Summarizing Data

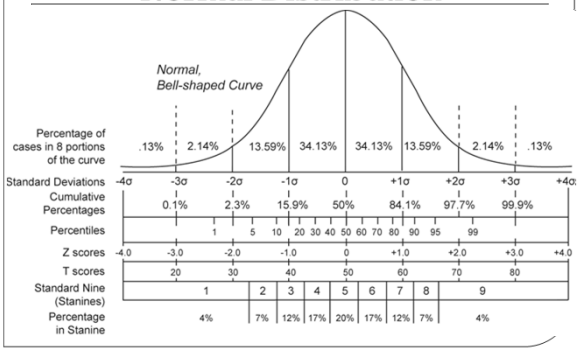
Description With Statistics

- Aspects or characteristics of data that we can describe are
  - Central Tendency (or Middle)
  - Dispersion (or Spread)
  - Skewness
  - Kurtosis
- Statistics that measure/describe central tendency are mean, median, and mode
- Statistics that measure/describe dispersion are range, variance, and standard deviation

Description With Statistics

- Central Tendency = middle, location, center
  - Measures of central tendency are mean, median, and mode (keywords)
- Dispersion = spread, variability
  - Measures of dispersion are range, variance, and standard deviation (keywords)
- Skewness = departure from symmetry
  - Positive skewness = tail of distribution (i.e., extreme scores) in positive direction
  - Negative skewness = tail of distribution (i.e., extreme scores) in negative direction
- Kurtosis = peakedness relative to normal curve

Normal Distribution



Describing Central Tendency

- "Central Tendency" is the aspect of data we want to describe
- We describe/measure the central tendency of data in a sample with the statistics:
  - Mean
  - Median
  - Mode
- We describe/measure the central tendency of data in a population with the parameter μ ('mu'); we usually do not know μ, so we estimate it with $\overline{X}$

Sample Mean

- The sample mean is the sum of the scores divided by the number of scores and it is symbolized by $\overline{X}$

$$\overline{X} = \frac{\Sigma X}{N}$$

- Example: 4, 1, 7
  - N=3
  - ΣX=12
  - $\overline{X}$ = ΣX/N = 12/3 = 4
- Characteristics:
  - $\overline{X}$ is the balance point
  - $\Sigma(X - \overline{X})=0$
  - $\overline{X}$ Minimizes $\Sigma(X - \overline{X})^2$ (Least Squares criterion)
    - Minimizes standard deviation
  - $\overline{X}$ is pulled in the direction of extreme scores

Sample Mean

- What is the mean for the following data:
  4, 1, 7, 6

- N=4
- $\Sigma X=18$
- $\overline{X} = \Sigma X/N = 18/4 = 4.5$

Sample Median

- The median is the middle of the ordered scores and it is symbolized as X_{50}
- Median position (as distinct from the median itself) is (N+1)/2 and is used to find the median
- Find the median of these scores: 4, 1, 7
  - N=3
  - Median position is (3+1)/2 = 4/2 = 2
  - Place the scores in order: 1, 4, 7
  - X_{50} is the score in position/rank 2
  - So $X_{50} = 4$

Sample Median

- Another example: 4, 1, 7, 6
  - N=4
  - Median position is (N+1)/2 = (4+1)/2 = 5/2 = 2.5
  - Place the scores in order: 1, 4, 6, 7
  - X_{50} is the score in position/rank 2.5
  - So $X_{50} = (4+6)/2 = 10/2 = 5$
- Characteristics:
  - Depends on only one or two middle values
  - For quantitative data when distribution is skewed
  - Minimizes $\Sigma|X-X_{50}|$
    - Minimizes absolute deviation

Sample Mode

- The mode is the most frequent score
- Examples:
  - 1 1 4 7
    - the mode is 1
  - 1 1 4 7 7
    - there are two modes: 1 and 7
  - 1 4 7
    - there is no mode
- Characteristics:
  - Has problems: more than one, or none; maybe not in the middle; little info regarding the data
  - Best for qualitative data (e.g., gender)
  - If it exists, it is always one of the scores
  - It is rarely used

Describing the Dispersion of Data

- "Dispersion" is the aspect of data we want to describe
- Any statistic that describes/measures dispersion should have these characteristics: it should…
  - Equal zero when the dispersion is zero
  - Increase as dispersion increases
  - Measure just dispersion, not central tendency

