Chapter 1
Review of Basic Concepts and Descriptive Statistics

© 2006 A. Karpinski

Review of Basic Concepts & Descriptive Statistics

1. Concepts and definitions

- What is statistics?
  - A branch of science concerned with methods for understanding and summarizing collections of numbers

  - Data (plural)
    - Observations made on the environment
    - The collection of numbers we wish to understand and summarize

  - Population
    - The complete set of data which we want to understand and summarize
    - Populations can be of any size and are completely determined by the researcher's interests
    - <u>Parameters</u> are quantitative summary characteristics of populations

  - Sample
    - Part of the population which we want to understand and summarize
    - Can be any subset of the population
    - <u>Statistics</u> are quantitative summary characteristics of samples

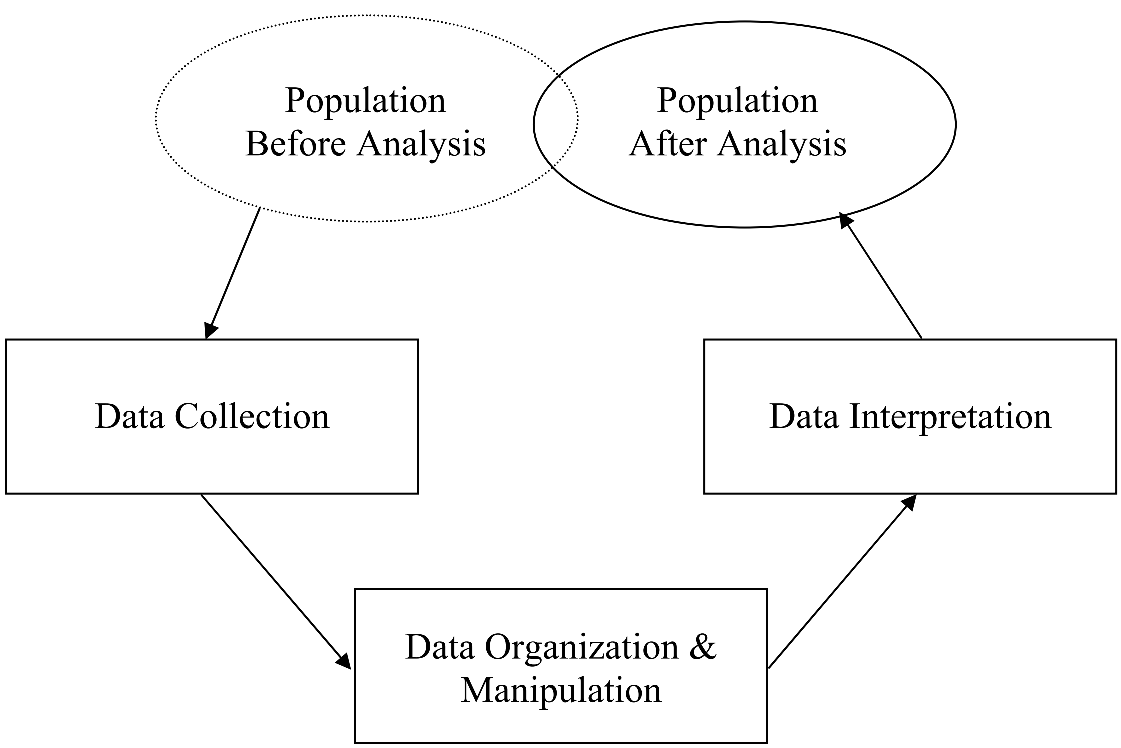
	Population Parameters	Sample Statistics	
	(Greek Symbols)	(Latin Symbols)	(Greek Symbols with hats)
Mean	μ	$\overline{X}$	$\hat{\mu}$
Variance	σ^2	s^2	$\hat{\sigma}^2$
Correlation	ρ	r	$\hat{\rho}$
Proportion	π	p	$\hat{\pi}$

 © 2006 A. Karpinski

- Random Sample
  - A sample in which each member of the population has an equal chance of being included and in which the selection of one member is independent from the selection of all other members
  - Are more likely to be similar to the population if the samples are large
  - All of the statistical procedures we develop will assume that the observations are randomly sampled from the population of interest

- Descriptive vs. Inferential Statistics
  - Descriptive –
    - Methods for describing and summarizing the data
    - Includes graphical and numerical techniques to summarize the distributional location (central tendency) and dispersion (variability), and the relationships between variables
    - Goal is data reduction

  - Inferential –
    - Methods for generalizing beyond the actual sample data and inferring properties of populations that were not observed
    - Requires the sample data to be representative of the population
    - Goal is to understand the characteristics of the population and the relationship between variables in the population



 © 2006 A. Karpinski

- Variables vs. Constants

  - Constant
    - A fixed number
    - A number that is the same for the entire population
    - We do not need to perform statistics on constants
    - For example: The speed of light, the number of chambers in a human heart

  - Variable
    - An attribute, characteristic, or property of some organism, object or event that can assume two or more values
    - Usually symbolized by an alphabetic letter toward the latter part of the alphabet
    - Subscripts are used to represent or stand for a unique observation on that variable

- Types of variables

  - Fixed Variable
    - A variable whose value(s) are pre-selected or manipulated by the researcher
    - For example: Gender, condition of the experiment, or whether a person has been diagnosed with depression or not
  - Random Variable
    - A variable whose value(s) are determined as a result of sampling
    - For example: self-esteem scores, number of puzzles solved, shock level given to a confederate

  - Independent Variable (IV)
    - In experiments, it is the variable that is manipulated or controlled
    - An IV can be either a fixed or random variable

  - Dependent Variable (DV)
    - The variable that is measured, and that the experimenter is attempting to predict or understand
    - A DV must be a random variable

 © 2006 A. Karpinski

- Discrete (or Categorical) Variable
  - A variable that takes on a finite number of values, usually whole numbers
  - Ethnicity is a discrete variable

- Continuous Variable
  - A variable that takes on an infinite number of values within some interval
  - Self-esteem scores are a continuous variable.

- Some notational issues: Consider verbal GRE scores for 7 people

  - '*N*' represents the overall sample size
    - N=7

  - A specific observation is represented by a lower case x. A subscript is used to match the observed score with a specific participant (or unit of observation)
    - Because verbal GRE scores is a random variable, we do not know a participant's score until we observe it. Sometimes, it is useful to represent a participant's score before we have observed its value. In this case, we simply use the subscripted lower case x.

      $$x_1$$

      And likewise we can represent the data from the entire sample:

      $$\{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$$

      Or more generally

      $$\{x_1, x_2, x_3, \ldots, x_N\}$$

    - Once we observe the data, then we can replace the random variables with their observed values

      If person 1 scored a 520, we write $x_1 = 520$

      And the data might look like:

      $$\{520, 590, 680, 750, 420, 630, 500\}$$

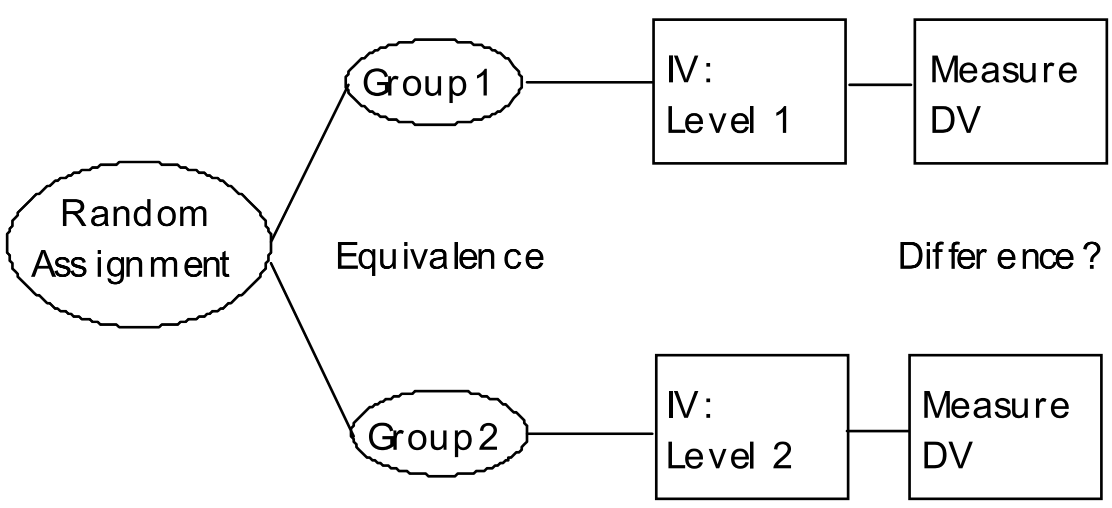
 © 2006 A. Karpinski

2. Research Strategies

- Experimental Research
  - Involves manipulation of the fixed, independent variables and control of all other variables by the experimenter to determine the effects on the random, independent variable
  - The IV may be an organismic variable (such as gender), and therefore controlled, but not manipulated (called a quasi-experiment)
  - In a true experiment all IVs are manipulated by the experimenter, and the experimenter may make causal claims

- Observational Research
  - Involves observation and measurement on one or more random variables without any experimental control
  - The notions of IV and DV are irrelevant
  - Definitive causal claims may never be made

- Hallmarks of experimental designs
  - Manipulation of an independent variable
  - Control of all other variables



© 2006 A. Karpinski

3. Measurement Scales

- Nominal Scale
  - Specific numbers assigned to an observation, but the numbers assigned are arithmetically meaningless
  - Example: Political Affiliation

1. = Democrat	1. = Republican
2. = Republican	2. = Independent
3. = Independent	3. = Democrat
4. = Other	4. = Other

  - Can not add, subtract, multiply, or divide these values
  - Also referred to as qualitative or categorical variables

- Ordinal Scale
  - The numerical values have a meaningful order (or ranking)

$$x_1 < x_2 < x_3 < x_4 < x_5$$

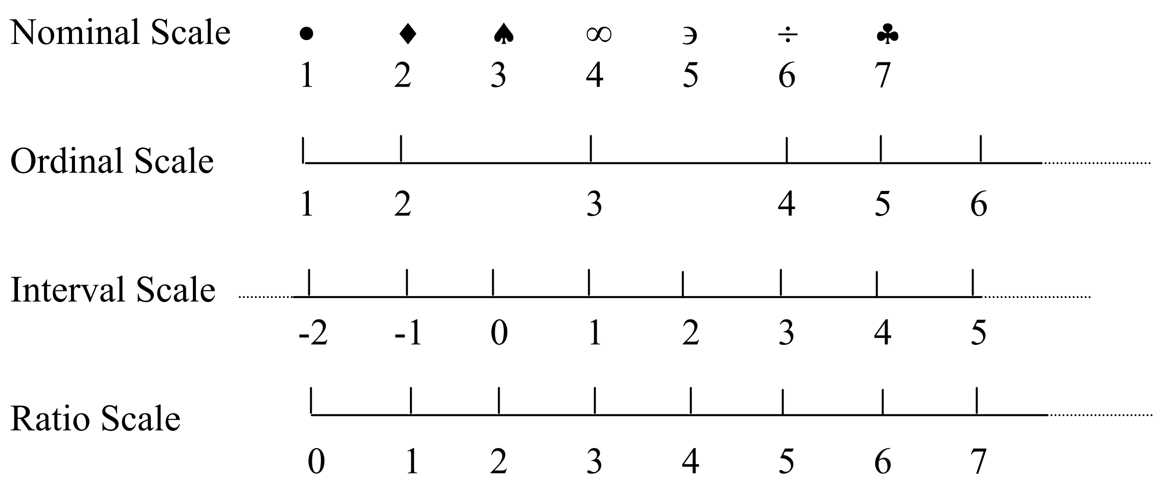
  - Example:
    - Birth Order
    - Grades of Eggs (A = 1, B = 2, C = 3; D = 4)
    - Ranking of 5 favorite fruits
  - Can perform any operation that preserves the order of the variables
  - Can not make a statement about the degree of difference on an ordinal scale

- Interval Scale
  - Equal differences on the scale have equal meaning
  - Often said to have arbitrary zero-point
  - Example:
    - Temperature
    - Calendar of years
  - Can perform any linear transformation on these variables
    - y = 12x – 6
  - Can not take ratios of values
  - We often assume that Likert scales are interval scales. It is more likely that these scales approximate an interval scale

© 2006 A. Karpinski

- Ratio Scale
  - A scale that preserves ratio of scale values
  - Often said that the zero-point is meaningful
  - Examples
    - Time
    - Length
    - Weight
    - Frequency count
  - Must be careful with transformations
    - Can only multiply by a positive number

- Recap of Measurement Scales



- In general,
  - Nominal scales can be used to make comparisons of frequency
  - Ordinal scales can be used to make comparisons of order
    $f(x) > f(y)$
  - Interval scales can be used to make comparisons of difference
    $[f(a) - f(b)] > [f(x) - f(y)]$
  - Ratio scales can be used to make multiplicative statements
    $f(a) = 2f(b)$

 © 2006 A. Karpinski

- Let's consider a simple example of what may happen when we violate the rules of operations for measurement scales
    - Suppose we have the ages of 6 family members (3 males and 3 females)

Males	Females
75	66
70	65
5	64
$\bar{X}_M = 50$	$\bar{X}_F = 65$

        - The females are older than the males

    - Now suppose we converted these data to ranks (birth order) and tried to perform the same operation

Males	Females
6	4
5	3
1	2
$\bar{R}_M = 4$	$\bar{R}_F = 3$

        - We might conclude that the males are older than the females
        - This error occurred because we attempted to make comparisons of difference on an ordinally scaled variable

- See Appendix A for a second example of measurement scales gone awry.

