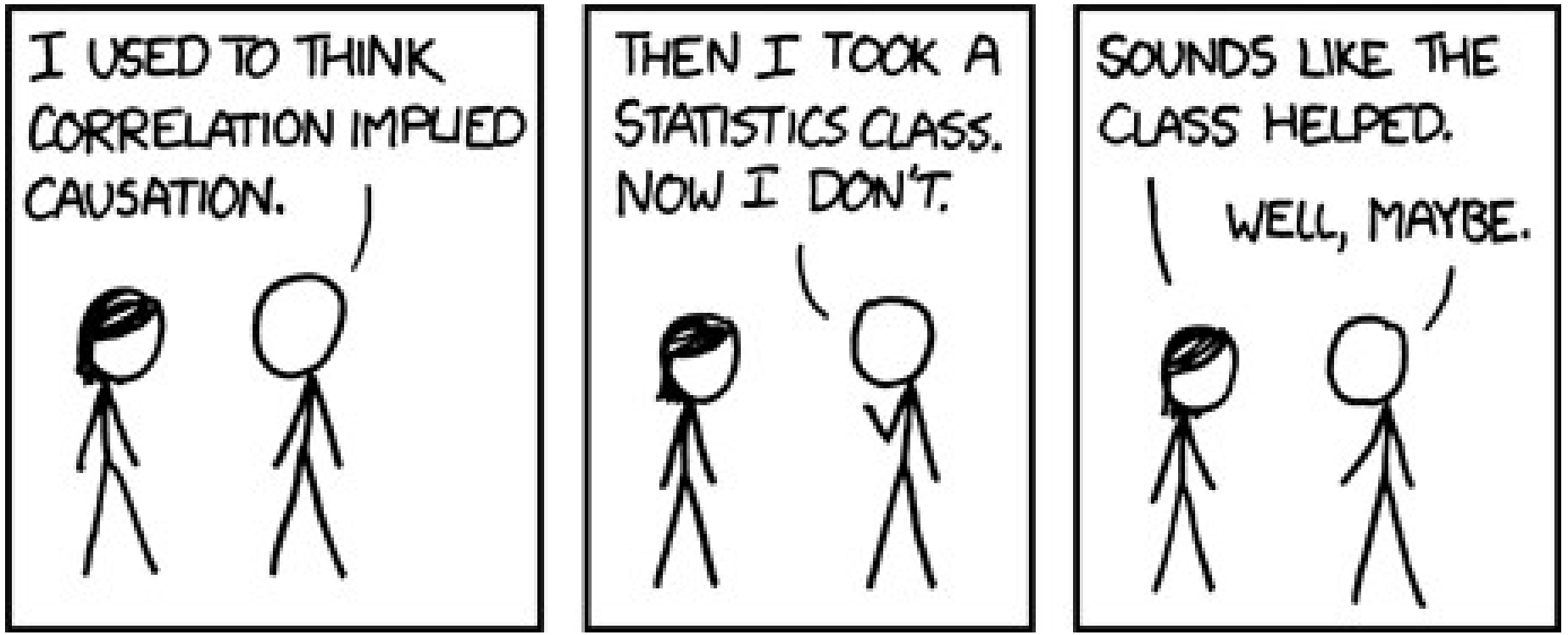
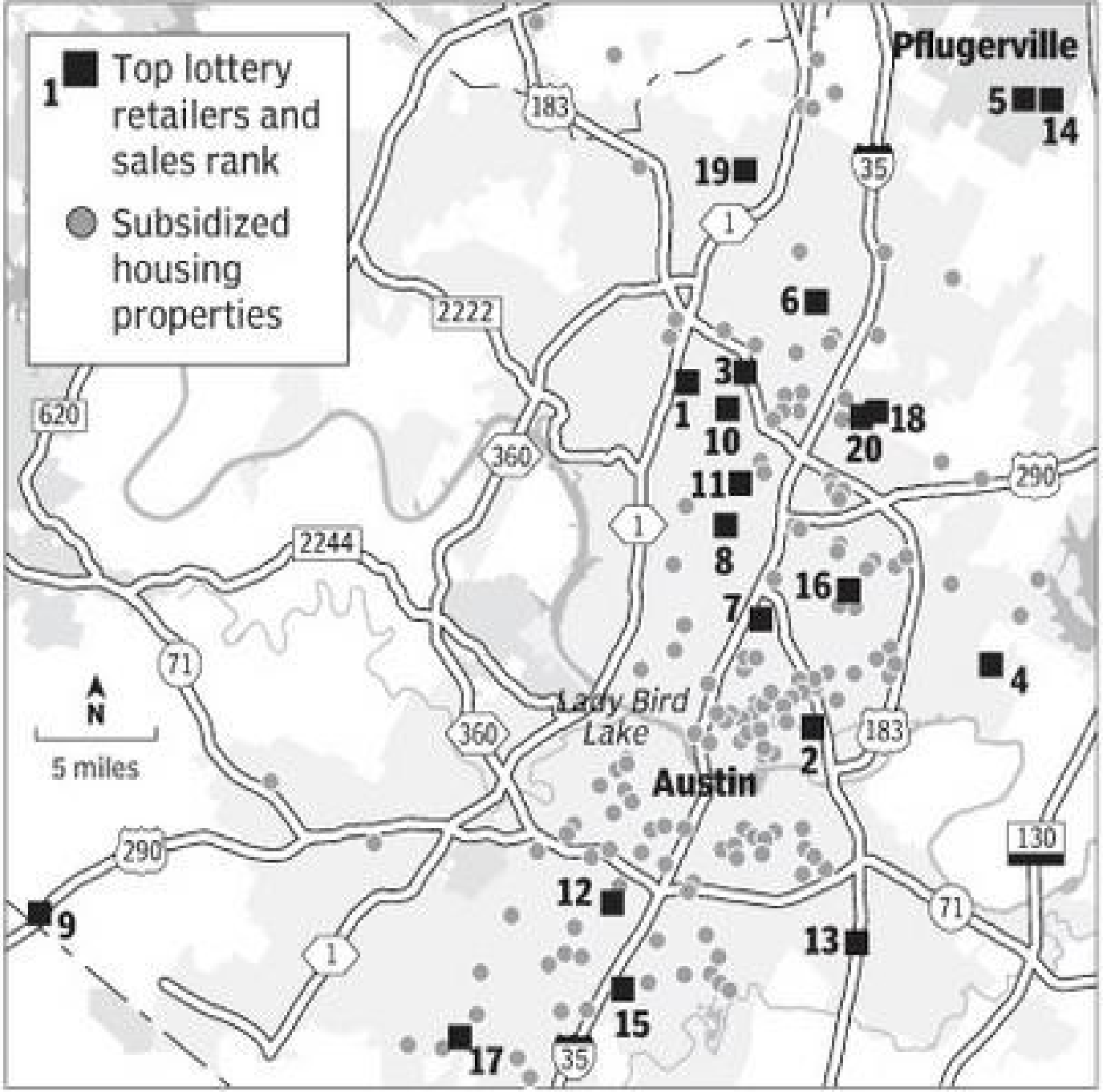
Inquiry 1 reports due T 7/26

