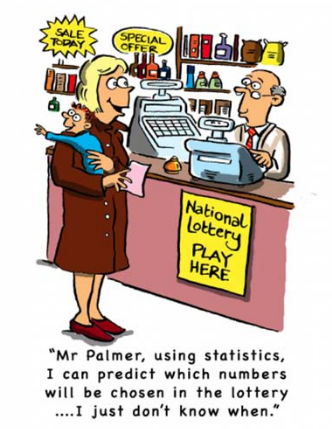
Descriptive Statistics



Today's Class

- Sampling
- Graphical Representation
- Measurement of Location
- Measurement of Variability

Sampling

- **Population (N):** the entire collection of objects or outcomes about which information is sought
- **Sample (n)**: a subset of a population, containing the objects or outcomes that are actually observed

A subset of the population.

Examples

Population	Sample
Diameters of all shafts in a lot	Diameters of the shafts that are actually measured
Employment status of all eligible adults in the US	Employment status of subjects who are interviewed
Lifetimes of the items made by a certain manufacturing process	Lifetimes of the subset of items tested

Simple Random Sample

- **Simple random sample** (SRS) of size n: a sample chosen by a method in which each collection of n population items is equally likely to comprise the sample

- **Independence:** the selection of one unit has no influence on the selection of other units
- **Lack of bias:** each unit has the same chance of being chosen

Types of Data

- **Numerical or quantitative:** when a numerical quantity is assigned to each item in the sample
  - Height (in cm, ft)
  - Weight (in kg, lb)
  - Age (in years)
- **Categorical or qualitative:** when sample items are placed into categories and category names are assigned to the sample items
  - Hair color
  - Country of origin
  - Location of accidents

How to Visualize Data?

- Histograms
- Box and whisker plot
- Stem-and-leaf plot
- Scatter plot
- Time-series
- Others

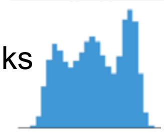
"Doesn't matter where they're posted, those are not *BAR* graphs."

Histogram

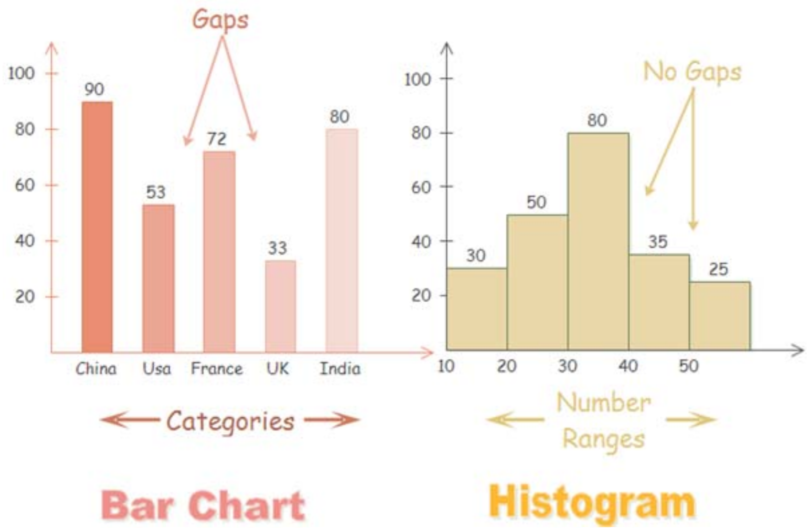
- Histogram: a graphical display of data using bars of different heights
- Unimodal
  - A histogram with only one peak

- Bimodal
  - A histogram has two peaks

- Multimodal
  - A histogram has more than two peaks



Histogram Vs. Bar Graph



Example: Histogram

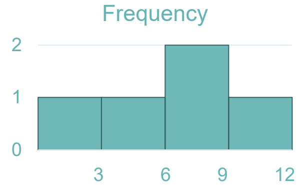
- Consider a small set of n=5 data points, corresponding for example to the number of hours of YouTube each of five people watch in a week