 © 2006 A. Karpinski

4. Describing Distributions

- There are four population parameters we can use to describe any distribution:
  - Mean – The central value about which the observations scatter
  - Variance – How far the observations scatter around the central value
  - Skewness – How symmetric the distribution is around the central value
  - Kurtosis – How far rare observations scatter around the central value

- Let's consider some common estimators of these population parameters

i. Measures of Central Tendency

- (Arithmetic) Mean
  - The average – the sum of *n* scores, divided by *n*

Population mean: Discrete Continuous

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \quad \text{or} \quad \int_{\infty}^{-\infty} x_i f(x_i)dx$$

Sample mean:

$$\bar{X} = \frac{\sum_{i=1}^{N} x_i}{N}$$

  - Takes into account the numerical value of each and every observation
  - Known as the "center of gravity" of the data

- Median
  - The score at or near the 50^{th} percentile
  - The value above and below which 50% of the observations fall

  - Insensitive to extreme values in a distribution
  - Difficult to work with mathematically

 © 2006 A. Karpinski

- Mode
  - Most common value appearing in the data set
  - May be multiple modes in a data set

- Using the mean, median, and mode, we can get a sense of the distribution of the data.

ii. Measures of Variability

- Variance
  - The average of the squared deviations around the mean

<u>Population Variance</u>: $\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$

Numerator = Sum of squares around the mean
Denominator = Number of observations

- Used only if we have a census

<u>Sample Variance</u>: $s^2 = \frac{\sum (x_i - \overline{X})^2}{N - 1}$

Numerator = Sum of squares around the mean
Denominator = Degrees of Freedom

- If we divide by *N* rather than *N*-1, then we underestimate the true sample variance, and we obtain a biased estimate of the variance

 © 2006 A. Karpinski

- Standard Deviation
  - Square root of the variance
  - Places the variability measure back on the original units of measurement

Population Standard Deviation: $\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$

Sample Standard Deviation: $s = \sqrt{\frac{\sum (x_i - \bar{X})^2}{N-1}}$

- Interquartile Range (IQR) or Semiquartile Range (SQR)

$$IQR = 75^{th} \text{ percentile} - 25^{th} \text{ percentile}$$

  - Very easy to visualize and interpret
  - The interval in which the middle 50% of the data fall

  - Some people prefer to use the semi-quartile range, SQR, which is half of the IQR

$$SQR = \frac{IQR}{2}$$

    - The SQR is interpreted as the typical distance of an observation from the median (although this is a lazy method of calculating this quality – a better measure is the MAD)

- Median Absolute Deviation (MAD)
  - The median absolute deviations around the median

$$MAD = Median\{|x_i - median_x|\}$$

  - In other words
    - For each value, calculate the deviation from the median
    - Take the absolute value of these deviations
    - Find the median of this set of values

 © 2006 A. Karpinski

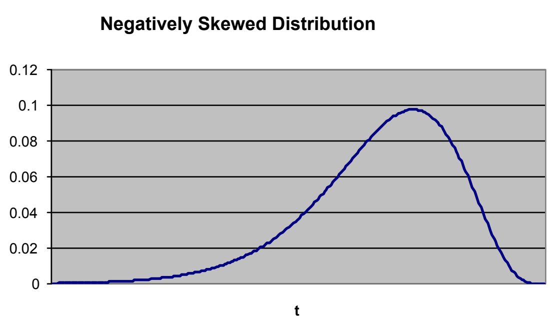
- The MAD can be properly interpreted as the typical distance of an observation from the median

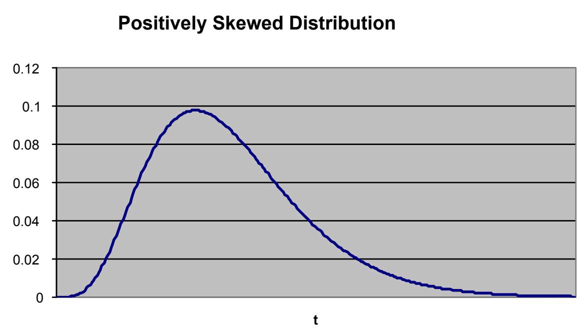
- Some authors refer to the <u>mean</u> absolute deviation. To avoid confusion, we will abbreviate this quantity AD

$$AD = \frac{\sum_{i=1}^{N} |x_i - Median_x|}{N}$$

iii. Measures of Skewness

- Measures of the extent to which the distribution departs from symmetry
  - Positive skew – there exists a long tail to the right of the distribution
  - Negative skew – there exists a long tail to the left of the distribution





- Calculating a coefficient of skewness:

  - γ_1 ratio

$$\gamma_1 = \frac{m_3}{m_2^{3/2}} \qquad \text{where} \quad m_j = \frac{\sum (x_i - \bar{X})^j}{N}$$

$$m_1 = \frac{\sum (x_i - \bar{X})}{N} \qquad m_2 = \frac{\sum (x_i - \bar{X})^2}{N} \qquad m_3 = \frac{\sum (x_i - \bar{X})^3}{N}$$

  - Interpretation: A value of zero $\Rightarrow$ Symmetry
    Values greater than zero $\Rightarrow$ Positive Skew
    Values less than zero $\Rightarrow$ Negative Skew

 © 2006 A. Karpinski

iv. Measures of Kurtosis

- Measures the relative proportion of observations in the tails of the distribution. It gives us a measure of the distance from the mean of the extreme values of the distribution

  - Mesokurtic: the kurtosis of the normal distribution

  - Leptokurtic – a relatively large proportion of the observations are located in the tails of the distribution
    Distribution looks skinny and peaked

  - Platykurtic – a relatively small proportion of the observations are located in the tails of the distribution
    Distribution looks flat

- To estimate Kurtosis, we use γ_2 ratio:

$$\gamma_2 = \frac{m_4}{m_2^2} - 3$$

  - Interpretation of γ_2 ratio:

0	$\Rightarrow$	Mesokurtic
Values greater than 0	$\Rightarrow$	Leptokurtic
Values less than zero	$\Rightarrow$	Platykurtic

  - Note: SPSS uses a more complex (and better) formula to compute measures of skewness and kurtosis, but the interpretation is the same.

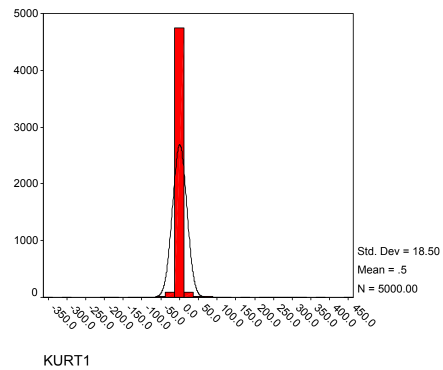
$$Kurtosis = \frac{N(N+1)}{(N-1)(N-2)(N-3)}\sum\left(\frac{x_i - \overline{X}}{s}\right)^4 - \frac{3(N-1)^2}{(N-2)(N-3)}$$

© 2006 A. Karpinski

- Example of a Leptokurtic distribution

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std.	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
KURT1	5000	-338.40	441.43	.4997	18.50169	8.405	.035	248.976	.069
Valid N (listwise)	5000								



- Compared to the normal distribution, this distribution is
    - Peaked/Pointy
    - Has too many observations in the tail
    - Leptokurtic

- For a normal distribution with a mean of zero and a standard deviation of 18.5

    - .13742% of the observations should be beyond 3SD from the mean
    - For a sample of 5000, that comes out to 6-7 observations
    - But in this sample, there are 43 observations beyond 3SD from the mean

```
compute outlier = 0.
if (kurt1> 55.5)  outlier = 1.
if (kurt1 < -55.5)  outlier = 1.
execute.
freq var = outlier.
```

OUTLIER

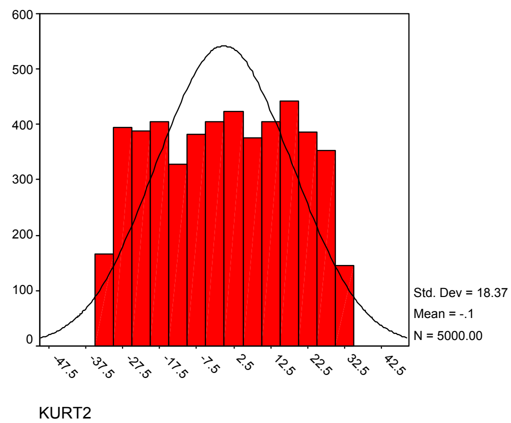
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	4957	99.1	99.1	99.1
	1.00	43	.9	.9	100.0
	Total	5000	100.0	100.0	

© 2006 A. Karpinski

- Example of a platykurtic distribution

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std.	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
KURT2	5000	-31.98	32.00	-.0672	18.37366	-.028	.035	-1.190	.069
Valid N (listwise)	5000								



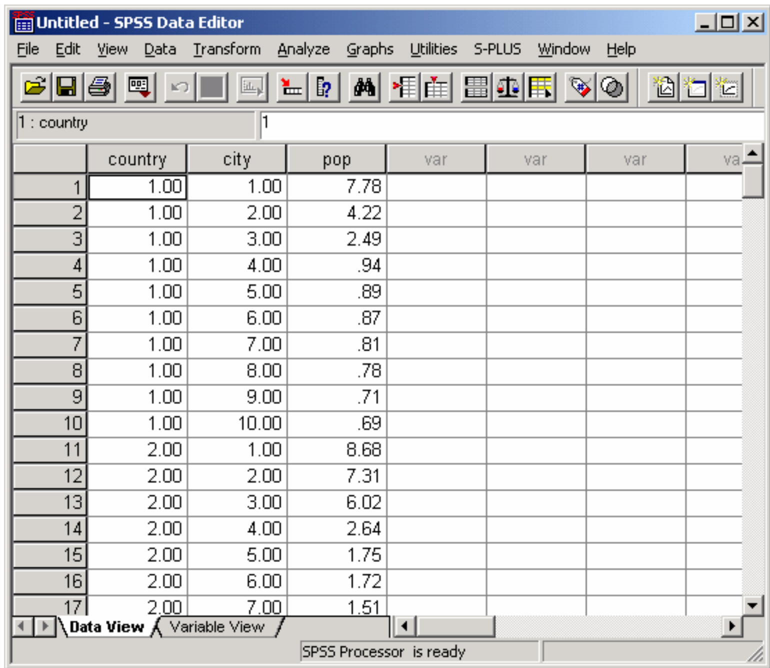
- Compared to the normal distribution, this distribution is
  - ⇒ Flat
  - ⇒ Has too few observations in the tail
  - ⇒ Platykurtic

 © 2006 A. Karpinski

v. Obtaining descriptive statistics in SPSS

- A quick primer on entering data in SPSS

    - You can enter data directly into the SPSS Data Editor



    - Or you can enter data in text format using SPSS syntax

```
DATA LIST FREE
  /country city pop.
BEGIN DATA.
1 1 7.78
1 2 4.22
.
.
.
16 9 15.00
16 10 11.13
END DATA.
```

    - Both give the same end result, but there are many reasons to prefer the syntax-based method. If you will be an SPSS user, you should be familiar with both methods.