Today: Analyzing Data and Statistics



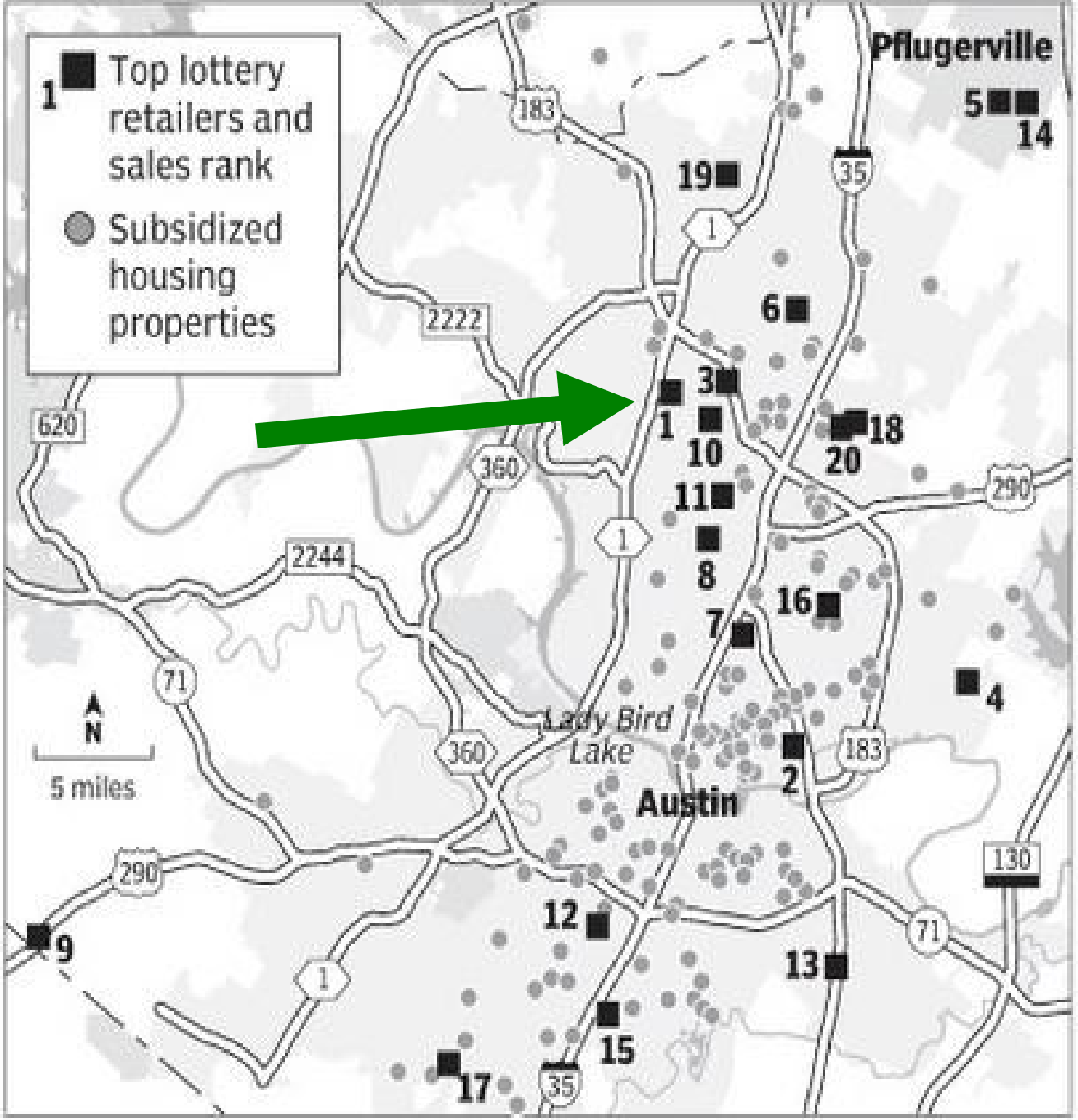
xkcd.com



statesman.com 9/5/10

AMERICAN-STATESMAN

"Austin's top lottery outlets are surrounded by low-income housing."



statesman.com 9/5/10

AMERICAN-STATESMAN

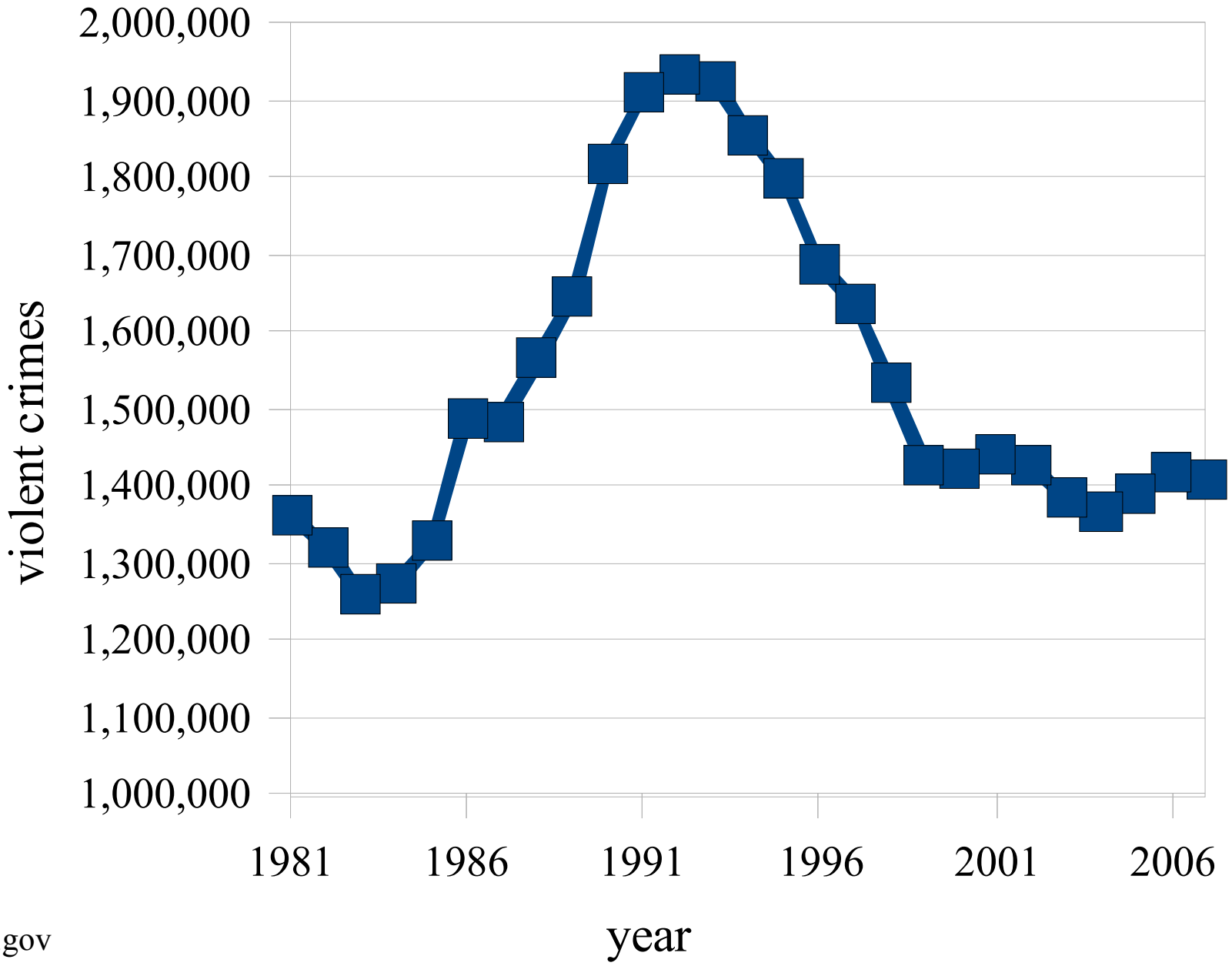
"the best-selling lottery outlet is the Zip-N, on Shoal Creek Boulevard at Anderson Lane, an area with a mix of upscale homes, older subdivisions and apartment complexes."

