BIOS 600: Principles of Statistical Inference

Introduction, Types of Data, and Data Presentation

Fall 2015

Reading (optional)

- Pagano and Gauvreau, Chapters 1 and 2 (or any material from an introductory text on types of variables and basic graphical displays)
- DataCamp Lab 0: Introduction to R (R users)
- Getting started in Stata video (Stata users)
- Stata graphics video (Stata users)
- Data Camp Lab 1: Introdution to Data (R users)
- Video about boxplots; may be particularly interesting for nutrition students
- Video about histograms
- 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*.

What is Biostatistics?

Biostatistics is the science of obtaining, analyzing and interpreting data in order to understand and improve human health. Biostatisticians forge advances in science that benefit human health through innovations in biostatistical methodology and theory as well as the thoughtful implementation of biostatistical methods in practice.

Welcome to BIOS 600!

What have I gotten into?

BIOS 600

- is an introductory course in probability theory and statistical inference, moving from tools for describing data to tools for using data to help inform scientific conclusions
- provides a tour of basic statistical methods commonly encountered in public health and biomedical research
- places emphasis on understanding of basic statistical methods, use of the methods to evaluate evidence from studies, and communication of statistical results to non-statisticians
- requires use of standard statistical software (e.g., Stata or R)

Learning Objectives

By December, students successfully completing the course will

- have a basic working knowledge of important statistical topics including descriptive statistics and probability, inference on means and proportions, regression methods, and nonparametrics
- understand how to evaluate which methods are appropriate in answering a research question for a given study design
- be able to evaluate data using modern statistical software and interpret analysis results

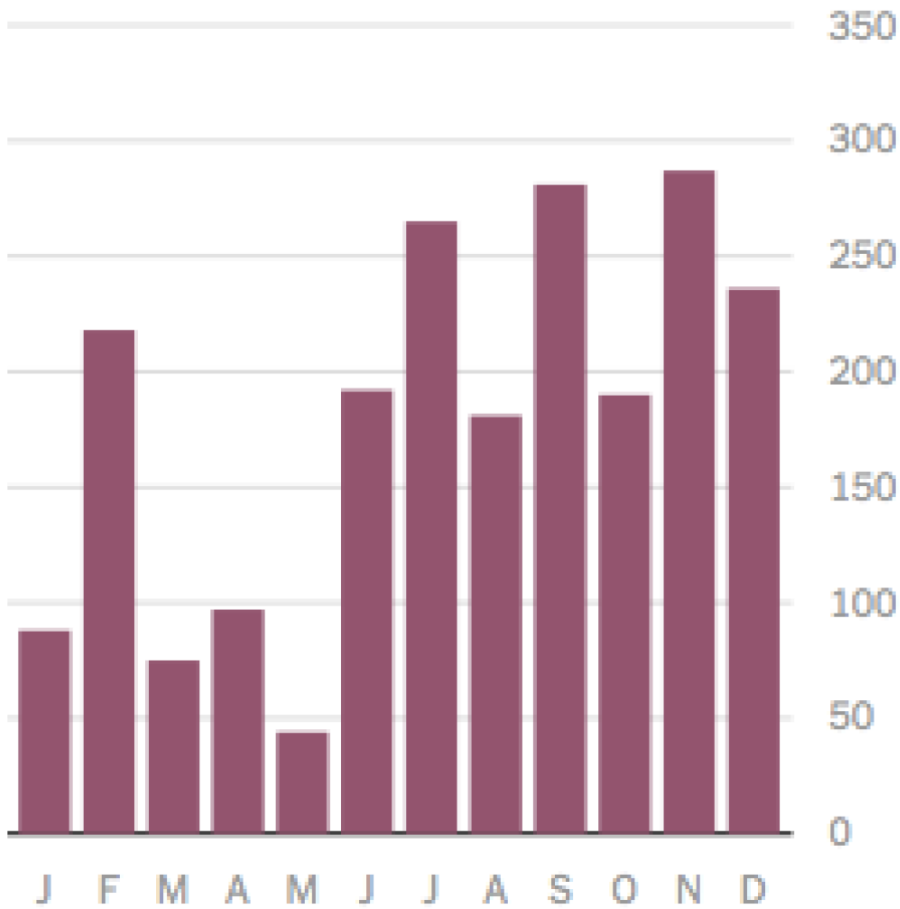
Learning Objectives

By December, students successfully completing the course will

- be able to evaluate straightforward statistical usage in public health and medicine, with a focus on relevant research publications
- have the tools to interact knowledgeably with biostatisticians in planning, conducting, analyzing, and reporting public health and medical research (and know how to determine when a biostatistician should be consulted)

Learning Objectives

Suppose the following graphic depicts the number of new cases of Ebola by month worldwide. What does this mean?



Big Picture

Statistics is the process by which we convert data into useful information. As part of this process, we

- collect data
- summarize data
- interpret the results

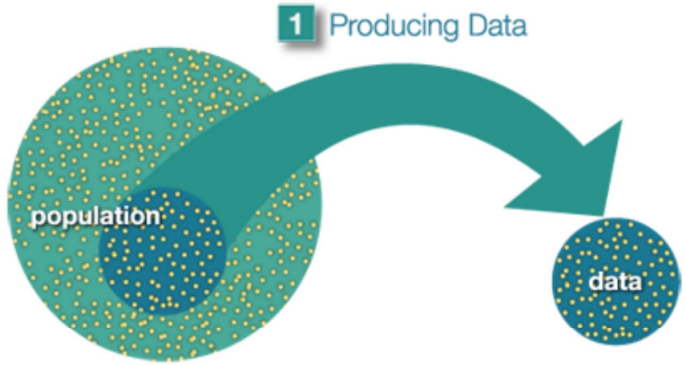
What is the Population?

First we identify the group we would like to learn about. This group is called the *population*. This population could be, for example, all babies born in sub-Saharan Africa, all breast cancer patients in North Carolina, or all adults in the United States.

Graphic from the CMU Open Learning Initiative.

Sampling from the Population

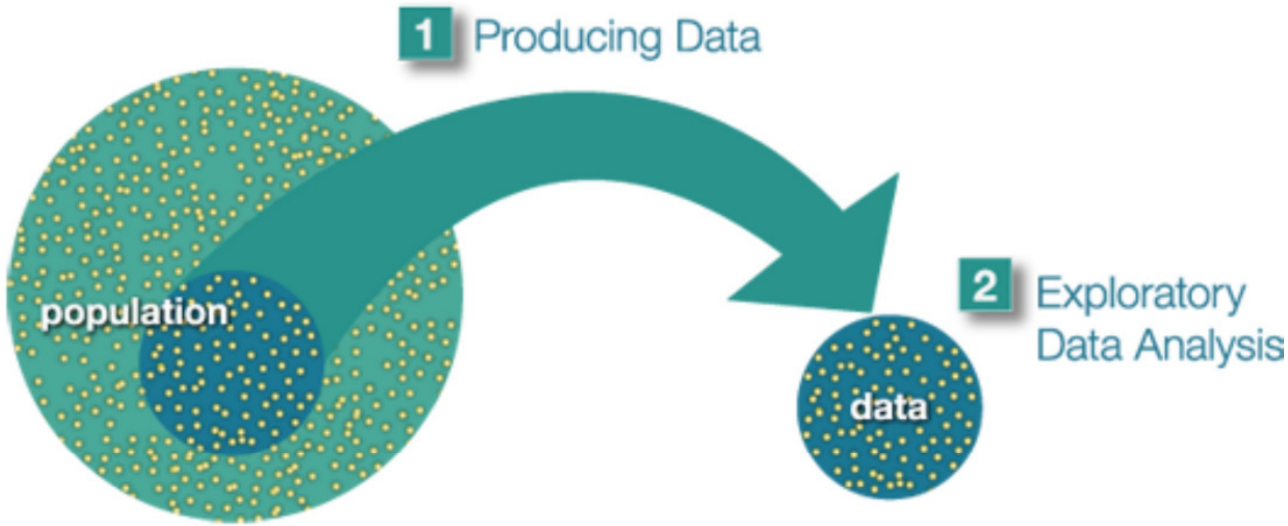
Suppose we wish to know what adults in the U.S. think about universal health care. It would be virtually impossible to ask every adult in the U.S. this question (though the Census still tries to do this, despite good advice!). Usually we have to compromise by taking a *sample* of people from the population for further study. We have to be careful that our *sample* is a representative one – for example, we would have biased results if our sample consisted only of Tea Party members.