 © 2006 A. Karpinski

* Example #1: Data on the population of the 10 largest cities in 16 countries (from 1960)

  * Method #1: DESCRIPTIVES

```
DESCRIPTIVES VARIABLES=pop
  /STATISTICS=MEAN STDDEV VARIANCE MIN MAX
     KURTOSIS SKEWNESS.
```

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std.	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
POP	160	.69	110.21	12.0680	15.84678	251.121	3.181	.192	12.824	.381
Valid N (listwise)	160									

$\overline{X} = 12.07$

$s^2 = 251.12$

$s = 15.85$

$Skew = 3.18$

$Kurt = 12.82$

  * Method #2: EXPLORE

```
EXAMINE VARIABLES=pop
  /PLOT NONE
  /STATISTICS DESCRIPTIVES.
```

Descriptives

			Statistic	Std. Error
POP	Mean		12.0680	1.25280
	95% Confidence Interval for Mean	Lower Bound	9.5937	
		Upper Bound	14.5423	
	5% Trimmed Mean		9.6625	
	Median		7.2150	
	Variance		251.121	
	Std. Deviation		15.84678	
	Minimum		.69	
	Maximum		110.21	
	Range		109.52	
	Interquartile Range		10.1425	
	Skewness		3.181	.192
	Kurtosis		12.824	.381

© 2006 A. Karpinski

- Method #3: Use EXCEL

Population	
Mean	12.068
Standard Error	1.25279833
Median	7.215
Mode	2.49
Standard Deviation	15.8467847
Sample Variance	251.120585
Kurtosis	12.824188
Skewness	3.1805543
Range	109.52
Minimum	0.69
Maximum	110.21
Sum	1930.88
Count	160

- Example #2: Speeding Data
  - A police officer sets up a speed-trap in a 55mph zone. He obtains the following data

Speed (in MPH)				
45	52	58	65	75
48	54	59	66	75
49	55	60	69	79
49	57	63	72	82
51	58	63	74	85

  - Method #1: DESCRIPTIVES

```
DESCRIPTIVES VARIABLES=mph
 /STATISTICS=MEAN STDDEV VARIANCE
   KURTOSIS SKEWNESS.
```

Descriptive Statistics

	N	Mean	Std.	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
MPH	25	62.5200	11.35826	129.010	.380	.464	-.859	.902
Valid N (listwise)	25							

© 2006 A. Karpinski

- Method #2: EXPLORE

```
EXAMINE VARIABLES=mph
  /PLOT NONE
  /STATISTICS DESCRIPTIVES.
```

Descriptives

			Statistic	Std. Error
MPH	Mean		62.5200	2.27165
	95% Confidence Interval for Mean	Lower Bound	57.8315	
		Upper Bound	67.2085	
	5% Trimmed Mean		62.2444	
	Median		60.0000	
	Variance		129.010	
	Std. Deviation		11.35826	
	Minimum		45.00	
	Maximum		85.00	
	Range		40.00	
	Interquartile Range		20.0000	
	Skewness		.380	.464
	Kurtosis		-.859	.902

- Method #3: Use EXCEL

MPH	
Mean	62.52
Standard Error	2.27165138
Median	60
Mode	49
Standard Deviation	11.3582569
Sample Variance	129.01
Kurtosis	-0.8589291
Skewness	0.37952576
Range	40
Minimum	45
Maximum	85
Sum	1563
Count	25

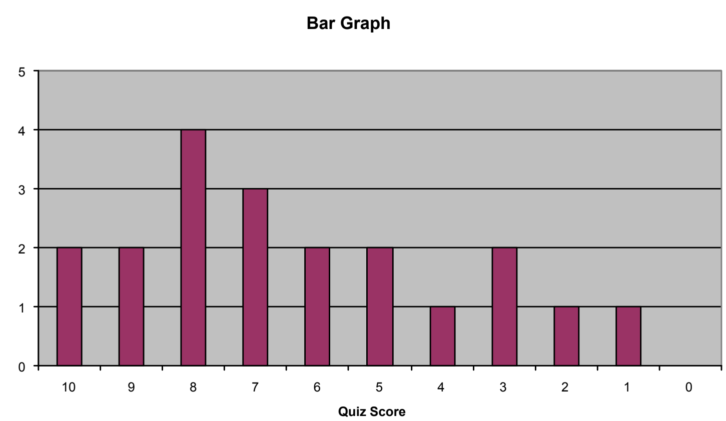
$$\bar{X} = 62.52$$
$$s^2 = 129.01$$
$$s = 11.36$$
$$Skew = 0.38$$
$$Kurt = -0.86$$

 © 2006 A. Karpinski

5. Exploratory data analysis techniques

- Exploratory data analysis is a set of techniques that graphically displays the data (rather than relying on numbers and statistics).

- Bar graph
  - If your data are discrete, then you can create a bar graph.
  - For each level of *X*, the height of the bar is the number of observations at that level
  - In a bar graph, some people insist that the bars not touch each other (to highlight the fact that the DV is discrete)

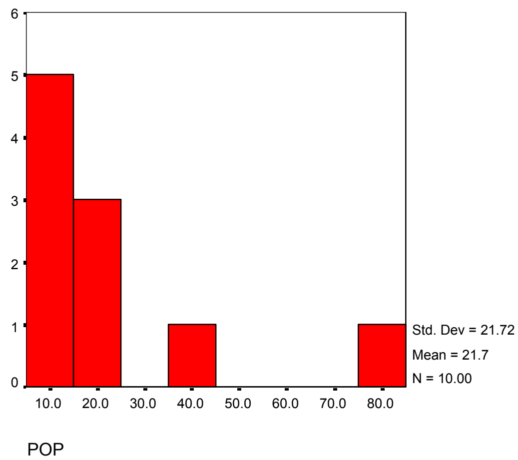


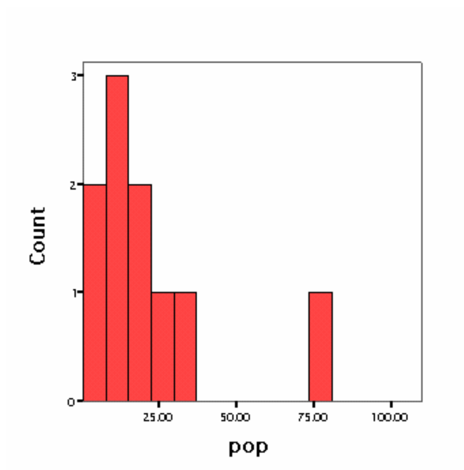
- Histogram
  - A histogram is the equivalent of a bar graph. Technically, a bar graph is used for discrete data, and a histogram is used for continuous data.
  - The continuous variable must be divided into *b* equal 'bins'. A graph is then created that represents the number of observations that fall into the bin.
  - The tricky part of creating a histogram is determining the number of bins.
    - If you have too many bins, then each bin will only have one observation
    - If you have too few bins, then you will not have enough bins to represent the data.

 © 2006 A. Karpinski

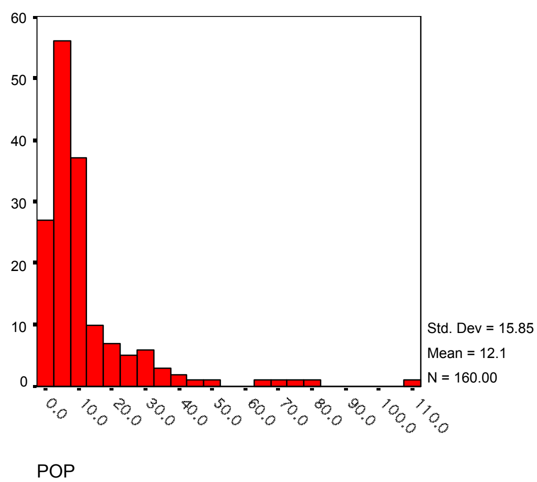
- Example #1: The population of the 10 largest cities in 16 countries (in 1960)

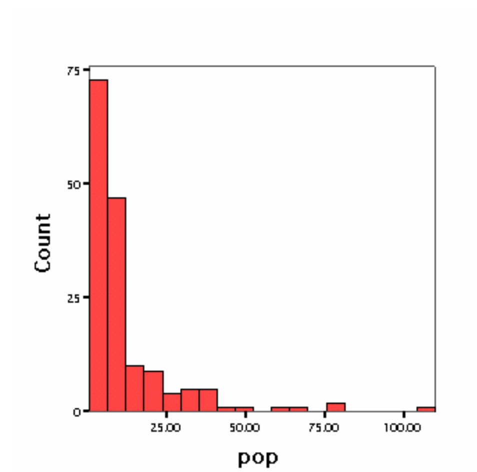
  - USA data only





  - World data

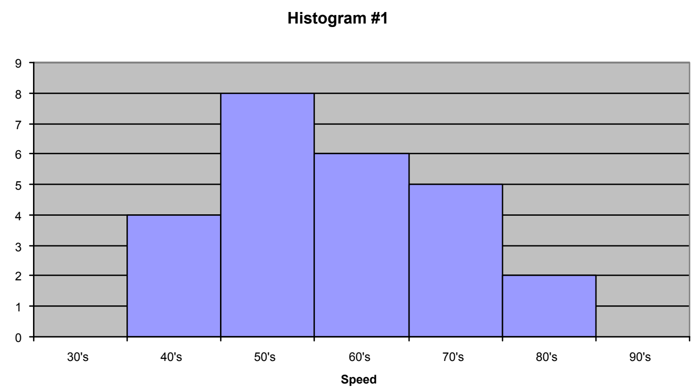


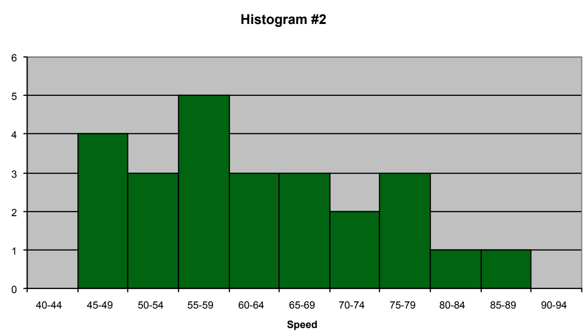


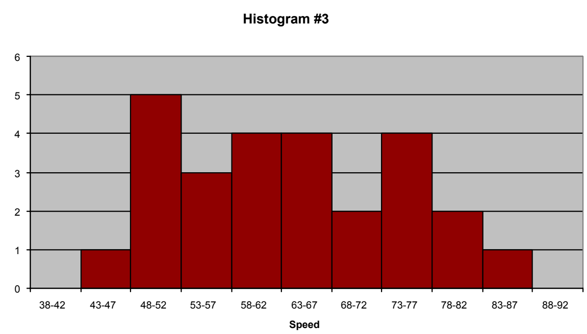
  - SPSS has two ways of obtaining a histogram (GRAPH and IGRAPH). They do not draw the same histogram!

© 2006 A. Karpinski

- Example #2: Speeding Data (in Excel)







- Each of these histograms is of the same data, but the "bin width" is different.
- From Histogram #2, we might think that the distribution is right skewed, but that interpretation is not supported in the other graphs.

