Assignment 2
Organizing and Producing Data

Math 363

September 17, 2009

1. The life span in days of 88 wildtype and 99 transgenic mosquitoes is given in `mosquitoes.txt`. Download these data using

```
> mosquitoes<-read.delim("http://math.arizona.edu/~jwatkins/mosquitoes.txt")
```

   (a) Give a summary of the life span of both types of mosquitoes.

   (b) Give side by side box plots of the life span of both types of mosquitoes.

   (c) The **empirical survival function** gives, for each time t, the fraction of the life spans greater than t. Give a mathematical relationship between the empirical survival function and the empirical cumulative distribution function.

   (d) Divide the data using `wildtype<-mosquitoes[1:88,1]` and `transgenic<-mosquitoes[,2]` and make side by side empirical survival functions by using the command `par(mfrow = c(1, 2))`. Resize them appropriately.

   (e) One genotype of mosquito lives longer, on average, than the other. Explain how this can be seen in the boxplots and in the survival function.

2. The Chesapeake Bay is the largest estuary in the United States. During the 1970s, concerned over the health of the Chesapeake Bay led to a long term study of the environmental conditions of the bay. They study reported the following annual average salinity for the first 5 years of the study.

year	1	2	3	4	5
% salinity	13.9	14.8	13.5	15.0	15.3

   (a) Compute by hand the mean and variance of the length of the year and the percent salinity for these five years. Show your work.

   (b) Find the covariance of salinity and year.

   (c) Compute the covariance and correlation of year and percent salinity. Show your work.

   (d) Find the least squares regression line using the year as the explanatory variable.

   (e) Give a residual plot for the regression line.

   (f) Estimate the salinity for years 6 and 10.

3. Recall the data on the global consumption of oil. Download these data into R using the command

   `> oil<-read.delim("http://math.arizona.edu/~jwatkins/oil.txt")`

   (a) To examine worldwide oil consumption before the oil embargo of the 1970s, save the first 22 entries with `> year<-oil[1:22,1]`. Do the same for `barrels` and make a scatterplot of years vs. millions of barrels of oil.

   (b) Make a scatterplot of years vs. the logarithm of millions of barrels of oil.

   (c) Give the equation for the regression line of years vs. the logarithm of millions of barrels of oil.

   (d) Give the instantaneous rate of growth for this time frame.

   (e) Display the residual plot and describe any structure that you see in the residuals.

   (f) Predict oil use in 1988 under continued exponential growth from the period up to 1974 and give the difference between this predicted use and actual use.

4. Download the data set `mammals` by calling for `library("MASS")`

   (a) Enter `mammals` and describe the data set.

   (b) Plot the data with `plot(mammals)` and describe the plot.

   (c) Give appropriate transformations of one or both of the variables that gives a scatterplot suitable for linear regression. Make a scatterplot of the transformed data and describe the plot.

   (d) Give the coefficients in the regression line of the transformed variables.

   (e) On average, how does brain size change with a doubling of body size.

   (f) Use `data.frame` with `mammals` and the residuals to give examples of mammal species with unusually large and unusually small brain size. (The `order` command can be used to order the residuals from small to large keeping the rows of the data intact.)

5. Watch the video

   `http://www.ted.com/index.php/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen.html.`

   (a) There are many ways to visualize and present data (e.g., via a table, graph, animation, statistic). Describe briefly three different techniques/strategies Rosling uses or discusses to convey information about data (including 'boring statistics'). Briefly explain why or why not you think these approaches are effective.

   (b) Pick one of the animation sequences Rosling presents. Discuss how his approach to presenting the data yielded some new insight that might not have been obvious via a more traditional approach (e.g., a simple bar graph and lots of hand-waiving). In particular, think about how he might have chosen to show certain inter-relationships and why (e.g., evolution of some particular statistic across countries with respect to time).

   (c) True or False (with a brief explanation as to why): Data visualization is an art in itself. One (philosophical) approach is to 'present the data in such a way that it speaks for itself'.

   (d) Why do you think Rosling chose to plot many of the plots he showed on logarithmic axes?

6. Do 1 and 2 in Exercise A on the HIP trial, page 13 in Freedman.

7. Do 3 in Exercise A on the Snow's study of cholera, page 13 in Freedman.