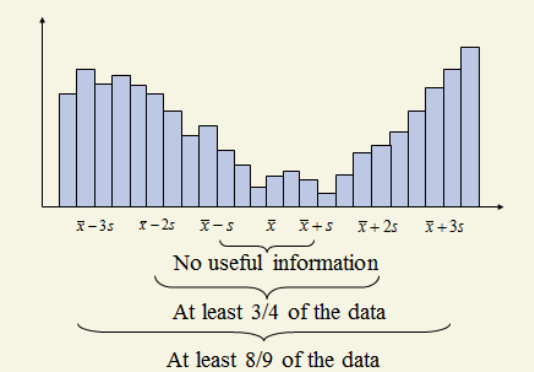
Chapter 2.5 Interpreting Standard Deviation

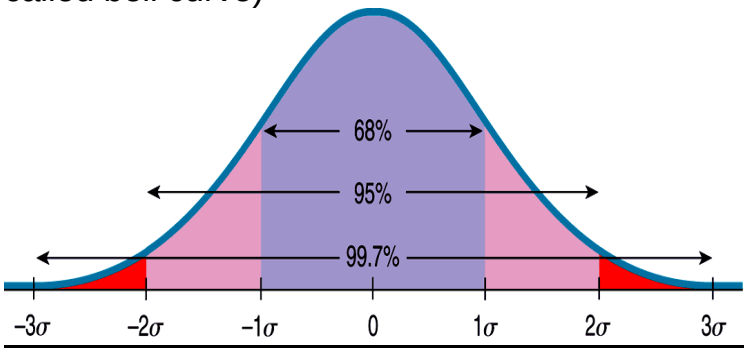
- Chebyshev Theorem
- Empirical Rule

Chebyshev Theorem says that for ANY shape of data distribution

- at least 3/4 of all data fall no farther from the mean than 2 standard deviations away,
- at least 8/9 of all data fall within 3 standard deviations from the mean,
- In general, for any number k>1, the interval $(\bar{x}-ks, \bar{x}+ks)$ contains at least a fraction $1-\frac{1}{k^2}$ of all measurements.



Empirical Rule (68-95-99.7 rule) applies ONLY to Normal Distribution (modeled by so called bell curve)



In a Normal model:

- about 68% of the values fall within one standard deviation of the mean;
- about 95% of the values fall within two standard deviations of the mean; and,
- about 99.7% (almost all!) of the values fall within three standard deviations of the mean.

Ch. 2.5 p.85 #78 Using Chebyshev

2.78 Blogs for *Fortune* 500 firms. Refer to the *Journal of Relationship Marketing* (Vol. 7, 2008) study of the prevalence of blogs and forums at *Fortune* 500 firms with both English and Chinese

Web sites, Exercise 2.9 p.49. In a sample of firms that provide blogs and forums as marketing tools, the mean number of blogs/forums per site was 4.25, with a standard deviation of 12.02.

a. Provide an interval that is likely to contain the number of blogs/forums per site for at least 75% of the *Fortune* 500 firms in the sample.

b. Do you expect the distribution of the number of blogs/forums to be symmetric, skewed right, or skewed left? Explain.

2.80 p. 85 Using 68-95-99 rule for normal model

2.80 Motivation of drug dealers. Researchers at Georgia State University investigated the personality characteristics of drug dealers in order to shed light on their motivation for participating in the illegal drug market (*Applied Psychology in Criminal Justice*, Sep. 2009). The sample consisted of 100 convicted drug dealers who attended a court-mandated counseling program. Each dealer was scored on the Wanting Recognition (WR) Scale, which provides a quantitative measure of a person's level of need for approval and sensitivity to social situations. (Higher scores indicate a greater need for approval.) The sample of drug dealers had a mean WR score of 39, with a standard deviation of 6. Assume the distribution of WR scores for drug dealers is mound-shaped and symmetric.

a. Give a range of WR scores that will contain about 95% of the scores in the drug dealer sample.

b. What proportion of the drug dealers had WR scores above 51?

c. Give a range of WR sores that contain nearly all the scores in the drug dealer sample.

