Lecture 1: Describing and Exploring the Data

• 1. 1 Descriptive and Inferential Statistics

- Descriptive statistics are used when one wants to describe some characteristic of a sample

- It involves:

  - collecting, organizing, and summarizing data

  - but also … loss of information

- Inferential statistics are used when one wants to infer something about the population given information about a sample

- It includes:

  - making inferences

  - testing hypothesis

  - examining relationships

  - making predictions

- Examples of inferential statistics

  - high school GPA scores are used to predict performance in college

  - a sample of public attitudes about religion is used to infer the attitude of the country

• 1. 2 Population vs. Samples

- A population is a group of observational units that includes all possible such units. Theoretically, a population can be finite or infinite

- A sample is a subset of a population

• 1. 3 Parameters vs. Statistics

- A parameter is a value of interest in the population

  - population mean, population median, population standard deviation, population variance, population correlation coefficient

  - they are typically denoted by Greek letters (e.g., μ, σ, σ^2, ρ)

- A statistic is a value computed using samples

  - sample mean, sample median, sample standard deviation, sample variance, sample correlation coefficient

  - they are typically denoted by English letters (e.g., $\overline{x}$, s, s^2, r)

• 1. 4 Measurement

- Measurement is the attempt to capture – via quantification – attributes in a construct

- One of the least achieved– and most difficult – goals in psychology

- The measurement numbers [quantification] we obtain must be good reflections of the true quantities, so that information about magnitudes or amounts of the property can be at least inferred from the values observed

- Some kinds of measurement allow one to make good inferences about differences among the true magnitudes – and actual differences – of properties between objects. Other measurements only allow one to make weak inferences

- Currently, an issue of heated debate in some fields of psychology

• 1. 5 Measurement Scales

- Measurement at the ***nominal scale*** level (or categorical scale) refers to grouping individual observations into qualitative classes

  - No numerical measurement or information about quantity

  - No numerical equivalence among classes

  - Examples: Big Five (*OCEAN*), attachment styles, college major, gender

- Measurement at the ***ordinal scale*** level refers to when individual observations are grouped into classes, between which one can establish an (arbitrary) order or ranking

  - Distance between values is not necessarily equal

  - Examples: order of finish in a race, items with responses “very little, little, neutral, much, a lot;” school grades

- Measurement at the ***interval scale*** level occurs when the numbers denote magnitude of differences among observations

    - The measurement is some linear function of the true magnitude

    - Interval scales allow comparisons between objects in terms of their magnitude

    - Examples: physiological measures (e.g., temperature; the difference between 30 and 50 degrees is exactly the same as between 100 and 120 degrees

    - No logical zero; zero is no more meaningful than any other value

- Measurement at the ***ratio scale*** level occurs when, in addition to the properties described for the interval scale there is a logical zero point

    - The measurement is some linear function of the true magnitude

    - Ratio scales allow interpreting ratios of numerical measurements as ratios of magnitudes of objects

$$m(o_i) / m(o_j) = t(o_i) / t(o_j)$$

    - Ratio scales have a nonarbitary zero

    - Examples: age, height, weight, speed

- Scales (Summary)

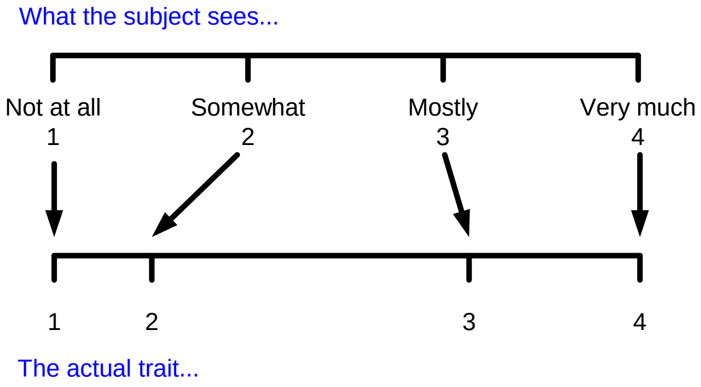
    - Nominal: classification

    - Ordinal: classification, order

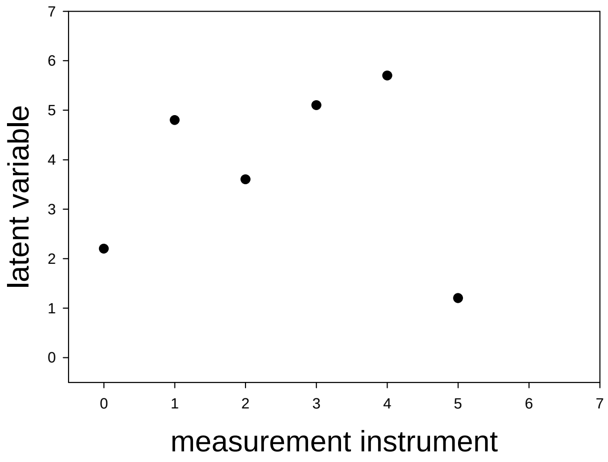
    - Interval: classification, order, equal intervals

    - Ratio: classification, order, equal intervals, true zero

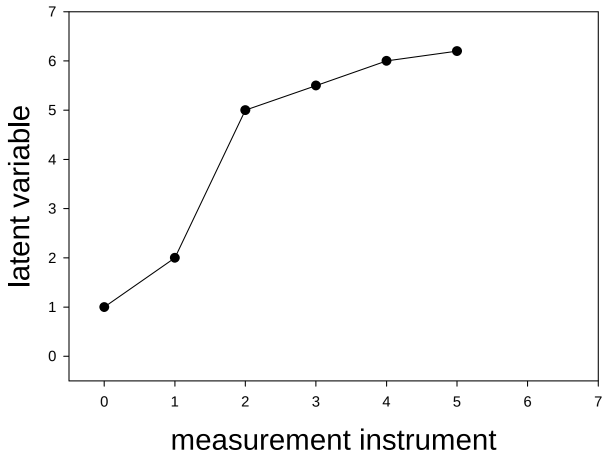
- **1. 6 Relationship Between Measurement and Construct**