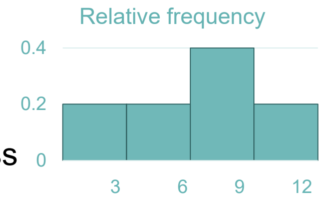
	X1	X2	X3	X4	X5
Hrs	6	7	3	12	7

  - Draw a histogram for frequency
  - Draw a histogram for relative frequency

Creating Histogram

- Choose boundary points for the class intervals
- Compute the frequencies
  - the number of observations in each interval
- Compute the relative frequencies
  - frequencies / the total number of observations
- Draw a rectangle for each class whose height is equal to the frequencies or relative frequencies





Example
: Measure of Location

YouTube

- Consider a small set of n=5 data points, corresponding for example to the number of hours of YouTube each of five people watch in a week

	X1	X2	X3	X4	X5
Hrs	6	7	3	12	7

  - Mean?
  - Median?
  - Mode?

Example
: Measure of Location

- Consider a small set of n=5 data points, corresponding for example to the number of hours of YouTube each of five people watch in a week

	X1	X2	X3	X4	X5
Hrs	6	7	3	12	7

  - Mean: $\overline{x} = \frac{6+7+3+12+7}{5} = 7$
  - Median: 3 5 7 7 12 $\rightarrow$ 7
  - Mode: 7

Measure of Location

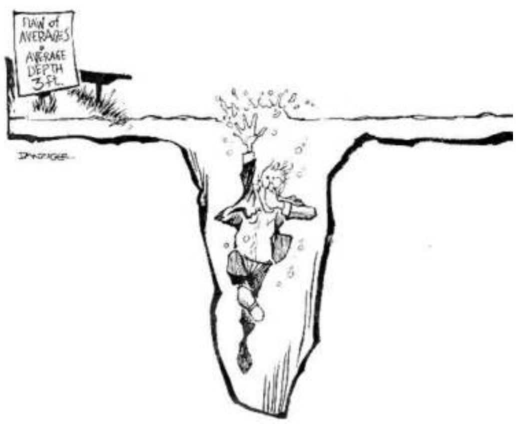
- **Mean** $$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$
- **Median**
  - Order the n data points from smallest to largest
    $\tilde{X}$ is the number in position $\frac{n+1}{2}$ if n is odd
    $\tilde{X}$ is the average of the numbers in positions $\frac{n}{2}$ and $\frac{n}{2}+1$ if n is even
- **Mode**
  - The value that has the highest frequency

Symmetry and Skewness

- A histogram is symmetric if its right half is a mirror image of its left half
  - Mean ≅ Median
- Histograms that are not symmetric are referred to as skewed
  - skewed to the left, or negatively skewed
    - a histogram with a long left-hand tail
    - the mean < the median
  - skewed to the right, or positively skewed
    - a histogram with a long right-hand tail
    - the mean > the median

(Navidi, 2010)

The Flaw of the Averages



http://www.danzigercartoons.com/

Example
Measure of variability

- In the previous example, find the variance and the sample standard deviation

	X1	X2	X3	X4	X5
Hrs	6	7	3	12	7

  - Population Variance?
  - Population Standard Deviation?

Example
Measure of variability

- In the previous example, find the variance and the sample standard deviation

	X1	X2	X3	X4	X5
Hrs	6	7	3	12	7

  - Population Variance:

$$\sigma^2 = \frac{(6-7)^2 + (7-7)^2 + (3-7)^2 + (12-7)^2 + (7-7)^2}{5} = 8.4$$

  - Standard Deviation:

$$\sigma = \sqrt{8.4} = 2.9$$

Example
Measure of variability

- In the previous example, find the sample variance and the sample standard deviation.

	X1	X2	X3	X4	X5
Hrs	6	7	3	12	7

  - Sample Variance?
  - Sample Standard Deviation?

Example
Measure of variability

- In the previous example, find the sample variance and the sample standard deviation.