Comparing the estimations by Chebyshev and Empirical

The 50 companies' percentages of revenues spent on R&D are below (sorted already)

5.2	5.6	5.9	6.0	6.5	6.5	6.5	6.6	6.8	6.9	6.9	6.9	7.1	7.1	7.2	7.2	7.4	7.5
	7.5	7.7	7.7	7.8	7.9	8.0	8.0	8.1	8.2	8.2	8.2	8.4	8.5	8.8	9.0	9.2	9.4
	9.5	9.5	9.6	9.7	9.9	10.1	10.5	10.5	10.6	11.1	11.3	11.7	13.2	13.5	13.5		

1. Calculate the range and use it to obtain a rough approximation of s.

   Ans: ..... $\leq s \leq$ .....

2. Compute $\overline{x}$ = ... and s = ...
3. Calculate the intervals $[\overline{x} - ks, \overline{x} + ks]$ for k=1, 2, 3, and for each interval give
   a. Percentage estimated by the Chebyshev's Rule
   b. Percentage estimated by the Empirical Rule
   c. The actual percentage of observations in the interval. Compare them with a. and b.

k	$[\overline{x} - ks, \overline{x} + ks]$	Chebyshev' s Rule	Empirical Rule	Actual
1				
2				
3				

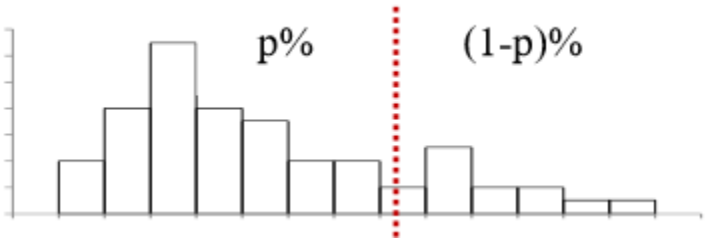
Chapter 2.6 Numerical Measures of Relative Standing

- Percentile Ranking
- The z-score
- Using Empirical Rule to describe relative standing

Measures of position

Percentiles

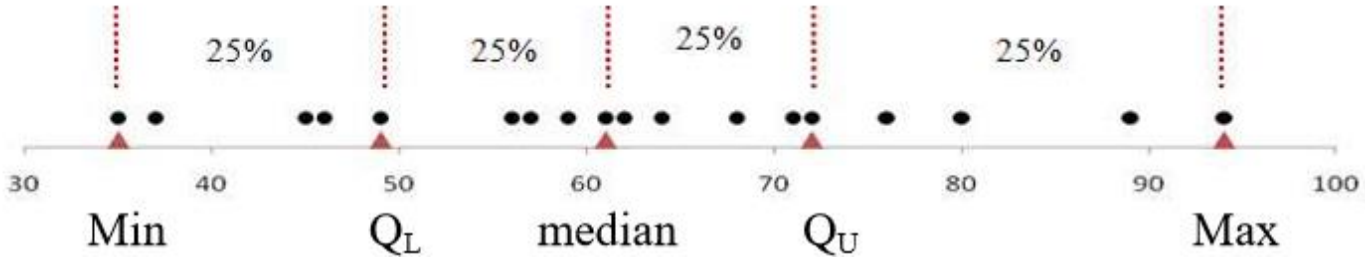
The ***p*th percentile** is a number such that *p*% of the data falls below it and (100 – *p*)% falls above it



Example: You scored 560 on the GMAT exam. This score puts you in the 58th percentile.

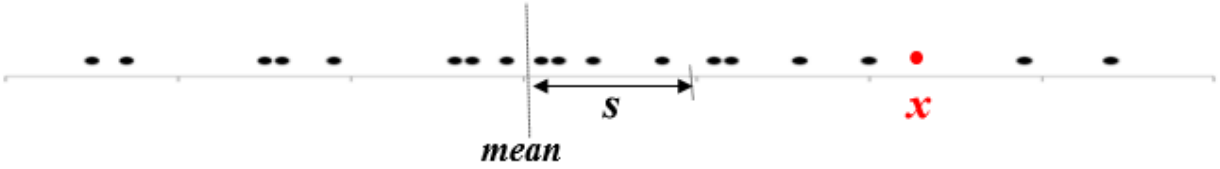
- What percentage of test takers scored lower than you did?
- What percentage of test takers scored higher than you did?

Median = 50th percentile
The lower quartile QL = 25th percentile
The upper quartile QU = 75th percentile



Z-scores

z-score = number of standard deviation that x is above (if positive) or below (if negative) of the mean



For a sample: $z = \frac{x - \bar{x}}{s}$

For the population $z = \frac{x - \mu}{\sigma}$

The mean of z-scores is always 0,
The standard deviation of z-scores is always 1,
For mound-shaped distribution
1. Approximately 68% of the measurements will have a z-score between -1 and 1.
2. Approximately 95% of the measurements will have a z-score between -2 and 2.
3. Approximately 99.7% (almost all) of the measurements will have a z-score between -3 and 3.

