Stat 400 Data Analysis Nov. 4, 2005

Today: A data set from Indonesia. "Ubi" generic name for sweet potato tubers.

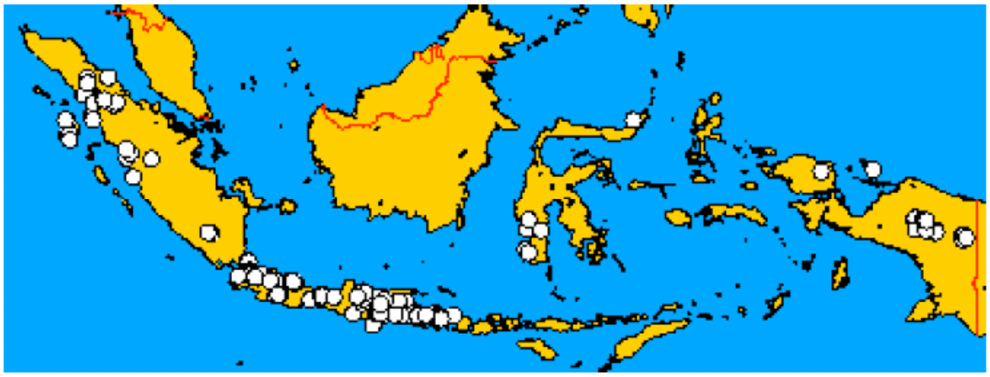


Figure 1. Sites of collecting sweetpotato germplasm in Indonesia. This map was generated by MapQuest through Internet.

More info re "ubi" research in Irian Jaya, Indonesia:
http://papuaweb.org/dlib/tema/ubi/index.html

The "ubi" data set (reprinted below):

This data was collected as an exercise for some Indonesian stats instructors in Manokwari, Irian Jaya. Eastern Indonesia University Development Project.
More about EIUDP (http://www.sfu.ca/mediapr/sfnews/2000/July27/angerilli.html)

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. See schematic next page.

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 8th node to the 9th 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.

"Ubi" Data

species	north	west	length	width	internod	diameter	team	northplt	westplt
1	2	4	151	28	2.5	0.45	1	5	1
1	2	7	106	33	4.5	0.44	2	5	1
1	1	5	166	36.4	2.2	0.5	3	5	1
1	9	2	154	40.5	6	0.496	4	5	1
1	6	2	109.5	46	4.8	0.49	5	5	1
2	2	4	176	33	6.8	0.54	1	5	2
2	2	7	159	35.2	2.5	0.48	2	5	2
2	8	1	166.7	25.4	4.6	0.5	3	5	2
2	9	2	92	27	2.4	0.52	4	5	2
2	2	8	150	35	5	0.51	5	5	2
3	2	4	128	33	7.5	0.49	1	4	1
3	2	7	53.5	15.5	3.4	0.26	2	4	1
3	3	1	81	23.8	2.2	0.26	3	4	1
3	9	2	139	16	1.5	0.48	4	4	1
3	2	6	122.5	28	4.6	0.45	5	4	1
4	2	4	177	42	8.4	0.68	1	4	2
4	2	7	189	28.5	5.8	0.55	2	4	2
4	6	5	431	35	6.1	0.72	3	4	2
4	9	2	465	36	8	0.655	4	4	2
4	3	4	346.5	27.2	5.5	0.6	5	4	2
5	2	4	194	42	5.1	0.65	1	3	1
5	2	7	117.5	38	3.7	0.45	2	3	1
5	6	3	147	36.4	5.5	0.48	3	3	1
5	9	2	96	31.5	3.5	0.442	4	3	1
5	5	4	202	40	5	0.5	5	3	1
6	2	4	260	27.5	4	0.56	1	3	2
6	2	2	218.3	28	5.9	0.45	2	3	2
6	6	1	354.6	41.3	9.5	0.53	3	3	2
6	9	2	205	40	5.5	0.525	4	3	2
6	8	2	403	47	11	0.51	5	3	2
7	2	4	166	34	4.3	0.51	1	2	1
7	2	2	205	33	5.6	0.4	2	2	1
7	8	5	351.3	36.3	5	0.47	3	2	1
7	9	2	127	23.8	3.3	0.388	4	2	1
7	4	5	228	26.8	4	0.4	5	2	1
8	2	4	99	20	5.1	0.49	1	2	2
8	2	2	107.2	25	8.3	0.46	2	2	2
8	8	3	150	20.4	7.4	0.32	3	2	2
8	9	2	320	21.8	5	0.335	4	2	2
8	4	3	327.5	22.4	5.9	0.4	5	2	2
9	2	4	250	29	10.2	0.36	1	1	1
9	2	2	196	24	8.5	0.28	2	1	1
9	8	6	522.3	23.3	8.3	0.44	3	1	1
9	9	2	212	22.9	7.7	0.32	4	1	1
9	2	5	137.5	27.4	5.6	0.3	5	1	1
10	2	4	295	40	7.5	0.57	1	1	2
10	2	2	211.3	26.2	7	0.46	2	1	2
10	1	5	403	39.8	4.8	0.55	3	1	2
10	9	2	245	34.5	5	0.31	4	1	2
10	5	2	116	38.5	4.3	0.51	5	1	2

