Foundations of Educational Measurement

Chapter 6

Discussion Topics

- Educational measurement
- Descriptive statistics
  - Central tendency
  - Variation
  - Relationships
- Validity of measurement
- Reliability of measurement

Educational Measurement

- Measurement: assignment of numbers to differentiate values of a variable
- Evaluation: procedures for collecting information and using it to make decisions for which some value is placed on the results
- Assessment - multiple meanings
  - Measurement of a variable
  - Evaluation
  - Diagnosis of individual difficulties
  - Procedures to gather information on student performance

Educational Measurement

- Purpose of measurement for research
- Obtain information about the variables being studied
- Provide a standard format for recording observations, performances, or other responses of participants
- Provide for a quantitative summary of the results from many participants

Educational Measurement

- Four measurement scales
  - Nominal – categories
    - Race, gender, types of schools (e.g., public, private, parochial)
  - Ordinal - ordered categories
    - Finishing position in a race, grade levels
  - Interval - equal intervals between numbers on the scale
    - Test scores, achievement levels
  - Ratio - equal intervals and an absolute zero (0)
    - Height, weight, time

Descriptive Statistics

- Statistics: procedures that summarize and analyze quantitative data
  - Descriptive statistics: statistical procedures that summarize a set of numbers in terms of central tendency, variation, or relationships
  - Important for understanding what the data tells the researcher

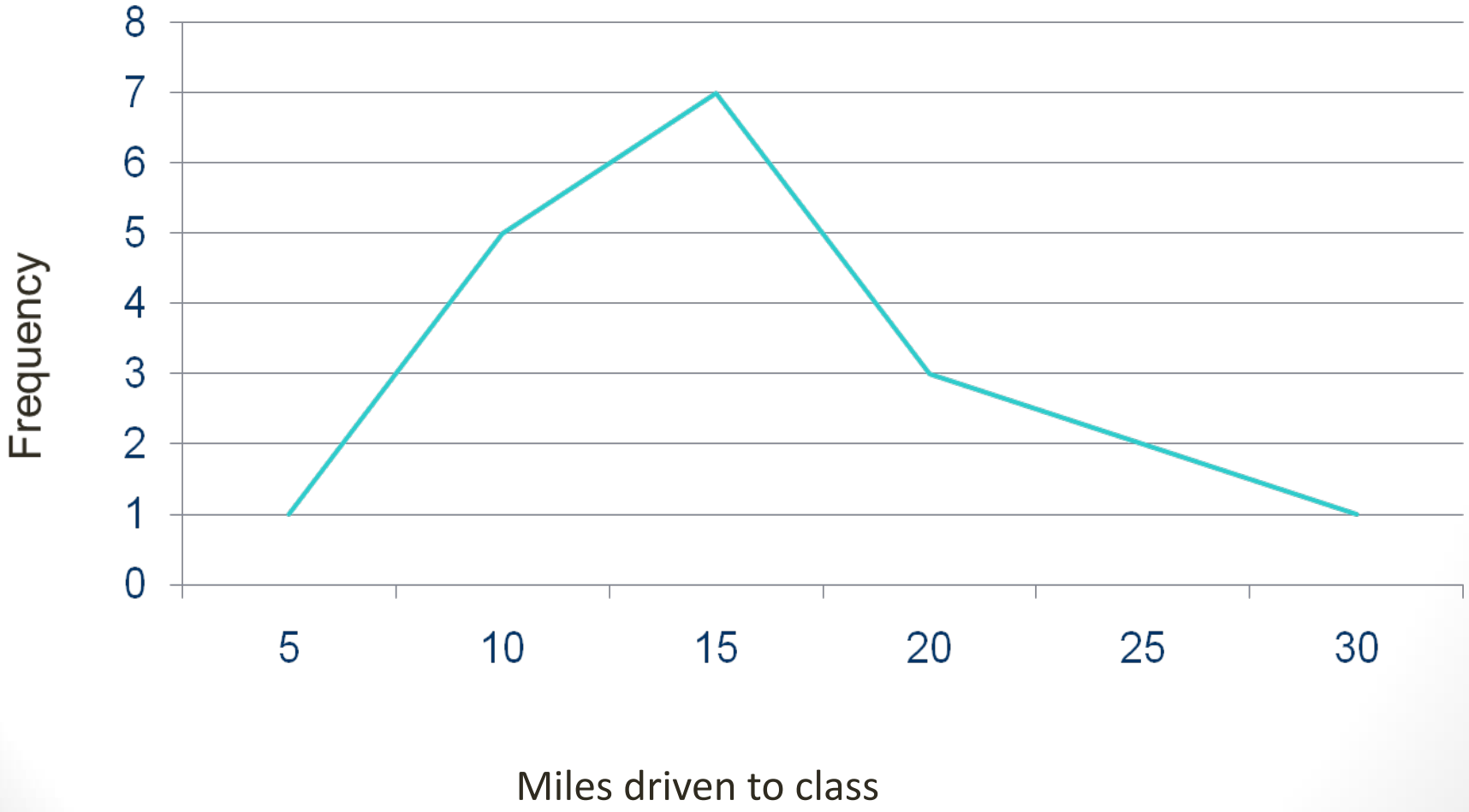
Descriptive Statistics

- Frequency distributions
  - An organization of the data set indicating the number of times (i.e., frequency) each score was present
  - Types
    - Frequency table
    - Frequency polygon
    - Histogram