2007 crime clock statistics:

Every 22.4 seconds	**One Violent Crime**
Every 31.0 minutes	One Murder
Every 5.8 minutes	One Forcible Rape
Every 1.2 minutes	One Robbery
Every 36.8 seconds	One Aggravated Assault
Every 3.2 seconds	**One Property Crime**
Every 14.5 seconds	One Burglary
Every 4.8 seconds	One Larceny-theft
Every 28.8 seconds	One Motor Vehicle Theft

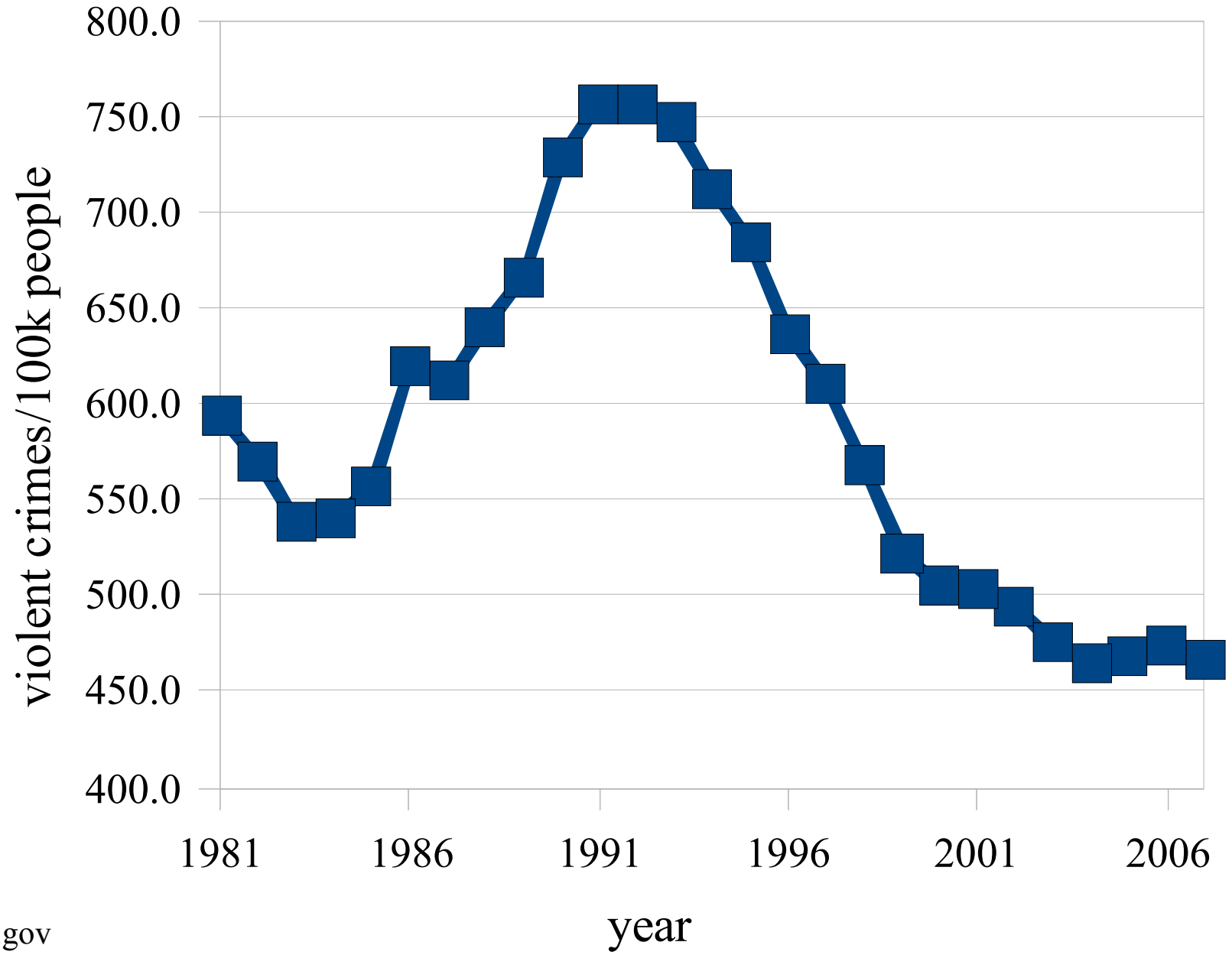
From FBI.gov

U.S. Violent Crime



from fbi.gov

U.S. Violent Crimes per Capita



from fbi.gov

http://www.wunderground.com/cgi-bin/findweather/getForecast?query=78705

Average height in RM class

Average height in RM class
On average are men taller than women?

Average height of Wed. lab vs. Thur. lab

… OR... Average height of male vs female RM students

Most statistical work can be done, and more easily done, by computer using programs such as:

MS Excel

Open Office

SPSS

SAS

Most statistical work can be done, and more easily done, by computer using programs such as:

MS Excel is the most common.

Available from UT for cheap, ~$30.

If you have not used it, start practicing now.

Most statistical work can be done, and more easily done, by computer using programs such as:

Open Office is a free alternative.

If you have not used it, start practicing now.

The Basics:
mean, median, and mode

Mean- aka the average.

Sum of all numbers divided by the number of data points.

$(14+17+7+6+4+11+8)/7 = 9.57$

Median- the middle number of a group of ordered numbers

1 17 7 6 4 11 8

4 6 7 8 11 14 17 median is 8

Median- the middle number of a group of ordered numbers

1 17 7 6 4 11 8

2 6 7 8 11 14 17 median is 8

What about 4 6 7 11 14 17?

Median- the middle number of a group of ordered numbers

1 17 7 6 4 11 8

2 6 7 8 11 14 17 median is 8

What about 4 6 7 11 14 17?

Median is 9.

Mode- the most common value in a group.

9, 8, 3, 4, 5, 2, 4, 5, 2, 3, 6, 1, 6, 2, 3, 9, 2, 6

Mode is 2

Why are there 3 ways to analyze a group of numbers?

Why are there 3 ways to analyze a group of numbers?

The mean is the most common form of analysis.

Why are there 3 ways to analyze a group of numbers?

The mean is the most common form of analysis.
2, 3, 2, 4, 2, 7, 2, 5, 3, 2, 5, 4, 3, 5, 6, 121, 130

Mean = 18

Why are there 3 ways to analyze a group of numbers?

2, 3, 2, 4, 2, 7, 2, 5, 3, 2, 5, 4, 3, 5, 6, 121, 130

Mean = 18

Is this an accurate representation of these numbers?

Why are there 3 ways to analyze a group of numbers?

