25), type = "l", lty = 6, col = "red", lwd = 2)
dev.off()
```
Std Normal Distribution

The graph shows a standard normal distribution.

- The x-axis represents the value of x.
- The y-axis represents the probability density function f(x).

The distribution is bell-shaped with a mean of 0 and a standard deviation of 1.
R has extensive color capability
library(dichromat)
par(mar=c(1,2,1,1))
layout(matrix(1:6,ncol=1))
image(1:10,1,matrix(1:10, ncol=1),
col=colorschemes$BrowntoBlue.10,
main="Brown to Blue (10)", axes=FALSE)
image(1:100,1,matrix(1:100 ,ncol=1),
col=colorRampPalette(colorschemes$BrowntoBlue.10,space="Lab")(100),
main="Brown to Blue Ramp", axes=FALSE)
image(1:10,1,matrix(1:10, ncol=1),
col=dichromat(colorschemes$BrowntoBlue.10),
main="Brown to Blue (10) -- deuteranopia", axes=FALSE)
image(1:12,1,matrix(1:12, ncol=1),
col=colorschemes$Categorical.12,
main="Categorical (12)", axes=FALSE)
image(1:12,1,matrix(1:12, ncol=1),
col=dichromat(colorschemes$Categorical.12),
main="Categorical (12) -- deuteranopia", axes=FALSE)
image(1:12,1,matrix(1:12, ncol=1),
col=dichromat(colorschemes$Categorical.12, "protan"),
main="Categorical (12) -- protanopia", axes=FALSE)
What have I not covered?

• Lots!
• Much on Categorical Variables
  – R is extensively equipped to address these analyses
  – Graphics are plentiful, e.g., mosaic plots
• Spatial Analyses
  – see bibliography and links file
A few illustrations from SigmaPlot
Inheritance of Ear Length in Corn

- **P**
  - Distribution: Short
  - Frequency: 60

- **F₁**
  - Distribution: Short and Long
  - Frequency: 54

- **F₂**
  - Distribution: Short and Long
  - Frequency: 8 x 54

Scanned from: Russell, iGenetics: Pearson Education, Inc., publishing as Benjamin Cummings.
Quantitative Variation:

- Disinhibitory Effects of Alcohol in Mice
- Looks a lot like the corn illustration
Verbal Acknowledgment Type:
- I - Independent
- P - Prompted
- A - Absent

Follow-up
Percent Successful

Response/Initiation Ratio

Verbal Acknowledgment Type:
I - Independent
P - Prompted
A - Absent

Chutes and Ladders
Legos

Follow-up
Sometimes, Tables are Best

Table 1
Rhesus Macaque Brain/Body Characteristics and Regressions of Brain Size on Body Size

A. Descriptive Characteristics

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean (S.E.)&lt;sup&gt;‡&lt;/sup&gt; Body Weight (Kg)</th>
<th>Mean (S.E.)&lt;sup&gt;‡&lt;/sup&gt; Brain Size(cc)</th>
<th>r&lt;sup&gt;‡,4&lt;/sup&gt;</th>
<th>Relative Brain Size&lt;sup&gt;°&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Females</td>
<td>44</td>
<td>6.25 (.47)</td>
<td>94.66 (1.11)</td>
<td>.40</td>
<td>1.51</td>
</tr>
<tr>
<td>Males</td>
<td>39</td>
<td>6.85 (.63)</td>
<td>102.85 (1.41)</td>
<td>.54</td>
<td>1.50</td>
</tr>
</tbody>
</table>

B. Regressions

<table>
<thead>
<tr>
<th></th>
<th>Least Squares</th>
<th>Reduced Major Axis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Intercept(S.E.)</td>
<td>Slope(S.E.)&lt;sup&gt;‡&lt;/sup&gt;</td>
</tr>
<tr>
<td>Females</td>
<td>88.7 (2.3)</td>
<td>.95 (.33)</td>
</tr>
<tr>
<td>Males</td>
<td>94.6 (2.4)</td>
<td>1.20 (.31)</td>
</tr>
</tbody>
</table>

<sup>‡</sup> Standard Error
<sup>‡</sup> Brain Size was measured as a volume (cc; see 1) and 1cc was assumed to be equivalent to 1g
<sup>‡</sup> Both slopes and correlations were significant (t-tests, df=81, p's < .007) for each sex.
<sup>‡</sup> Pearson Correlation Coefficient
<sup>°</sup> Relative brain size in percentage calculated from the mean brain and body sizes for each sex.
<sup>‡</sup> Females differed significantly from males (t(81) = 4.61, p < .001)
<sup>‡</sup> Bootstrap standard errors were derived from N=1000 sampling for each sex. Tests of RMA parameters are described in the text.
Often, the Online Supplement is Best
For Review Papers, Effective Schematic Summary Graphics are Useful

Hurles, et al., 2008
Which is a more effective way of presenting these data, comparing outcome values for two categories?

SPSS Interactive graphics option is much more extensive than the old legacy graphs option.
A few slides on some visual perception issues

• based on work from J. Heer, Berkeley
Preattentive Processing

• Preattentive Visual Perception
  – attention focusing not required
  – stimuli perceived immediately
  – what properties facilitate discrimination
  – what properties can mislead
Preattention: Color Selection

Based on figs from: http://www.csc.ncsu.edu/faculty/healey/PP/PP.html
Preattention: Shape

quick/accurate detection
older citation see healey or see web search preattentive visual

Preattentive Visual Dimensions

many preattentive dimensions of visual modality
- hue
- shape
- texture
- length
- width
- size
- orientation
- curvature
- intersection
- intensity
- flicker
- direction of motion
- stereoscopic depth
- lighting direction

[Chris Healey, Preattentive Processing, www.csc.ncsu.edu/faculty/healey/PP/PP.html]
Pre-attentive Processing

• < 200 - 250ms
  – eye movements > 200ms
  – some processing is faster, implying another process
Example: Conjunction of Features

The target has two features that are present in the distractors (both color and shape). Sequential Searching Required.