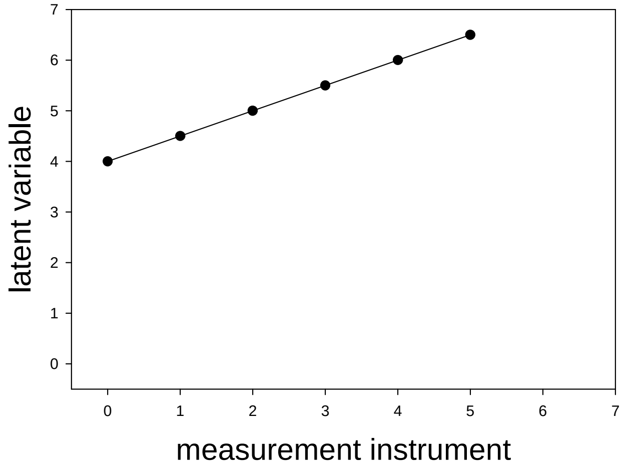
Nominal



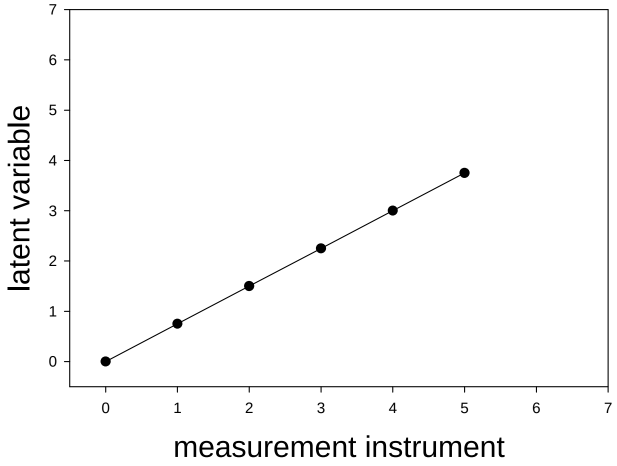
Ordinal



Interval



Ratio



• 1. 7 Frequency Distributions

- An account of classes and their frequencies is called ***frequency distribution***

- It is a way to summarize and display raw data

- It loses information from individual scores but provides insights about the whole sample

- When frequencies are assigned to intervals or groupings of categories, the frequencies are termed ***grouped frequency distributions***. The groupings are then termed ***class intervals***

- Example: class intervals for age, IQ

- ***Interval size*** (or width) refers to the difference between the largest and the smallest number within a class interval

- A possible rule: i = (highest score – lowest score) / number of class intervals

- A class interval has ***real*** and ***apparent limits***

- Example: in the class interval ages 20 – 25, the real limits are 19.5 to 25.5

- The ***number*** of class intervals will depend on the data and the balance between detail and summary. Typically, no more than 10 to 20 class intervals are necessary

- The ***midpoint*** of a class interval is the number that falls exactly halfway among the numbers in the interval

- Intervals can be ***open*** and have unequal size, depending on the characteristics of the data

Theoretical Frequency Distribution

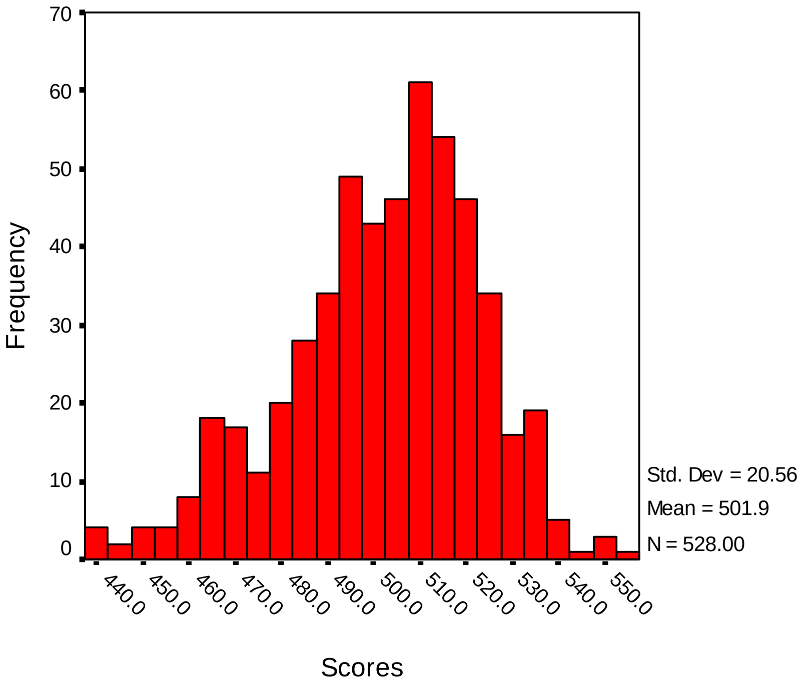
Height (inches)	f
78-82	2
73-77	21
68-72	136
63-67	682
58-62	136
53-57	21
48-52	2
	1,000 = N