2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 6, 7, 121, 130

Median = 4

Mean = 18

Median can be more accurate when there are a few especially large or small numbers.

What is your favorite color?

What is your favorite color?

Mode can be used with non-numerical data.

Is there a numerical way to determine the accuracy of our analysis?

2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 6, 7, 121, 130

Median = 4

Mean = 18

Is there a numerical way to determine the accuracy of our analysis?

2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 6, 7, 121, 130

Median = 4

Mean = 18

Standard Deviation is a measure of variability.

Standard deviation is a measure of variability. The standard deviation is the root mean square (RMS) deviation of the values from their arithmetic mean.

$$s = \sqrt{\frac{\sum (X - M)^2}{n - 1}}$$

where $\sum$ = Sum of

X = Individual score

M = Mean of all scores

N = Sample size (Number of scores)

Is there a numerical way to determine the accuracy of our analysis?

2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 6, 7, 121, 130

Mean = 18

Standard deviation = 40.5

Standard deviation is a measure of variability.

Is there a numerical way to determine the accuracy of our analysis?

2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 6, 7

Mean = 3.67

Standard deviation = 1.6

Standard deviation is a measure of variability.

Is there a numerical way to determine the accuracy of our analysis?

2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 6, 7
(121, 130)

Mean = 3.67

Median was 4

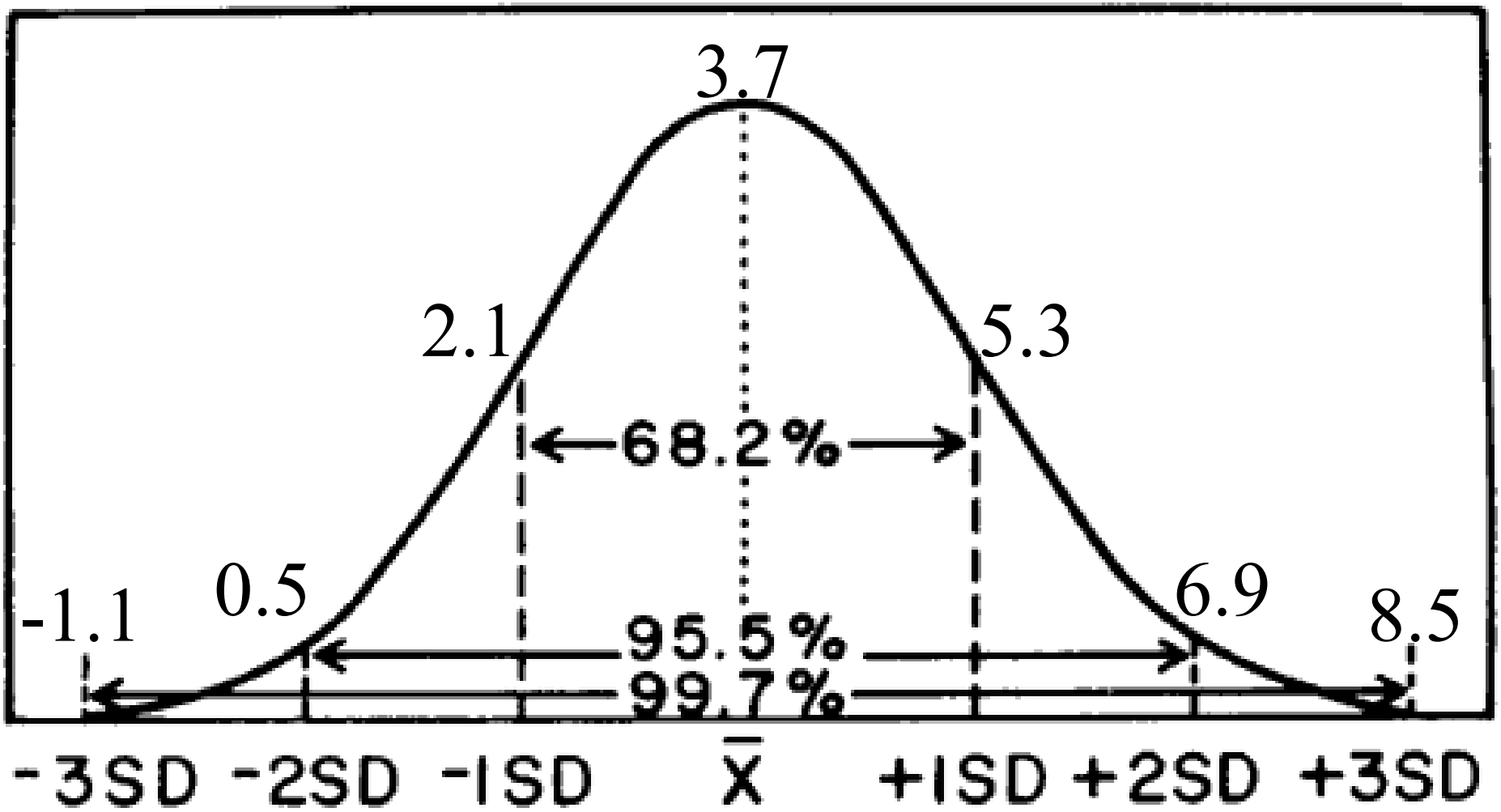
Is there a numerical way to determine the accuracy of our analysis?

2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 6, 7

Mean = 3.67 ±1.6

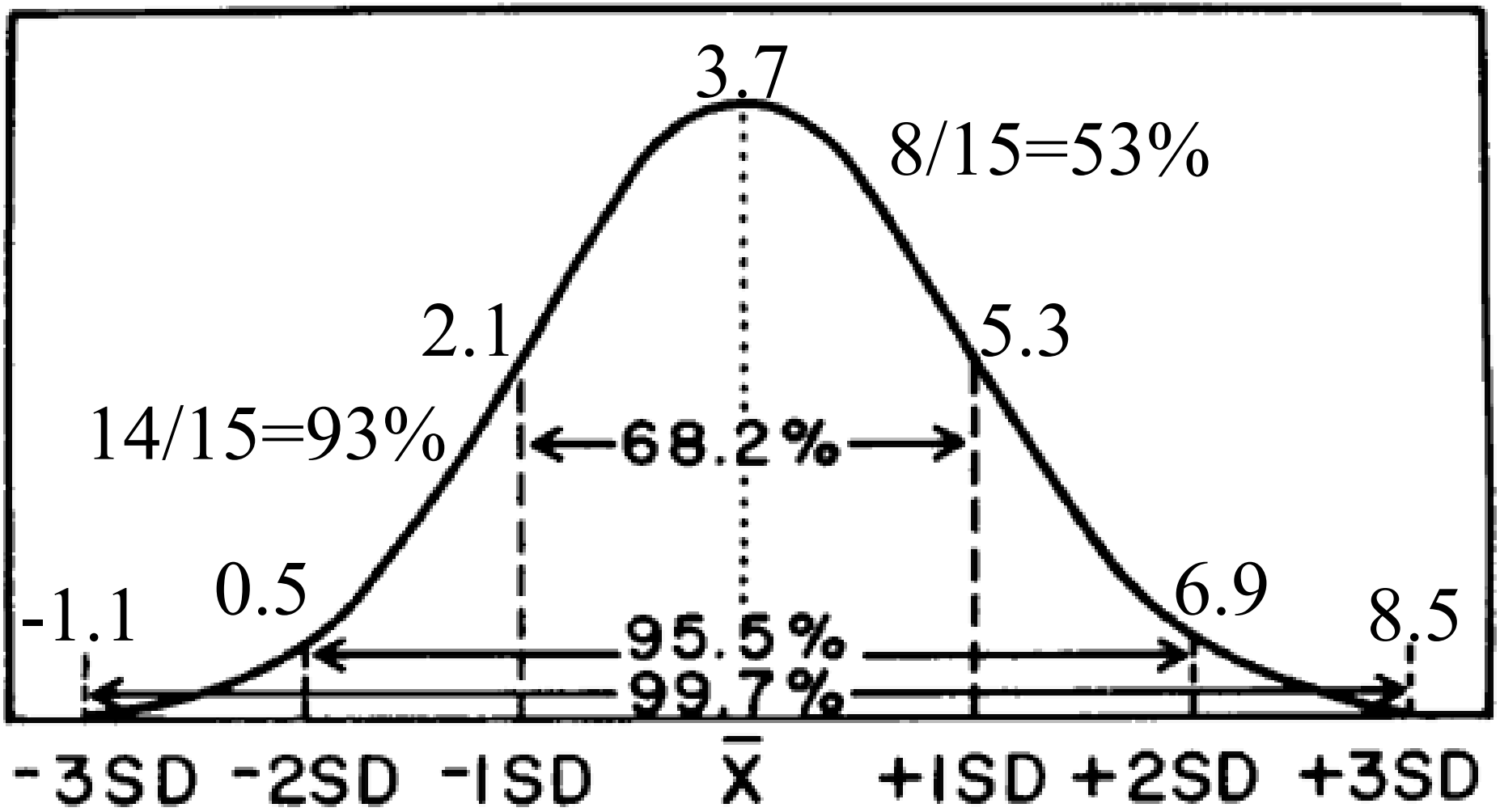
Standard deviation is a measure of variability.