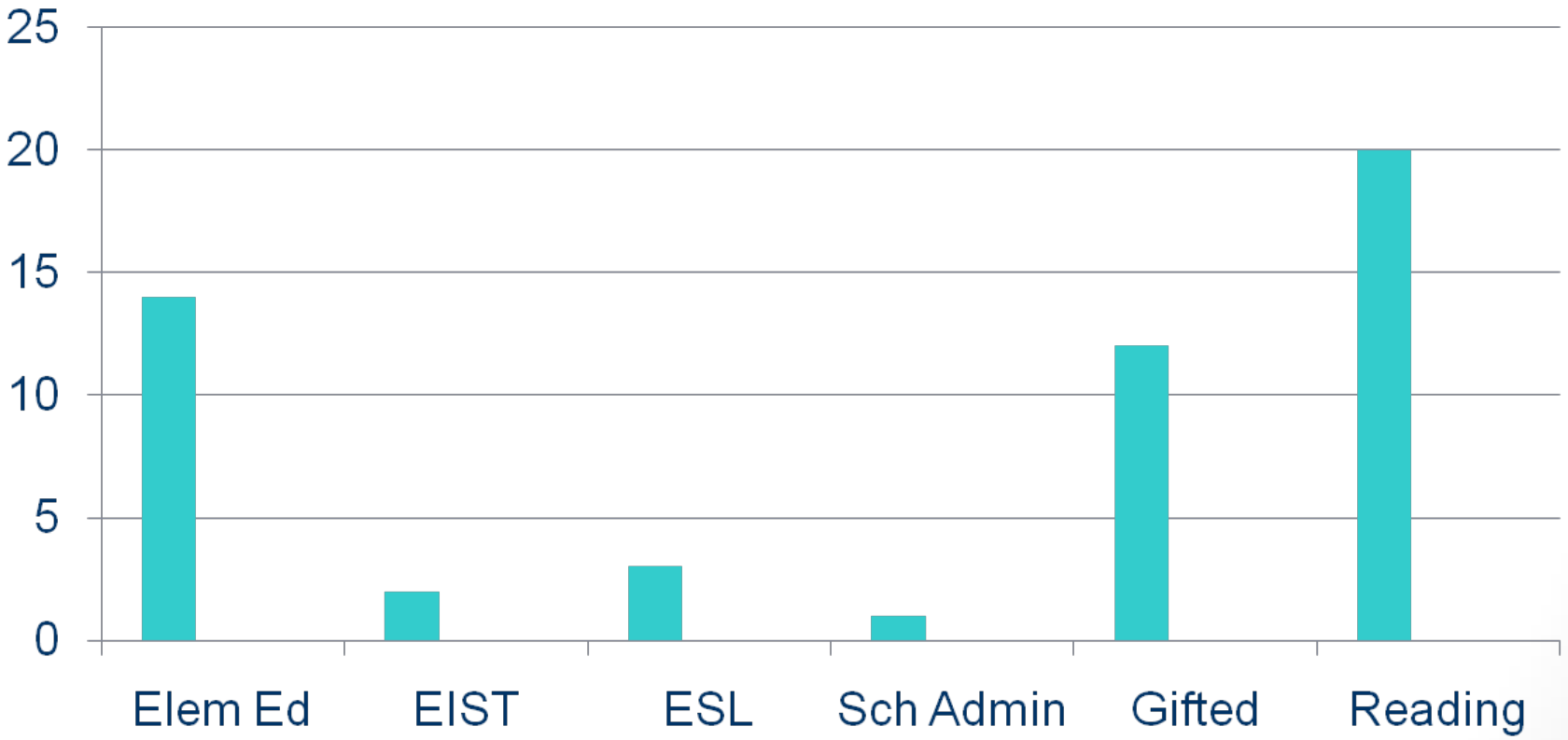
Frequency Table

Program	F	%
Elem Ed	14	26.9
EIST	2	3.8
ESL	3	5.8
Sch Admin	1	1.9
Gifted	12	23.1
Reading	20	38.5
Total	52	100.0