	X1	X2	X3	X4	X5
Hrs	6	7	3	12	7

  - Sample Variance:

$$s^2 = \frac{(6-7)^2 + (7-7)^2 + (3-7)^2 + (12-7)^2 + (7-7)^2}{5-1} = 10.5$$

  - Standard Deviation:

$$s = \sqrt{10.5} = 3.24$$

Measure of Variability

- Deviations from the mean

$$x_i - \bar{x}$$

- Sum of the deviations

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

Measure of Variability

	Population	Sample
Variance	$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$	$S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$
Standard Deviation	$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2}$

Why divide by n-1?

- Consider a population with three elements {1,2,3}
- The mean of the population

$$\mu = \frac{1+2+3}{3} = 2$$

- The variance of the population

$$\sigma^2 = \frac{(1-2)^2 + (2-2)^2 + (3-2)^2}{3} = \frac{2}{3}$$

Why divide by n-1?
Sample Accuracy

- Given the population in the previous slide, suppose all we can take is a sample of two elements taken with repetition to learn about the population
- We would like the sample to accurately estimate the mean and variance values of the population

Why divide by n-1? Sample Accuracy, Cont'd

Samples	Sample mean $\bar{x}$	divided by n	divided by n – 1
{1,1}	1	$\frac{(1-1)^2+(1-1)^2}{2} = 0$	$\frac{(1-1)^2+(1-1)^2}{2-1} = 0$
{2,2}	2	0	0
{3,3}	3	0	0
{1,2}	1.5	$\frac{(1-1.5)^2+(2-1.5)^2}{2} = 0.25$	$\frac{(1-1.5)^2+(2-1.5)^2}{2-1} = 0.5$
{2,1}	1.5	$\frac{(2-1.5)^2+(1-1.5)^2}{2} = 0.25$	$\frac{(2-1.5)^2+(1-1.5)^2}{2-1} = 0.5$
{1,3}	2	$\frac{(1-2)^2+(3-2)^2}{2} = 1$	$\frac{(1-2)^2+(3-2)^{2^2}}{2-1} = 2$
{3,1}	2	$\frac{(3-2)^2+(1-2)^2}{2} = 1$	$\frac{(3-2)^2+(1-2)^{2^2}}{2-1} = 2$
{2,3}	2.5	$\frac{(2-2.5)^2+(3-2.5)^2}{2} = 0.25$	$\frac{(2-2.5)^2+(3-2.5)^2}{2-1} = 0.5$
{3,2}	2.5	$\frac{(3-2.5)^2+(2-2.5)^2}{2} = 0.25$	$\frac{(3-2.5)^2+(2-2.5)^2}{2-1} = 0.5$
Avg	2	1/3 $(\neq \sigma^2)$	2/3 $(= \sigma^2)$

Example: Quartiles

- In the previous example, find the lower and upper quartiles

	X1	X2	X3	X4	X5
Hrs	6	7	3	12	7

- Sort the data
- The lower quartile?
  - Q1: the median of the smallest half
- The upper quartile?
  - Q3: the median of the largest half
- Fourths spread (Interquartile range)?
  - f_s(IQR) = Q3 - Q1

Example: Quartiles

- In the previous example, find the lower and upper quartiles

	X1	X2	X3	X4	X5
Hrs	6	7	3	12	7

  - Sort the data (Size n=5): 3, 6, 7, 7, 12
  - To find the lower quartile (Q1)
    - Q1: the median of the smallest half {3, 6, 7} = 6
  - To find the upper quartile (Q3)
    - Q3: the median of the largest half {7, 7, 12} = 7
  - To find fourths spread
    - f_s(IQR) = Q3 - Q1 = 7 – 6 = 1

Quartiles and Percentiles

Quartiles	Percentiles
Lower quartile (Q1)	25th percentile
Median (Q2)	50th percentile
Upper quartile (Q3)	75th percentile