• 1. 8 Graphs of Distributions

- A ***histogram*** is a display of data in which each class interval is represented by a portion of the *x*-axis, and their frequencies are represented by the height in the *y*-axis

- Alternative forms of displaying data are the pie charts and bar charts

- It loses information from individual scores but provides insights about the whole sample



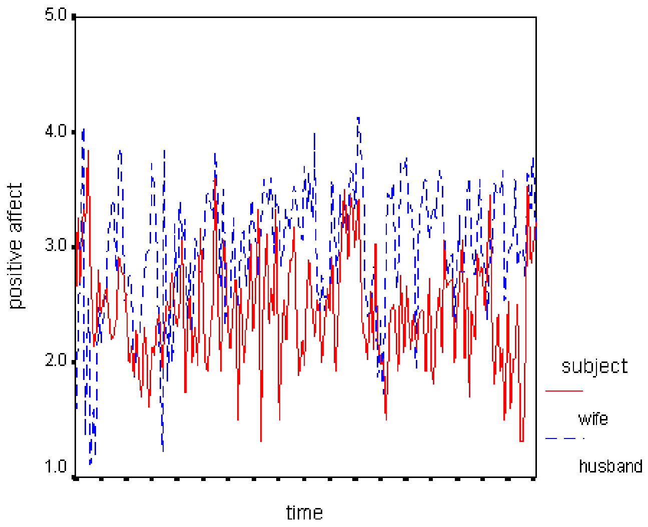
- A ***frequency polygon*** is a display of data in which the midpoints of the intervals and their frequency are joined by straight lines.

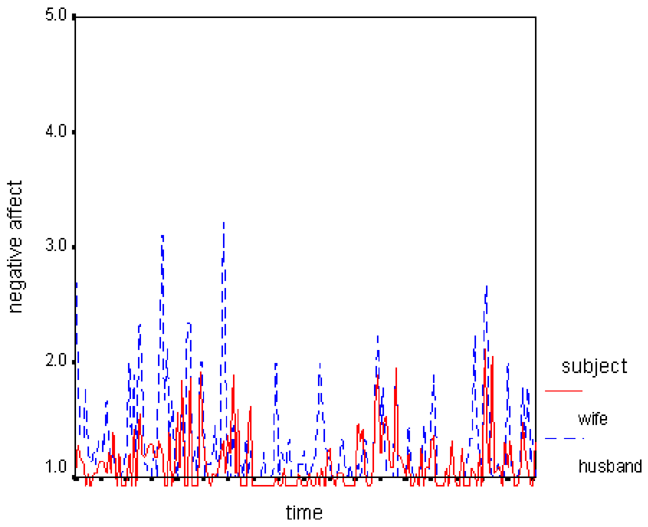
  - It is particularly helpful in the case of numerous intervals for yielding a smooth curve

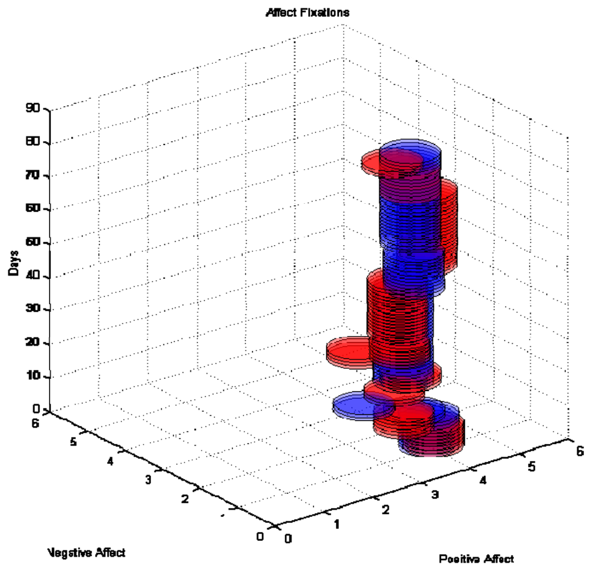
- A ***cumulative frequency polygon*** is a display of data representing the relation between a class interval and the frequency at or below its upper limit

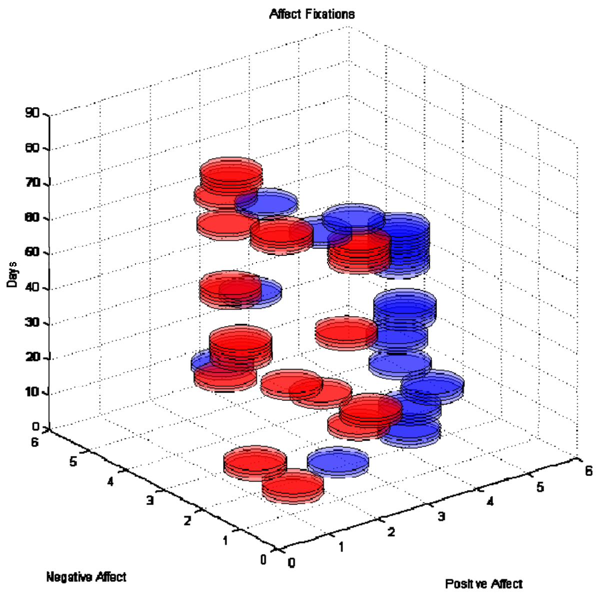
http://www.itl.nist.gov/div898/handbook/index.htm

