

BIOS 600: Principles of Statistical Inference

Data Presentation

Fall 2012

Reading

- ▶ Pagano and Gauvreau, remainder of Chapter 2
- ▶ Warm-up activity: spread of a particularly virulent form of the plague
- ▶ Want to know more? Tufte's classic, *The Visual Display of Quantitative Information*, has wonderful illustrations that show why *a picture is worth a thousand words*.

Overview

- ▶ Graphs are pictorial representations of data and are a very important part of descriptive statistics. They appeal to visual memory in ways that tables or frequency counts cannot.

Frequency Distributions

A *frequency distribution* consists of a set of classes or categories along with the corresponding numerical counts. If data are continuous or discrete with many categories, the range of values is often broken down into a smaller set of distinct, nonoverlapping intervals. Creating frequency distributions in Stata is straightforward using `tabulate` .

Frequency Distributions

Table 1: Reported Cases of Polio, 2011, World Health Organization

WHO Region	Case Count
Africa	16,647
Americas	1,844
Southeast Asia	65,529
Europe	1,544
Eastern Mediterranean	11,692
Western Pacific	7,319
Total	104,575

Frequency Distributions

Table 2: Singleton Birthweight in U.S., 2009, CDC

Weight (g)	Number of Births
< 1000	29,027
1000-1499	30,890
1500-1999	65,603
2000-2499	211,227
2500-2999	767,266
3000-3499	1,618,454
3500-3999	1,090,696
4000-4499	271,307
4500-4999	37,909
5000+	4258
Total	4,126,637

It is easier to make comparisons if all the categories have equal width. Is that the case here? Why or why not?

Resting Pulse Rate

Take your resting pulse rate and submit to poll everywhere, keeping track of the exact number for later.

official US time

Relative Frequency

The *relative frequency* for an interval is the proportion of the total number of observations appearing in that interval. This is calculated by dividing the number of values within an interval by the total number of values in the table.

Table 3: Reported Cases of Polio, 2011, World Health Organization

WHO Region	Case Count	Relative Frequency (%)
Africa	16,647	
Americas	1,844	
Southeast Asia	65,529	
Europe	1,544	
Eastern Mediterranean	11,692	
Western Pacific	7,319	
Total	104,575	

Relative Frequency

The *relative frequency* for an interval is the proportion of the total number of observations appearing in that interval. This is calculated by dividing the number of values within an interval by the total number of values in the table.

Table 4: Reported Cases of Polio, 2011, World Health Organization

WHO Region	Case Count	Relative Frequency (%)
Africa	16,647	$\frac{16,647}{104,575} = 0.159 = 15.9\%$
Americas	1,844	
Southeast Asia	65,529	
Europe	1,544	
Eastern Mediterranean	11,692	
Western Pacific	7,319	
Total	104,575	

Relative Frequency

The *relative frequency* for an interval is the proportion of the total number of observations appearing in that interval. This is calculated by dividing the number of values within an interval by the total number of values in the table.

Table 5: Reported Cases of Polio, 2011, World Health Organization

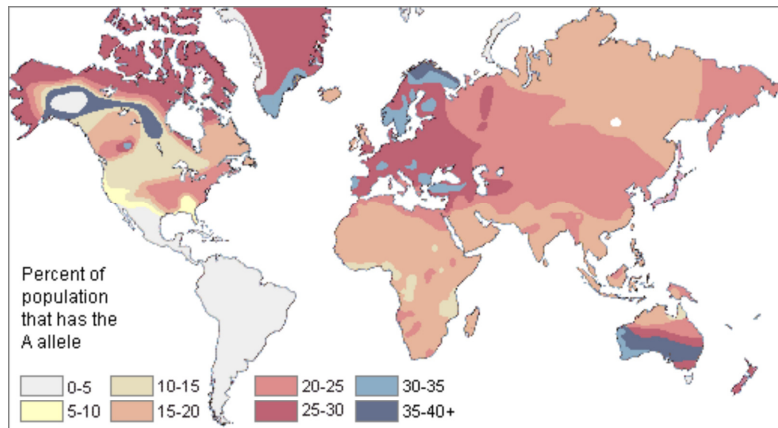
WHO Region	Case Count	Relative Frequency (%)
Africa	16,647	15.9
Americas	1,844	1.8
Southeast Asia	65,529	62.7
Europe	1,544	1.5
Eastern Mediterranean	11,692	11.1
Western Pacific	7,319	7.0
Total	104,575	100

