

Univariate Analysis Exercise

1. Four measurements were made of male Egyptian skulls from five different time periods ranging from 4000 B.C. to 150 A.D. The data are found in `eskulls.dta`. We want to determine if there are any differences in the skull sizes between the time periods and if they show any changes with time. The researchers theorize that a change in skull size over time is evidence of the interbreeding of the Egyptians with immigrant populations over the years.

- a. Construct a histogram of maximal head breadth. How would you interpret the histogram? Look for overall pattern and striking deviation from that pattern. Describe its shape, center, and spread.
- b. Now construct separate histograms for the five different time periods (hint: use the `by()` option to do it in a single command). What new insights do you gain from this new comparison?
- c. For the same data, produce a boxplot of maximal head breadth over the five time periods. What are the strengths and weaknesses of this display, as compared to the histograms in part (b)?
- d. Calculate and report the mean and standard deviation of head breadth for the five time periods.

2. Between September 15, 1986 and October 18, 1987 two anthropologists recorded when each resident of the Gwaimasi village (in a rain forest of New Guinea) left and returned to the village. Upon their return from a hunting trip, the amount of edible food of various types captured by each resident was recorded. If more than one person was involved in the capture, credit was split evenly. The dataset (`gwaimasi.dta`) also contains information on marital status and family size.

- a. Examine the distribution of how much food each resident brought to the village (`total`). Produce an appropriate graph to visualize the distribution. How would you interpret the graph? Again, look for overall pattern and striking deviations from that pattern. Describe its shape, center, and spread.
- b. Calculate and report the mean and median of the distribution. Which statistic is a more appropriate estimate of central tendency, and why?
- c. We have data on the amount of fish each villager caught with a spear, with a hook, and with some other means. Calculate and report summary statistics of central tendency and dispersion for each variable (the level and spread of each batch, as Drennan would say). Fill in the following table (hint: it can be done with a single command in Stata), and interpret what you see:

	Spear	Hook	Other
Mean			
Median			
Range			
IQR			
SD			

d. Drennan argues that the trimmed mean is sometimes a useful way to estimate central tendency. He also says that most statpacks won't calculate a trimmed mean, you can do so in Stata with the help of a user-written command. If you are connected to the Internet, type `findit robmean` in the command window. In the viewer window that appears, click on the link to the `robmean` function from UCLA, and then click on the link to install the program. Once installation is complete, type `help robmean` in the Stata command window to learn about what `robmean` does. Now, calculate a 10% and a 20% trimmed mean for `fspear`. How do the results compare to what you reported in part (c)? What are the potential advantages of the trimmed mean over the median (see Drennan)?

3. Samples of Romano-British pottery were taken at four sites in the United Kingdom. A chemical analysis of the pottery was performed to measure the percentage of five metal oxides present in each sample. The purpose of the analysis was to determine if different sites produced pottery with different chemical compositions. Find the data in `pottery.dta`.

Examine the distributions of aluminum oxide in each of the four sites. Report what you believe are the most appropriate graphical and numerical summaries of the distributions. Provide an appropriate interpretation.

4. As you know, Franz Boas's immigrant data includes anthropometric data for seven immigrant groups. The dataset you created last week is organized so that each row is a descendant of immigrants; parents' attributes are given as column variables.

a. Produce a frequency table that shows how many descendants of immigrants are in each group, and what proportion of the sample each group represents.

b. Now, create a horizontal bar chart that displays the frequency distribution of the same data.

c. Any subsequent analysis of Boas's data will have to take sex differences into account, so it's important to know how the sex distribution breaks down across the immigrant groups. Create a new horizontal bar chart that shows the sex distribution within each immigrant group.

d. Redisplay the graph from part (c) as a dot plot. Which of the two graphs do you prefer, and why?