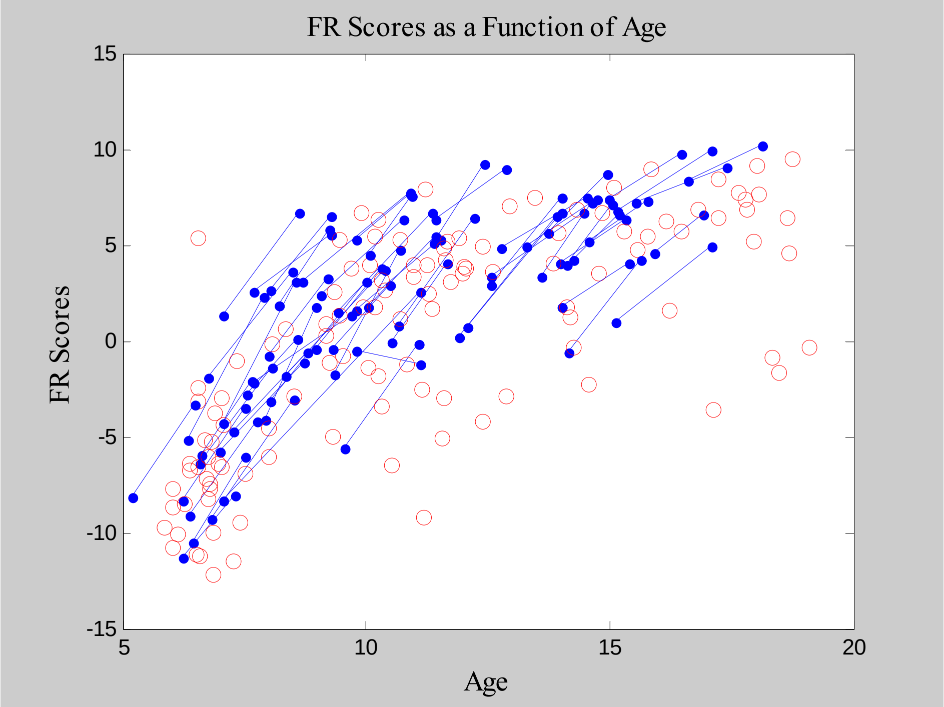
http://www.youtube.com/watch?v=jbkSRLYSojo

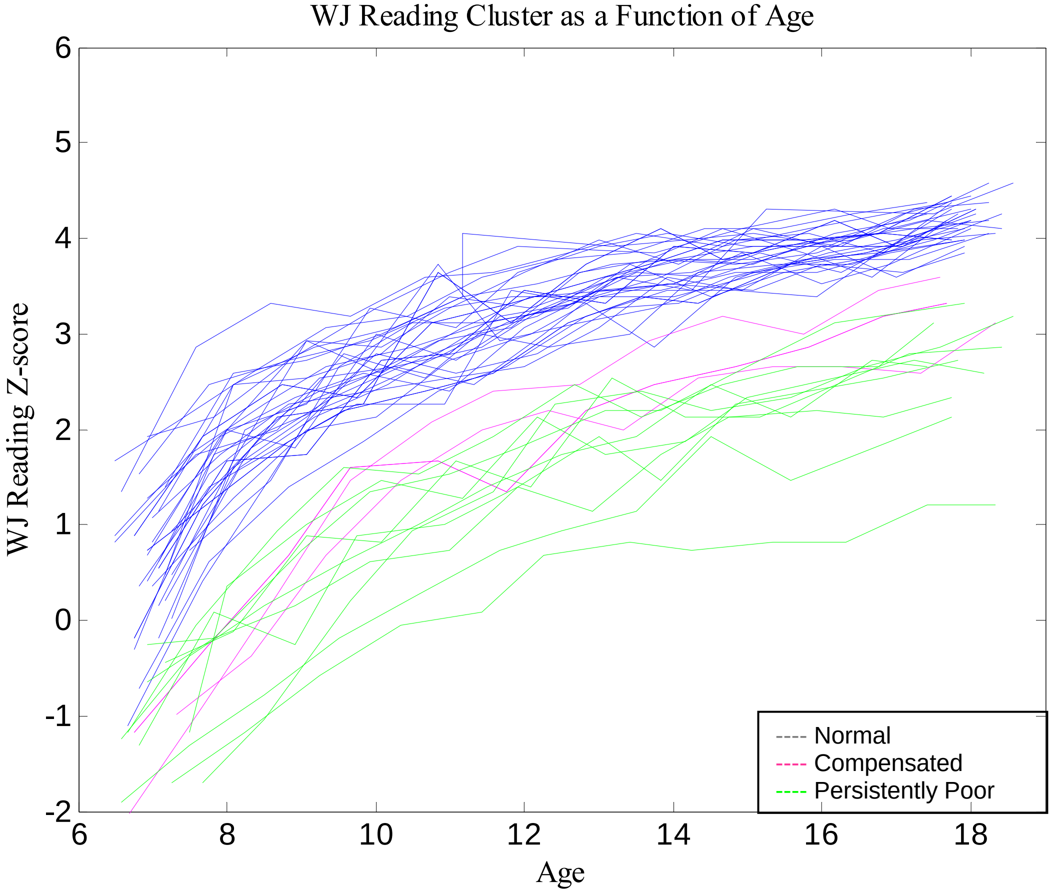

5.0
4.0
3.0
2.0
1.0
positive affect
time
subject
wife
husband


5.0
4.0
3.0
2.0
1.0
negative affect
time
subject
wife
husband


Affect Fixations
Days
90
80
70
60
50
40
30
20
10
0
6
5
4
3
2
0
0
1
2
3
4
5
6
Negtive Affect
Positive Affect


Affect Fixations
Days
90
80
70
60
50
40
30
20
10
0
6
5
4
3
2
1
0
0
1
2
3
4
5
6
Negative Affect
Posit ve Affect