Graphic from the CMU Open Learning Initiative.

Collecting Data

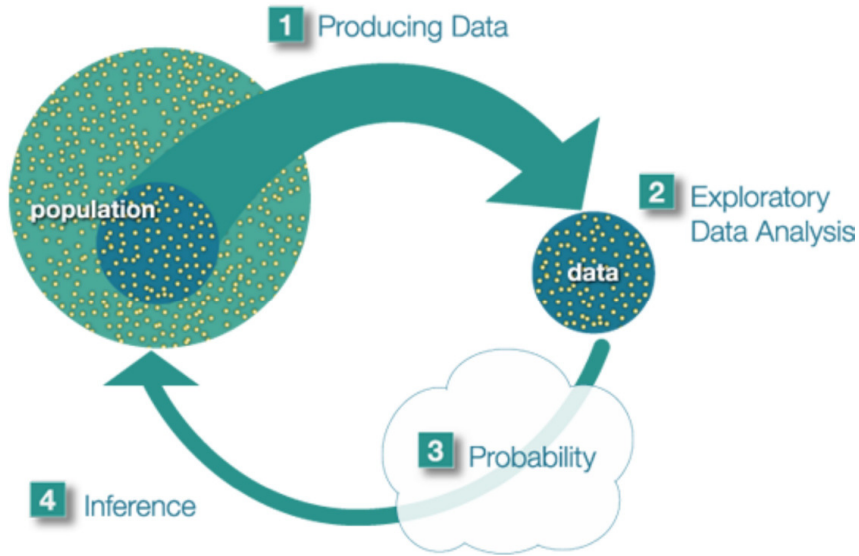
Suppose our sample consists of 5000 U.S. adults, and we ask each adult whether or not they support universal health care. We'll then need to take the 5000 answers and summarize them in some way. For example, we can calculate the % of those in our sample who support universal health care – say it's 65%.



Graphic from the CMU Open Learning Initiative.

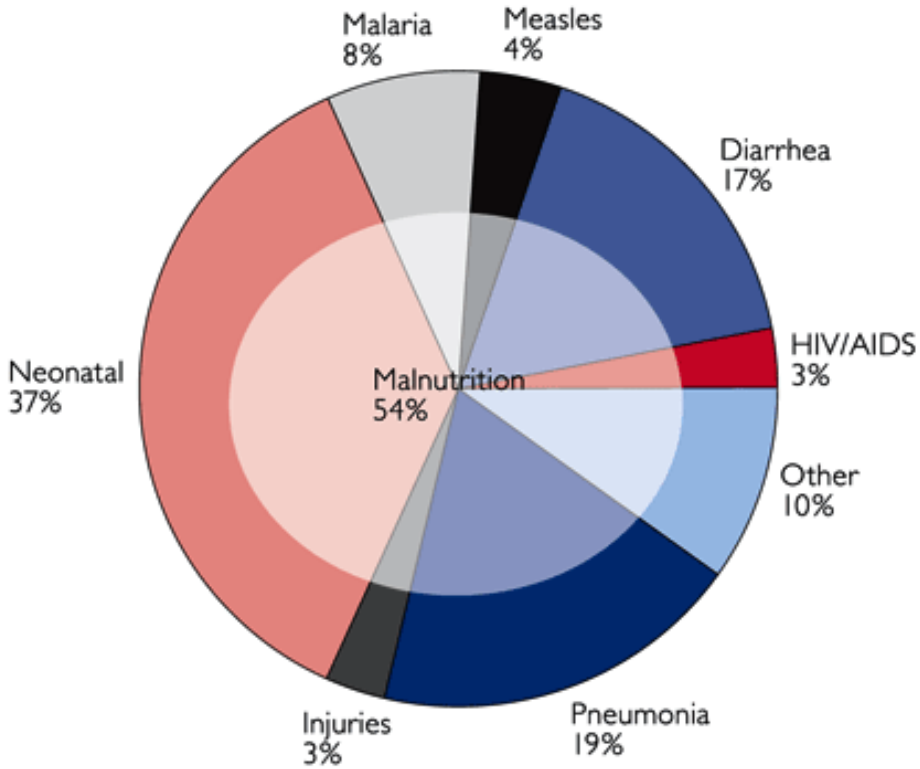
Drawing Conclusions

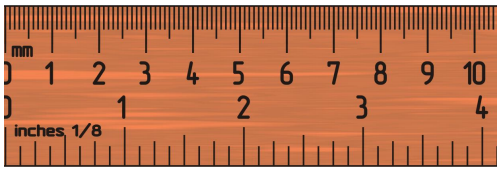
We then use *probability* and *inference* theory to help us determine whether there is significant support of universal health care and to characterize the uncertainty in our estimate (65%). We then draw conclusions about our original population (all U.S. adults)



Graphic from the CMU Open Learning Initiative.

Types of Data





Nominal Data

Classify into named categories without numeric meaning, e.g.

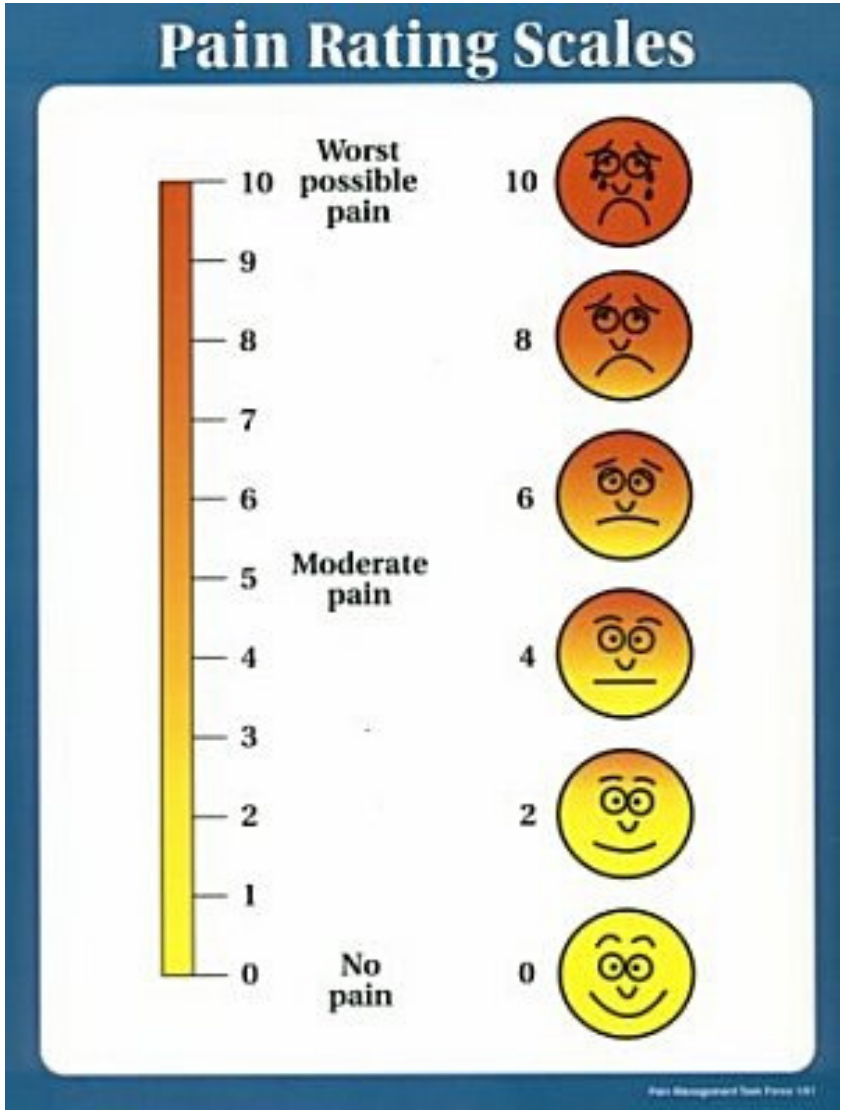
- college graduate (0=not college graduate, 1=college graduate; 'N'=no, 'Y'=yes) – this variable is binary or dichotomous (two possible values)
- health insurance provider (0=uninsured, 1=medicaid, 2=medicare, 3=private)
- blood type (A, B, AB, O)
- whether or not you have colon cancer
- also called categorical data