Describing the Dispersion of Data

- We describe/measure the dispersion of data in a sample with the statistics:
  - Range = high score-low score
  - Sample variance, s^{*2}
  - Sample standard deviation, s^*
  - Unbiased variance estimate, s^2
  - Standard deviation, s
- We describe/measure the dispersion of data in a population with the parameter σ ('sigma') or σ^2; we usually do not know σ or σ^2, so we estimate them with one of the statistics

Range

- Formula is high score – low score.
- Example: 4 1 5 3 3 6 1 2 6 4 5 3 4 1, N = 14
  - Arrange data in order: 1 1 1 2 3 3 3 4 4 4 5 5 6 6
  - Range = high score – low score = 6 – 1 = 5

Sample Variance (s^{*2})

- Definitional formula: $s^{*2} = \frac{\Sigma(X-\overline{X})^2}{N}$

  the average squared deviation from $\overline{X}$
- Example: 1 2 3
  - N=3, $\overline{X} = \Sigma X/N = 6/3 = 2$
  - $\Sigma(X-\overline{X})^2 = (1-2)^2+(2-2)^2+(3-2)^2 = 1+0+1 = 2$
  - $s^{*2} = 2/3 = .6667$
- Computational formula: $s^{*2} = \frac{[N\Sigma X^2-(\Sigma X)^2]}{N^2}$
  - $\Sigma X^2 = 1^2+2^2+3^2 = 1+4+9 = 14$, $\Sigma X = 6$, $N = 3$
  - $s^{*2} = [3(14)-(6)^2]/3^2 = [42-36]/9 = 6/9 = 2/3 = .6667$
- S^{*2} is in squared units of measure
- This gives you the AVERAGE SQUARED DEVIATION AROUND THE MEAN

Sample Standard Deviation (s*)

- Formula: $s^* = \sqrt{s*2}$
- Example: 1 2 3
  - N=3, $\overline{X} = \Sigma X/N = 6/3 = 2$
  - $\Sigma(X-\overline{X})^2 = (1-2)^2+(2-2)^2+(3-2)^2 = 1+0+1 = 2$
  - $s^{*2} = 2/3 = .6667$
  - $s^* = \sqrt{.6667} = .8165$
- s* is in original units of measure
- s* is the typical distance of scores from the mean (i.e., the average deviation of scores from the mean)

Unbiased Variance Estimate (s^2)

- Definitional formula: $s^2 = \frac{\Sigma(X-\overline{X})^2}{(N-1)}$
- Example: 1 2 3
  - N=3, $\overline{X}$ = ΣX/N=6/3=2
  - $\Sigma(X-\overline{X})^2 = (1-2)^2+(2-2)^2+(3-2)^2=1+0+1=2$
  - $s^2=2/2=1.0$
- Computational formula:

  $s^2 = \frac{[N\Sigma X^2-(\Sigma X)^2]}{[N(N-1)]}$
  - $\Sigma X^2 = 1^2+2^2+3^2=1+4+9=14$, ΣX=6, N=3
  - $s^2=[3(14)-(6)^2]/[3(2)]=[42-36]/6=6/6=1.0$
- s^2 is in squared units of measure
- The only difference between s^{*2} and s^2 is the "-1" in the denominator of the formula for s^2

Standard Deviation (s)

- Formula: $s=\sqrt{s^2}$
- Example: 1 2 3
  - N=3, $\overline{X}$ = ΣX/N=6/3=2
  - $\Sigma(X-\overline{X})^2 = (1-2)^2+(2-2)^2+(3-2)^2=1+0+1=2$
  - $s^2=1.0$
  - $s=\sqrt{1}=1.0$
- s is in original units of measure

Why do we care about measures of central tendency and dispersion?

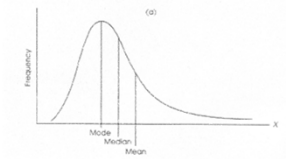
- Once we have collected data, the first step is usually to organize the information using simple descriptive statistics (e.g., measures of central tendency and dispersion)
- Measures of central tendency are AVERAGES
  - Mean, median, and mode are different ways of finding the one value that best represents all of your data
- Measures of dispersion tell us how much scores DIFFER FROM ONE ANOTHER

Why do we have two formulae for variance and standard deviation?