FR Scores as a Function of Age
FR Scores
Age
15
10
5
0
-5
-10
-15
5
10
15
20


WJ Reading Cluster as a Function of Age
WJ Reading Z-score
Age
6
5
4
3
2
1
0
-1
-2
6
8
10
12
14
16
18
Normal
Compensated
Persistently Poor

• 1. 9 Random Variables

- ***Variables*** are properties of objects or events that take on different values (e.g., *age*, gender)

- ***Discrete variables*** take on a small set of possible values (e.g., gender, marital status)

- ***Continuous variables*** can take on any reasonable value

- ***Independent variables*** (IVs) are those manipulated by the experimenter

- ***Dependent variables*** (DVs) are the outcome variables

- Let X be a function that associates a real number with each and every elementary event in some sample space ***S***. Then X is called a ***random variable*** on the sample space ***S***

- A random variable X represents values that are associated with elementary events, so that particular values of X occur when the appropriate elementary events occur, however the association might be

- Random variables can be specified in three ways: listing all possible numerical events and their associated probability, graphing this relationship, or expressing a rule of the probability for each value

- Typical notation is to use capital letters, such as X, Y, Z to denote random variables and lowercase letters, such as x, y, z to denote particular values of the random variable

Central Tendency and Variability

• 1. 10 Measures of Central Tendency

- Measures of central tendency are ways to describe the average value of the distribution

- Measures of variability are ways to describe the dispersion or spread of the data

- The ***mode*** is the midpoint or class name of the most frequent measurement class

  - the highest peak in a graphical depiction of the distribution

  - a distribution may have more than one modal class

  - it is very sensitive to the size and number of class intervals (in the case of numerical events)

  - it is a very unreliable source of information about the basic probability distribution

  - it is typically used when the measures are on nominal scale

- The ***median*** is the point exactly midway between the top and bottom halves of the distribution

  - when ***N*** is odd, the median is the score of individual number $(N + 1)/2$

    X = [1 2 3 **4** 5 6 7] – Median is 4

  - when ***N*** is even, the median is the value midway between the scores for individual (*N*/2) and (*N*/2) + 1 (half way between two middle scores)

    X = 1 2 3 4 | 5 6 7 8 – Median is 4.5

  - for grouped frequency distributions, the median is the point at or below which exactly 50% of the cases fall

$$median = \text{lower real limit} + i\left(\frac{.50N - cf \text{ below lower limit}}{f \text{ in interval}}\right)$$

if the interval frequency = 0, the median is the midpoint of the interval

- Example

Class	f	cf
74-78	10	200
69-73	18	190
64-68	16	172
59-63	16	156
54-58	11	140
49-53	27	129
44-48	17	102
39-43	49	85
34-38	22	36
29-33	6	14
24-28	8	8
	200	

$$\textit{median} = \text{lower real limit} + i\left(\frac{.50N - cf \text{ below lower limit}}{f \text{ in interval}}\right)$$

$$\textit{median} = 43.5 + 5\left(\frac{.50(200) - 85}{17}\right) = 43.5 + 5\left(\frac{15}{17}\right) = 43.5 + 4.4 = 47.9$$

- The ***mean*** is the arithmetic average of the distribution

- for raw data, the mean can be expressed as $\overline{X} = \sum_{i=1}^{N} \frac{x_i}{N}$

- for grouped distribution scores, the mean can be expressed as $\overline{X} = \frac{1}{N}\sum_{j=1}^{J} x_j f_j$ where

x_j is the midpoint of any interval j, f_j is the frequency corresponding to that interval, and the sum is taken over all J of the intervals

- Example

Class	*x*	*f*	*Xf*
74-78	76	10	761
69-73	71	18	1278
64-68	66	16	1056
59-63	61	16	976
54-58	56	11	616
49-53	51	27	1377
44-48	46	17	782
39-43	41	49	2009
34-38	36	22	792
29-33	31	6	186
24-28	26	8	208
		200	10,040

$$\overline{X} = \frac{1}{N}\sum_{j=1}^{J} x_j f_j = \frac{10{,}040}{200} = 50.2$$

- the mean is the score in a distribution about which deviations in one direction exactly equal deviations in the other

- the sum of the signed deviations about the mean is 0 in any distribution of numerical values

$d_i = (x_i - \overline{x})$, so

$$\sum_i di = \sum_i x_i - \sum_i \overline{x} = \sum_i x_i - N\overline{x} = Nx - Nx = 0$$

- Comparison among mean, median, and mode

- Mean

    Used for inference as well as description; best estimator of the parameter

    Based on all data in the distribution

    Generally preferred except for “bad” distribution. Most commonly used statistic for central tendency

- Median

    Good for “bad” distributions

    Good for distributions with arbitrary ceiling or floor

- Mode

    Good for nominal variables

    Good if you need to know most frequent observation

    Quick and easy

• 1. 11 “Best Guess” Interpretations of Central Tendency

- Picking the mean as the “best guess” score from a distribution produces an error *d*

- Over all possible cases, the average error would be $\sum_i \frac{d_i}{N}$ so, on average, if picking the mean as the best guess score from a distribution, the amount of error will be 0

- But what about if just a single pick and you want to be absolutely right?
Mode; the size of the error doesn’t matter

- What about if just a single pick and want to make the smallest absolute error?
Median; the typical score: it is closest on average to all the scores in the distribution