Ordinal Data

Categories are ordered, but differences between values not easily measured; only relative comparisons are made about differences between levels

- Colon cancer stage 0, I, IIA, IIB, IIC, IIIA, IIIB, IIIC, IVA, IVB
- Likert scale: 5=strongly agree, 4=agree, 3=neutral, 2=disagree, 1=strongly disagree
- Frequency of product use: daily or more; several times/week; weekly; several times/month; monthly; rarely; never

Pain Scale


Pain Rating Scales
10
9
8
7
6
5
4
3
2
1
0
Worst possible pain
Moderate pain
No pain
10
8
6
4
2
0

Rank and Count Data

Count data: counted observations

Rank data: ranked from least to greatest or greatest to least

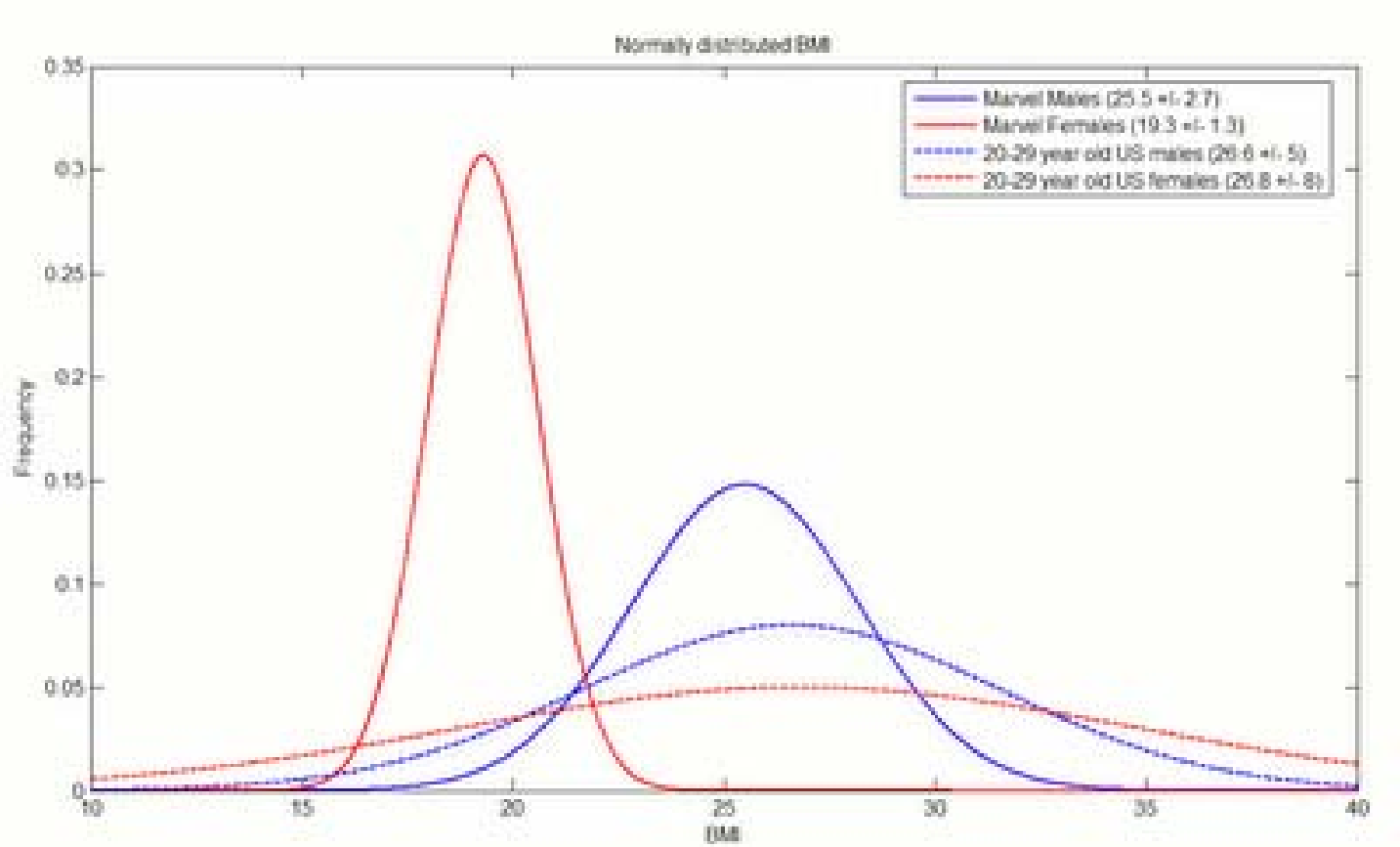
Causes of Death for Children Under Age 5 Globally

Cause	Count	Rank
Diarrhoeal diseases	1,064,000	2
HIV/AIDS	152,000	7
Injuries	228,000	6
Malaria	684,000	4
Measles	76,000	8
Neonatal death	3,040,000	1
Noncommunicable diseases	304,000	5
Pneumonia	988,000	3
Other	988,000	
Total	7.6 million	

Over 70% of these deaths are in Africa and SE Asia (source: WHO)

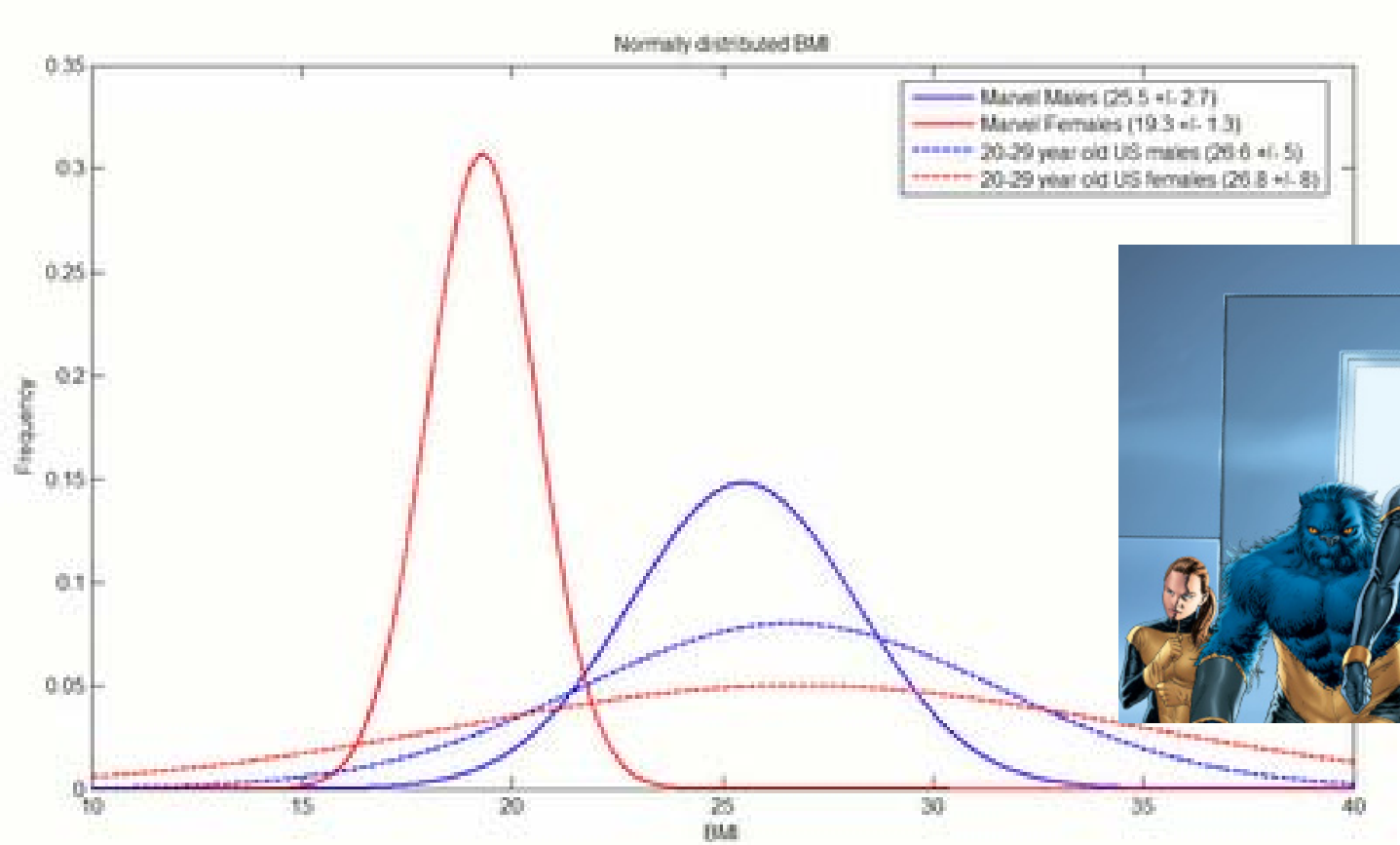
Continuous Data

Data representing measurable quantities in which the difference between any two possible data values can be arbitrarily small, e.g. birth weight, ppm ozone, BMI (plot from Healey et al.)