Class Exercises: Ch. 2.6

2.93 Compare the *z*-scores to decide which of the following *x* values lie the greatest distance above the mean or the greatest distance below the mean.

a. x=100,μ=50,σ=25

b. x=1,μ=4,σ=1

c. x=0,μ=200,σ=100

d. x=10,μ=5,σ=3

2.99 Lead in drinking water. The U.S. Environmental Protection Agency (EPA) sets a limit on the amount of lead permitted in drinking water. The EPA *Action Level* for lead is .015 milligrams per liter (mg/L) of water. Under EPA guidelines, if 90% of a water system's study samples have a lead concentration less than .015 mg/L, the water is considered safe for drinking. I (coauthor

Sincich) received a recent report on a study of lead levels in the drinking water of homes in my subdivision. The 90th percentile of the study sample had a lead concentration of .00372 mg/L. Are water customers in my subdivision at risk of drinking water with unhealthy lead levels? Explain.

2.102 Blue- vs. red-colored exam study. In a study of how external clues influence performance, professors at the University of Alberta and Pennsylvania State University gave two different forms of a midterm examination to a large group of introductory students. The questions on the exam were identical and in the same order, but one exam was printed on blue paper and the other on red paper (*Teaching Psychology*, May 1998). Grading only the difficult questions on the exam, the researchers found that scores on the blue exam had a distribution with a mean of 53% and a standard deviation of 15%, while scores on the red exam had a distribution with a mean of 39% and a standard deviation of 12%. (Assume that both distributions are approximately mound-shaped and symmetric.)

a. Give an interpretation of the standard deviation for the students who took the blue exam.

b. Give an interpretation of the standard deviation for the students who took the red exam.

c. Suppose a student is selected at random from the group of students who participated in the study and the student's score on the difficult questions is 20%. Which exam form is the student more likely to have taken, the blue or the red exam? Explain.

Chapter 2.7 **Methods for Detecting Outliers: Box Plots and *z*-Scores**

- Five-Numbers-Summary
- Box Plot
- Outliers

The first (**lower**) quartile is the 25^{th} percentile. It is the point below which lie 1/4 of the data.
The second quartile (the median) is the 50^{th} percentile. It is the point below which lie 1/2 of the data.
The third (**upper**) quartile is the 75^{th} percentile. It is the point below which lie 3/4 of the data.

The **interquartile range (IQR)** is the difference between the first and the third quartiles. It is, beside the range, variance and standard deviation, a **measure of spread** (dispersion) of the data.

Box Plot

The **five-number summary** of a distribution is a list of five numbers: the minimum, the first quartile, the median, the third quartile and the maximum.

Example: Data (sorted!):
35 37 45 46 49 56 57 57 59 **61** 62 64 68 71 72 76 80 89 110

Find the five number summary

……..
The **outliers**: unusually large or small observation

Rules of Thumb for Detecting Outliers:

- Box Plot Method:
  - Observations falling beyond the inner fences are called **outliers**.
  - Observations falling between the inner fences and the outer fences are called **suspect outliers**
  - Observations falling beyond the outer fences are deemed **highly suspect outliers.**

- z-scores: Observations with z-scores greater than 3 in absolute value are considered outliers.

A **boxplot** is a graphical display of the five-number summary. Helps to detect the shape of the distribution and the outliers.

Constructing Boxplots (professional way, so called modified boxplot)

1. Draw a single line, horizontal (or vertical) axis, and mark the full range of the data. Draw short perpendicular to the axis lines at the lower and upper quartiles and at the median. Then connect them to form a box (see example).
2. Erect “fences” around the main part of the data.
   a) The upper fence is 1.5 IQRs above the upper quartile.
   b) The lower fence is 1.5 IQRs below the lower quartile.
   c) Note: the fences only help with constructing the boxplot and should not appear in the final display.
3. Use the fences to grow “whiskers.”
   a) Draw lines from the ends of the box up and down to the *most extreme data values found within the fences*.
   b) If a data value falls outside one of the fences, we do *not* connect it with a whisker.