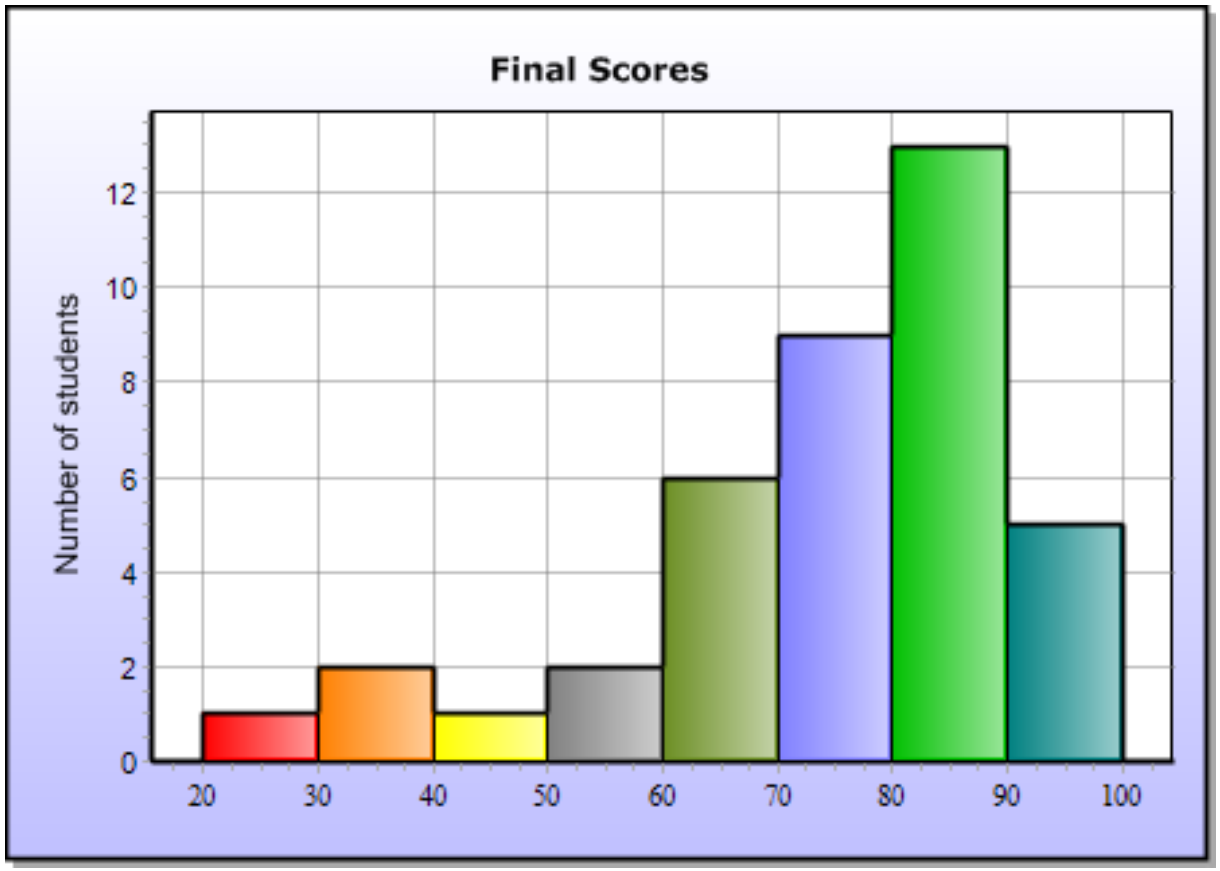
Polygon



Bar Chart Categorical Variable



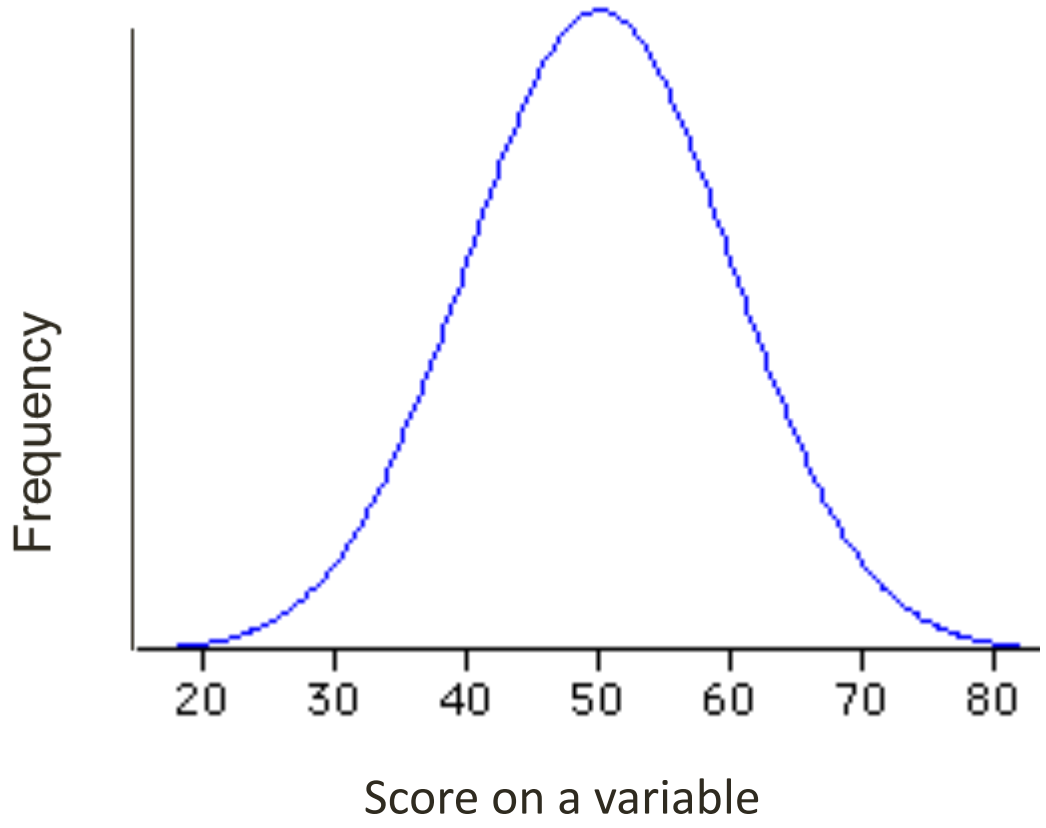
Histogram Continuous variable



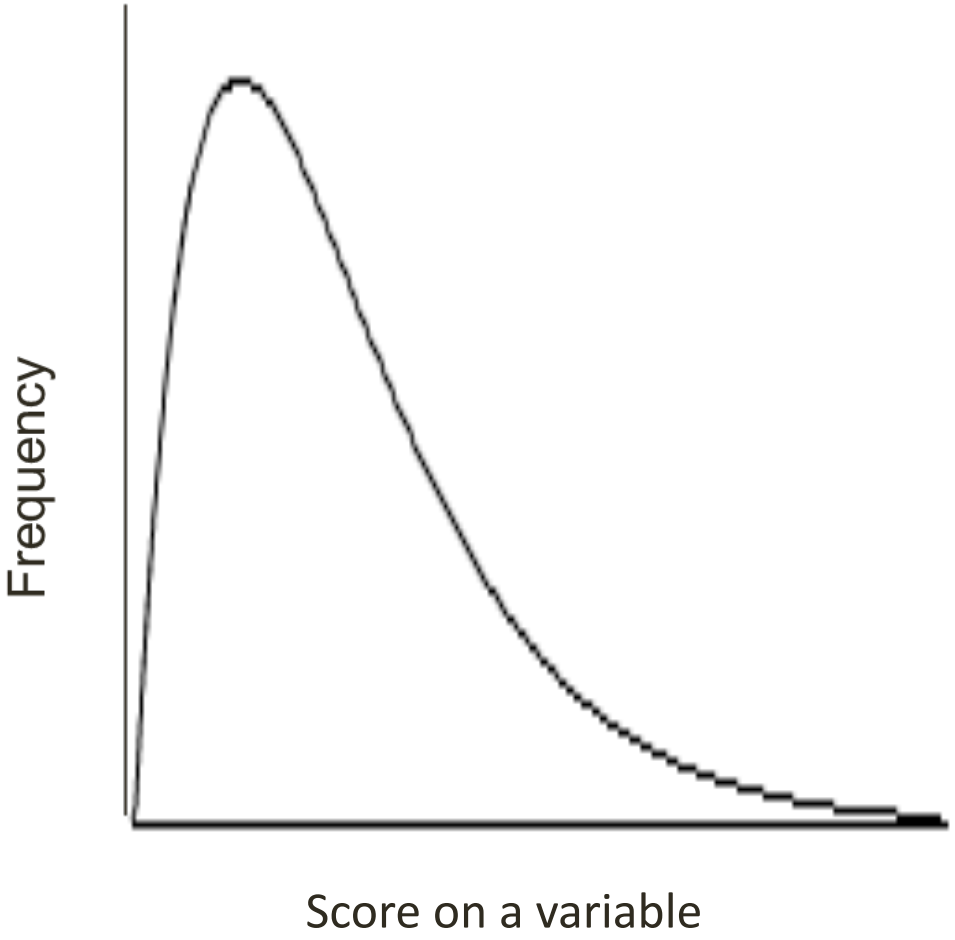
Descriptive Statistics

- Frequency distributions
  - Shapes (see Figure 6.3)
    - Normal - scores are equally distributed around the middle
    - Positively skewed - the set of scores is characterized by a large number of low scores and a small number of high scores
    - Negatively skewed - the set of scores is characterized by a large number of high scores and a small number of low scores