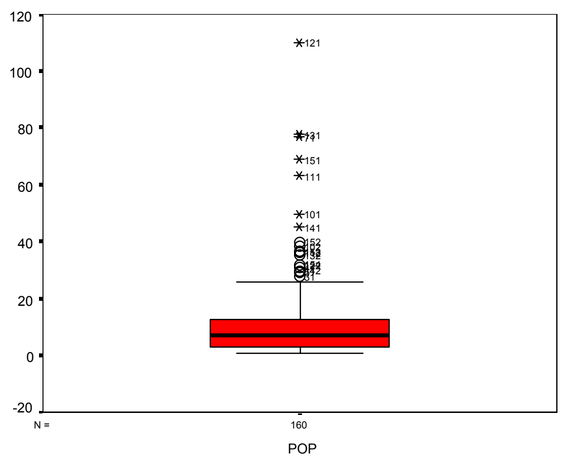
- Note that none of these are incorrect. They all accurately display the same data. However, you can arrive at very different conclusions about the data from these graphs. Interpret histograms with caution!

 © 2006 A. Karpinski

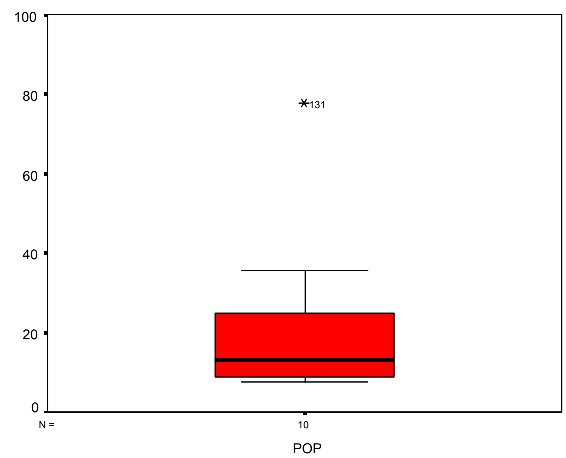
- Boxplot (or box-and-whisker plot)
    - The boxplot is a nice graph to identify the central aspects of the data, as well as the extreme observations, or outliers.
    - The boxplot consists of several components:
        - The median is the measure of central tendency used for a box plot
        - The IQR forms the “box” around the median
        - Each “whisker” extends to the largest (or smallest) observation no more than 1.5 IQRs from the “box”
        - Observations beyond the whiskers are identified as outliers

    - Example #1: Population data

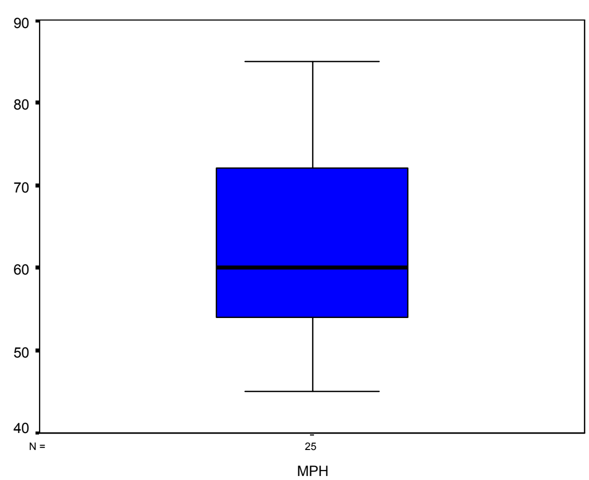
All 16 countries



USA only



- Example #2: Speeding data



 © 2006 A. Karpinski

- Stem-and-leaf plot
  - A stem-and-leaf plot is similar to a histogram in that you get a sense of the shape of the distribution
  - But for a stem-and-leaf plot, you can see the actual value of each observation (or at least the first two places of each observation)
  - To construct a stem-and-leaf plot
    - Convert the data into two-digit numbers (round any remaining digits)
    - Construct the stem – the left-most of the digits

      (In this case, the tens digit)

STEM	LEAF
4	
5	
6	
7	
8	

    - Next, fill in the leaves (in this case, the tens digit)

      Each observation gets its own leaf

STEM	LEAF
4	5899
5	12457889
6	033569
7	24559
8	25

© 2006 A. Karpinski

- In SPSS, use the EXAMINE command

```
EXAMINE VARIABLES=mph
  /PLOT STEMLEAF.
```

  - Beware! SPSS will often divide stems into 2 or more steps (with the same stem digit)

```
MPH Stem-and-Leaf Plot
 Frequency    Stem &  Leaf

      .00        4 .
     4.00        4 .  5899
     3.00        5 .  124
     5.00        5 .  57889
     3.00        6 .  033
     3.00        6 .  569
     2.00        7 .  24
     3.00        7 .  559
     1.00        8 .  2
     1.00        8 .  5

 Stem width:     10.00
 Each leaf:       1 case(s)
```

- For the population example, SPSS divided each stem into 5 stems

```
EXAMINE VARIABLES=pop
  /PLOT STEMLEAF.
```

```
POP Stem-and-Leaf Plot
 Frequency    Stem &  Leaf

    18.00        0 .  000000011111111111
    37.00        0 .  2222222222222222222222222333333333333
    15.00        0 .  444444445555555
    21.00        0 .  666666677777777777777
    14.00        0 .  88888899999999
    15.00        1 .  000000111111111
     2.00        1 .  33
     2.00        1 .  55
     6.00        1 .  666667
     2.00        1 .  88
     5.00        2 .  00111
     2.00        2 .  23
     3.00        2 .  445
    18.00 Extremes    (>=28)

 Stem width:     10.00
 Each leaf:       1 case(s)
```


© 2006 A. Karpinski

6. Resistant statistics

- When we collect real data, they often are not as well behaved as the data you find in textbook examples. We often observe outliers – observations that are very different (i.e., much larger or much smaller) from the main body of the data.
- One challenge we face is how to analyze data with outliers. We do not want our conclusions to be determined or influenced by one or two deviant observations.

- Unfortunately, many of the statistics we use are greatly influenced by outliers. For example, consider the sample mean:

    - Case 1: $\{3,4,5,6,7\}$ $\quad \overline{X} = 5$
    - Case 2: $\{3,4,5,6,107\}$ $\quad \overline{X} = 25$

- There are several terms that are used to describe statistics:
    - A <u>resistant statistic</u> is a statistic that is insensitive to localized misbehavior in the data (such as an outlier).
    - A <u>robust statistic</u> is a statistic that is insensitive to departures from statistical assumptions required for the tests we conduct.

    - We have just seen that the sample mean is not resistant.
    - On the other hand, the sample median is highly resistant to outlying values:

        - Case 1: $\{3,4,5,6,7\}$ $\quad Median = 5$
        - Case 2: $\{3,4,5,6,107\}$ $\quad Median = 5$

    - We would like to have a set of statistics that is resistant so that we can accurately estimate values and model parameters regardless of the underlying shape and characteristics of the data.

 © 2006 A. Karpinski

- Resistant estimators of central tendency

  - The median is not at all influenced by extreme observations. However, the median can be overly sensitive to the values of the middle one or two observations.

  - M-estimators are a class of estimators of central tendency that are obtained by solving the following equation for θ, the M-estimator:

$$\sum_{i=1}^{N}\psi\left(\frac{x_i-\theta}{\delta}\right)=0 \qquad (1\text{-}1)$$

where ψ is a function (subject to certain constraints)
θ is the M-estimator
δ is an (optional) scale estimator

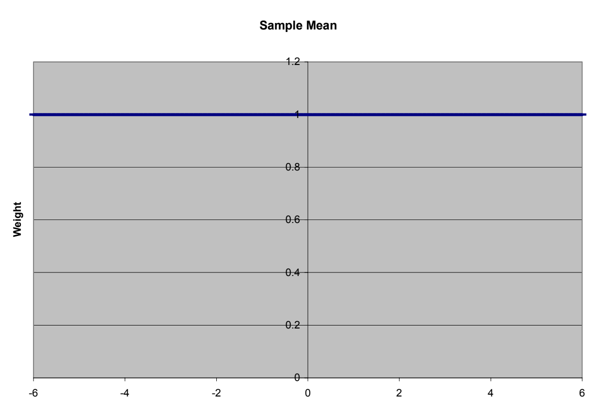
A more intuitive understanding of M-estimators can be obtained by re-writing equation (1-1) as a set of weights applied to the observations

$$\theta_i=\frac{\sum_{i=1}^{N}w_ix_i}{\sum_{i=1}^{N}w_i} \qquad (1\text{-}2)$$

- Example #1: If we let $\psi=\left(\frac{x_i-\theta}{\delta}\right)^2$ and let $\delta=\sigma$, then we need to solve:

$$\sum_{i=1}^{N}\left(\frac{x_i-\theta}{\delta}\right)^2=0$$

With a little calculus, we can show $\hat{\theta}=\overline{X}$. If we re-write θ as in equation (1-2), we find that $w_i=1$ for all observations



 © 2006 A. Karpinski

- The trick is to find a function ψ that gives resistant estimators of the central tendency of a distribution

- There are two M-estimators that have been shown to be robust and resilient:
  - Tukey's bisquare estimator
  - The Huber estimator

- Tukey's bisquare estimator
  (Which Tukey denies he discovered!)

  - Let's call $\frac{(x_i - \theta)}{\delta} = u$

    where θ is the M-estimator of central tendency
    δ is a scale estimator (in this case the MAD)

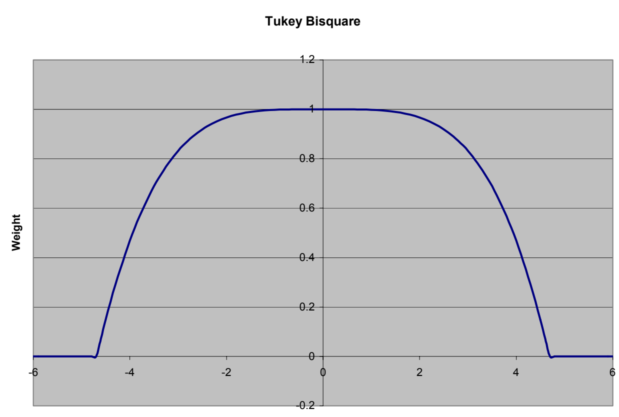
    Then u can be interpreted like a standardized score

  - Now we define a weight function w_i such that

$$w_i = \begin{cases} \left(1 - \frac{u_i^2}{4.685^2}\right)^2 & \text{for} \quad |u_i| \le 4.685 \\ 0 & \quad\quad\quad |u_i| > 4.685 \end{cases}$$

  - $\hat{\theta}$ is the Tukey bisquare estimator. We cannot solve for $\hat{\theta}$ directly – we must use an iterative process

  - To understand the Bisquare, let's examine the weight function



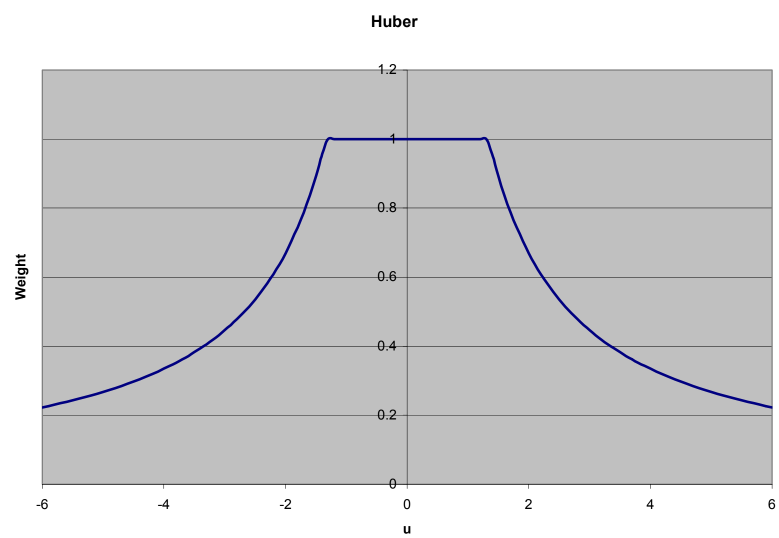
 © 2006 A. Karpinski

- The Huber estimator
  - Again, let $u = \frac{(x_i - \theta)}{\delta}$ with $\delta = MAD$