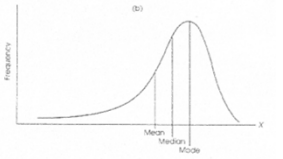
- Remember that our statistics are ESTIMATES of the parameters in the population
- When we use N as the denominator (as in s^{*2} & s^*), we produce a biased estimate (it is too small)
- We are trying to be good scientists so we will be conservative and use the unbiased estimate of the variance (s^2) and its associated standard deviation (s)
- We will address the idea of 'bias' later in the semester and this will be our introduction to the concept

Skewness

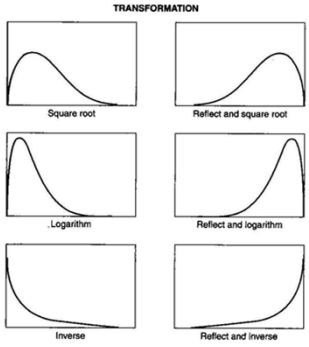
Positive Skewness



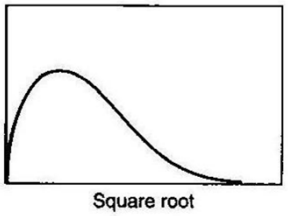
Negative Skewness



Common Data Transformations



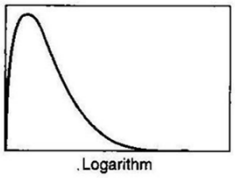
Square Root Transformation



Moderate Positive Skew

SPSS Syntax
compute new_exam1=sqrt(exam1).
execute.

Logarithmic Transformation

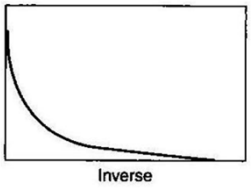


Substantial Positive Skew

SPSS Syntax
compute new_exam1=lg10(exam1).
execute.

If zero is a value, then use…
compute new_exam1=lg10(exam1+constant).
execute.
--"constant" is a numeric value added to each score so that the lowest value is 1.

Inverse Transformation

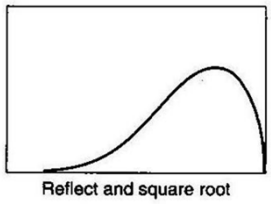


Severe Positive Skew

SPSS Syntax
compute new_exam1=1/exam1.
execute.

If zero is a value, then use…
compute new_exam1=1/(exam1+constant).
execute.
--"constant" is a numeric value added to each score so that the lowest value is 1.

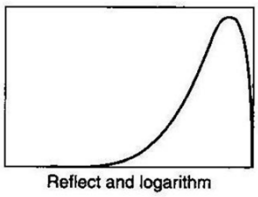
Reflected Square Root Transformation



Moderate Negative Skew

SPSS Syntax
compute new_exam1=sqrt(constant-exam1).
execute.
--"constant" is a numeric value from which each score is subtracted so that the smallest score is 1 (usually equal to the largest score +1)

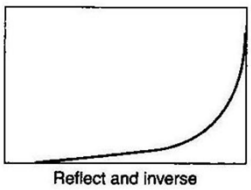
Reflected Logarithmic Transformation



Substantial Negative Skew

SPSS Syntax
compute new_exam1=lg10(constant-exam1).
execute.
--"constant" is a numeric value from which each score is subtracted so that the smallest score is 1 (usually equal to the largest score +1)

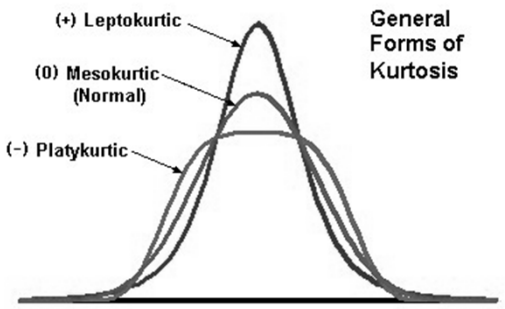
Reflected Inverse Transformation



Severe Negative Skew

SPSS Syntax
compute new_exam1=1/(constant-exam1).
execute.
--"constant" is a numeric value from which each score is subtracted so that the smallest score is 1 (usually equal to the largest score +1)

Kurtosis



SPSS Basics

- Three windows
  - Data editor (where we enter data)
  - Syntax editor (where we create and store syntax)
  - SPSS viewer (where we can see the output/results of our analyses)
- Two primary interfaces
  - Graphical user interface (point-and-click)
    - Very easy to use
    - Preferred for simple operations
  - Syntax
    - Takes a bit longer to learn
    - More flexible
    - Preferred for creating scores in a data file
    - Preferred for complex operations

Reading Assignment

Read the following chapters in Aspelmeier and Pierce for the next class session:

-Chapter 1: Introduction to SPSS: A user-friendly approach
-Chapter 2: Basic operations
-Chapter 3: Finding sums
-Chapter 4: Frequency distributions and charts
-Chapter 5: Describing distributions