Continuous Data

BMI (drawing by Marvel Comics)



Example: Delivery Outcomes

Consider data on outcomes of deliveries at UNC Hospital. What types of variables are these?

- Payment type (private, Medicaid, federal/state exchange, emergency Medicaid)
- Mother's perceived length of labor (values reported ranged from 15 minutes to just over 2 weeks, though vast majority $<$ 2 days)
- Infant weight (kg)
- Gravidity (# pregnancies)
- Degree of perineal tearing (none, 1st, 2nd, 3rd, 4th)
- Degree of perineal tearing ($<$ 3rd degree or $\geq$ 3rd degree)

Recap: Measurement Scales

Take a few minutes and identify variables from each scale based on your own research interests.

- Nominal
- Ordinal
- Count
- Continuous

Not so easy

While it is often practical to classify variables according to different scales to facilitate analysis, sometimes this is an oversimplification, even for variables some may consider quite basic.

- gender and sexual identity
- race
- exposure to cigarette smoke

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 and in R using the ftable and table functions.

Frequency Distributions

Table: Reported Cases of Polio (WPV), 2013, 2014, and 2015 through August 12, from World Health Organization

Country	*Case Count*		
	2013	2014	2015
Pakistan	22	306	29
Afghanistan	3	28	7
Nigeria	40	6	0
Somalia	95	5	0
Equatorial Guinea	0	5	0
Iraq	0	2	0
Cameroon	0	5	0
Syria	0	1	0
Ethiopia	0	1	0

Frequency Distributions

Table: 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?

Relative Frequency

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

Table: Reported Cases of Polio through August 12, 2015 (WHO)

Country	*Case Count*	*Relative Frequency (%)*
Pakistan	29	$\frac{29}{29+7+0} = 0.759 = 75.9\%$
Afghanistan	7	
Nigeria	0	
Somalia	0	
Equatorial Guinea	0	
Iraq	0	
Cameroon	0	
Syria	0	
Ethiopia	0	
Total	36	100%

Relative Frequency

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

Table: Reported Cases of Polio through August 12, 2015 (WHO)

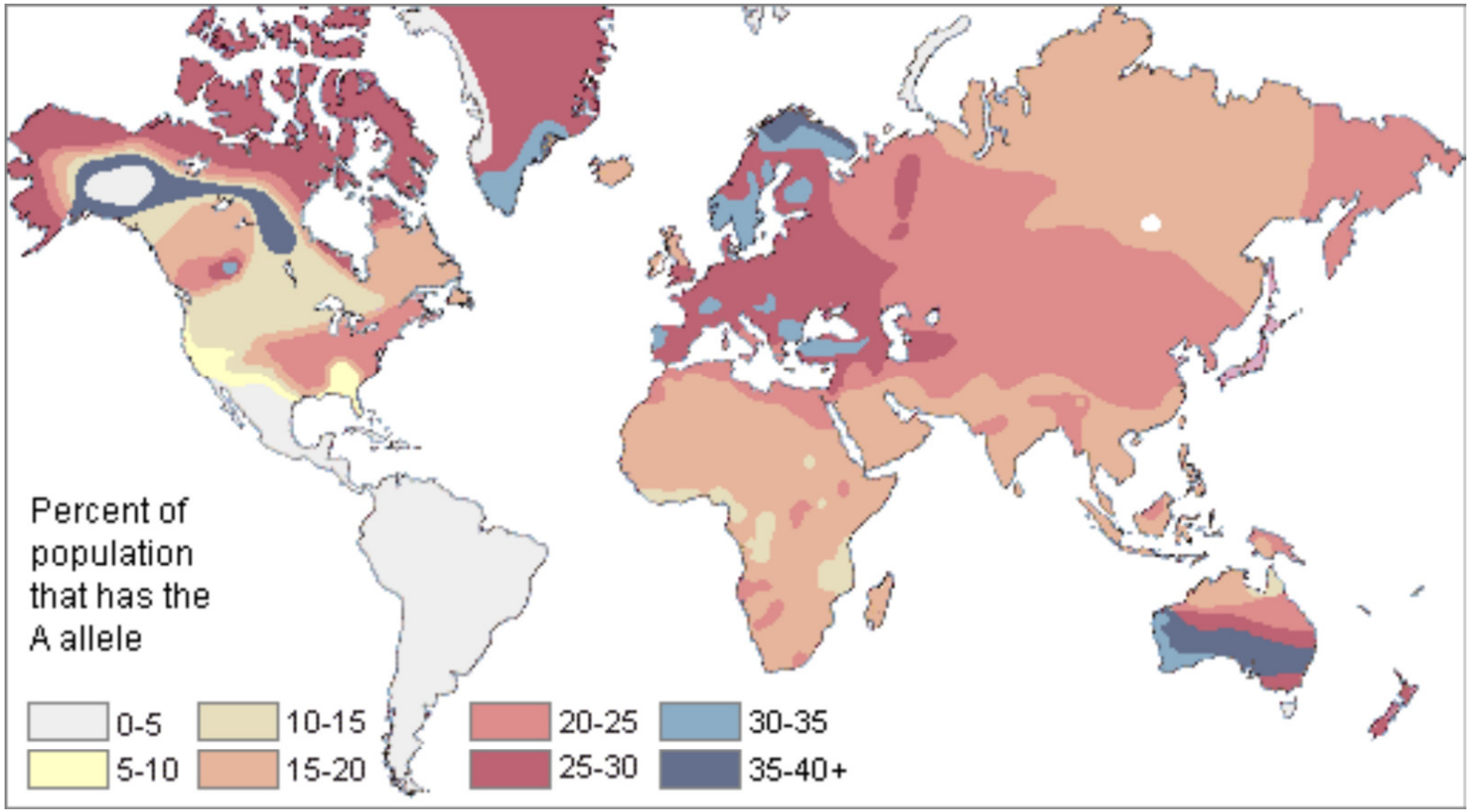
Country	*Case Count*	*Relative Frequency (%)*
Pakistan	29	75.9%
Afghanistan	7	24.1%
Nigeria	0	0%
Somalia	0	0%
Equatorial Guinea	0	0%
Iraq	0	0%
Cameroon	0	0%
Syria	0	0%
Ethiopia	0	0%
Total	36	100%