Frequency Distributions

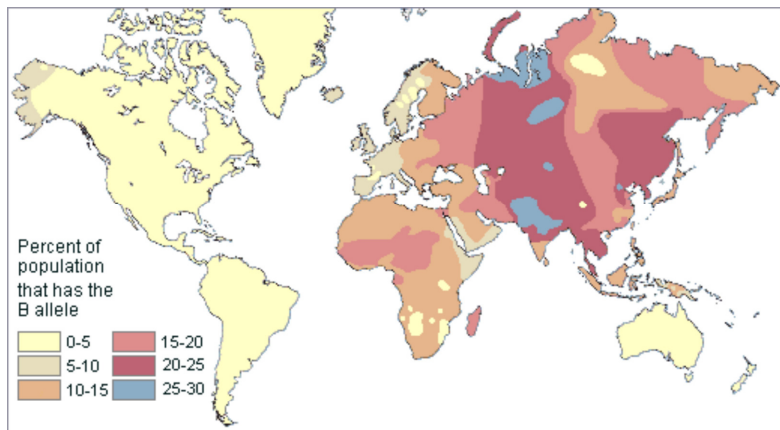
Table 6: Singleton Birthweight in U.S., 2009, CDC

Weight (g)	Number of Births	Relative Frequency (%)
< 1000	29,027	0.70
1000-1499	30,890	0.75
1500-1999	65,603	1.59
2000-2499	211,227	5.12
2500-2999	767,266	18.59
3000-3499	1,618,454	39.22
3500-3999	1,090,696	26.43
4000-4499	271,307	6.57
4500-4999	37,909	0.92
5000+	4258	0.10
Total	4,126,637	

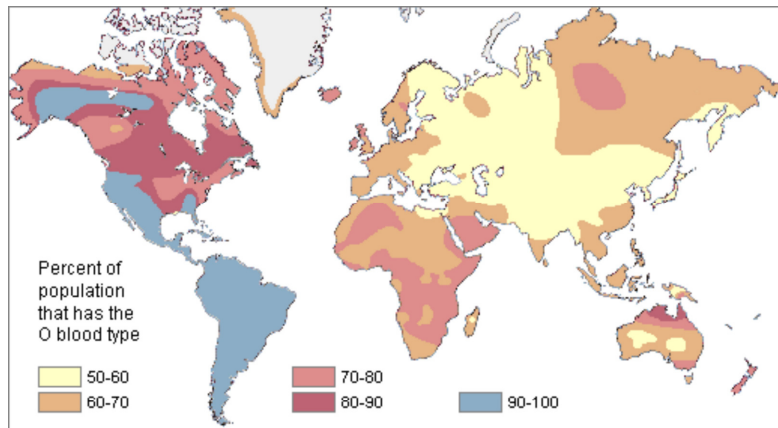
Graphical Relative Frequency Distribution: Blood Allele A



Graphical Relative Frequency Distribution: Blood Allele B



Graphical Relative Frequency Distribution: Blood Allele O



Cumulative Relative Frequency

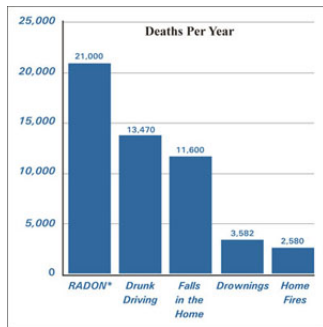
Table 7: Singleton Birthweight in U.S., 2009, CDC

Weight (g)	Number of Births	Relative Frequency (%)	Cum. Rel. Freq. (%)
< 1000	29,027	0.70	0.70
1000-1499	30,890	0.75	1.45
1500-1999	65,603	1.59	3.04
2000-2499	211,227	5.12	8.16
2500-2999	767,266	18.59	26.75
3000-3499	1,618,454	39.22	65.97
3500-3999	1,090,696	26.43	92.40
4000-4499	271,307	6.57	98.98
4500-4999	37,909	0.92	99.90
5000+	4258	0.10	100.00
Total	4,126,637		