    - Outlier scores – scores that distort findings because they are so different from the other scores in the sample

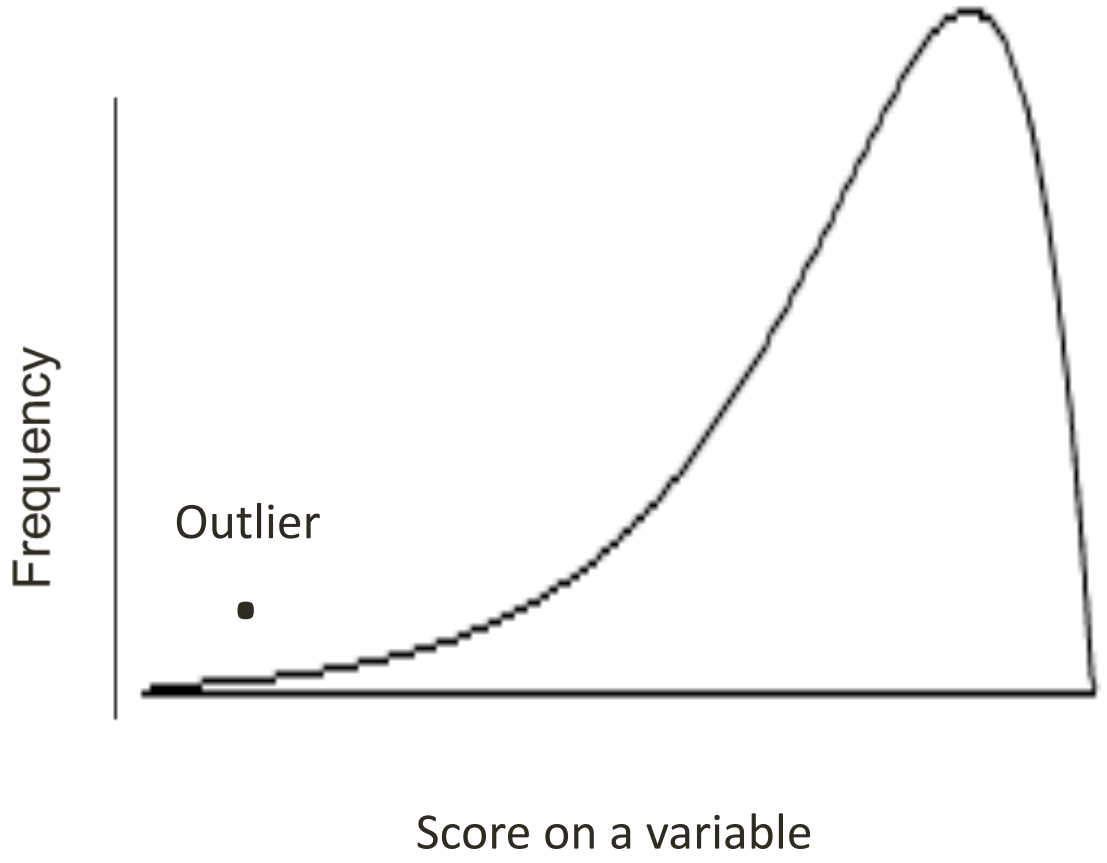
A normal distribution



Positive skew



Negative skew



Descriptive Statistics

- Central tendency
  - What is the typical score?
  - Three measures
    - Mode: the most frequently occurring score
    - Median: the score above and below which one-half of the scores occur
    - Mean
      - The arithmetic average of all scores
      - Statistical properties make it very useful
      - Concerns related to outlying scores