Frequency Distributions

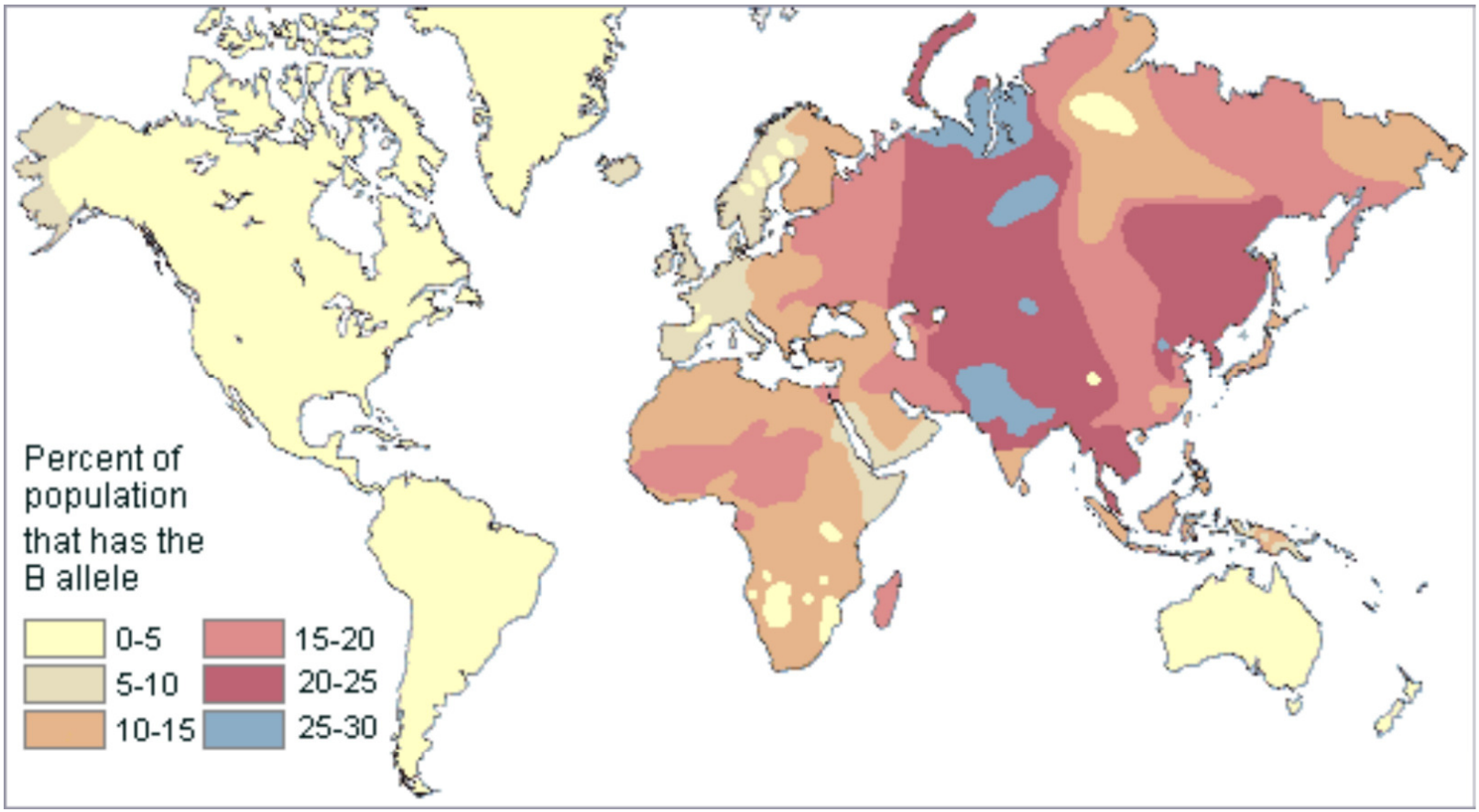
Table: Singleton Birthweight in U.S., 2009, CDC

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	

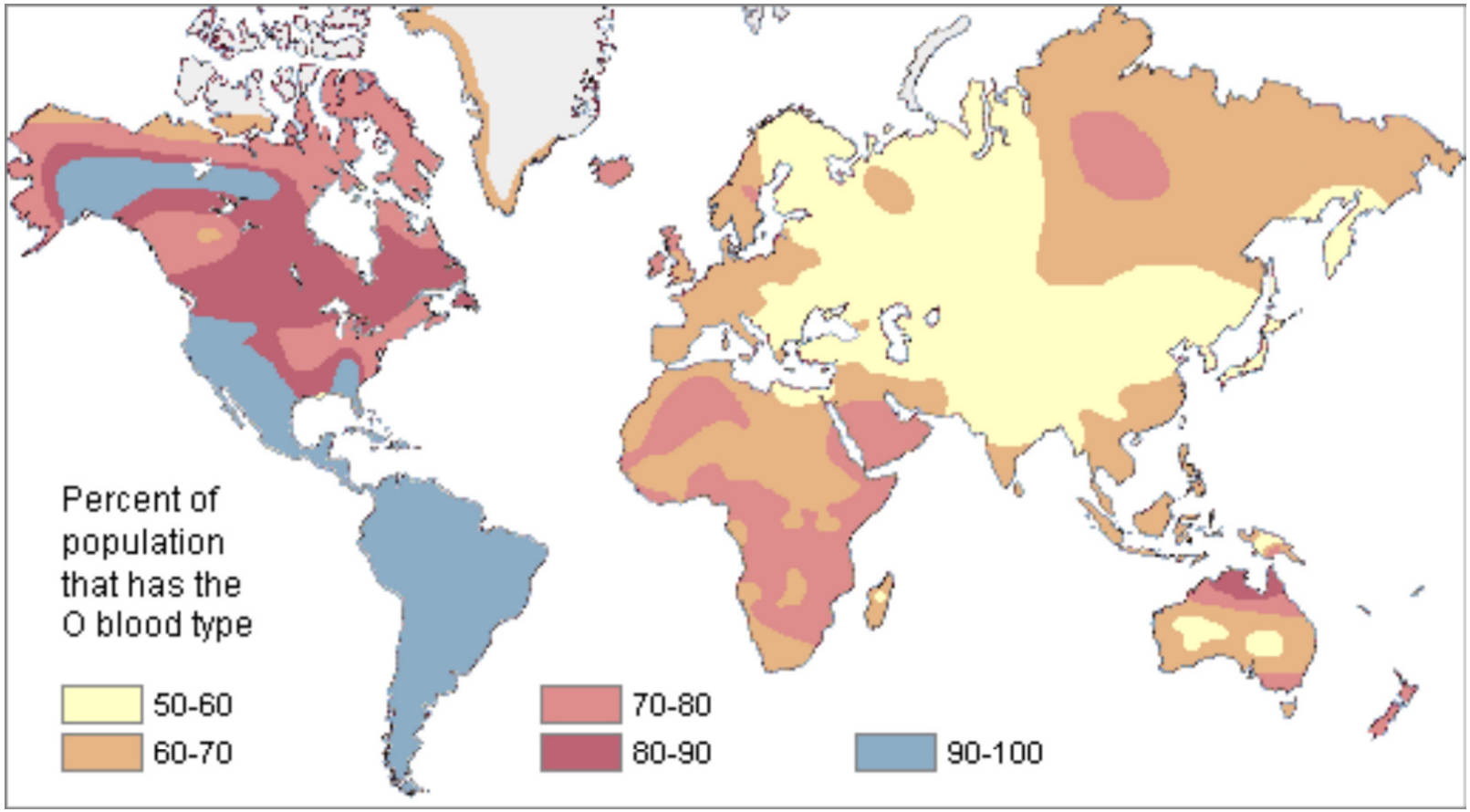
Graphical Relative Frequency Distribution: Blood Allele A



Graphical Relative Frequency Distribution: Blood Allele B



Graphical Relative Frequency Distribution: Blood Allele O



Cumulative Relative Frequency

Cumulative relative frequency is the sum of relative frequencies below a given level. It is simple to calculate in R (table) and Stata (tabulate).

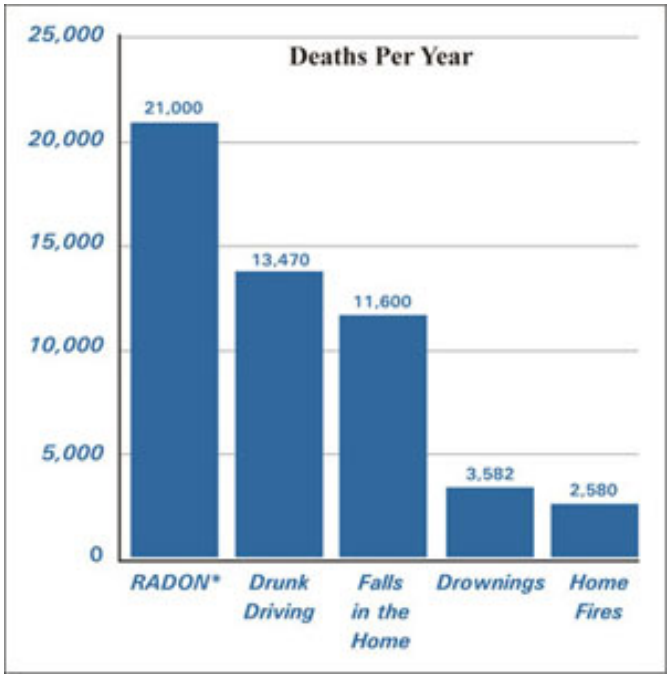
Table: Singleton Birthweight in U.S., 2009, CDC