Percent of data at 1, 2, or 3 standard deviations from the mean



http://www.westgard.com/lesson34.htm

2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 6, 7



15/15=100%

http://www.westgard.com/lesson34.htm

How significant of a difference is this?

Set 1= 2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 6, 7

Mean = 3.67 ±1.6

And

Set 2= 8, 6, 7, 8, 9, 5, 6, 7, 9, 8, 9, 5

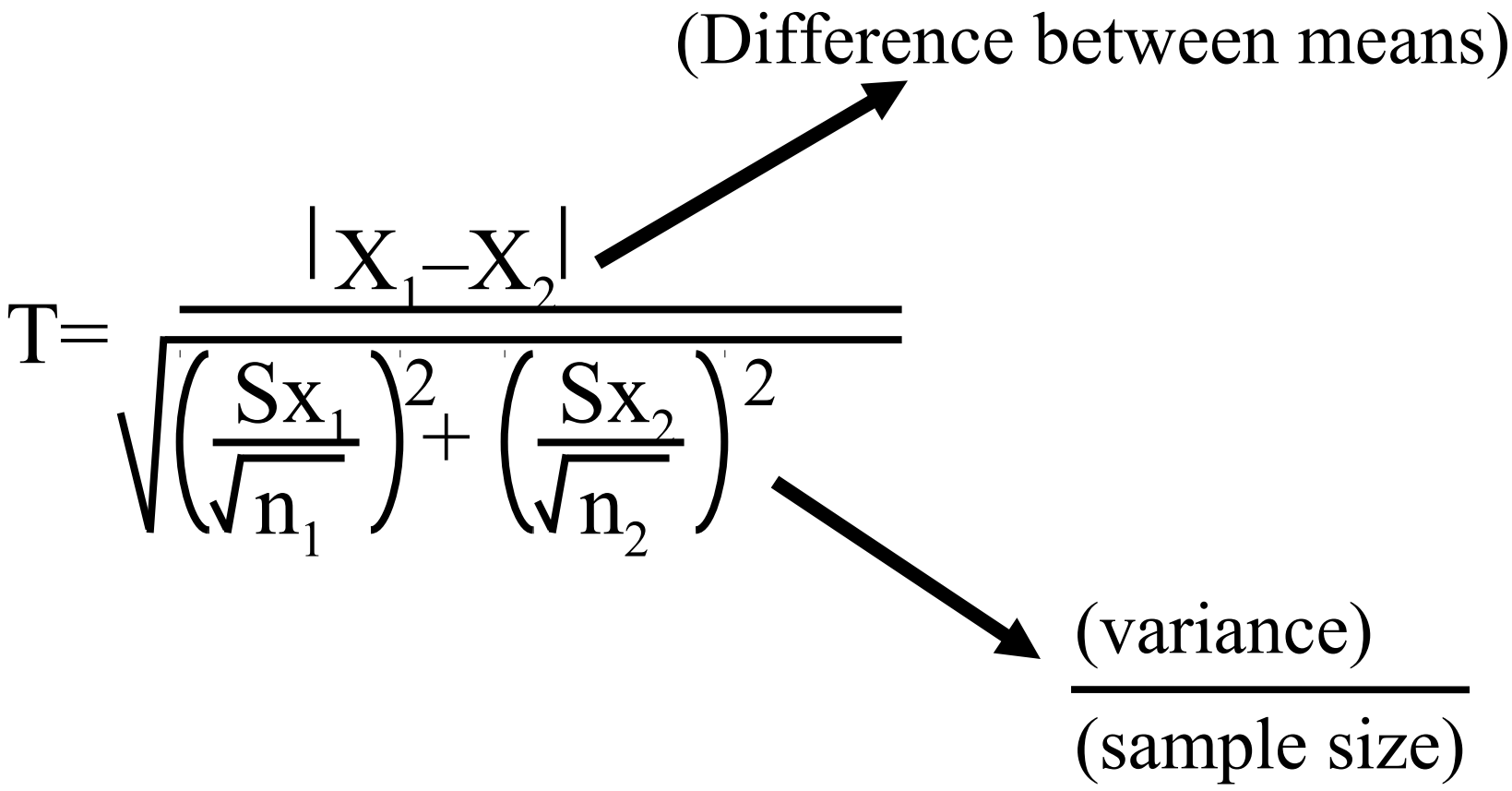
Mean = 7.25 ± 1.48

The ‘Students’ T-test is a method to assign a numerical value of statistical difference.

The ‘Students’ T-test is a method to assign a numerical value of statistical difference.

$$T = \frac{|X_1 - X_2|}{\sqrt{\left(\frac{Sx_1}{\sqrt{n_1}}\right)^2 + \left(\frac{Sx_2}{\sqrt{n_2}}\right)^2}}$$

The ‘Students’ T-test is a method to assign a numerical value of statistical difference.



The ‘Students’ T-test is a method to assign a numerical value of statistical difference.

$$T = \frac{|X_1 - X_2|}{\sqrt{\left(\frac{Sx_1}{\sqrt{n_1}}\right)^2 + \left(\frac{Sx_2}{\sqrt{n_2}}\right)^2}}$$

T is then used to look up the P-value from a table. Also need ‘degrees of freedom’ $= (n_1+n_2)-1$.

Partial table for determining P from T

	P-value		
Df	**0.05**	**0.02**	**0.01**
1	12.71	31.82	63.66
2	4.303	6.965	9.925

T

How significant of a difference is this? Using a speadsheet to get a P value = 3.44×10^{-6}.