Calculating Central Tendency

25
24
23
23
23
22
21
21
20
19

Mean 25+24+23+23+23+22+21+ 21+20+19 =221

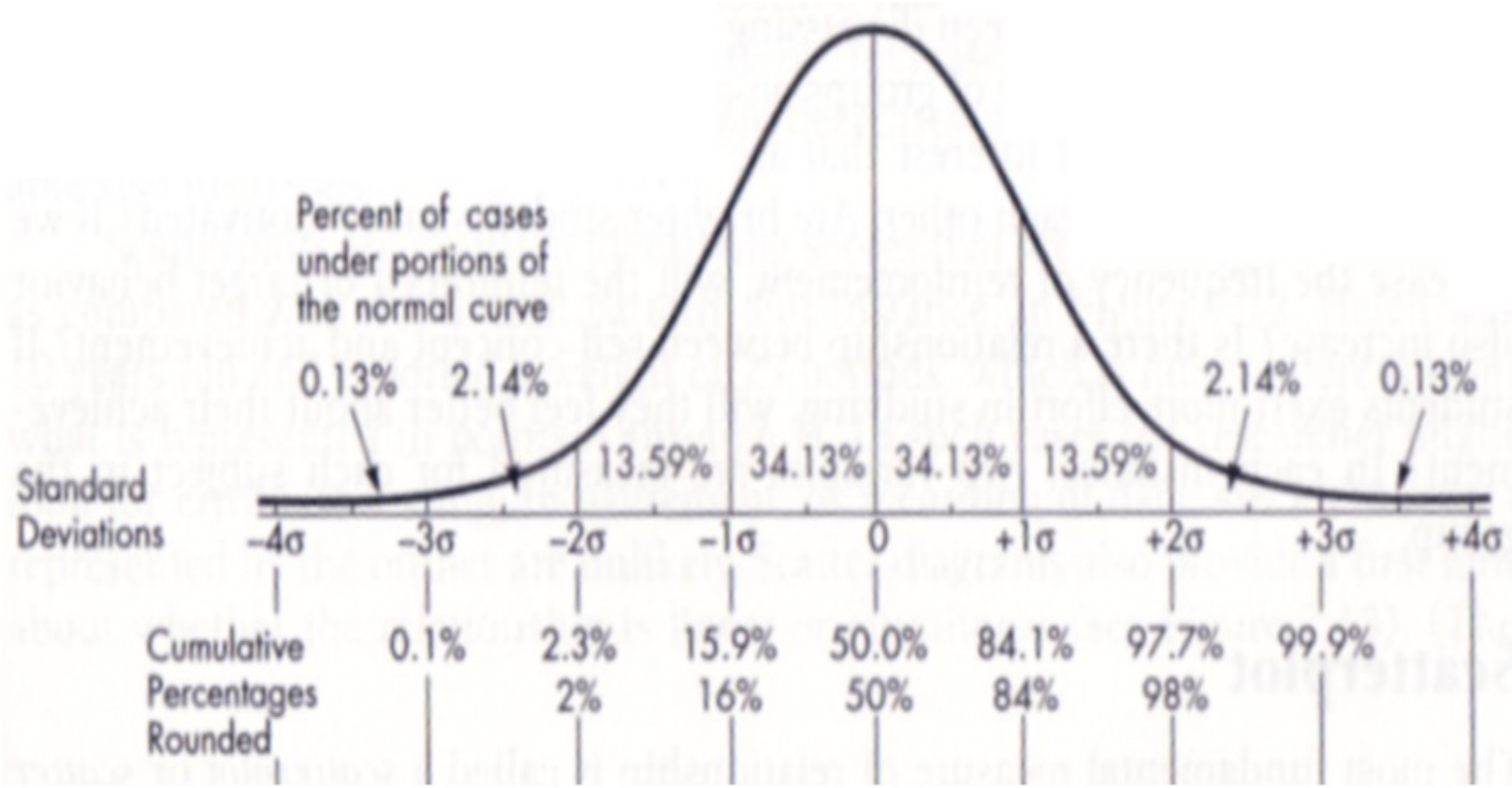
221/10=22.1

Median the middle point of the 10 scores =22.5

Mode most frequently occurring score is 23

Descriptive Statistics

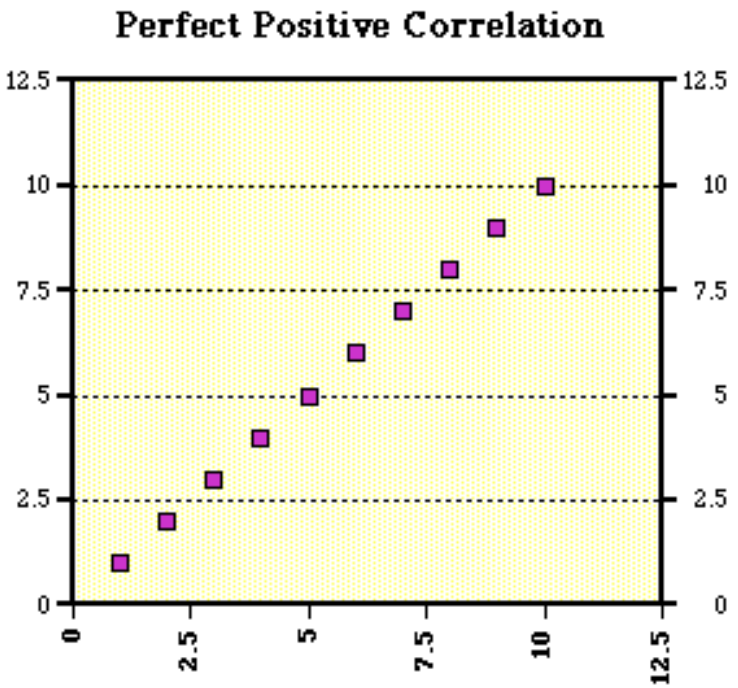
- Variability
  - How different are the scores?
  - Two types
    - Range: the difference between the highest and lowest scores
    - Standard deviation
      - The average distance of the scores from the mean
      - The relationship to the normal distribution
        - ±1 SD = 68% of all scores in a distribution