4. Add the outliers by displaying any data values beyond the fences with special symbols.

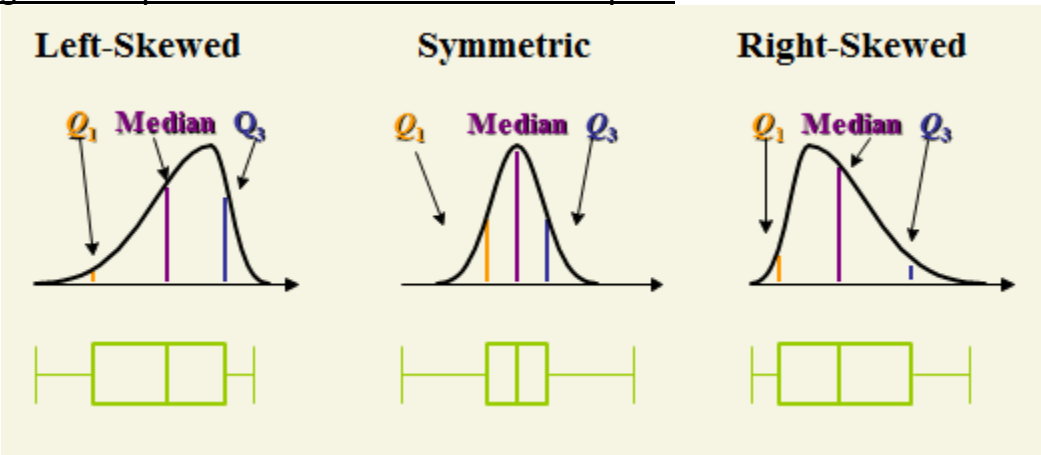
We often use a different symbol for “far outliers” that are farther than 3 IQRs from the quartiles.

The numbers outside of inner fences are suspected outliers. The outliers between the inner and outer fences are called **mild**, and those outside outer fences are called **extreme.**

Exercise: draw the boxplot for the data above (Using a ruler helps to avoid distortions)

Example: (the same – again) Use the TI-83/84calculator to draw a box plot.

Finding the shape of distribution from the box plot:



In case of symmetric, one-peak distribution, the data with z-scores more than 3 or less than -3 are considered the outliers.

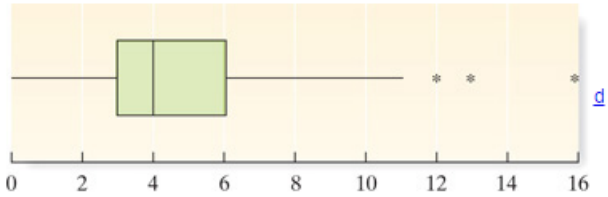
NOTE:

If the distribution is skewed, or outliers are present, the median and IQR are more appropriate measures of center and dispersion (spread) than the mean and standard deviation.

Rule of Thumb: if we want a fast approximation of the standard deviation, we use a quarter of the range.

Class Exercise: 2.108 p. 98

NW 2.108 Consider the horizontal box plot shown below.



a. What is the median of the data set (approximately)?

b. What are the upper and lower quartiles of the data set (approximately)?

c. What is the interquartile range of the data set (approximately)?

d. Is the data set skewed to the left, skewed to the right, or symmetric?

e. What percentage of the measurements in the data set lie to the right of the median? To the left of the upper quartile?

f. Identify any outliers in the data.

NW 2.110 Treating psoriasis with the "Doctorfish of Kangal." Psoriasis is a skin disorder with no known cure and no proven effective pharmacological treatment. An alternative treatment for psoriasis is ichthyotherapy, also known as therapy with the "Doctorfish of Kangal." Fish from the hot pools of Kangal, Turkey, feed on the skin scales of bathers, reportedly reducing the symptoms of psoriasis. In one study, 67 patients diagnosed with psoriasis underwent 3 weeks of ichthyotherapy (*Evidence-Based Research in Complementary and Alternative Medicine*, Dec. 2006). The Psoriasis Area Severity Index (PASI) of each patient was measured both before and after treatment. (The lower the PASI score, the better is the skin condition.) Box plots of the PASI scores, both before (baseline) and after 3 weeks of ichthyotherapy treatment, are shown in the accompanying diagram.

