Stat 400 Data Analysis Sept. 3, 2003

Some points to note from Chapter 1 (Cleveland text)

p 5 visualization + prior knowledge -> conclusions
p 7 multiway dot plot - note choice of species orders and location orders
p 8 histogram not so good for comparing univariate distributions
p 9 scatter plots of two variables may have and x -> y asymmetry
analyze these by fit + residual charts
p 10 Can a 3-D relationship be seen in 2-D? Not by simple scatter plots.
p 12 "Sometimes visualization can fully replace the need for probabilistic inference"
visualization useful for analysis as well as for summary
p 14 direct manipulation refers to altering the graphical representation of data by changing a parameter of the display. Usually the display changes immediately on a computer screen, but the same principles apply to sequences of graphs.

A data set to get started with …

The "ubi" data set (reprinted below):

This data was collected as an exercise for some Indonesian stats instructors in Manokwari, Irian Jaya.
Ten participants worked in 5 pairs (teams) to collect some measurements of yam plants in a research plot. The plot included 10 square sub-plots of 10 m x 10 m each with a different species of yam. Each subplot had hundreds of plants so random points in each subplot were used to select the plants to be measured. The variables “north” and “east” give the coordinates in each subplot relative to the south-west corner of each subplot. The plots themselves were in a grid and the variables “northplt” and “eastplt” give the coordinates of the subplots relative to the south-west corner of the plot.

The four measured variables (all in cm) are

Length – the length of the plant (it had a single stem)
Width - the maximum width of the plant
Internod – The distance from the 8^{th} node to the 9^{th} node along the stem
Diameter – the width of the stem at the base of the plant.

The objective was to rank the species according to the degreee that they were “bushy” as opposed to “stringy”. (The reason was that it was thought this would correlate with yam Productivity, and we did not want to dig up plants since these were being studied in another

Project). The variables length and internod should be small for a bushy plant, and width and diameter should be large.

The plot was on the side of a hill, and it was thought quite likely that there would be a Fertility and/or moisture gradient over the plot. This very much complicates the analysis, since the design of the study did not allow for this gradient. So the statistical problem is how to allow for this design error in ranking the species.

We need to define what is meant by "bushiness" in terms of the available data. We have to be careful to consider transformations of the raw data in forming this index. We need to decide if there is a team bias that needs to be adjusted for. We need to figure out how to capture the plot gradient even though it seems confounded with species.

HW for Friday – be prepared to discuss this data set and strategies for analysis.

Species	north	east	length	width	internod	diameter	team	northplt	eastplt
1	2	4	151.0	28.0	2.5	0.450	1	1	0
1	2	7	106.0	33.0	4.5	0.440	2	1	0
1	1	5	166.0	36.4	2.2	0.500	3	1	0
1	9	2	154.0	40.5	6.0	0.496	4	1	0
1	6	2	109.5	46.0	4.8	0.490	5	1	0
2	2	4	176.0	33.0	6.8	0.540	1	1	1
2	2	7	159.0	35.2	2.5	0.480	2	1	1
2	8	1	166.7	25.4	4.6	0.500	3	1	1
2	9	2	92.0	27.0	2.4	0.520	4	1	1
2	2	8	150.0	35.0	5.0	0.510	5	1	1
3	2	4	128.0	33.0	7.5	0.490	1	2	0
3	2	7	53.5	15.5	3.4	0.260	2	2	0
3	3	1	81.0	23.8	2.2	0.260	3	2	0
3	9	2	139.0	16.0	1.5	0.480	4	2	0
3	2	6	122.5	28.0	4.6	0.450	5	2	0
4	2	4	177.0	42.0	8.4	0.680	1	2	1
4	2	7	189.0	28.5	5.8	0.550	2	2	1
4	6	5	431.0	35.0	6.1	0.720	3	2	1
4	9	2	465.0	36.0	8.0	0.655	4	2	1
4	3	4	346.5	27.2	5.5	0.600	5	2	1
5	2	4	194.0	42.0	5.1	0.650	1	3	0
5	2	7	117.5	38.0	3.7	0.450	2	3	0
5	6	3	147.0	36.4	5.5	0.480	3	3	0
5	9	2	96.0	31.5	3.5	0.442	4	3	0
5	5	4	202.0	40.0	5.0	0.500	5	3	0
6	2	4	260.0	27.5	4.0	0.560	1	3	1
6	2	2	218.3	28.0	5.9	0.450	2	3	1
6	6	1	354.6	41.3	9.5	0.530	3	3	1
6	9	2	205.0	40.0	5.5	0.525	4	3	1
6	8	2	403.0	47.0	11.0	0.510	5	3	1
7	2	4	166.0	34.0	4.3	0.510	1	4	0
7	2	2	205.0	33.0	5.6	0.400	2	4	0
7	8	5	351.3	36.3	5.0	0.470	3	4	0
7	9	2	127.0	23.8	3.3	0.388	4	4	0
7	4	5	228.0	26.8	4.0	0.400	5	4	0

8	2	4	99.0	20.0	5.1	0.490	1	4	1
8	2	2	107.2	25.0	8.3	0.460	2	4	1
8	8	3	150.0	20.4	7.4	0.320	3	4	1
8	9	2	320.0	21.8	5.0	0.335	4	4	1
8	4	3	327.5	22.4	5.9	0.400	5	4	1
9	2	4	250.0	29.0	10.2	0.360	1	5	0
9	2	2	196.0	24.0	8.5	0.280	2	5	0
9	8	6	522.3	23.3	8.3	0.440	3	5	0
9	9	2	212.0	22.9	7.7	0.320	4	5	0
9	2	5	137.5	27.4	5.6	0.300	5	5	0
10	2	4	295.0	40.0	7.5	0.570	1	5	1
10	2	2	211.3	26.2	7.0	0.460	2	5	1
10	1	5	403.0	39.8	4.8	0.550	3	5	1
10	9	2	245.0	34.5	5.0	0.310	4	5	1
10	5	2	116.0	38.5	4.3	0.510	5	5	1