Set 1= 2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 6, 7

Mean = 3.67 ±1.6

And

Set 2= 8, 6, 7, 8, 9, 5, 6, 7, 9, 8, 9, 5

Mean = 7.25 ± 1.48

How significant of a difference is this?

P value = 3.44×10^{-6}. So the chance that these 2 sets of data are **not** significantly different is 3.44×10^{-6}

Set 1= 2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 6, 7

Mean = 3.67 ± 1.6

And

Set 2= 8, 6, 7, 8, 9, 5, 6, 7, 9, 8, 9, 5

Mean = 7.25 ± 1.48

How significant of a difference is this?

P value = 3.44×10^{-6}. So the chance that these 2 sets of data are significantly different is

$1 - 3.44 \times 10^{-6}$ or 0.999996559

We can be 99.9996559% certain that the difference is statistically significant.

Set 1= 2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 6, 7

Mean = 3.67 ±1.6

Set 2= 8, 6, 7, 8, 9, 5, 6, 7, 9, 8, 9, 5

Mean = 7.25 ± 1.48

Generally a P-value of 0.05 or less is considered a statistically significant difference.

20% random difference : 80% confidence

10% random difference : 90% confidence

5% random difference : 95% confidence

1% random difference : 99% confidence

0.1% random difference : 99.9% confidence

T-test is one valid and accurate method for determining if 2 means have a statistically significant difference, or if the difference is merely by chance.

Spreadsheet T-test-

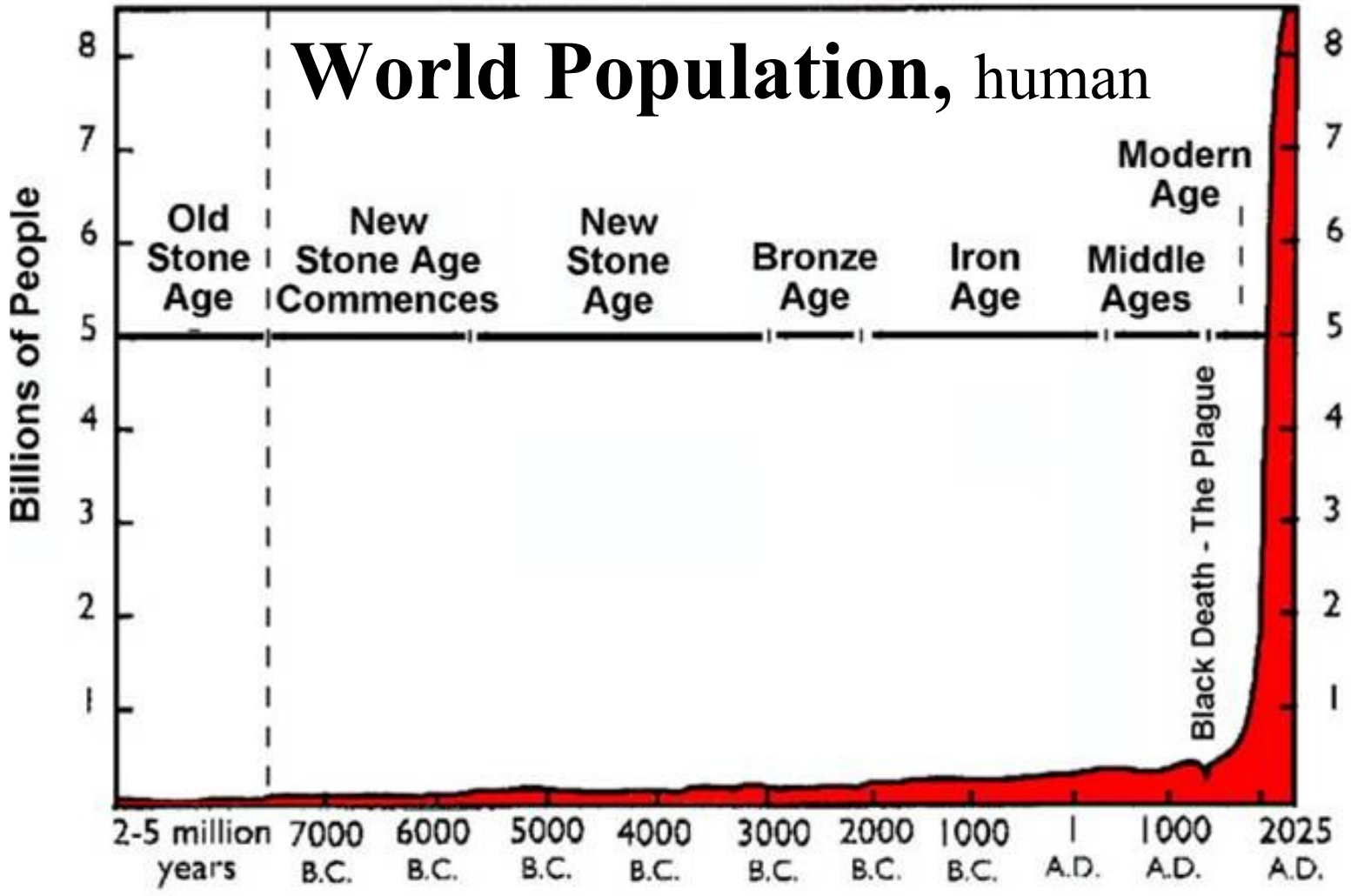
mode/tails- depends on your prediction about the direction of the difference between the groups. If you predicted group A would be lower than group B, pick 1 tail. If you predicted group B would lower than group A, pick 1 tail. If you didn't predict which would be higher, use 2 tails. You can't change your mind after the data are gathered.

type/variance- There are three types of T test you can do. One type compares the same subjects in 2 different conditions. Like if you test whether heart rate increased after drinking a cup of hot sauce or whether plant growth would increase after adding fertilizer to pots of soil. In these cases you would be comparing the heart rate of the same people, or the growth of the same pot of plants, before and after the treatment. This requires a "paired" or "dependent" T test. Excel and Open Office call this a "type 1" test.

If you are comparing different subjects, this is an independent T-test. If you want to know whether nursing students consume more coffee than do biology students. You would then have two groups of test subjects rather than taking 2 measurements on each person. Now you would use an "unpaired" or "independent" T-test. Excel and Open Office call these "type 2" or "type 3" tests. Now the tricky part is to decide which of these to use. Are the standard deviations about the same for both groups, or are they different? If in doubt, go with "type 3" for unequal variances.