        - ±2 SD = 97% of all scores in a distribution
        - Use of percentile ranks - the percentage of scores at or below a specified score

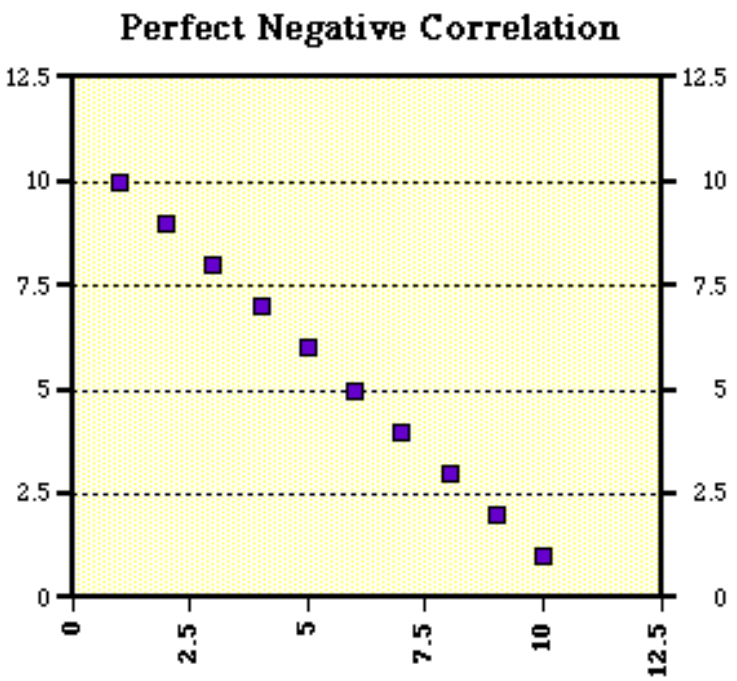


Descriptive Statistics

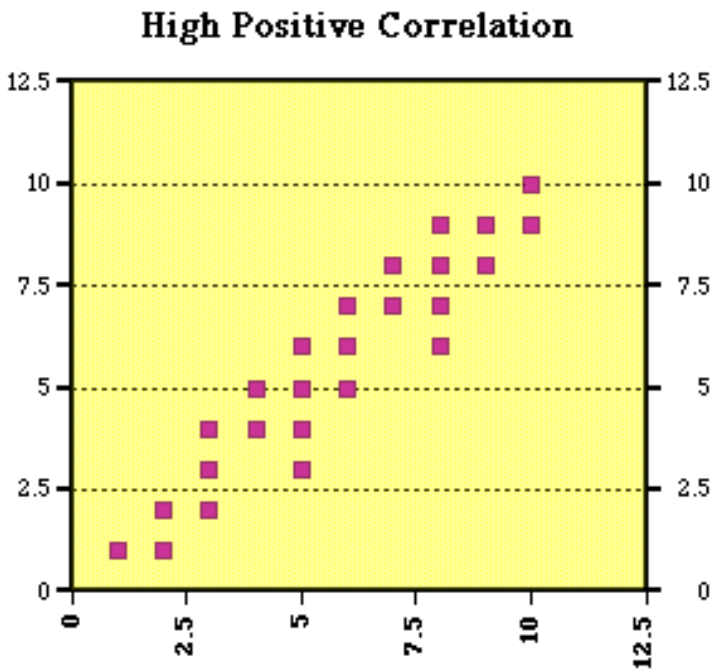
- Relationship
  - How do two sets of scores relate to one another?
  - Correlation
    - A measure of the relationship between two variables
      - Strength - 0.00 to 1.00
      - Direction - positive (+) or negative (-)
    - Scatterplots – graphic depictions of correlations
      - Scores of each of the two variables for each individual