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

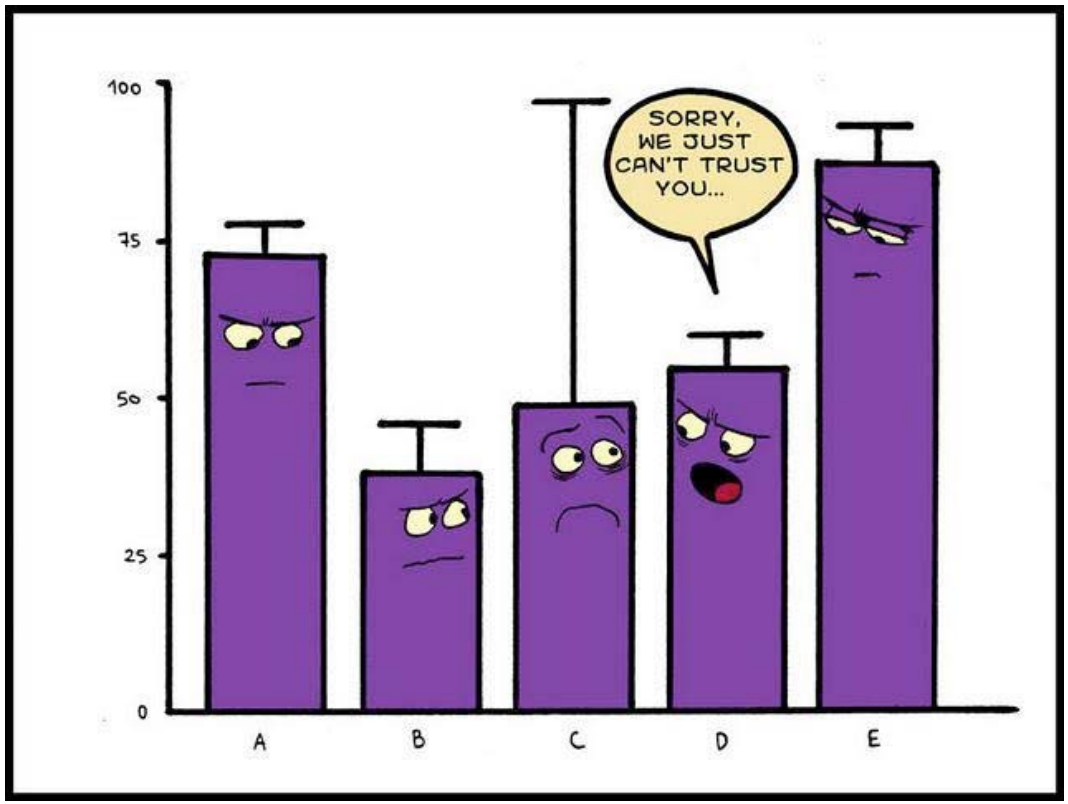
Bar charts can be used to display a frequency distribution for nominal or ordinal data. The various categories of interest are presented along the horizontal (x) axis, and a vertical bar is drawn above each category with the height of the bar representing either the frequency or relative frequency of 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 or R barplot).

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



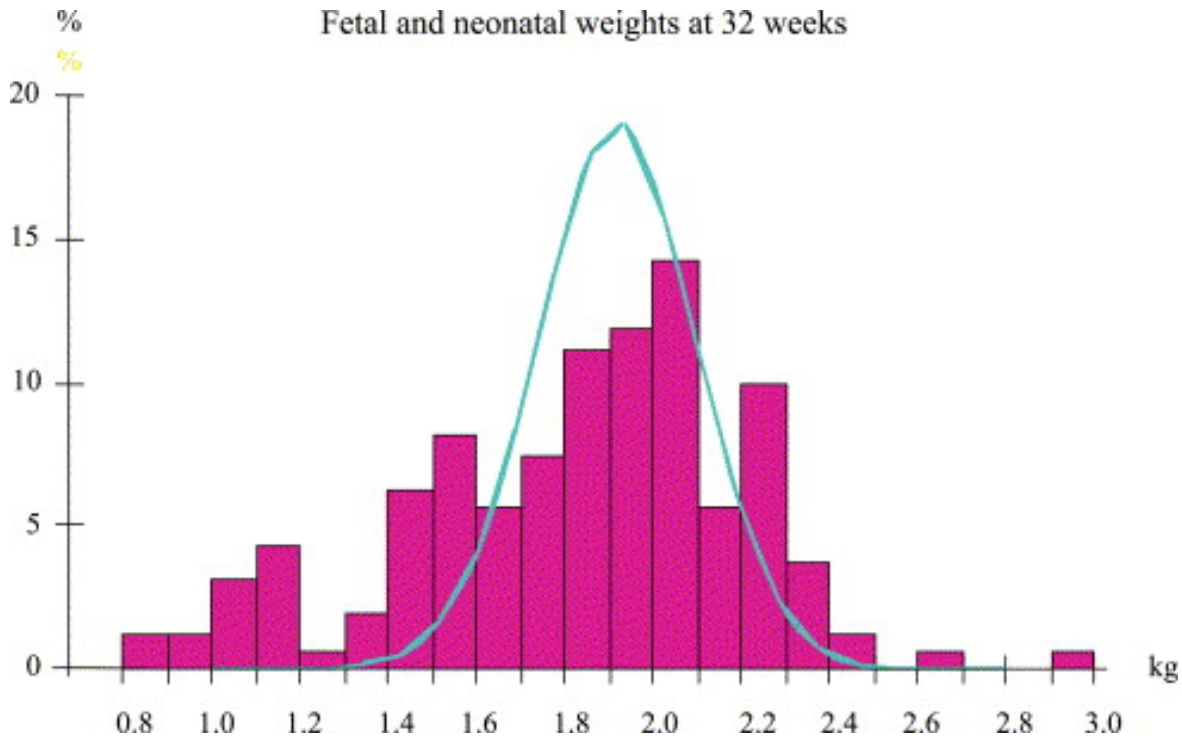
Bar Chart

This bar chart has error bars to indicate variability.