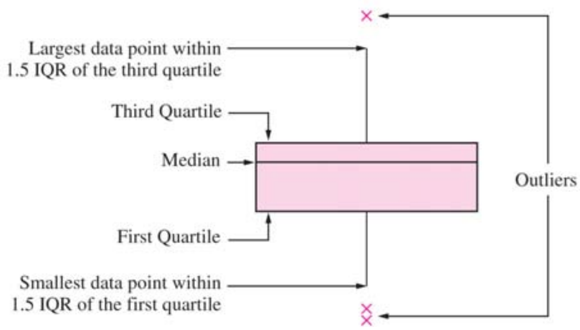
- How to find quartiles
  - Sort *n* observations in ascending order
  - Separate them by half (including the median in both halves if *n* is odd)
  - Lower quartile: median of the first half
  - Upper quartile: median of the second half
  - Fourths spread: f_s (IQR) = Q3 - Q1

Outliers

- Outliers are points that are much larger or smaller than the rest of the sample points
- Observations farther than $1.5f_s$ from closest fourth
- Extreme outlier is farther than $3f_s$ from the closest fourth, otherwise considered mild

Box and whisker plots

- A boxplot is a graphic that presents the median, the first and third quartiles, and any outliers present in the sample



(Navidi, 2010)

Example: Quartiles

YouTube

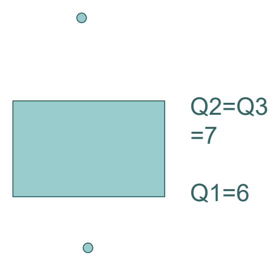
- In the previous example,

	X1	X2	X3	X4	X5
Hrs	6	7	3	12	7

  - Q1 = 6
  - Q2 = 7
  - Q3 = 7
  - $f_s = 1$, $1.5\ f_s = 1.5$
  - Min & Max Whiskers?

    6,7
  - Outliers?

    3,12

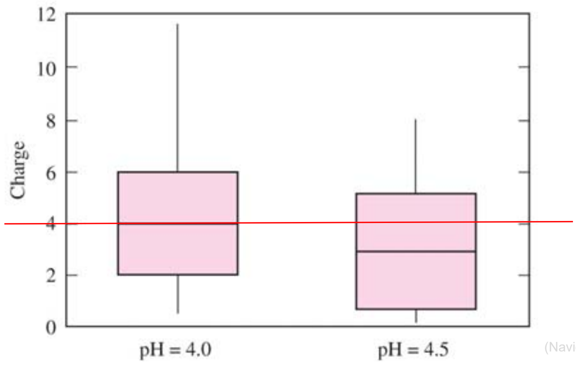


Comparative Boxplots

- Sometimes we want to compare more than one sample
- We can place the boxplots of the two (or more) samples side-by-side
- This will allow us to compare how the medians differ between samples, as well as the first and third quartile
- It also tells us about the difference in spread between the two samples

Comparative Boxplots

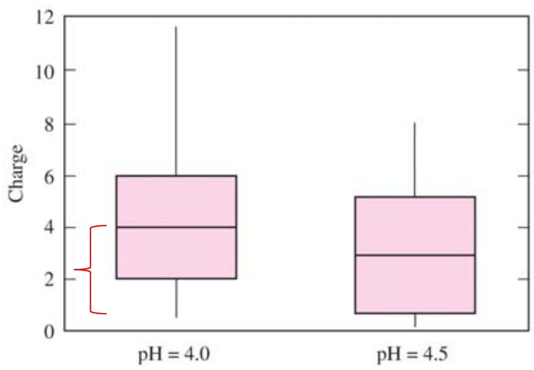
- The median charge for the pH of 4.0 is greater than the 75th percentile of charge for the pH of 4.5. T or F?



(Navidi, 2010)

Boxplot Example, Cont'd

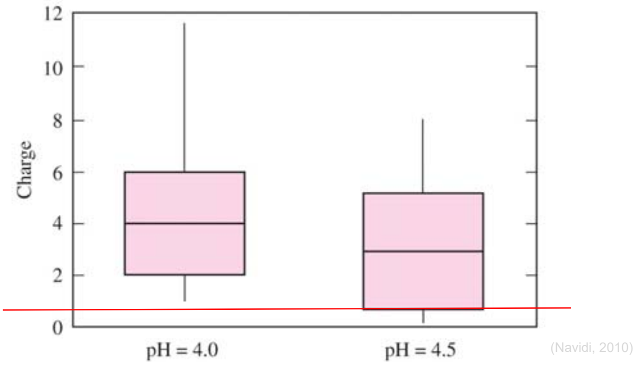
- About half of the sample values of pH 4.0 are between 2 and 4. T or F?