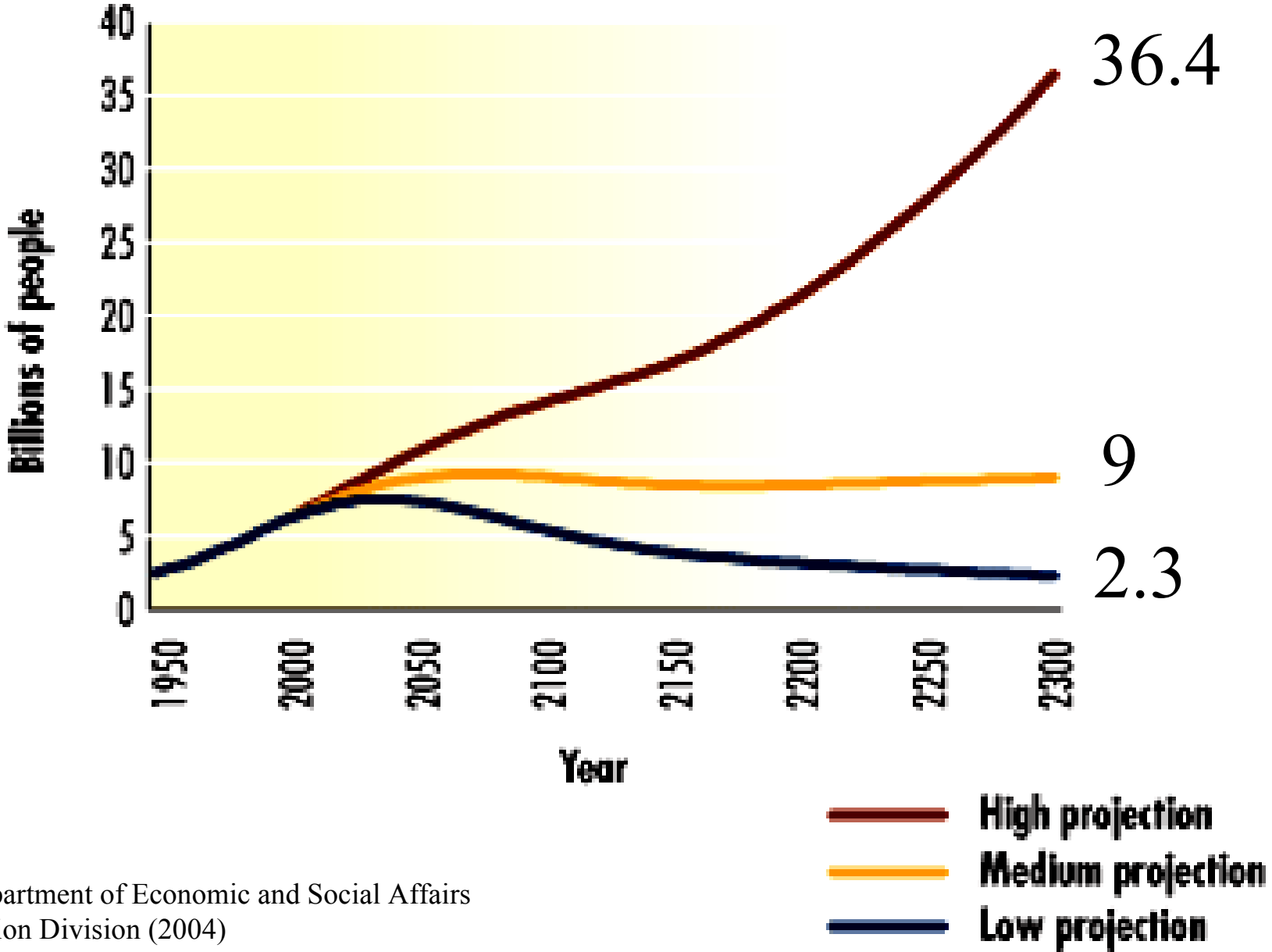
Information for mode and type adapted from Alverno College (http://depts.alverno.edu/nsmt/stats.htm)

We can have uncertainty about past events... or future events



From "World Population: Toward the Next Century," copyright 1994 by the Population Reference Bureau

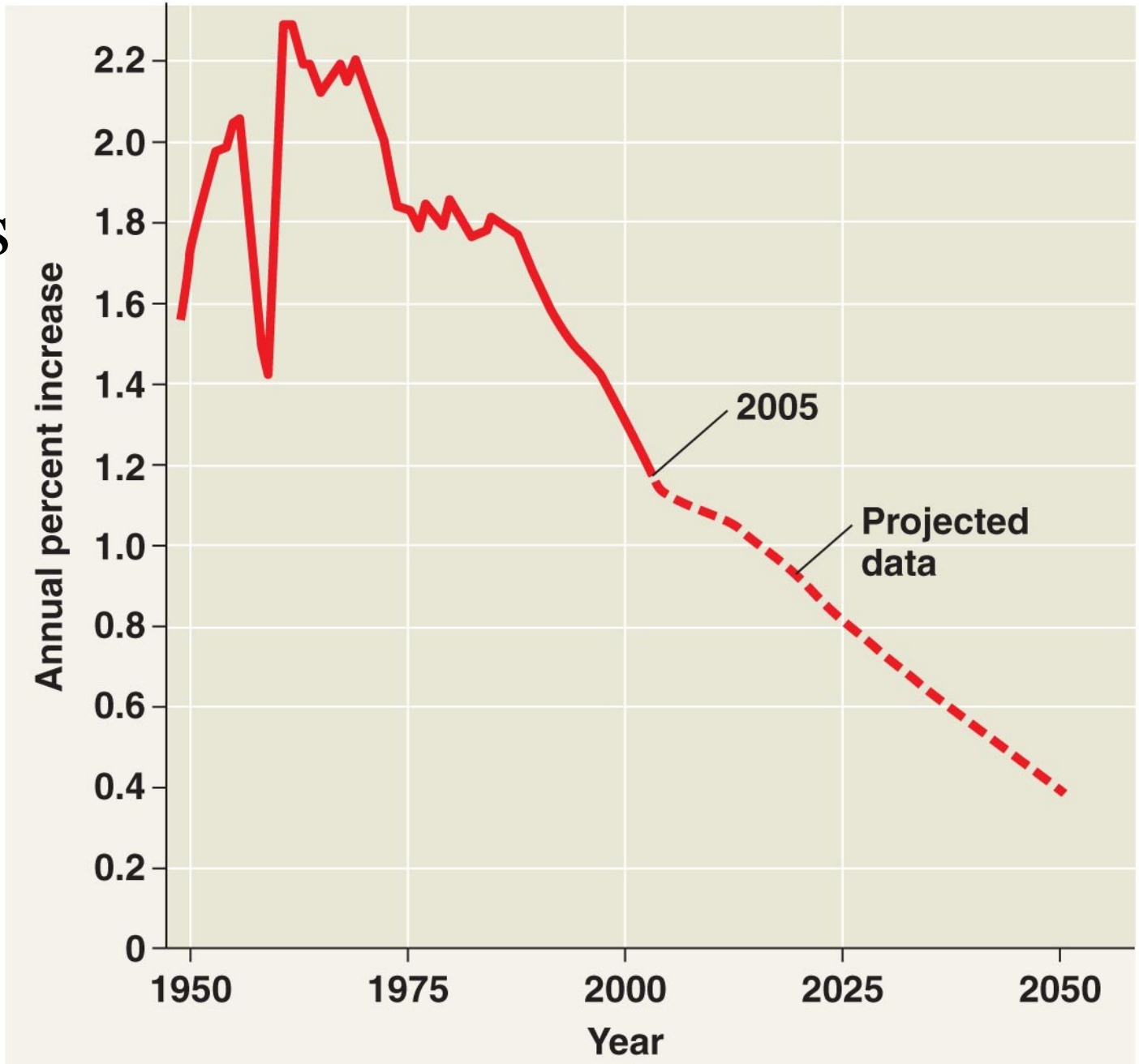
World Population Trends and Projections, 1950 – 2300



UN Department of Economic and Social Affairs
Population Division (2004)

Annual % increase in global population

Predictions are based on current conditions and trends



Campbell's "Biology fig. 53.23

The Chi Square Test

This was not covered in class, but I am leaving it in the presentation if someone needs it.