    And define a weight function w_i such that

$$w_i = \begin{cases} 1 & \text{for } |u| \leq 1.339 \\ \frac{1.339}{u_i} sign(u_i) & \text{for } |u| > 1.339 \end{cases}$$

  - $\hat{\theta}$ is the Huber estimator. Again, we cannot solve $\hat{\theta}$ directly – we must use an iterative process

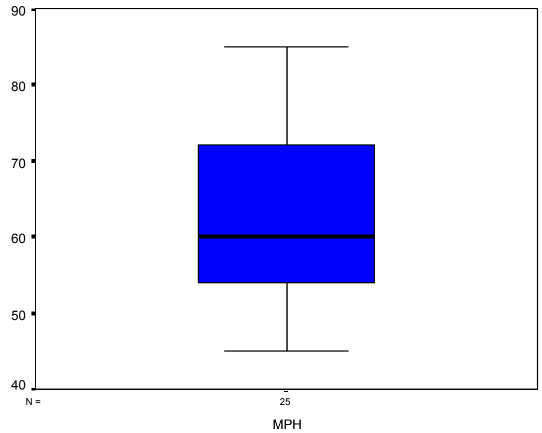


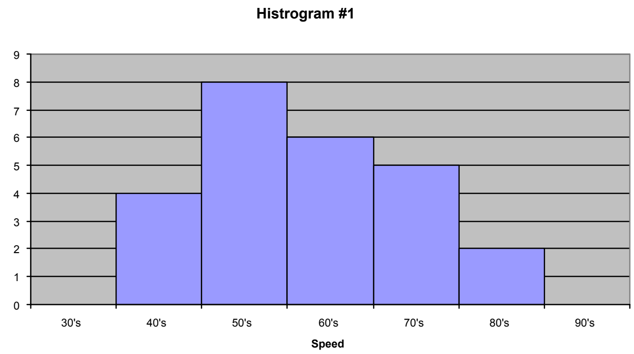
- Unfortunately, there is not a rule of thumb explaining when to use the mean, the median, the bisquare, and Huber estimators to describe the central tendency of a distribution. Ideally, they should all be considered, along with a boxplot and/or histogram of the data.

- In general, M-estimators are very difficult to calculate. They cannot be solved directly; iterative methods must be used. Luckily, SPSS will calculate them for us!

© 2006 A. Karpinski

- Example #1: Speeding example






```
EXAMINE VARIABLES=mph
  /MESTIMATORS  HUBER(1.339) TUKEY(4.685).
```

Descriptives

		Statistic	Std. Error
MPH	Mean	62.5200	2.27165
	Median	60.0000	

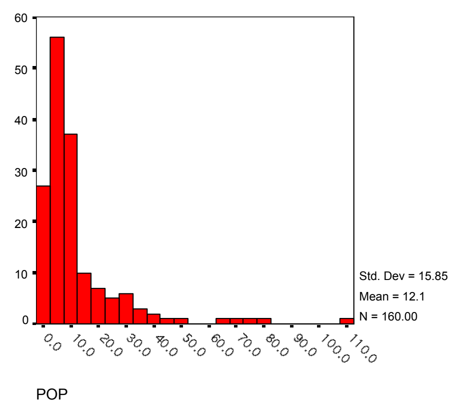
M-Estimators

	Huber's M-Estimator[a]	Tukey's Biweight[b]
MPH	61.5818	61.7566

a. The weighting constant is 1.339.

b. The weighting constant is 4.685.

- Example #2: Population of world's cities example




```
EXAMINE VARIABLES=pop
  /MESTIMATORS  HUBER(1.339) TUKEY(4.685).
```

Descriptives

		Statistic	Std. Error
POP	Mean	12.0680	1.25280
	Median	7.2150	

M-Estimators

	Huber's M-Estimator[a]	Tukey's Biweight[b]
POP	7.4134	6.2263

a. The weighting constant is 1.339.

b. The weighting constant is 4.685.

 © 2006 A. Karpinski

- I do not expect you to calculate M-estimators by hand, but you should be able to:
  - Explain why Bisquare and Huber estimators are more resistant to outliers than the sample mean
  - Explain how and why Bisquare and Huber estimators are different than the sample mean

7. Properties of Estimators

- Thus far, we have developed several measures for each of the population parameters. Which ones are the best? Let's consider some desirable properties for any statistic

- There are three desirable properties for any estimator:

  - Unbiasedness: We would like an estimate of the population parameter to be, on average, equal to the population parameter. That is, we want our estimates to be correct on average

  - Consistency: We would like our estimates of the population parameters to increase in their accuracy as the size of the sample increases. That is, the more data we have collected the better our estimate of the population parameters should be

  - Efficiency: When comparing two possible estimators of a population parameter, we prefer the estimate with the smaller variance. That is, we prefer estimates that tend to cluster around their average value to estimates that are more dispersed.

- In order to evaluate these properties, we need to understand some properties regarding the 'average value' and the variance of a random variable

  - The 'average value' of a random variable is known as its expected value, $E(X)$

© 2006 A. Karpinski

- There are three important properties of the expectation. Let's let X be a random variable, and let a and b be constants

$$E(a) = a$$
$$E(aX) = aE(X)$$
$$E(X + b) = E(X) + b$$

or

$$E(aX + b) = aE(X) + b$$

    - In English:
      ⇒ Adding a constant to every value in the data, increases the mean by that constant.
      ⇒ If you multiply every value in the data by a constant, you multiply the mean by that constant.

  Thus, we can easily obtain the expected value of a linear transformation of any random variable.

- The amount of variability of a random variable around its expected value is known as its <u>variance</u>, $Var(X)$

- There are three important properties of the variance. Let's let X be a random variable, and let a and b be constants

$$Var(a) = 0$$
$$Var(aX) = a^2 Var(X)$$
$$Var(X + b) = Var(X)$$

or

$$Var(aX + b) = a^2 Var(X)$$

    - In English:
      ⇒ Adding a constant to every value in the data, does not change the variance.
      ⇒ If you multiply every value in the data by a constant, you multiply the variance by that constant squared.

  Now we can easily obtain the variance of a linear transformation of any random variable.

 © 2006 A. Karpinski

- With these simple formulas, you can calculate the mean and variance of <u>any</u> linear transformation (so long as you know the mean and variance of the original variable).
- Note that we did not make any distributional assumptions, so you do not have to check any assumptions to calculate the new mean and variance.

- For example, suppose you have temperature in degrees Fahrenheit and you want to convert it to degrees Celsius. We know the following formula does the conversion:

$$T_C = \frac{5}{9} * (T_F - 32)$$

  - We need to write this equation in the form of a linear transformation

$$Y = aX + b$$

$$T_C = \frac{5}{9} T_F - 32 * \left(\frac{5}{9}\right)$$

$$a = \frac{5}{9} \qquad\qquad b = -\frac{160}{9}$$

  - Now, we can calculate the expected value of the temperature in Celsius

$$E(aX + b) = aE(X) + b$$

$$E(T_C) = E\left(\frac{5}{9} T_F - \frac{160}{9}\right)$$

$$= \frac{5}{9} E(T_F) - \frac{160}{9}$$

  - If we know that the average temperature in a dataset is $212^\circ F$, we can use the expected value of a linear transformation to find the average temperate in Celsius

$$E(T_C) = \frac{5}{9}(212) - \frac{160}{9} = 100$$

 © 2006 A. Karpinski

- Likewise, we can also determine the variance of the Celsius scale:

$$Var(aX+b) = a^2 Var(X)$$

$$Var(T_C) = Var\left(\frac{5}{9}T_F - \frac{160}{9}\right)$$

$$= \left(\frac{5}{9}\right)^2 Var(T_F) = .309\sigma_F^2$$

  - If we know the variance of a distribution of temperatures in Fahrenheit is $\sigma_F^2 = 30$, then we can use the variance of a linear transformation to find the variance of the Celsius distribution

$$\sigma_C^2 = .309\sigma_F^2 = .309 * 30 = 9.26^\circ$$

- In practice, we are most interested in estimates of the mean and the variance of the population (it turns out that the Normal distribution can be completely described using only these two parameters). So let's find the best estimators of the mean and the variance.

  - Claim: The sample mean is an <u>unbiased</u> estimator of the population mean.

$$E(\overline{X}) = \mu$$

$$\overline{X} = \frac{x_1 + x_2 + ... + x_N}{N}$$

$$E(\overline{X}) = E\left(\frac{x_1 + x_2 + ... + x_N}{N}\right)$$

$$= E(x_1 + x_2 + ... + x_N) * E\left(\frac{1}{N}\right)$$

$$= [E(x_1) + E(x_2) + ... + E(x_N)] * \left(\frac{1}{N}\right)$$

$$= (\mu + \mu + ... + \mu) * \left(\frac{1}{N}\right)$$

$$= \frac{N\mu}{N} = \mu$$

 © 2006 A. Karpinski

- We could also show that the sample mean is consistent and efficient in normally distributed data.
- In other words, the sample mean satisfies all the desirable properties for an estimator of the population mean.

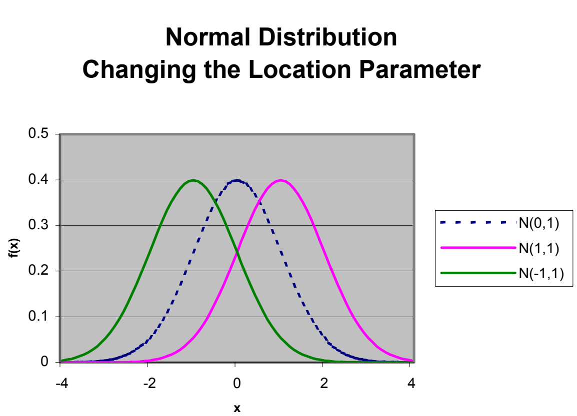
- We can also show that the sample variance is an unbiased, consistent, and efficient estimator of the population variance (again, in normally distributed data).

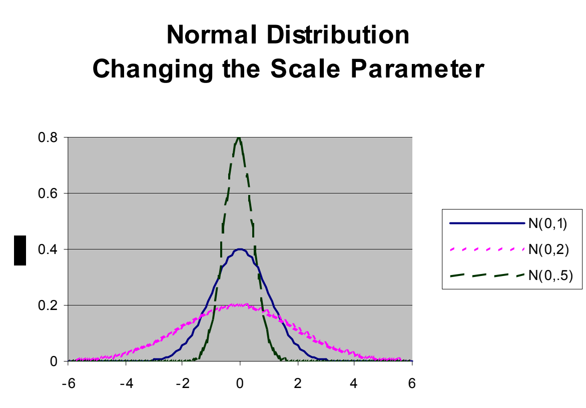
• As a consequence, when we want to estimate the population mean and variance, we usually use the sample mean and variance.