Bar Charts

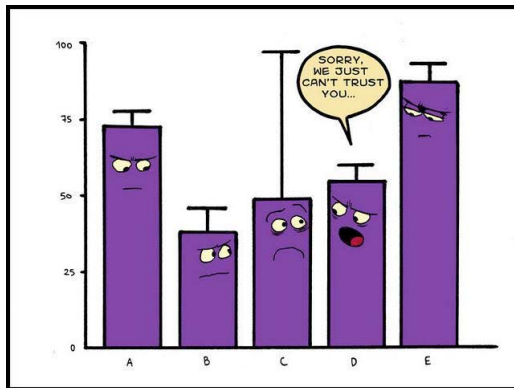
Bar charts can be used to display a frequency distribution for nominal or ordinal data. The various categories of interest are presented along the horizontal (x) axis, and a vertical bar is drawn above each category with the height of the bar representing either the frequency or relative frequency of the observations in that class. The bars should have equal width and be separated from each other so they do not imply continuity (see Stata **graph bar**).

Figure 1: EPA estimates that radon causes thousands of cancer deaths in the U.S. each year



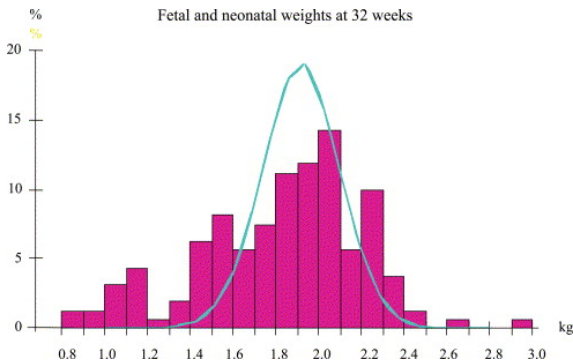
Bar Chart

This bar chart has error bars to indicate variability.



Histograms

The *histogram* displays a frequency distribution for discrete or continuous data. The area of each bar is proportional to the frequency (or relative frequency) of the categories. The bars may touch, indicating the underlying data are continuous. Here is a histogram of weights of infants born at 32 weeks of gestation. (See [Stata histogram](#).)



Histogram Comparing Two Groups (100m Reaction Time)

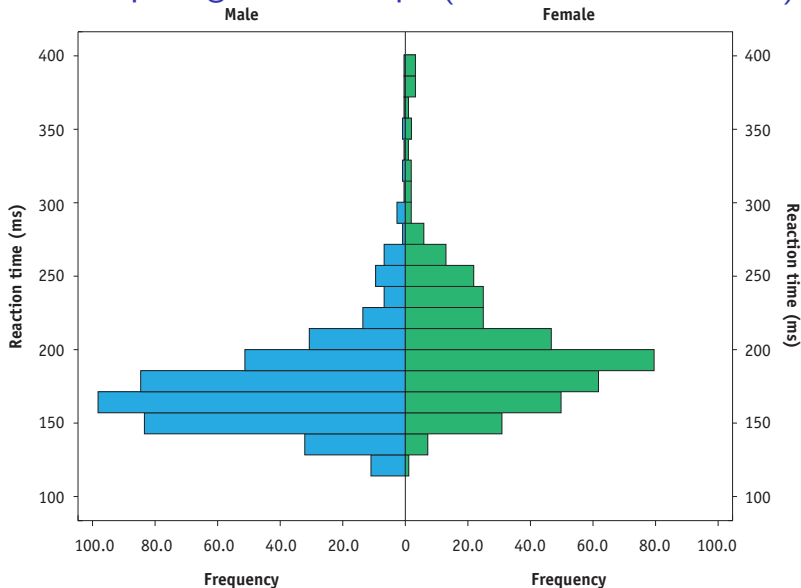


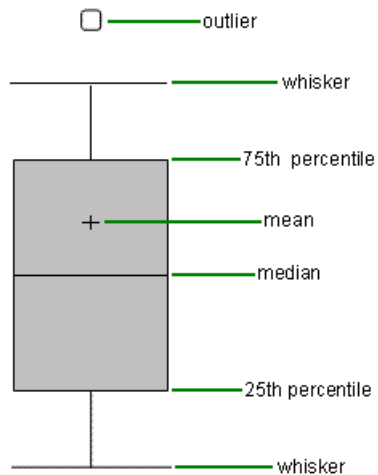
Figure 2. Reaction times of 425 sprinters at the Beijing Olympics. Mean female reaction times are 23 ms slower than for men