species #1	species #2
species #3	species #4
species #5	species #6
species #7	species #8
species #9	species #10

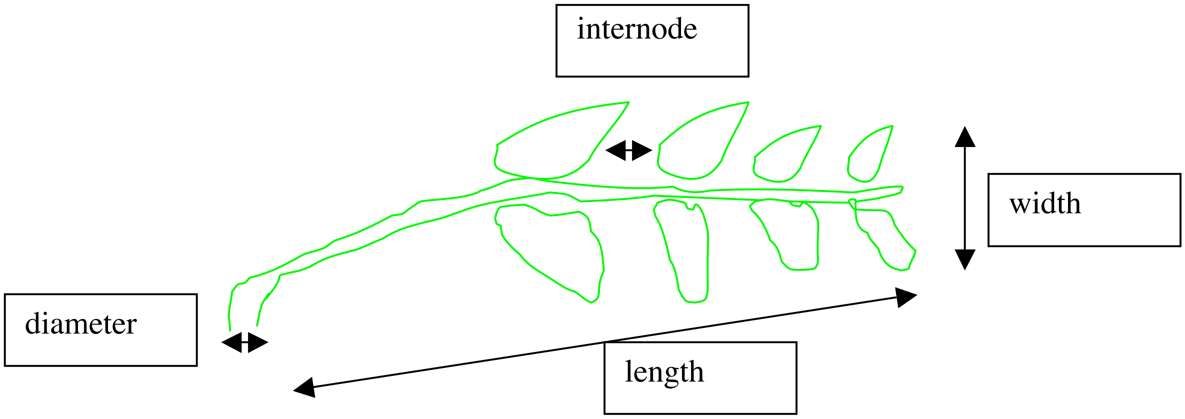
There are several issues to consider:

0. What would we like to learn from this data (imagining ourselves as ubi researchers!)?
1. Since the species is completely confounded with location, is there any hope of detecting species differences?
2. How should we examine the data for the data-screening phase?
3. How should "bushiness" be measured based on the available data?
4. Are there any other features of the species worth reporting?

Exercise for hand-in Wednesday Nov 16: Analyze ubi data with above questions in mind. Report your analysis, and your findings. Your report should read like a "story", with a beginning, middle and end! I suggest

1. Objective of the analysis (if you were the researcher).
2. Analysis: Data Screening (just say what you did and only report details (graphs, tables) if they are important to your story.)
Data Analysis – estimation and testing
3. Summary: Graphical summary for key features of your story
Verbal Summary of your findings
4. Concluding remarks This is where you could comment on the design of the study or other aspects outside of your story.

Here is a rough schematic of the ubi plant and the measurements taken.