8. The Normal Distribution

• The normal distribution is a 2-parameter distribution:
  - The location parameter – μ
  - The scale parameter – σ

  Once these two parameters are known, the entire distribution is defined
  - Both the skewness and kurtosis coefficients are zero for a normal distribution



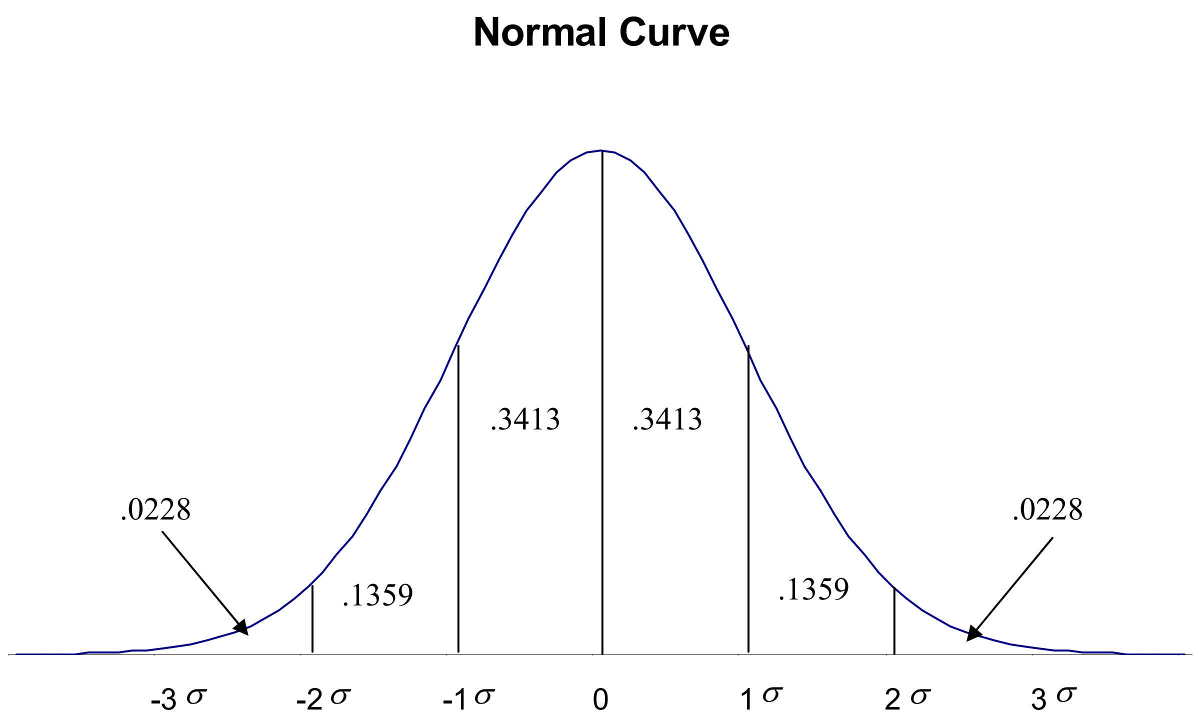


• The density function of the normal distribution

$$f(x_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-\mu)^2 / 2\sigma^2}$$

 © 2006 A. Karpinski

- Properties of the normal curve
  - 68.26% of the observations lie between -1σ and $+1\sigma$
  - 95.44% of the observations lie between -2σ and $+2\sigma$



 © 2006 A. Karpinski

- Standardized z-scores
  - Motivation for the use of standardized z-scores:
    Person 1: IQ of 135 on a test distributed $N(110,18)$
    Person 2: IQ of 95 on a test distributed $N(50,30)$
    Who has the higher intelligence score?

    - It is very difficult to answer this question because of the differences in the two normal distributions
    - A solution is to transform both distributions into a new distribution with a common mean and standard deviation

- A standardized normal distribution has $\mu = 0$ and $\sigma = 1$: $N(0,1)$
  - Also referred to as the Gaussian Distribution
  - We can easily convert a value from any normal distribution to a value on the standard normal curve

- To convert raw normal scores into z-scores

$$z = \frac{x_i - \mu}{\sigma}$$

  - This transformation maintains the shape of the distribution
  - Note that this is a linear transformation of the raw data

- Using the properties of the expectation and variance of a random variable, we can prove that z-scores are distributed $N(0,1)$

  - Let X be any normally distributed random variable: $X \sim N(\mu_X, \sigma_X)$

    - The z-transformation $z = \dfrac{x_i - \mu_X}{\sigma_X}$ is a linear transformation with

$$a = \frac{1}{\sigma_X} \quad \text{and} \quad b = -\frac{\mu_X}{\sigma_X}$$

    - What is the mean and variance of the z-distribution?

 © 2006 A. Karpinski

$$\begin{aligned}\mu_Z &= aE(X)+b\\ &= \left(\frac{1}{\sigma_X}\right)\mu_X - \frac{\mu_X}{\sigma_X}\\ &= \frac{\mu_X}{\sigma_X} - \frac{\mu_X}{\sigma_X}\\ &= 0\end{aligned}$$

$$\begin{aligned}\sigma_Z^2 &= \left(\frac{1}{\sigma_X}\right)^2 * \sigma_X^2\\ &= \frac{\sigma_x^2}{\sigma_x^2}\\ &= 1\end{aligned}$$

- Now we have shown that $Z \sim N(0,1)$

- To return to our IQ example:

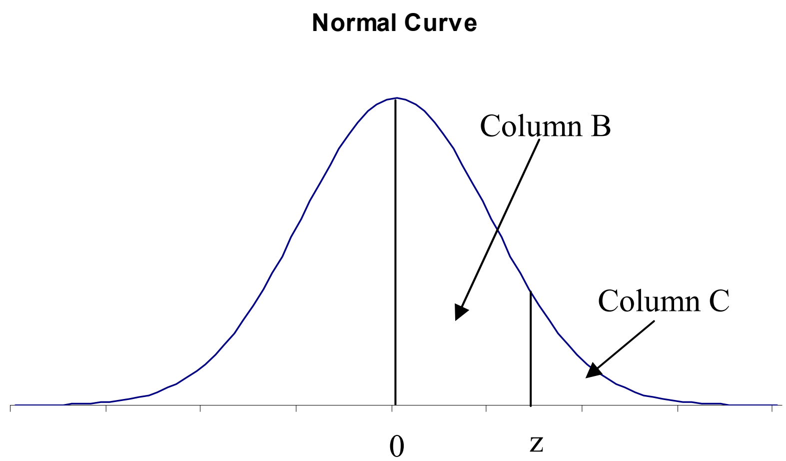
$$IQ_1 = \frac{(135-110)}{18} = 1.387 \qquad IQ_2 = \frac{(95-50)}{30} = 1.50$$

- Converting z-scores into percentiles
  - Can only be used if the data are normally distributed and have been converted into z-scores
  - If an observation is at the x^{th} percentile, it means that 45% of the observations have a value of x less than the observation.

 © 2006 A. Karpinski

- A common Z-table

z	Area between 0 and z	Area beyond z
0.00	.0000	.5000
0.01	.0040	.4960
1.00	.3413	.1587
1.38	.4162	.0838
1.39	.4177	.0823
1.50	.4332	.0668
2.00	.4772	.0228
3.00	.4987	.0013



- For Person A $(z = 1.39)$, we find that his/her score is at the 91.77^{th} percentile
- For Person B $(z = 1.50)$, we find that his/her score is at the 93.32^{th} percentile
- Be sure you can convert and interpret negative z-scores!

 © 2006 A. Karpinski

- Another z-score example:
  Suppose that the body weight for 18 year-old American men is normally distributed: $N(150,21)$. One case is chosen at random from the distribution.

  - What is the probability that this case weighs between 145lbs and 155 lbs?

    - We first need to convert 145 and 155 lbs to z-scores

    $$z_{145} = \frac{(145-150)}{21} = -.2381 \qquad z_{155} = \frac{(155-150)}{21} = +.2381$$

    - Next, we need to look up the probability associated with each z-value

    $$z_{145} = -.2381, p = .4052 \qquad z_{155} = +.2381, p = .5948$$

    - Finally, the probability we are interested in, $p(145 < x < 155)$ is equal to

    $$p(x < 155) - p(x < 145)$$
    $$.5948 - .4052 = .1896$$

    So the probability of the observation weighing between 145 and 155 lbs is 18.96%

  - What is the probability that the case weighs more than 1.5 standard deviations from the mean (in either direction)?

    - We already know the z-scores, $z_1 = -1.5$ and $z_2 = +1.5$

    - Because of the symmetry of the normal distribution, we only need to look up the probability associated with one z-value

    $$z_1 = -1.5, p = .0668$$

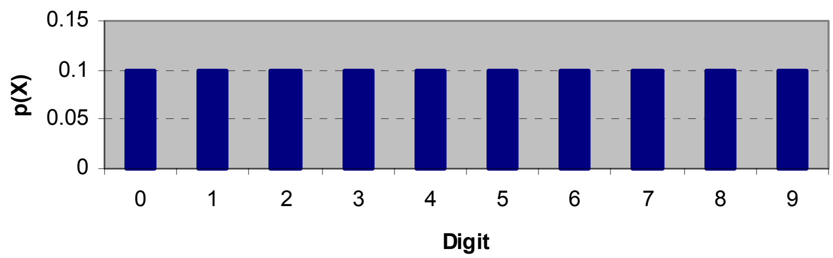
    - Finally, the probability we are interested in is the probability of being 1.5 standard deviations from the mean *in either direction*. So we need to double the *p*-value we obtained.

    So the probability of the observation weighing more than 1.5 standard deviations from the mean is 13.36%

 © 2006 A. Karpinski

9. Sampling Distributions
    - A sampling distribution is a distribution of all the possible values that a sample statistic can assume

    - Example #1 : X_1 = a random digit



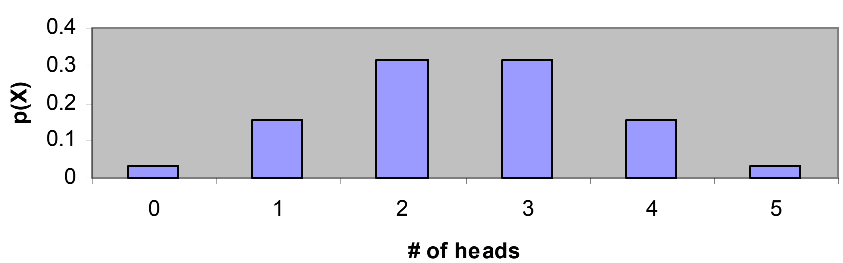
    - Example #2: X_2 = the number of heads in 5 tosses of a fair coin

        - Using the binomial theorem (see Appendix B), we can calculate the probability of each possible outcome:

Probability of zero heads:	$p(0;5,.5) = .031$
Probability of one head:	$p(1;5,.5) = .156$
Probability of two heads:	$p(2;5,.5) = .313$
Probability of three heads:	$p(3;5,.5) = .313$
Probability of four heads:	$p(4;5,.5) = .156$
Probability of five heads:	$p(5;5,.5) = .031$

Total Probability: 1.00

        - And so the sampling distribution for the number of heads on 5 flips of a fair coin is:



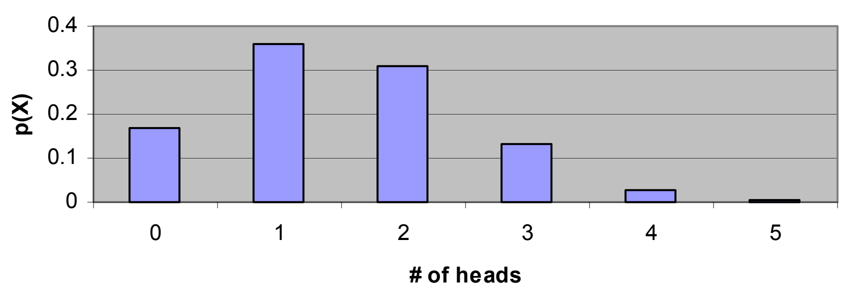
© 2006 A. Karpinski

- Example #3: X_3 = the number of heads in 5 tosses of a biased coin with probability of heads equal to 0.3

    - Using the binomial theorem, we can again calculate the probability of each possible outcome:

Probability of zero heads:	$p(0; 5,.3) = .168$
Probability of one head:	$p(1; 5,.3) = .360$
Probability of two heads:	$p(2; 5,.3) = .309$
Probability of three heads:	$p(3; 5,.3) = .132$
Probability of four heads:	$p(4; 5,.3) = .028$
Probability of five heads:	$p(5; 5,.3) = .002$