The Chi Square Test

- A statistical method used to determine goodness of fit
  - Goodness of fit refers to how close the observed data are to those predicted from a hypothesis

This was not covered in class, but I am leaving it in the presentation if someone needs it.

The Chi Square Test

- The general formula is

$$\chi^2 = \Sigma \frac{(O - E)^2}{E}$$

- where
  - O = observed data in each category
  - E = observed data in each category based on the experimenter's hypothesis
  - Σ = Sum of the calculations for each category

This was not covered in class, but I am leaving it in the presentation if someone needs it.

Two flies with different traits are bred together

- Out of 352 offspring
  - 193 straight wings, gray bodies
  - 69 straight wings, ebony bodies
  - 64 curved wings, gray bodies
  - 26 curved wings, ebony bodies

This was not covered in class, but I am leaving it in the presentation if someone needs it.

According to our hypothesis, there should be a 9:3:3:1 ratio of fly offspring

Phenotype	Expected probability	Expected number
straight wings, gray bodies	9/16	9/16 X 352 = 198
straight wings, ebony bodies	3/16	3/16 X 352 = 66
curved wings, gray bodies	3/16	3/16 X 352 = 66

Apply the chi^2 formula

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3} + \frac{(O_4 - E_4)^2}{E_4}$$

$$\chi^2 = \frac{(193 - 198)^2}{198} + \frac{(69 - 66)^2}{66} + \frac{(64 - 66)^2}{66} + \frac{(26 - 22)^2}{22}$$

$$\chi^2 = 0.13 + 0.14 + 0.06 + 0.73$$

$$\chi^2 = 1.06$$

This was not covered in class, but I am leaving it in the presentation if someone needs it.

- Interpret the chi square value
  - The calculated chi square value can be used to obtain probabilities, or P values, from a chi square table
    - These probabilities allow us to determine the likelihood that the observed deviations are due to random chance alone

- If the chi square value results in a probability that is less than 0.05 (ie: less than 5%)
  - The hypothesis is rejected

This was not covered in class, but I am leaving it in the presentation if someone needs it.

- Interpret the chi square value
  - Before we can use the chi square table, we have to determine the degrees of freedom (*df*)
    - The *df* is a measure of the number of categories that are independent of each other
    - $df = n - 1$
      - where n = total number of categories
    - In our experiment, there are four categories
      - Therefore, df = 4 – 1 = 3

This was not covered in class, but I am leaving it in the presentation if someone needs it.

TABLE 2.1

Chi Square Values and Probability

						Null Hypothesis rejected	
Degrees of Freedom	$P = 0.99$	0.95	0.80	0.50	0.20	0.05	0.01
1.	0.000157	0.00393	0.0642	0.455	1.642	3.841	6.635
2.	0.020	0.103	0.446	1.386	3.219	5.991	9.210
3.	0.115	0.352	1.005	1.06 2.366	4.642	7.815	11.345
4.	0.297	0.711	1.649	3.357	5.989	9.488	13.277
5.	0.554	1.145	2.343	4.351	7.289	11.070	15.086
6.	0.872	1.635	3.070	5.348	8.558	12.592	16.812
7.	1.239	2.167	3.822	6.346	9.803	14.067	18.475
8.	1.646	2.733	4.594	7.344	11.030	15.507	20.090
9.	2.088	3.325	5.380	8.343	12.242	16.919	21.666
10.	2.558	3.940	6.179	9.342	13.442	18.307	23.209
15.	5.229	7.261	10.307	14.339	19.311	24.996	30.578
20.	8.260	10.851	14.578	19.337	25.038	31.410	37.566
25.	11.524	14.611	18.940	24.337	30.675	37.652	44.314
30.	14.953	18.493	23.364	29.336	36.250	43.773	50.892

From Fisher, R. A., and Yates, F. (1943) *Statistical Tables for Biological, Agricultural, and Medical Research.* Oliver and Boyd, London.

- Interpret the chi square value
  - With *df* = *3,* the chi square value of 1.06 is slightly greater than 1.005 (which corresponds to P= 0.80)
  - A P = 0.80 means that values equal to or greater than 1.005 are expected to occur 80% of the time based on random chance alone
  - Therefore, it is quite probable that the deviations between the observed and expected values in this experiment can be

This was not covered in class, but I am leaving it in the presentation if someone needs it

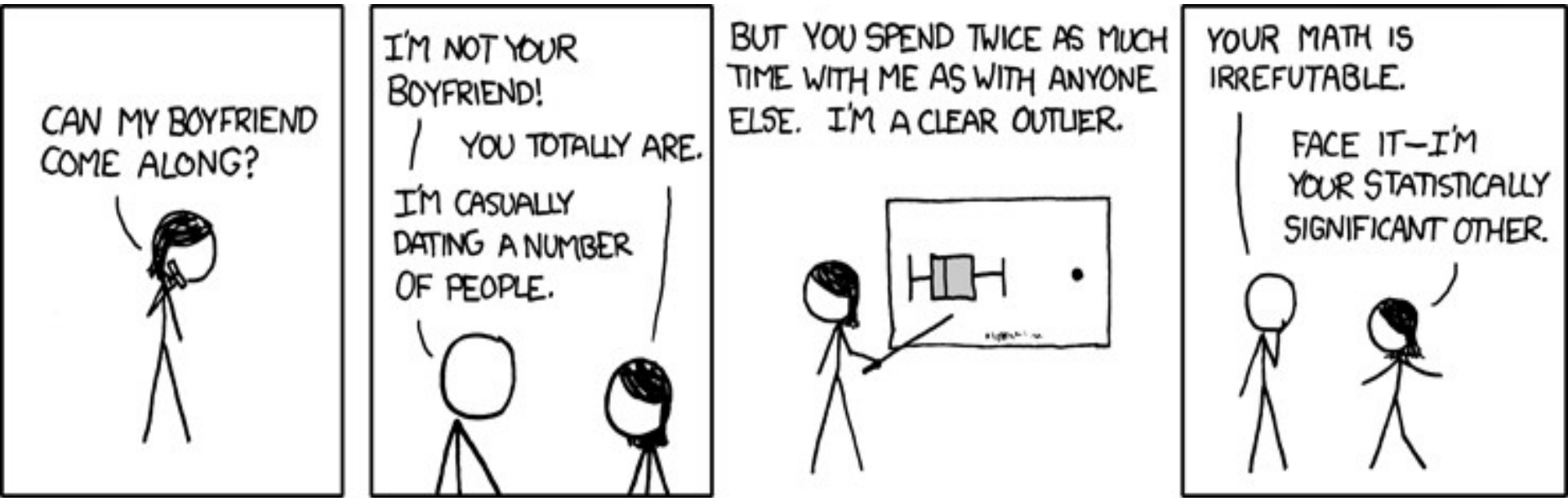
Spreadsheet applications will compute chi^2

Is the male:female ratio in the CNS different from the general population?

observed	expected	
40	50	male
60	50	female
Chi-sq =	0.0455	

This was not covered in class, but I am leaving it in the presentation if someone needs it.

Inquiry 1 Reports due T 7/26



xkcd.com