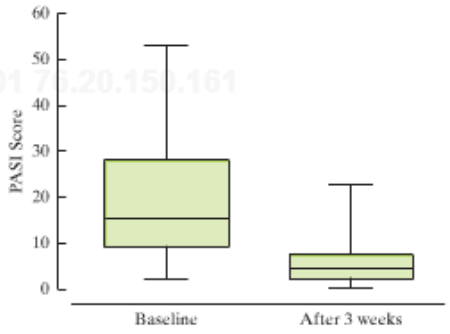
a. Find the approximate 25th percentile, the median, and the 75th percentile for the PASI scores before treatment.

b. Find the approximate 25th percentile, the median, and the 75th percentile for the PASI scores after treatment.

c. Comment on the effectiveness of ichthyotherapy in treating psoriasis.

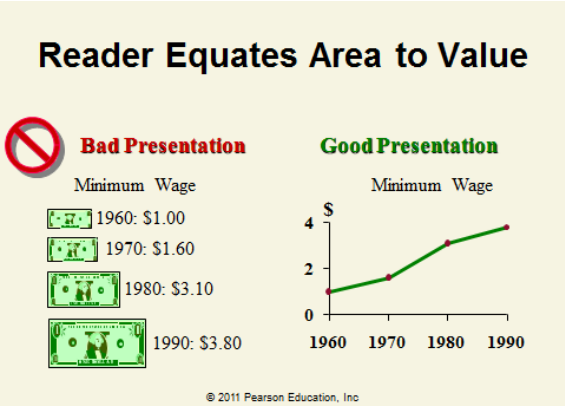
Comment also on the shape of both graphs. Do you suspect any outliers? Why yes or why not? Which measure of center and dispersion should be reported?

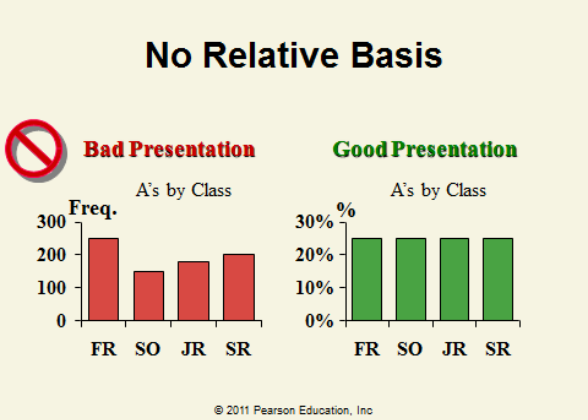
Output for Exercise 2.108

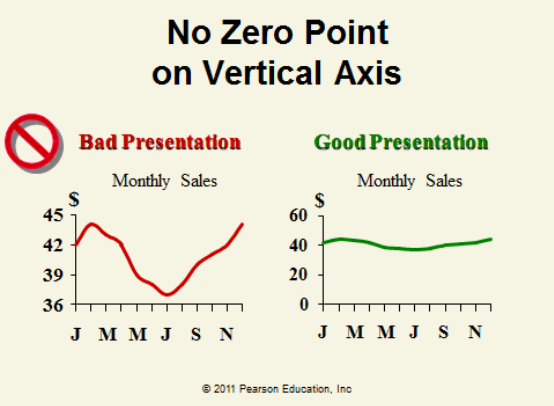


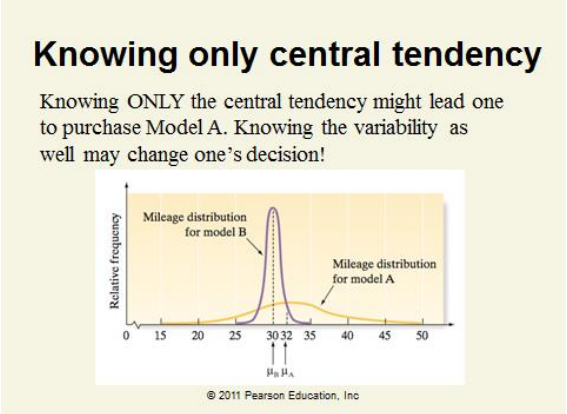
Source: Grassberger, M. and Hoch, W. "Ichthyotherapy as alternative treatment for patients with psoriasis: A pilot study." *Evidence-Based Research in Complementary and Alternative Medicine*, Vol. 3, No. 4, Dec. 2006, pp. 483–488 (Figure 3). Copyright © The Author (2006). Published by Oxford University Press. All rights reserved.