Histograms

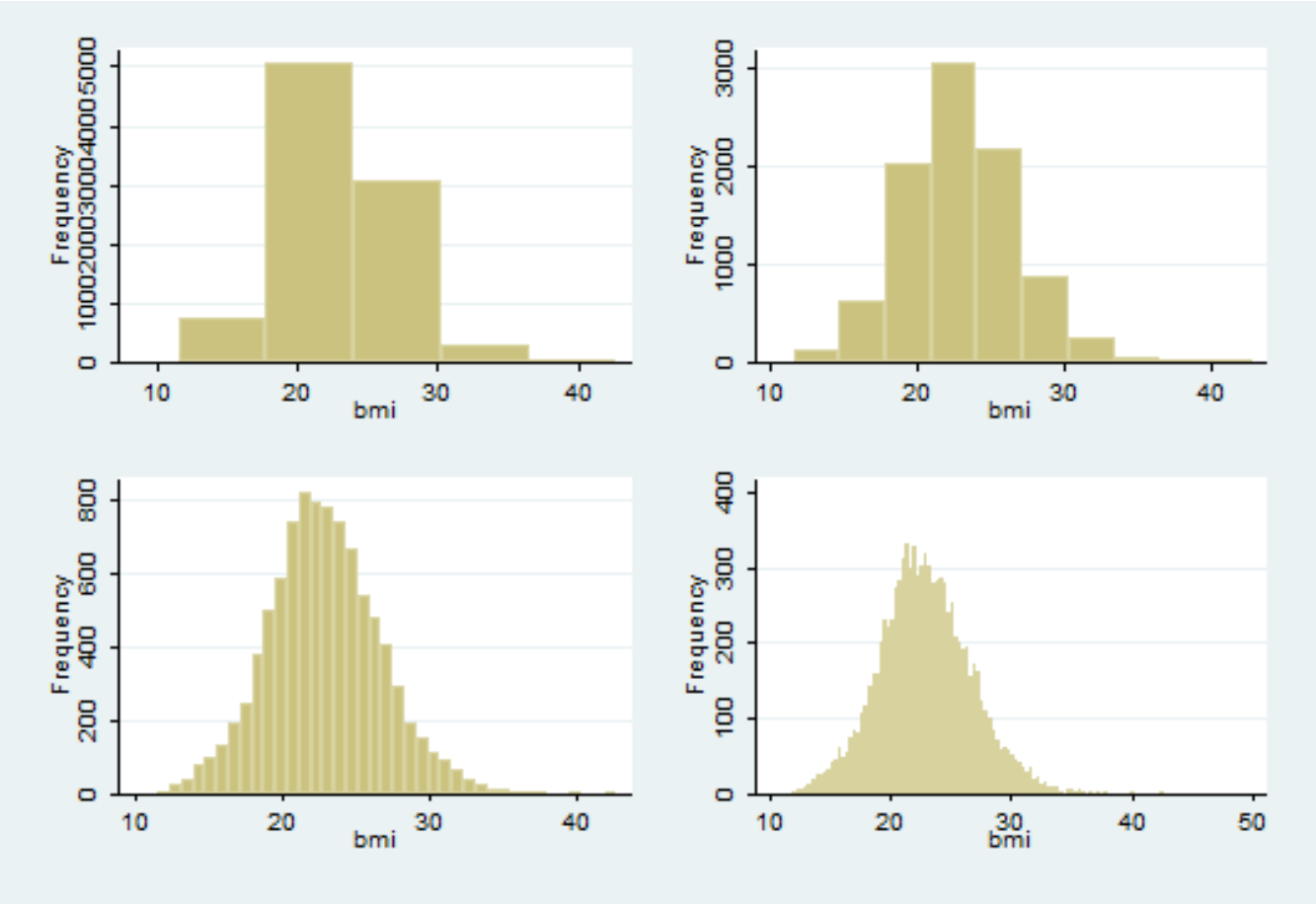
The *histogram* displays a frequency distribution for discrete or continuous data. The area of each bar is proportional to the frequency (or relative frequency) of the categories. The bars may touch, indicating the underlying data are continuous. Here is a histogram of weights of infants born at 32 weeks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



Tick marks represent numeric values of interval limits. Width of the intervals (bins) can be modified.

Histogram

BMI of subjects in China Health & Nutrition Survey, varying the bin width.



Histogram Comparing Two Groups (100m Reaction Time)

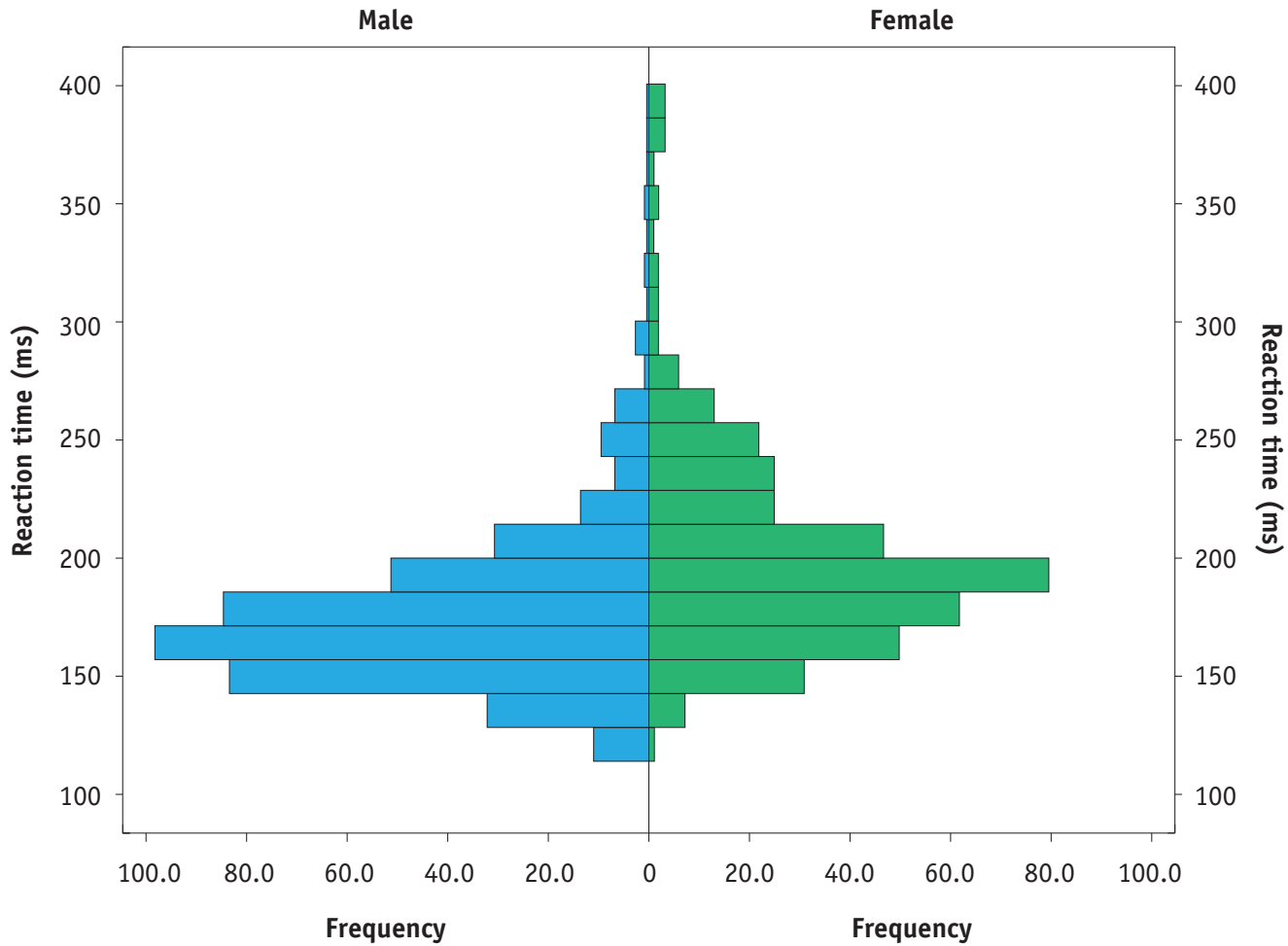


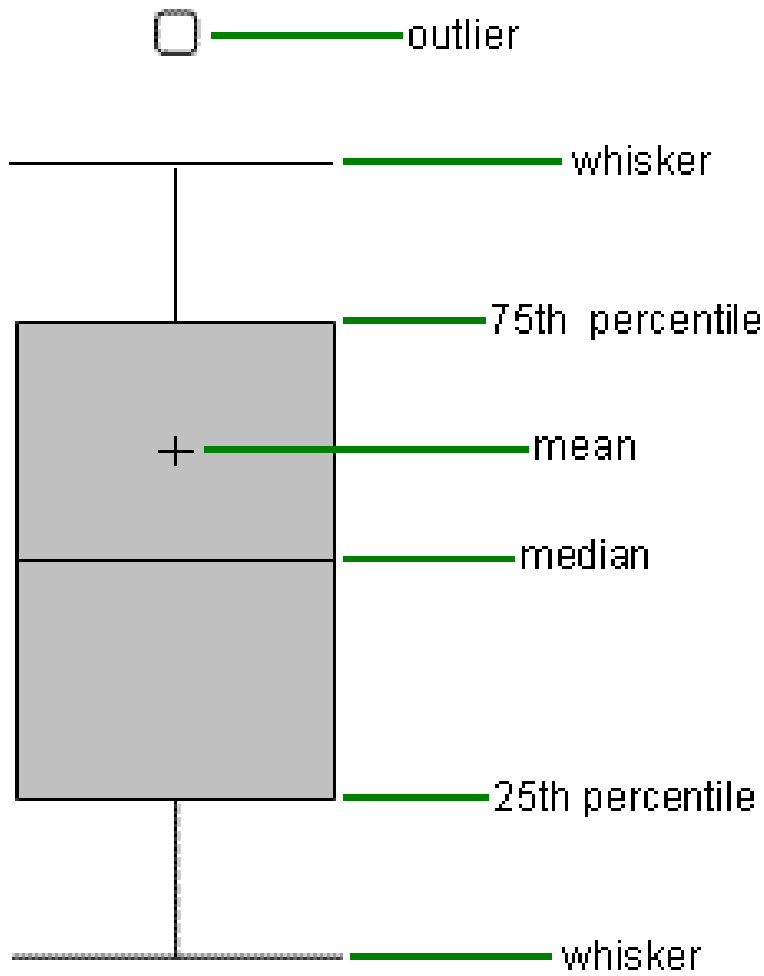
Figure 2. Reaction times of 425 sprinters at the Beijing Olympics. Mean female reaction times are 23 ms slower than for men

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. We will discuss these summary statistics in greater detail soon.