Total Probability: 1.00



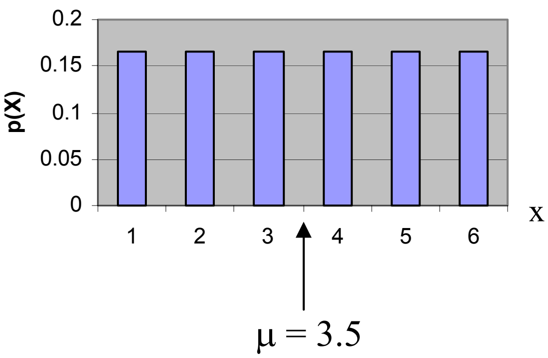
- Sampling Distribution of the Mean
    - In virtually all cases we compute statistics, we are interested in making inferences about the mean of a distribution
    - When we collect data, we observe a sample mean, a sample variance, and the sample size
    - From this information (and with a few assumptions) we can calculate the sampling distribution of the mean. That is, we can calculate the entire set of possible values the mean might assume, and the probability associated with each of those values
    - The sampling distribution of the mean can be thought of as the pattern of means that would result if we repeatedly took a random sample of size *n* from the population

 © 2006 A. Karpinski

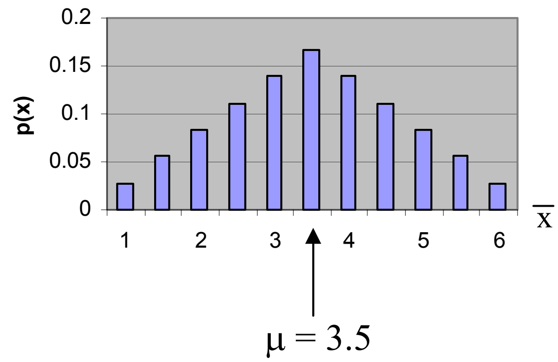
- Example #1: Let X be the role of a fair six-sided die, and let $n = 2$

  - The probability of each side occurring on any one roll is 1/6
  - Imagine drawing a sample of two from this population
  - There are 36 possible pairs of numbers for n=2
  - Let's calculate the mean of the pairs
  - This distribution of possible means is the sampling distribution of the mean

Population Probability Distribution



Sampling Distribution of the Mean (n=2)



  - Notice that the mean of x in the population distribution is the same as the mean of the sampling distribution, but the variability is greatly decreased in the sampling distribution!
  - This is an example of the fact we previously proved – that $\bar{x}$ is an unbiased estimator of μ
  - It is not true that the observed mean will always be identical to the population mean!! (Why?)

  - It is also the case that the variance of the sampling distribution will be less than (or equal to) the variance of the population. This fact is known as the <u>law of large numbers</u>

Specifically:

$$\sigma_{\bar{x}}^2 = \frac{\sigma_\mu^2}{N} \qquad \text{or} \qquad \sigma_{\bar{x}} = \frac{\sigma_\mu}{\sqrt{N}}$$

© 2006 A. Karpinski

* Sampling from the Normal Distribution

  * Theorem: If X is distributed $N(\mu_x, \sigma_x)$ then the sampling distribution of the mean, $\bar{x}$, based on random samples of size *n*, will also be normally distributed, with:

    $\mu_{\bar{x}} = \mu_x$

    $\sigma_{\bar{x}} = \dfrac{\sigma_x}{\sqrt{n}}$

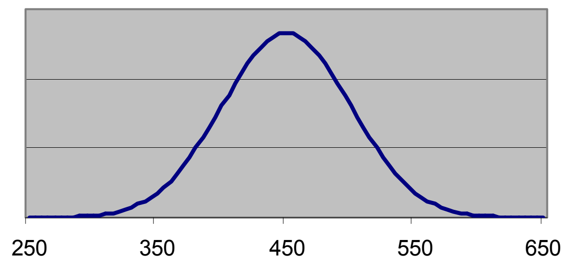
  * Key points:
    * If the parent distribution is normally distributed, the sampling distribution is normally distributed.
    * The expected mean of the sampling distribution is identical to the population mean.
    * The standard deviation of the sampling distribution is equal to the standard deviation of the population divided by the square root of the sample size. This number is referred to as the standard error of the mean.

      (In fact, any standard deviation of a sampling distribution is referred to as a standard error.)

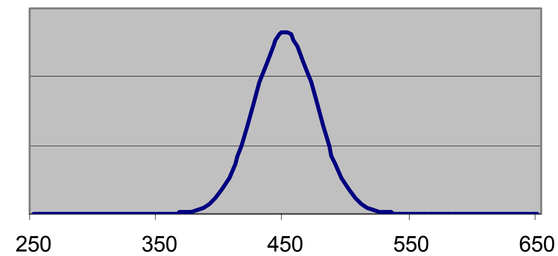
© 2006 A. Karpinski

- Let's consider an example of a sampling distribution from a parent population that is normally distributed. Let's assume that SAT scores are distributed $N(450, 50)$.

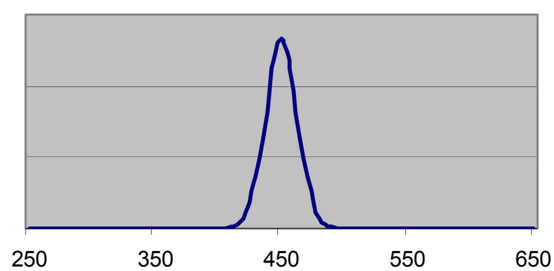
Population
$N(450,50)$



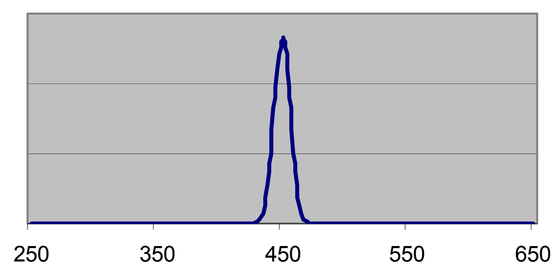
Sampling Distribution (n=4)
$N(450,25)$



Sampling Distribution (n=16)
$N(450,12.5)$



Sampling Distribution (n=64)
$N(450,6.25)$



© 2006 A. Karpinski

- Central Limit Theorem (CLT)
  - Just when you thought it could not get any better. . .

  - Theorem: If X comes from <u>any distribution</u> with mean = μ_x and standard deviation = σ_x then the sampling distribution of the mean, $\bar{x}$, based on random samples of size *n*, have the following properties:

    $\mu_{\bar{x}} = \mu_x$

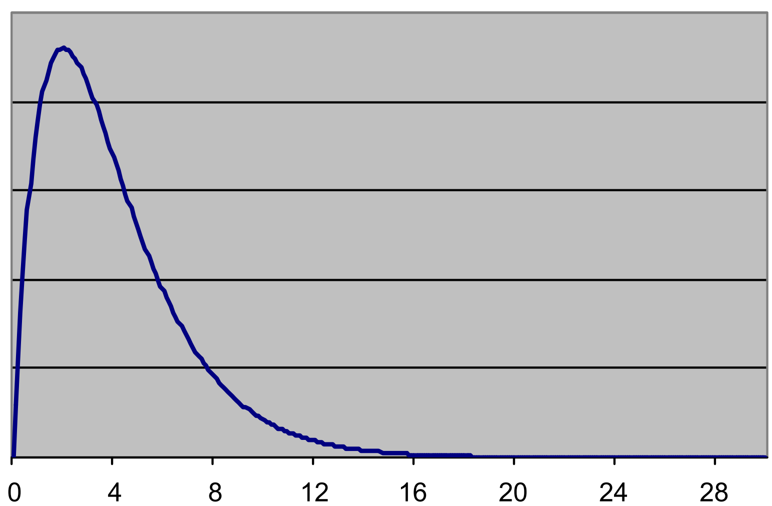
    $\sigma_{\bar{x}} = \dfrac{\sigma_x}{\sqrt{n}}$

    AND will tend to be <u>normally distributed</u> as the sample size becomes large.

  - Note that the CLT applies to the hypothetical sampling distribution of sample means, NOT to the sample distribution of observations!

  - An example with non-normally distributed data:

$\mu_x = 4$ and $\sigma_x = 2.828$

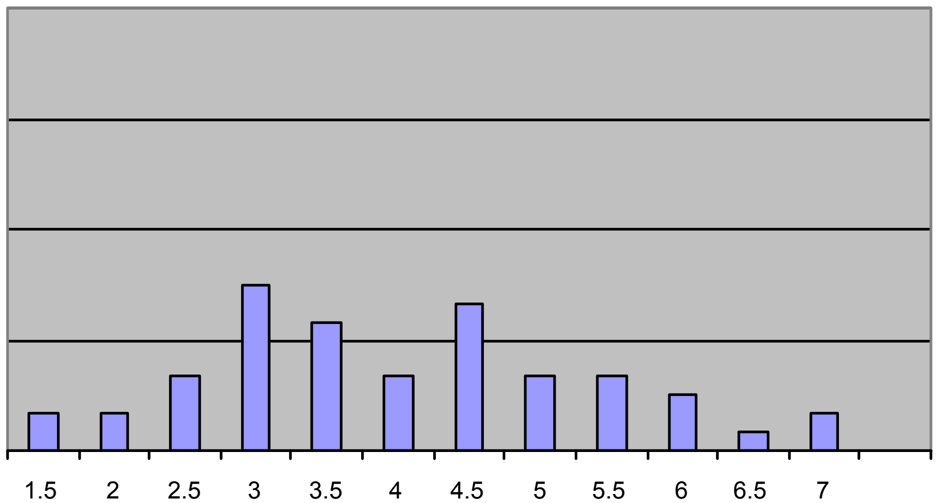


  - This distribution has what kind of skew?
  - Would its coefficient of skewness be positive or negative?

(For the curious, this is a gamma distribution)

 © 2006 A. Karpinski

- Let's start by taking random samples of size 5 (n=5) from the parent distribution.
- We'll take the mean of those samples and look at the resulting distribution of means. (For the super-curious, I took 50 random samples)



What should the mean and standard deviation of this distribution be?

Estimated

$\mu_{\bar{x}} = 4$

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{5}} = \frac{2.828}{2.236} = 1.265$$

Observed

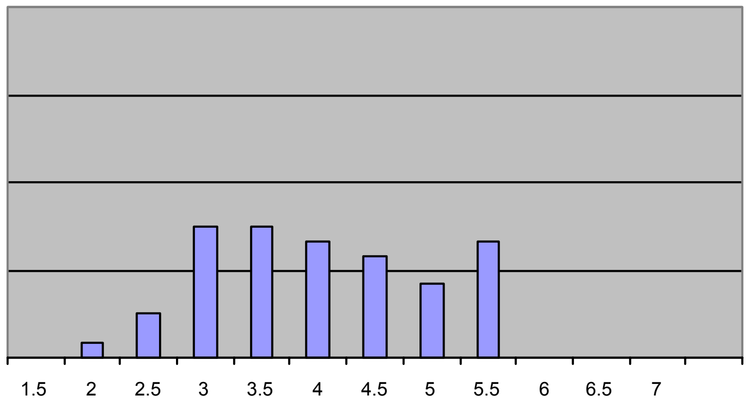
$\mu_{\bar{x}} = 3.74$

$\sigma_{\bar{x}} = 1.368$

Why the discrepancy?

© 2006 A. Karpinski

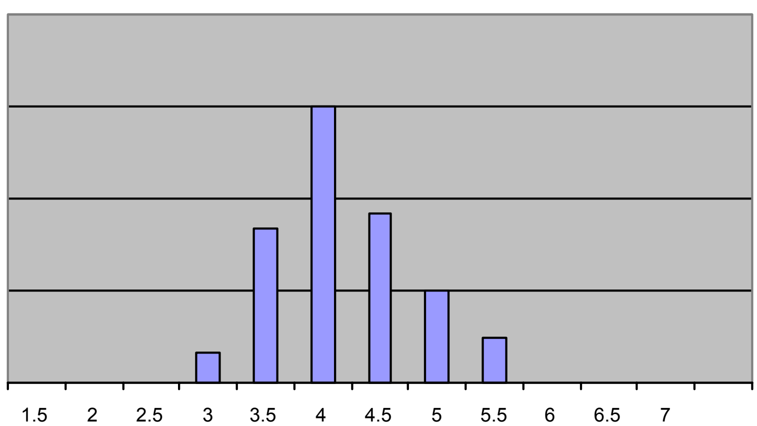
- Now let's take random samples of size 10 (n=10) from the parent distribution and look at the sampling distribution of the mean



What should the mean and standard deviation of this distribution be?