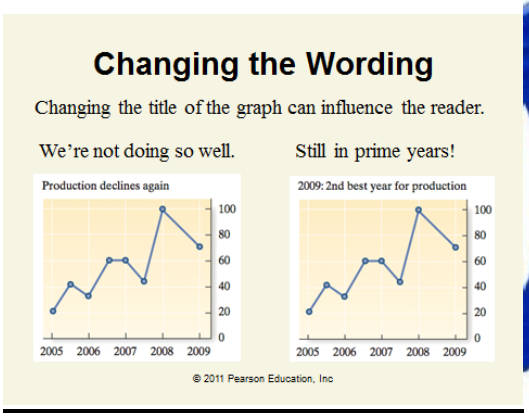
Chapter 2,10 Cheating with statistics Examples

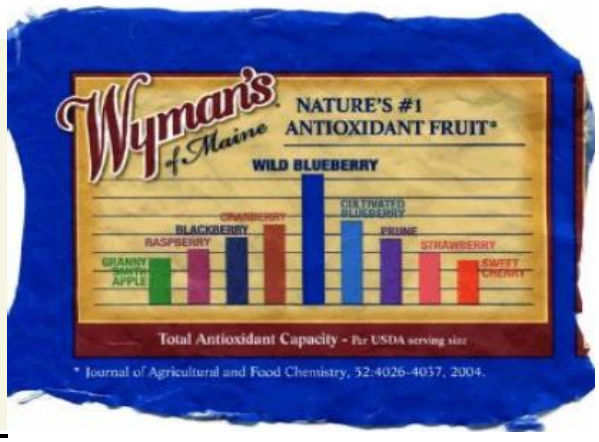


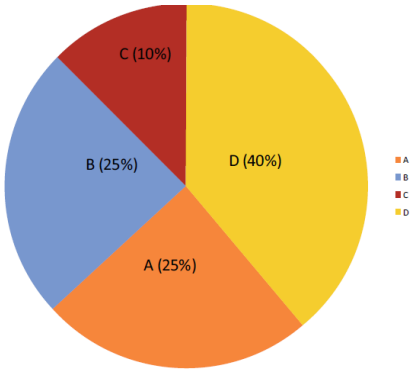


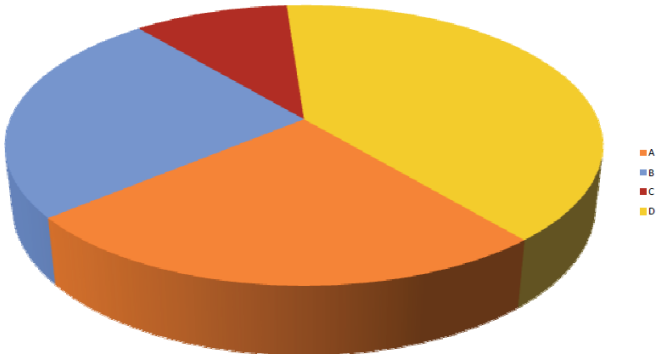




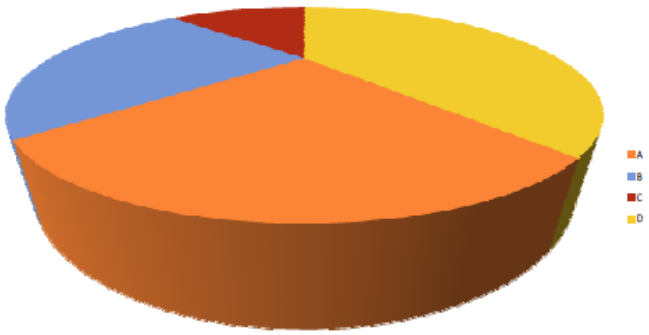


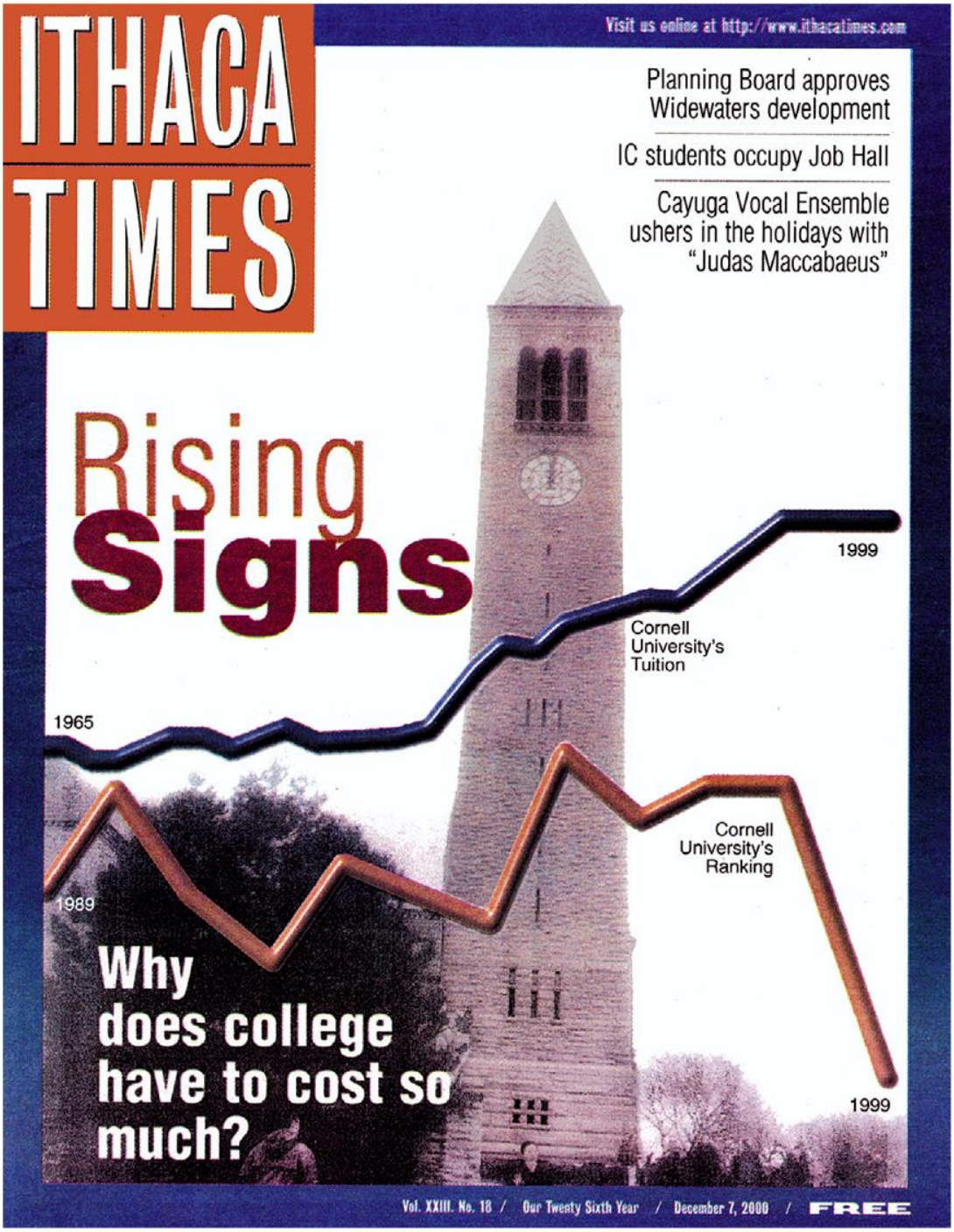






Want to emphasis slice A? It's easy: just make it 3D





END of Ch. 2. Review, study, do homework, and take a quiz.

Key Ideas

Describing Qualitative Data

1. Identify **category classes**
2. Determine **class frequencies**
3. **Class relative frequency** = (class freq)/*n*
4. **Graph** relative frequencies

Graphing Quantitative Data

1 Variable

1. Identify class intervals
2. Determine **class interval frequencies**
3. **Class relative relative frequency** = (class interval frequencies)/*n*
4. **Graph** class interval relative frequencies

Numerical Description of Quantitative Data

Measures of Central Tendency

- Mean
- Median
- Mode

Measures of Variation (Spread)

- Range
- Variance
- Standard Deviation
- Interquartile range

Measures of Relative standing

- Percentile score
- *z*-score

Rules for Detecting Quantitative Outliers

Interval	Chebyshev's Rule	Empirical Rule
$\bar{x} \pm s$	At least 0%	$\approx 68\%$
$\bar{x} \pm 2s$	At least 75%	$\approx 95\%$
$\bar{x} \pm 3s$	At least 89%	All

Rules for Detecting Quantitative Outliers

Method	Suspect	Highly Suspect
Box plot:	Values between inner and outer fences	Values beyond outer fences
z-score	$2 < \|z\| < 3$	$\|z\| > 3$