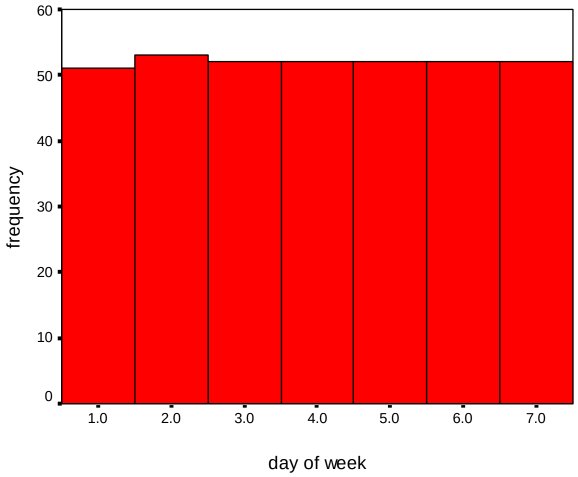
- Mean – average of signed error will be zero

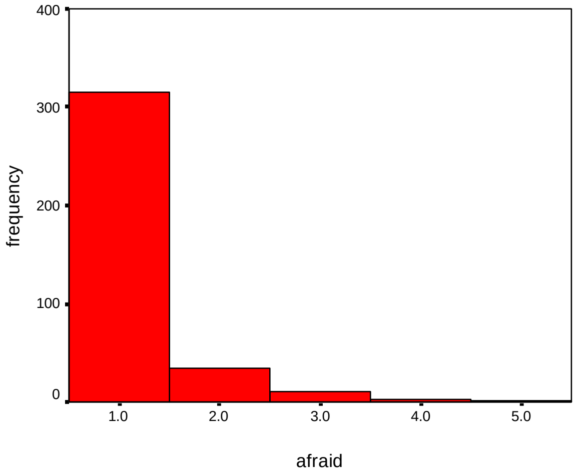
- Mode – will be absolutely right with greatest frequency

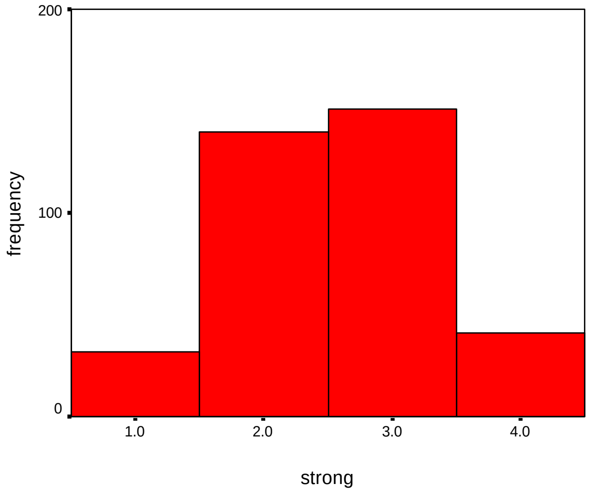
- Median – smallest absolute error

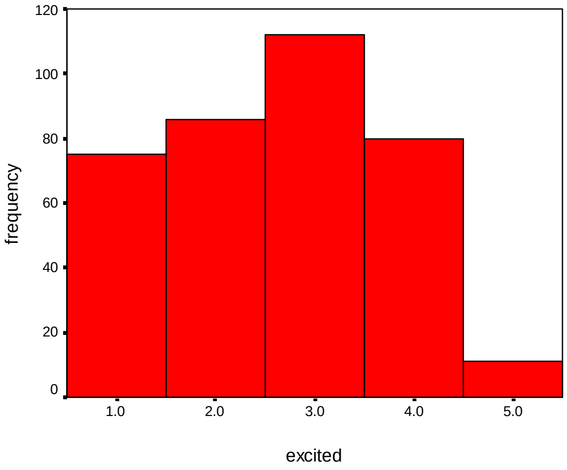
• 1. 12 Relations Between Central Tendency Measure and the Shapes of Distributions

- **Modality** of a distribution refers to its number of relative maximum points. Examples: (no modes, unimodal, bimodal, and multimodal distributions).