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; R: boxplot

Box Plot


outlier
whisker
75th percentile
mean
median
25th percentile
whisker

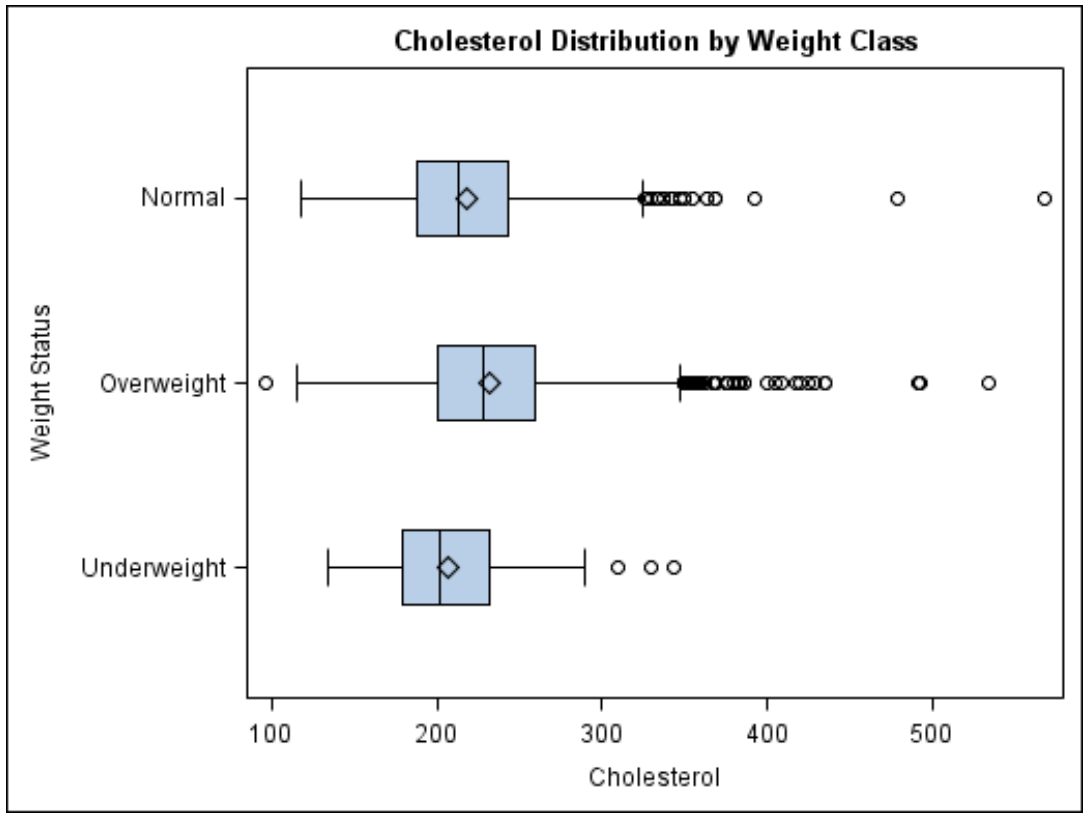
Box Plot

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


Cholesterol Distribution by Weight Class
Weight Status
Normal
Overweight
Underweight
100
200
300
400
500
Cholesterol

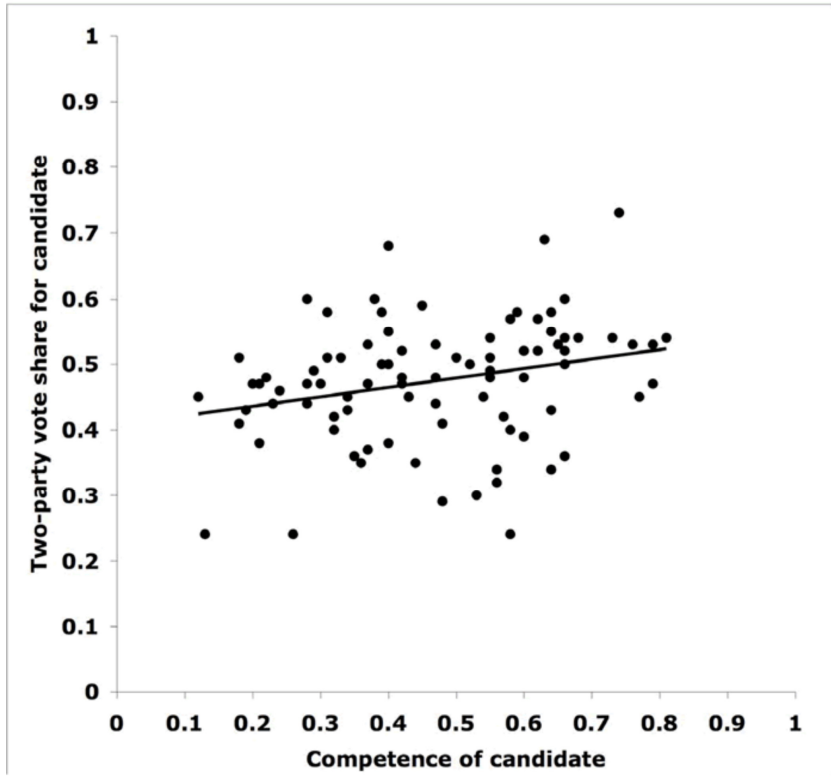
Two-Way Scatter Plots

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; R: plot

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.

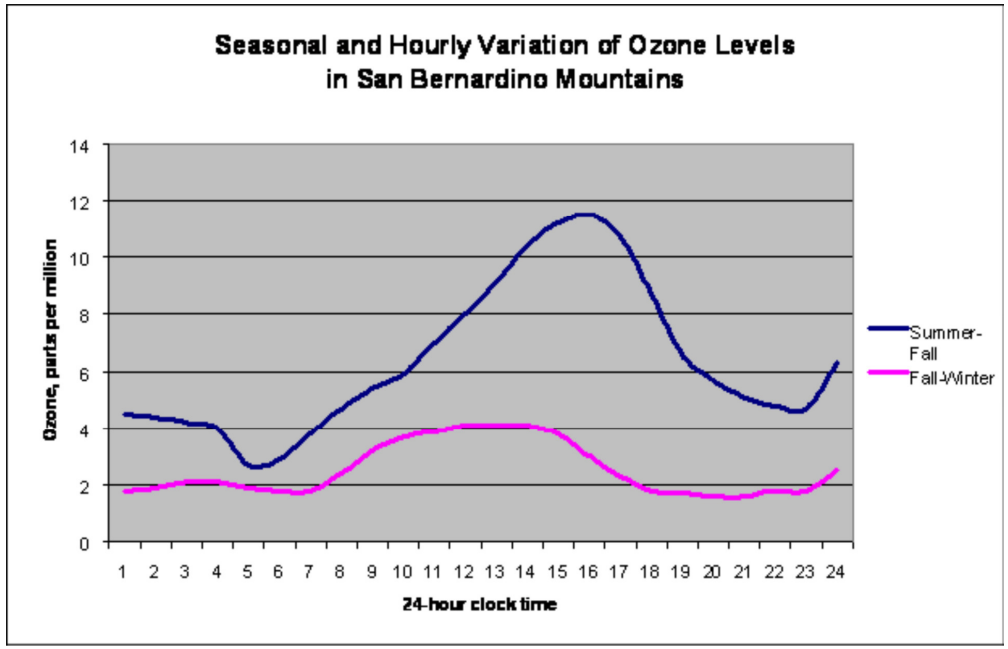


Life Expectancy and Wealth

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.



Reading for Next Time

- Pagano and Gauvreau, Chapter 3 (or web materials on mean, median, mode, variance, standard deviation – say from Wikipedia)
- Measures of Center
- Standard Deviation
- Mean, Median, and Mode (example)
- Measures of center and scale in Stata