Perfect correlations

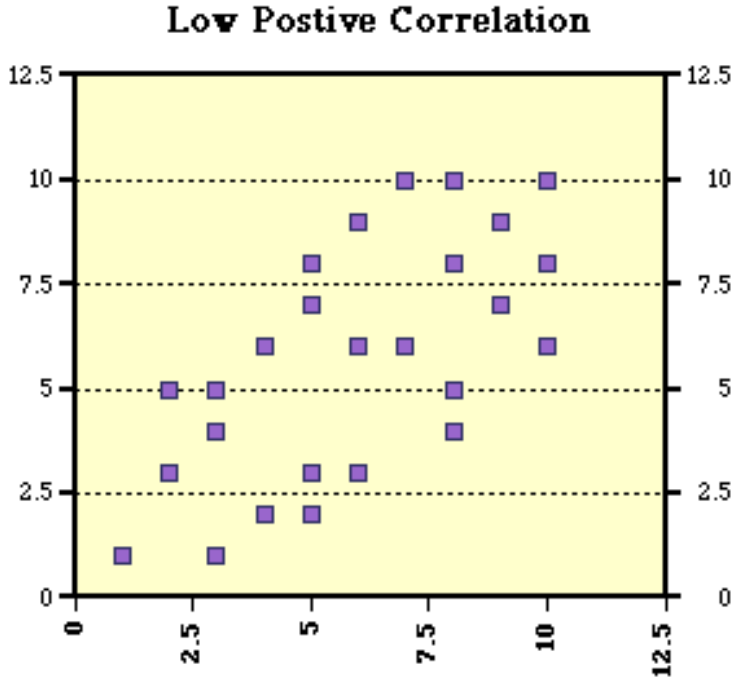




Positive correlations

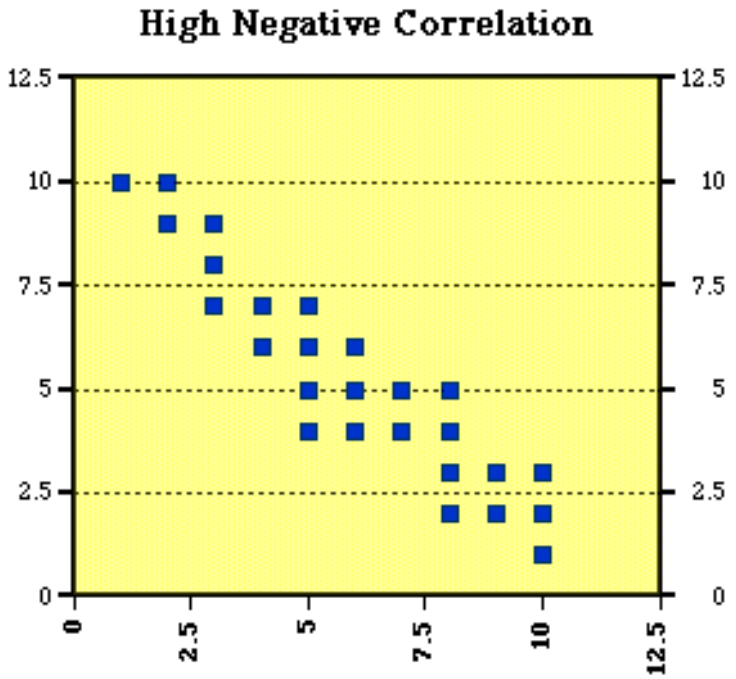


r = .88

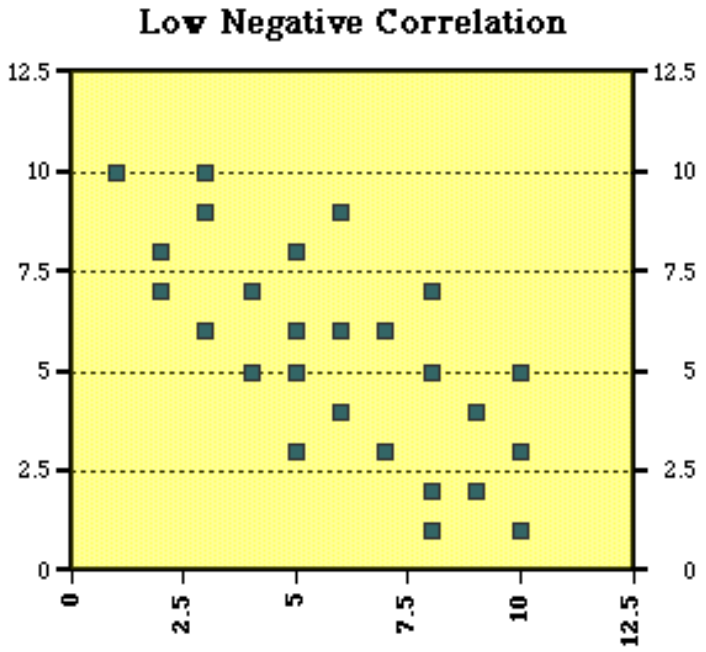


r = .35

Negative correlations



r = -.88



r = -.35

Interpreting Descriptive Statistics

Table 1
Scale Scores[1] for the Experimental and Control Groups

Scale	Experimental Group			Control Group		
	N	Mean	SD	N	Mean	SD
DSP	15	117.30	43.00	15	132.87	25.26
OSI	15	348.20	29.37	15	362.07	36.31
TSM	15	208.70	40.80	15	224.90	43.60
SCSI	15	40.00	10.38	15	48.47	9.56

[1] Scale scores reflect levels of stress, so lower scores are more desirable.

1. Which group had the higher scores on the DSP?
2. Which group had the better scores on the DSP?
3. Which group had more variation in their OSI scores?
4. Which group had less variation in their TSM scores?
5. Which group had more students on the SCSI?