Emergent Features

Preattentive since target has one distinctive feature
Emergent Features

Non-unique target features so sequential processing required; not preattentive
SUBJECT PUNCHED QUICKLY OXIDIZED
CERTAIN QUICKLY PUNCHED METHODS
SCIENCE ENGLISH RECORDS COLUMNS
GOVERN PRECISE EXAMPLE MERCURY
SUBJECT PUNCHED QUICKLY OXIDIZED
CERTAIN QUICKLY PUNCHED METHODS
SCIENCE ENGLISH RECORDS COLUMNS
TCEJBUS DEHCNUP YLKCIIQ DEZIDIXO
NIATREC YLKCIIQ DEHCNUP SDOHTEM
ECNEICS HSILGNE SDROCER SNMULOC
SNREVOG ESICERP ELPMAXE YRUCREM
SNREVOG ESICERP ELPMAXE YRUCREM
ECNEICS HSILGNE SDROCER SNMULOC
TCEJBUS DEHCNUP YLKCIIQ DEZIDIXO
NIATREC YLKCIIQ DEHCNUP SDOHTEM
ECNEICS HSILGNE SDROCER SNMULOC
# Preattentive Visual Properties

(Healey 96)

<table>
<thead>
<tr>
<th>Property</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>length</td>
<td>Triesman &amp; Gormican [1988]</td>
</tr>
<tr>
<td>width</td>
<td>Julesz [1985]</td>
</tr>
<tr>
<td>size</td>
<td>Triesman &amp; Gelade [1980]</td>
</tr>
<tr>
<td>curvature</td>
<td>Triesman &amp; Gormican [1988]</td>
</tr>
<tr>
<td>number</td>
<td>Julesz [1985]; Trick &amp; Pylyshyn [1994]</td>
</tr>
<tr>
<td>terminators</td>
<td>Julesz &amp; Bergen [1983]</td>
</tr>
<tr>
<td>intersection</td>
<td>Julesz &amp; Bergen [1983]</td>
</tr>
<tr>
<td>closure</td>
<td>Enns [1986]; Triesman &amp; Souther [1985]</td>
</tr>
<tr>
<td></td>
<td>Kawai et al. [1995]; Bauer et al. [1996]</td>
</tr>
<tr>
<td>intensity</td>
<td>Beck et al. [1983]; Triesman &amp; Gormican [1988]</td>
</tr>
<tr>
<td>flicker</td>
<td>Julesz [1971]</td>
</tr>
<tr>
<td>direction of motion</td>
<td>Nakayama &amp; Silverman [1986]; Driver &amp; McLeod [1992]</td>
</tr>
<tr>
<td>binocular lustre</td>
<td>Wolfe &amp; Franzel [1988]</td>
</tr>
<tr>
<td>stereoscopic depth</td>
<td>Nakayama &amp; Silverman [1986]</td>
</tr>
<tr>
<td>3-D depth cues</td>
<td>Enns [1990]</td>
</tr>
<tr>
<td>lighting direction</td>
<td>Enns [1990]</td>
</tr>
</tbody>
</table>
Gestalt Properties

Proximity

Why perceive pairs vs. triplets?
Gestalt Properties

Similarity

a

\[\begin{array}{cccccccccc}
\bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\
\bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\
\bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\
\bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\
\bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\
\bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\
\bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\
\bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\
\bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\
\bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\
\end{array}\]

b

\[\begin{array}{cccccccccc}
\times & \times & \times & \times & \times & \times & \times & \times & \times & \times \\
\times & \times & \times & \times & \times & \times & \times & \times & \times & \times \\
\times & \times & \times & \times & \times & \times & \times & \times & \times & \times \\
\times & \times & \times & \times & \times & \times & \times & \times & \times & \times \\
\times & \times & \times & \times & \times & \times & \times & \times & \times & \times \\
\times & \times & \times & \times & \times & \times & \times & \times & \times & \times \\
\times & \times & \times & \times & \times & \times & \times & \times & \times & \times \\
\times & \times & \times & \times & \times & \times & \times & \times & \times & \times \\
\times & \times & \times & \times & \times & \times & \times & \times & \times & \times \\
\times & \times & \times & \times & \times & \times & \times & \times & \times & \times \\
\end{array}\]
Gestalt Properties

Continuity

smooth not abrupt change
overrules proximity
Gestalt Properties

Connectedness

can overrule size, shape

a

b

c
d
Gestalt Properties

Closure

overrules proximity
Gestalt Properties

Symmetry

emphasizes relationships
Gestalt Laws of Perceptual Organization (Kaufman 74)

• Figure and Ground
  – Escher illustrations are good examples
  – Vase/Face contrast

• Subjective Contour
Colors for Labeling

• Ware recommends to take into account:
  – Distinctness
  – Unique hues
    • Component process model
  – Contrast with background
  – Color blindness
  – Number
    • Only a small number of codes can be rapidly perceived
  – Field Size
    • Small changes in color are difficult to perceive
  – Conventions
Ware’s Recommended Colors for Labeling – not good though for color blind individuals

Red, Green, Yellow, Blue, Black, White, Pink, Cyan, Gray, Orange, Brown, Purple.
The top six colors are chosen because they are the unique colors that mark the ends of the opponent color axes. The entire set corresponds to the eleven color names found to be the most common in a cross-cultural study, plus cyan (Berlin and Kay)