Box Plot (Box-and-Whisker Plot)

Box plots graphically depict numerical data through a five-number summary: the smallest observation (sample minimum), lower quartile, median, upper quartile, and largest observation (sample maximum). The boxplot may also indicate which, if any, observations might be considered outliers.

Box plots are non-parametric displays, and spacings between the different parts of the box describe the dispersion (spread) and skewness in the data. Stata: **graph box**.

Box Plot



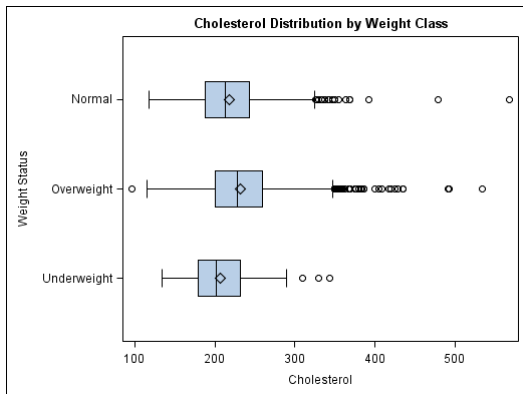
Box Plot

The ends of the whiskers have variable definitions, including the following

- ▶ Minimum and maximum values in the sample
- ▶ The lowest/highest data points within 1.5 times the interquartile range of the lower/upper quartile.
- ▶ One standard deviation above and below the mean
- ▶ The 9th and 91st percentiles
- ▶ The 2nd and 98th percentiles

Any data points that fall beyond the whiskers should be plotted as outliers with a dot or small circle.

Box Plot



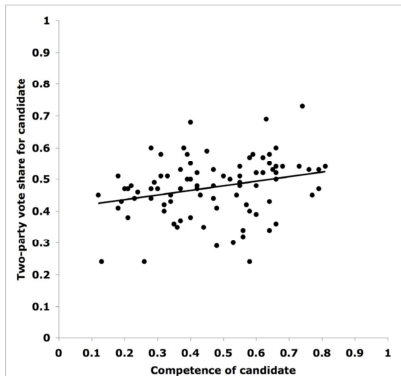
Two-Way Scatter Plots

A *two-way scatter plot* is used to depict the relationship between two different continuous measurements. Each point on the plot represents a pair of values; the scale for one variable is placed on the horizontal axis, and the scale for the other is on the vertical axis.

Stata: **twoway**.

Two-Way Scatter Plots

Ballew and Todorov (2007 PNAS) studied the association between snap judgments of the competence level of a gubernatorial candidate and the share of the vote that candidate received in an election.



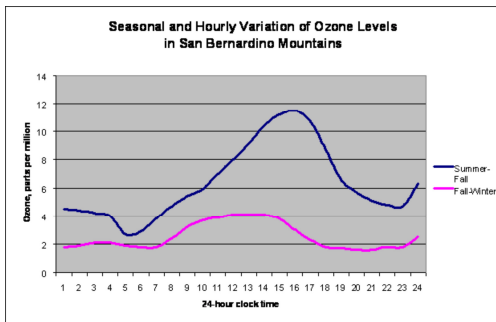
Life Expectancy and Wealth

Two Way Scatter Plot of Life Expectancy and Wealth

Line Graphs

A *line graph* is similar to a two-way scatter plot; however, each value on the horizontal axis has only one measurement on the vertical axis. The horizontal axis usually represents time, so that this type of plot allows us to view the change over time in the quantity on the vertical axis.

Page 1 of 1



Reading for Next Time

- ▶ Pagano and Gauvreau, Chapter 3
- ▶ Cheesy yet informative short song on mean, median, and mode
- ▶ Larson, 'Descriptive Statistics and Graphical Displays,' 2006
Circulation