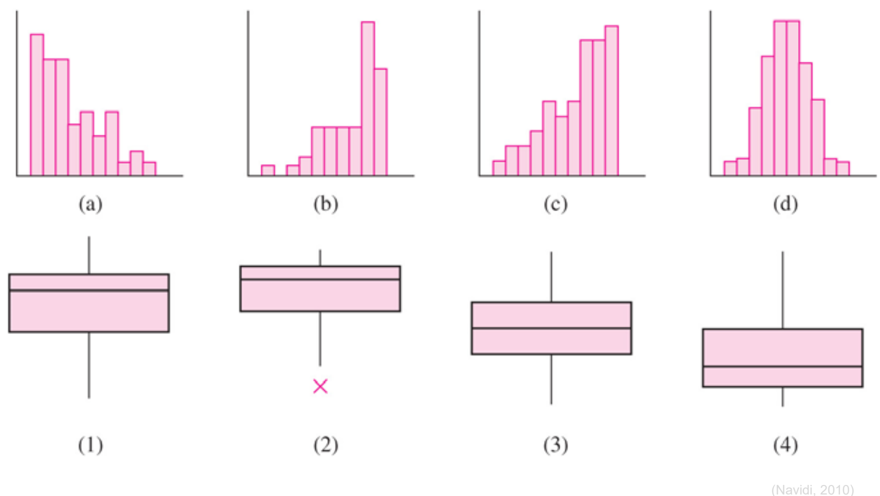
(Navidi, 2010)

Boxplot Example, Cont'd

- Approximately 25% of the charges for pH 4.5 are less than the smallest charge at pH 4.0. T or F?



Example



Stem-and-Leaf Plot

- A simple and compact way to summarize a data set
- Each item in the sample is divided into two parts:
  - Stem consisting of the leftmost one or two digits
  - leaf consisting of the next digit
- It also gives us some indication of the shape of our data.

Stem-and-Leaf Plot Example

- Duration of dormant periods of the Old Faithful Geyser in Minutes

4	259
5	0111133556678
6	067789
7	01233455556666699
8	000012223344456668
9	013

Example: Stem-and-Leaf Plot

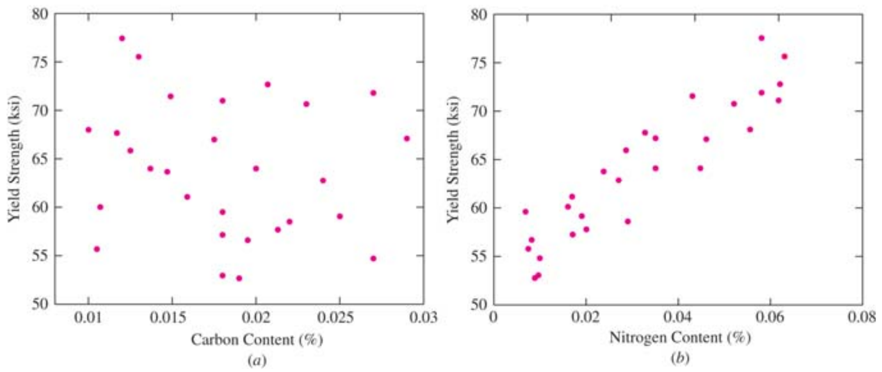
- Complete a stem-and-leaf plot for the following list of grades on a recent test:

73, 42, 67, 78, 99, 84, 91, 82, 86, 94

Stem	Leaf		
4	2		
5			
6	7		
7	3	8	
8	2	4	6
9	1	4	9

Scatterplot

- A scatterplot is a graph for bivariate data, for which items consists of a pair of values



Time Series

- Time series plots observations over time