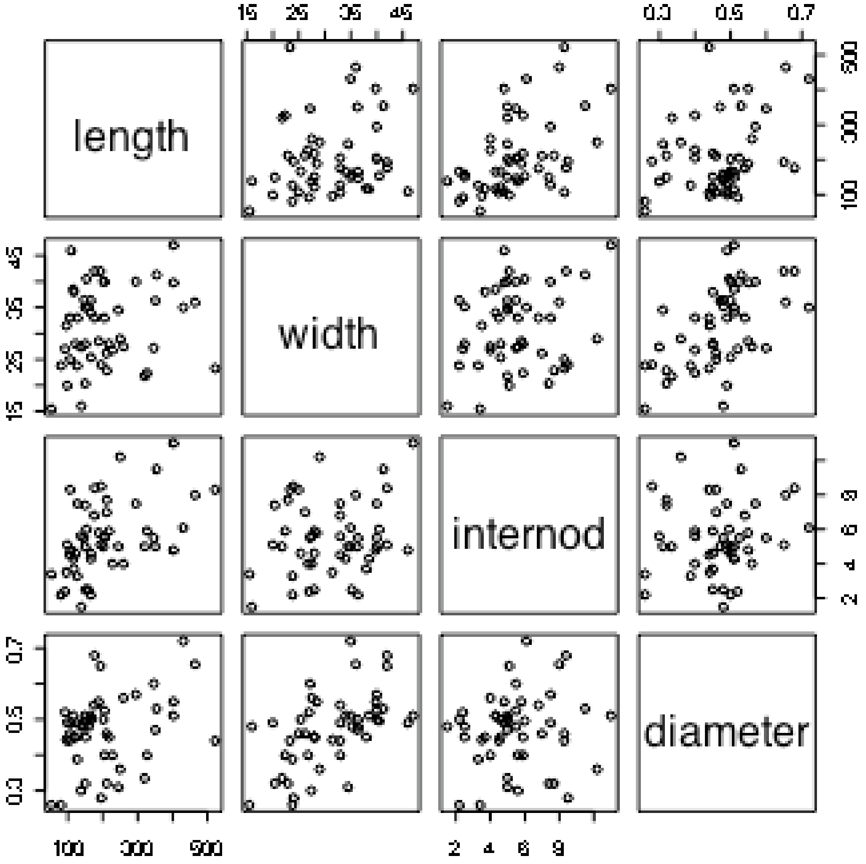


Note that length and internode seem to measure "elongation" in different ways, and width and diameter seem to measure "breadth" in different ways.

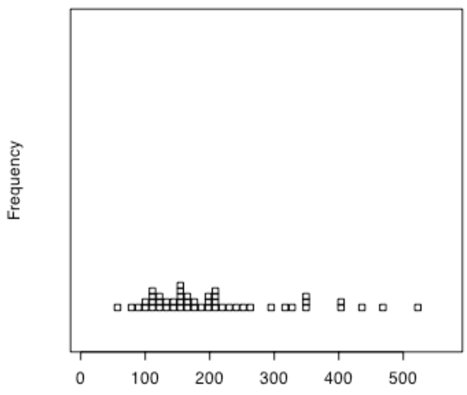
How would you measure "Squatness" which presumably is associated with a big breadth compared to elongation?

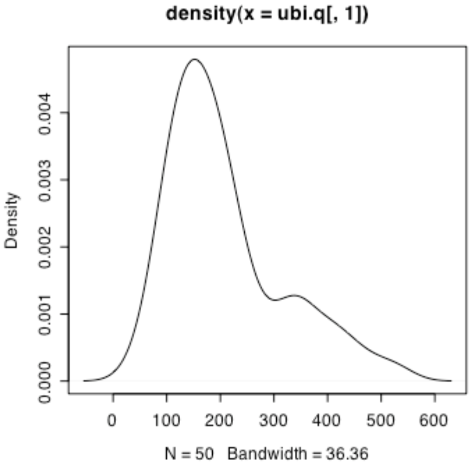
Some considerations
Standardization of variables?
Influence of outliers?
Symmetry of distributions?

```
ubi.q=cbind(ubi.df[,4:7])
> plot(ubi.q)
```



```
> my.dotplot(ubi.q[,1])
> density(ubi.q[,1])
```





Should we be looking at the distribution of the measurements across species, or within species? On what basis would we want to symmetrize the measurements?

In constructing our "squatness" index, how can we tell if we have a good index? What units should the index have?

Once we have an index, what do we do with it? Remember we want to find out the relative squatness of the ten species. In fact, a nice output would be a graphical description of the ten species average values of squatness (adjusted for location?).