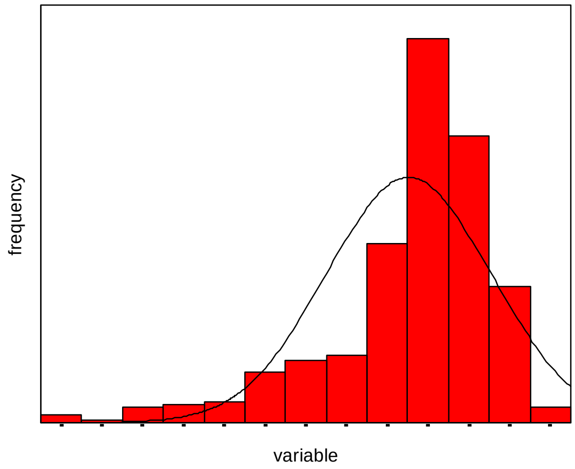


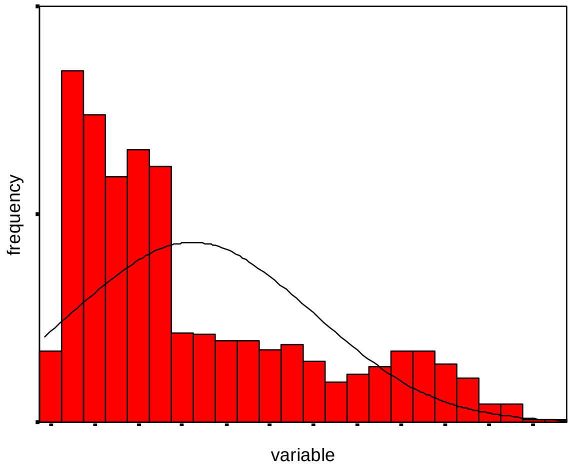


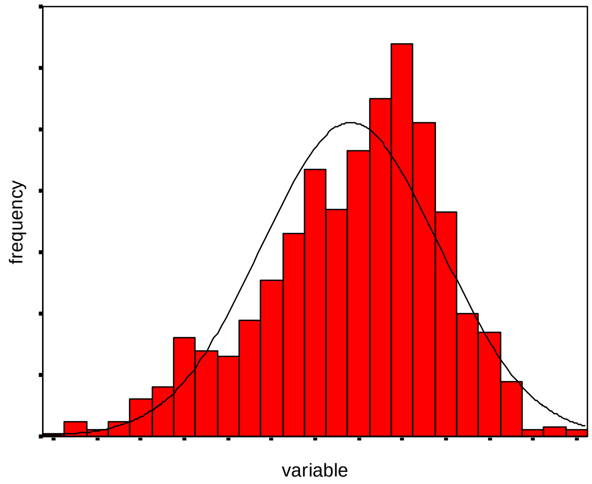


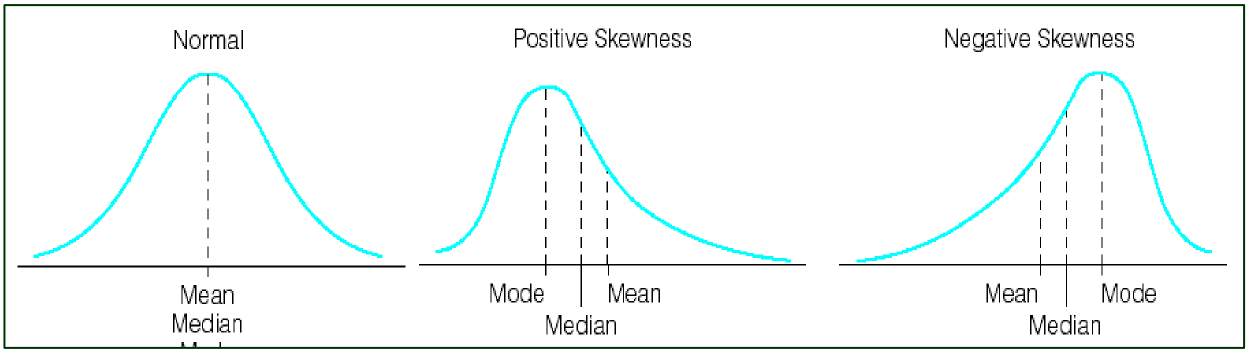


- **Skewness** refers to the symmetry of the distribution. Examples: (negatively skewed, positively skewed, and normal distributions).









- When a distribution is symmetric, the mean and the median are equal. But the opposite is not necessarily true

- When a distribution is negatively skewed, the bulk of the cases fall into the upper part of the distribution and the median is greater than the mean. The opposite is true in the case of a positively skewed distribution

- ***Resistance*** is the ability of a measure to be unaffected by changes in the values of a small number of cases.

> - The mean is extremely non-resistant. A change in the value of any case will affect the value of the mean. Outliers and highly skewed distributions affect the mean greatly
>
> - The median is highly resistant. It is only affected by a change in the value of a case that causes it to move from the top to the bottom half of the distribution or viceversa
>
> - The mode is highly resistant. It is only affected by changes that influence which score is the most frequent
>
> - Example: Given the following scores: [1, 3, 3, 6, 7, 7, 8, 8, 8, 9]
>
> Mo = 8, Md = 7, $\overline{x} = 6$
>
> If we introduce an outlier (60), Mo = 8, Md = 7, $\overline{x} = 11$.

- Description of a distribution in terms of its shape is illustrative but not informative about its essential (statistical) properties

• 1. 13 Measures of Dispersion

- ***Dispersion*** or ***spread*** refers to the tendency for observations to depart from central tendency

- ***Range*** is the spread of a distribution and is computed as

$range$ = (high score – low score)

Example: X = [12 14 14 16 16 18 20]; $range = 20 - 12 = 8$

- A ***deviation*** from the mean represents how much in error is the mean as a description of this case

$d_i = (x_i - \overline{x})$

- Similarly, a deviation from the median is the departure of a case from the median, or

$d'_i = (x_i - Md)$

- The ***variance*** as an index to reflect the spread of variability

- the average of the deviations about the mean is computed as

$\sum_{i=1}^{N}\frac{(x_i-\overline{x})}{N}$, but this will be zero so, instead, the squared deviations are computed

$$s^2=\sum_{i=1}^{N}\frac{(x_i-\overline{x})^2}{N}=\sum_{i=1}^{N}\frac{(d_i)^2}{N}$$

- the variance represents the average of the squared deviations from the mean

- the S^2 will be zero *if and only if* every case in the distribution shows the same value

- ***Average deviation*** – mean of absolute deviations from the median:

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

- The ***standard deviation*** of a distribution is the square root of the variance and is an index of variability in the original metric

$$s=\sqrt{s^2}$$

- An alternative way to compute the variance and standard deviation is as

$s^2=\sum_{i=1}^{N}\frac{{x_i}^2}{N}-\overline{x}^2$, and this comes from

$$s^2=\sum_{i=1}^{N}\frac{(x_i-\overline{x})^2}{N}=\sum_{i=1}^{N}\frac{(x_i-\overline{x})(x_i-\overline{x})}{N}=\sum_{i=1}^{N}\frac{({x_i}^2-2x_i\overline{x}+\overline{x}^2)}{N}=$$

$$=\sum_{i=1}^{N}\frac{{x_i}^2}{N}-2\sum_{i=1}^{N}\frac{x_i\overline{x}}{N}+\sum_{i=1}^{N}\frac{\overline{x}^2}{N}$$

$$=\sum_{i=1}^{N}\frac{{x_i}^2}{N}-2\overline{x}\sum_{i=1}^{N}\frac{x_i}{N}+N\frac{\overline{x}^2}{N}=\sum_{i=1}^{N}\frac{{x_i}^2}{N}-2\overline{x}(\overline{x})+(\overline{x}^2)$$

$$=\sum_{i=1}^{N}\frac{{x_i}^2}{N}-(\overline{x}^2)$$

- Example 1: Let X = [3, 4, 5, 6, 7]; $\bar{x} = 5$

$d_i = (x_i - \bar{x})$ = [-2, -1, 0, 1, 2] , deviations from the mean

$d_i^2 = (x_i - \bar{x})^2$ = [4, 1, 0, 1, 4] , squared deviations from the mean

$\sum (x_i - \bar{x})^2 = 10$, sum of squared deviations from the mean

$\sum \frac{(x_i - \bar{x})^2}{N} = \frac{10}{5} = 2$, average squared deviations from the mean (S^2)

S = $\sqrt{2}$ = 1.414, standard deviation

- Using the raw-score method $s^2 = \sum_{i=1}^{N} \frac{x_i^2}{N} - (\bar{x}^2)$,

x_i^2 = [9, 16, 25, 36, 49]

$\sum x_i^2$ = 135,

$S^2 = \frac{135}{5} - 25 = 27 - 25 = 2$

- Example 2: Let X = [11, 10, 9, 8, 6, -4, -5]; $\bar{x} = 5$

$d_i = (x_i - \bar{x})$ = [6, 5, 4, 3, 1, -9, -10] , deviations from the mean

$d_i^2 = (x_i - \bar{x})^2$ = [36, 25, 16, 9, 1, 81, 100] , squared deviations from the mean

$\sum (x_i - \bar{x})^2 = 268$, sum of squared deviations from the mean

$\sum \frac{(x_i - \bar{x})^2}{N} = \frac{268}{7} = 38.29$, average squared deviations from the mean (S^2)

S = $\sqrt{38.29}$ = 6.19, standard deviation

- Using the raw-score method $s^2 = \sum_{i=1}^{N} \frac{x_i^2}{N} - (\bar{x}^2)$,

x_i^2 = [121, 100, 81, 64, 36, 16, 25]

$$\sum x_i^2 = 443,$$

$$S^2 = \frac{443}{7} - 25 = 63.29 - 25 = 38.29$$

- The variance is calculated using deviations from the mean because the average squared deviation (variance) is always smaller than using any other value; that is, it minimizes the sum of squared deviations (*end of lecture)

- When the absolute difference is taken between a score and a measure of central tendency, the average absolute deviation is smallest when the median is used

• 1. 14 Standardized Scores

- A standardized score shows the relative status of a score in a distribution

- A standardized score, or ***z score***, expresses the deviation of a given score from the mean in standard deviation units as

$$z = \frac{x - \bar{x}}{S}$$

- When all the raw scores in a distribution are converted to *z* scores the new distribution has a mean of 0

$$\bar{z} = \sum_{i=1}^{N} \frac{z_i}{N} = \sum_{i=1}^{N} \frac{(x_i - \bar{x})}{NS}$$, because *N* and *S* are constant over the summation, then

$$\bar{z} = \frac{1}{NS} \sum_{i=1}^{N} (x_i - \bar{x}) = 0$$ (the sum of the deviations about the mean is always 0)

- When all the raw scores in a distribution are converted to *z* scores the new distribution has a standard deviation of 1.00

$$S_z{}^2 = \sum_{i=1}^{N} \frac{(x_i - \bar{x})^2}{NS^2} = \frac{S^2}{S^2} = 1$$, and S = 1

- Although creating standardized scores changes the $\bar{x}$ and *S*, it does not alter the form of the distribution and the frequencies of the *z* scores are the same as the frequencies of the corresponding raw scores

- Standardized scores have the same format in probability distributions than in frequency distributions

- the standardized value, relative to the probability distribution of *x* is

$$z = \frac{x - E(x)}{\sigma} = \frac{x - \mu}{\sigma}$$

- the mean of standardized scores for any probability distribution is always 0

$E(z) = 0$, as demonstrated by

$$E(z) = E\ \frac{X - E(X)}{\sigma} = \frac{E(X) - E(X)}{\sigma} = 0$$

- the standard deviation of standardized scores is always 1

$$\sigma^2 = E(z^2) - [E(z)]^2 = E\ \frac{[X - E(X)]^2}{\sigma 2}$$

$$= \frac{\sigma^2}{\sigma^2} = 1$$

(*)

$$f(d) = \sum_{i=1}^{N} (x_i - d)^2 \text{ ; where d} = \overline{x}$$

$$s^2 = \sum_{i=1}^{N} \frac{(x_i - \overline{x})^2}{N} = \sum_{i=1}^{N} \frac{(x_i - \overline{x})^2}{N-1}$$

$$f(t) = \sum_{i=1}^{N} (x_i - t)^2$$

$$f'(t) = \sum_{i=1}^{N} 2(x_i - t)(-1) = -2\sum_{i=1}^{N} (x_i - t)$$

$$f'(t) = 0 \rightarrow -2\sum_{i=1}^{N} (x_i - t) = 0$$

$$\sum x_i - nt = 0 \rightarrow tn \sum x_i \rightarrow \left[t = \frac{\sum x_i}{n} = \overline{x} \right]$$

It’s easy to derive finding the value that makes it zero.