Answers

- Higher DSP scores
  - Control group has a higher mean
- Better DSP scores
  - Experimental group had a lower mean (see note on table)
- More variation on OSI
  - Control group had a higher standard deviation
- Less variation on TSM
  - Experimental group had a lower standard deviation
- More students on SCSI
  - Both groups had the same number of participants

Correlation Table

	Neurotic	Bulimic	Perfectionist	Self-esteem
Neurotic	1.0			
Bulimic	.58	1.0		
Perfectionist	-.92	.87	1.0	
Self-esteem	-.33	.02	-.21	1.0

1. Which pair of variables has the highest, positive correlation?
2. Which pair of variables has the lowest, positive correlation?
3. Which pair of variables has the highest, negative correlation?
4. Which pair of variables has the lowest, negative correlation?

Answers

- Which pair of variables has the highest, positive correlation? Bulimic & Perfectionist
- Which pair of variables has the lowest, positive correlation? Bulimic & Self-Esteem
- Which pair of variables has the lowest, negative correlation? Perfectionist & Neurotic
- Which pair of variables has the lowest, negative correlation? Self-esteem & Perfectionist

Validity of Measurement

- Validity: the extent to which inferences are appropriate, meaningful, and useful
  - Refers to the interpretation of the results
  - A matter of degree
  - Specific to a particular use or interpretation
  - A unitary concept
  - Involves an overall evaluative judgment

Validity of Measurement

- Three sources of validity evidence
- Test content - evidence of the extent to which items on a test are representative of the larger domain of content or items from which they are drawn
- Internal structure - evidence of the extent to which the relationships between items and parts of the instrument are consistent with those reflected in the theoretical basis of the instrument or its intended use

Validity of Measurement

- Three sources of validity evidence
  - Relationships with other variables - evidence of the extent to which scores from an instrument are related to similar as well as different traits
    - Convergent evidence - scores correlate with measures of the same thing being measured at the same time
    - Discriminate evidence - scores do not correlate with measures of something different than that being measured at the same time
    - Predictability - the extent to which test scores predict future performance on a criterion variable

Validity of Measurement

- Importance of validity to research
- If the research results are to have any value, validity of the measurement of a variable must exist
- Use of established and “new” instruments and the implications for establishing validity
- Importance of establishing validity prior to data collection (e.g., pilot tests)

Validity of Measurement

- Importance of validity to research
- Validity as a matter of degree (i.e., the extent to which...)
- Judged on the basis of available evidence
- Varying levels of validity evidence are reported in articles

Reliability of Measurement

- Reliability
  - The extent to which scores are free from error
  - Error is measured by consistency
  - Sources of error
    - Test construction and administration
      - Ambiguous questions, confusing directions, changes in scoring, interrupted testing, etc.
    - Participants' characteristics
      - Test anxiety, lack of motivation, fatigue, guessing, etc.

Reliability of Measurement

- Reliability
  - Measurement
    - Reliability coefficients range from 0.00 to 1.00 regardless of the formula used to calculate them
    - 0.00 indicates no reliability or consistency
    - 1.00 indicates total reliability or consistency

Reliability of Measurement

- Five types of reliability evidence
  - Stability (i.e., test-retest)
    - Testing the same subject using the same test on two occasions
    - Limitation - carryover effects from the first to second administration of the test
  - Equivalence (i.e., parallel form)
    - Testing the same subject with two parallel (i.e., equal) forms of the same test taken at the same time
    - Limitation - difficulty in creating parallel forms

Reliability of Measurement

- Equivalence and stability
- Testing the same participants with two forms of the same test taken at different times
- Limitation - difficulty in creating parallel forms
- Internal consistency
- Testing the same subject with one test and “artificially” splitting the test into two halves
- Limitations - must have a minimum of ten (10) questions

Reliability of Measurement

- Internal consistency (continued)
  - Two forms
    - KR 20
      - Dichotomously scored (i.e., right or wrong) items
      - Typical of cognitive measures
    - Cronbach alpha
      - Non-dichotomously scored (e.g., strongly agree, agree, disagree, strongly disagree) items
      - Typical of non-cognitive measures

Reliability of Measurement

- Agreement
  - Used when traditional estimates such as stability, equivalence, equivalence and stability, or internal consistency are not applicable
  - Typically some form of agreement is used (e.g., raters agreeing with one another)

Reliability of Measurement

- Agreement (continued)
  - Situations in which this estimate is used
    - Observational measures - agreement between raters making the same observation
    - Insufficient numbers of test items on an instrument - agreement across the percentage of responses that are the same for several participants
    - Data with highly skewed distributions - percentage of agreement in the number of participants

Reliability of Measurement

- Importance of reliability
- If the results are to have any value, reliability of the measurement of a variable must exist
- Established prior to conducting the research (e.g., pilot study)
- Necessary but not sufficient condition for validity (i.e., to be valid, an instrument must be reliable, but a reliable instrument is not necessarily valid)

Reliability of Measurement

- Conditions affecting reliability
  - Length of the test (i.e., longer tests are typically more reliable)
  - Participants
    - Greater reliability with heterogeneous samples
    - Scores for older participants are typically more reliable than those for younger children
  - Trait being measured (i.e., cognitive traits are more reliable than affective characteristics)

Reliability of Measurement

- Enhancing reliability
  - Standardized administration procedures (e.g., directions, conditions, etc.)
  - Appropriate reading level
  - Reasonable length of the testing period
  - Counterbalancing the order of testing if several tests are being given