Estimated	Observed
$\mu_{\bar{x}} = 4$	$\mu_{\bar{x}} = 3.80$
$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{10}} = \frac{2.828}{3.162} = 0.894$	$\sigma_{\bar{x}} = 0.986$

- Finally let's take random samples of size 25 (n=25) from the parent distribution and look at the sampling distribution of the mean



Estimated	Observed
$\mu_{\bar{x}} = 4$	$\mu_{\bar{x}} = 3.95$
$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{25}} = \frac{2.828}{5} = 0.5656$	$\sigma_{\bar{x}} = 0.604$

© 2006 A. Karpinski

- The power of the central limit theorem (CLT) is that the underlying distribution of whatever you are studying does NOT have to be normally distributed for your statistics to be normally distributed, so long as your sample size is large

  (Remember, if your underlying distribution is normal, then the sampling distribution of the mean is automatically normally distributed, regardless of the sample size)

- A rule of thumb regarding the CLT is that you can generally count on normality to kick-in around n=30, but this is just an estimate, and it depends upon a number of factors.

- By looking at the formula for the standard error of the mean, we can see the law of large numbers. According to this law, the larger the sample size, the more likely it is that the sample mean will be close to the population mean.

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$$

n	$\sigma_{\bar{x}}$
1	$1^* \sigma_x$
5	$.447^* \sigma_x$
50	$.141^* \sigma_x$
100	$.1^* \sigma_x$
500	$.045^* \sigma_x$

  - This little demonstration shows the law of large numbers – the larger the sample, the more likely the mean of the sample is to reflect the mean of the general population

  - See Appendix C for an example of how people tend to process information about probabilities (it's not pretty!).

© 2006 A. Karpinski

Chapter 1: Appendices

A. Measurement scales gone awry: A second example

- A second example highlighting the importance of measurement scales and how complicated this issue can become:

    - Imagine that four participants rate how aggressive they find each of two film clips: a and b
    - Each person considers the input (the film clip), processes that input, and arrives at a decision:

      $f(a)$ and $f(b)$

    - What we would like to do is to average all the ratings of film a and compare them to the average ratings of film b

$$\frac{\sum f(a)}{N} > \frac{\sum f(b)}{N}$$

  - This equation represents the statement that "the average aggressiveness for movie a is greater than the average aggressiveness for movie b."
  - Note that a and b do not have subscripts because everyone watches the same movie

  - However, we assume that each participant is using and interpreting the rating scale in exactly the same manner. It seems likely that each person has his or her own scale:

$$\frac{\sum f_i(a)}{N} > \frac{\sum f_i(b)}{N}$$

 © 2006 A. Karpinski

* But the more plausible case presents big problems! Imagine that persons 1 and 2 have the following interpretation of the rating scale and that we can map these interpretations onto a single scale:

Person 1:	1	2	3			4			5	6	7
Person 2:	1			2	3	4	5	6			7
Combined:	1	2	3	4	5	6	7	8	9	10	11

* Suppose that we obtain the following data on the rating scale:

	Movie A	Movie B
Person 1	3	6
Person 2	6	2
	$\overline{X}_a = 4.5$	$\overline{X}_b = 4$

Movie A is more aggressive than Movie B

* But on the combined/actual scale, we obtain different results!

	Movie A	Movie B
Person 1	3	10
Person 2	8	2
	$\overline{X}_a = 5.5$	$\overline{X}_b = 6$

Movie B is more aggressive than Movie A

* In statistics, we assume that the data come from meaningful/interpretable scales. If not, we may misapply/misinterpret our statistics. These types of issues fall under the category of measurement theory

 © 2006 A. Karpinski

B. The Binomial Theorem

- The binomial theorem can be used to calculate probabilities of independent events whenever the outcome is dichotomous (typically referred to as a success or a failure).

- Example #1: X_2 = the number of heads in 5 tosses of a fair coin

  - Our goal is to list the probability of each possible outcome
  - One possible way to calculate these probabilities is to calculate them all by hand.

    - First, we need to determine the number of possible outcomes:

First coin =	Heads or tails	2 outcomes
Second coin =	Heads or tails	2 outcomes
Third coin =	Heads or tails	2 outcomes
Fourth coin =	Heads or tails	2 outcomes
Fifth coin =	Heads or tails	2 outcomes

2*2*2*2*2 = 32

    - Next, we need to list all 32 outcomes

	Coin 1	Coin 2	Coin 3	Coin 4	Coin 5		Coin 1	Coin 2	Coin 3	Coin 4	Coin 5
1	H	H	H	H	H	17	T	H	H	H	H
2	H	H	H	H	T	18	T	H	H	H	T
3	H	H	H	T	H	19	T	H	H	T	H
4	H	H	H	T	T	20	T	H	H	T	T
5	H	H	T	H	H	21	T	H	T	H	H
6	H	H	T	H	T	22	T	H	T	H	T
7	H	H	T	T	H	23	T	H	T	T	H
8	H	H	T	T	T	24	T	H	T	T	T
9	H	T	H	H	H	25	T	T	H	H	H
10	H	T	H	H	T	26	T	T	H	H	T
11	H	T	H	T	H	27	T	T	H	T	H
12	H	T	H	T	T	28	T	T	H	T	T
13	H	T	T	H	H	29	T	T	T	H	H
14	H	T	T	H	T	30	T	T	T	H	T
15	H	T	T	T	H	31	T	T	T	T	H
16	H	T	T	T	T	32	T	T	T	T	T

 © 2006 A. Karpinski

- Finally, we can tally the results and calculate the probabilities, for each possible outcome:

Total with 0 heads =	1/32	= 0.03125
Total with 1 head =	5/32	= 0.15625
Total with 2 heads =	10/32	= 0.31250
Total with 3 heads =	10/32	= 0.31250
Total with 4 heads =	5/32	= 0.15625
Total with 5 heads =	1/32	= 0.03125

- As you can see, this is a very tedious process that we would like to avoid at all costs!

o An alternative is to use the Binomial Theorem to calculate each of the possible outcomes:

$$p(x) = \binom{N}{x} p^x q^{(N-x)}$$

Where N = total number of trials
x = # of successes
p = probability of success
q = probability of failure (1-p)

And $$\binom{N}{x} = \frac{N!}{x!(N-x)!}$$

$p(0;5,.5) = .031$

$$p(0) = \binom{5}{0}(.5)^0(.5)^5 = \frac{5!}{0!*5!}*1*.03125 = .03125$$

$p(1;5,.5) = .156$
$p(2;5,.5) = .313$
$p(3;5,.5) = .313$
$p(4;5,.5) = .156$
$p(5;5,.5) = .031$

Total Probability: 1.00

o The binomial theorem saves us the trouble of listing all the possible outcomes

© 2006 A. Karpinski

- Example #2: A particular woman has given birth to 11 children. Assume the probability of having a boy and of having a girl is $p = .5$.

  a. What is the probability of having nine boys?
     - For this problem, $N = 11$, $x = 9$. $p = q = .5$

$$p(9) = \binom{11}{9}(.5)^9(.5)^2 = \frac{11!}{9!*2!}*.00195*.25 = 55*.00195*.25 = .0269$$

$p(9;11,.5) = .0269$

  b. What is the probability of having nine or more boys?
     - The probability of nine or more boys =
       Probability of having 9 boys +
       Probability of having 10 boys +
       Probability of having 11 boys +

$$p(9) = \binom{11}{9}(.5)^9(.5)^2 = \frac{11!}{9!*2!}*.00195*.25 = 55*.00195*.25 = .0269$$

$$p(10) = \binom{11}{10}(.5)^{10}(.5)^1 = \frac{11!}{10!*1!}*.000977*.5 = 11*.00977*.5 = .0054$$

$$p(11) = \binom{11}{11}(.5)^{11}(.5)^0 = \frac{11!}{11!*0!}*.5^{11}*1 = .5^{11} = .0005$$

$$p(x \geq 9) = .0269 + .0054 + .0049 = .0328$$

  c. What is the probability of having nine or more children of the same gender?
     - The probability of nine children of the same gender =
       Probability of having nine or more boys +
       Probability of having nine or more girls +

       $= .0328 + .0328 = .0656$

© 2006 A. Karpinski

- Example #3: Suppose the probability of having blue eyes is $p = .35$. Consider a sample of 15 people.

  a. What is the probability of three of the people having blue eyes?
     - For this problem, $N = 15$, $x = 3$. $p = .35$, $q = .65$

$$p(3) = \binom{15}{3}(.35)^3(.65)^{12} = \frac{15!}{12!*3!}*(.35)^3(.65)^{12} = 455*(.35)^3(.65)^{12} = .1110$$

  b. What is the probability of three of the people NOT having blue eyes?
     - If three people do not have blue eyes, then 12 people must have blue eyes. So we restate the problem as what
     - For this problem, $N = 15$, $x = 12$. $p = .35$, $q = .65$

$$p(3) = \binom{15}{12}(.35)^{12}(.65)^3 = \frac{15!}{3!*12!}*(.35)^{12}(.65)^3 = 455*(.35)^{12}(.65)^3 = .0004$$

C. A Psychological Perspective on Probability

- In general, people have a very poor intuitive understanding of probability and randomness

- Law of Small Numbers
  - o Small samples are assumed to be highly representative of the populations from which they are drawn
  - o In research settings, this leads to overconfidence in results of small samples, and overestimation of the replicability of results.
  - o Interpersonally, people think they "know" someone based on very small samples of behavior
  - o Abelson's second law of Statistics: Overconfidence abhors uncertainty. "Psychologically, people are prone to prefer false certitude to the daunting recognition of chance variability." (Abelson, 1995)

 © 2006 A. Karpinski

- Gambler's fallacy
    - After a long run of one outcome, people believe that the alternative outcome is more likely to occur
        - After 5 heads, a tail is more likely. . .
        - After 6 blacks in a row on the roulette wheel, people start betting on red with a greater frequency . . .
        - In a baseball game, when a .250 hitter goes 0 for 3, he is "due" for a hit in his fourth at bat
    - People think that random processes will correct themselves over the short haul.

- People exhibit an insensitivity to sample size
    - <u>An Example</u>: A certain town is served by two hospitals. In the larger hospital, about 45 babies are born each day, and in the smaller hospital, about 15 babies are born each day. As you know, about 50% of all babies are girls. However, the exact percentage varies from day to day. Sometimes it may be higher than 50%, sometimes it may be lower

      For a period of one year, each hospital recorded the number of days on which more than 60% of the babies born were girls. Which hospital do you think recorded more such days?
        - The larger hospital
        - The smaller hospital
        - About the same (that is, within 5% of each other)

      Results of how participants responded:

- The larger hospital	22%
- The smaller hospital	22%
- About the same	56%

        - But YOU know the real answer is the smaller hospital!
        - Again, in research settings, this finding suggests that researchers will be overconfident in results from small samples.

 © 2006 A. Karpinski

- People believe all chance events should "look random"
  - People believe that short sequences of random outcomes should be representative of the process that generated it.
  - Which sequence of heads and tails is more likely?
    - H-T-H-T-T-H
    - H-H-H-T-T-T

  - Which set of numbers is more likely to win the lottery?
    - 3,6,17,23,36,41
    - 1,2,3,4,5,6

  - Abelson's first Law of Statistics: Chance is lumpy.

- A small plug:

  Abelson, R. P. (1995). *Statistics as a principled argument*. Hillsdale, NJ: Lawrence Erlbaum.

- Abelson's laws of Statistics
  1. Chance is lumpy
  2. Overconfidence abhors uncertainty
  3. Never flout a convention just once
  4. Don't talk the Greek if you don't know the English translation
  5. If you have nothing to say, say nothing
  6. There is no free hunch
  7. You can't see the dust if you don't move the couch
  8. Criticism is the mother of methodology

© 2006 A. Karpinski