# Exploring your Data before Analysis: What Practitioners Should Know

The analysis before your analysis!

Dr. Kimberly Kostelis, Central Connecticut State University
Dr. Tracey Matthews, Springfield College
AAHPERD – Boston 2012

1

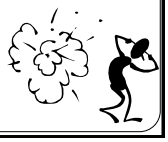
## Presentation Outline

- Data Exploration
- Basic Assumptions of Commonly Used Statistics
- Formulating Decisions After Exploration and Review of Assumptions

2

## Data Exploration – What are you looking for?

- What are major types or sources of error in data?
- These may be different depending on the type of data – self report, observation of behavior, physiological data, archival data, and so forth.



3

## Sources of Error – Self-Report

- Attitudes/Personality Inventories
  - Social desirability bias
  - Misunderstand the question
  - Not remembering events accurately
  - Accidentally skipping questions
- Self-report measures vs direct measures
  - Height and weight

4

## Sources of Error – Observer Ratings

- Reactivity
  - The presence of an observer may actually change the behavior that is being observed
  - Inter-observer reliability – training of observers

5

## Sources of Error - Physiologically

- Calibration of equipment
- Intra-rater reliability
- Inter-rater reliability
- Artifacts – interference of the signal

6

## Data Exploration

- Errors:
  - What types of errors can only be detected during data collection?
    - What quality control measures help to prevent errors during data collection?
  - What types of errors (e.g. unbelievable scores, inconsistent responses) can be detected and removed by examining the data set?

7

## Data Exploration can help to…

- Describe your sample and "get acquainted with your data"
- Identify extreme or impossible scores, response inconsistencies, etc
- Identify possible violations of assumptions

8

## Data Screening

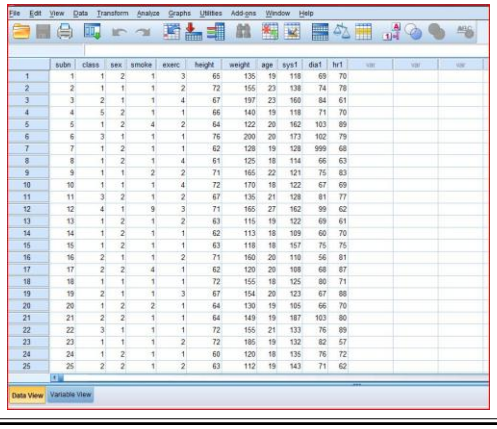
- Purposes:
  - Accuracy of data collected
    - Garbage in, Garbage out
  - Assess effect of and ways to deal with incomplete data
    - Equipment failure, not responding to items, not completing trials
  - Outliers or extreme values
  - Adequacy of fit between the data and assumptions of the specific procedures

9

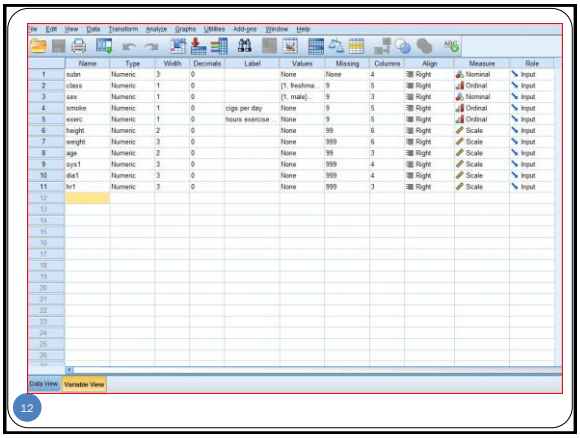
## Example of Data in SPSS

- Data set to show data screening procedures
- Effects of social stress on blood pressure (Mooney, 1990)

10



11



12

Dr. Tracey Matthews (tmatthews@springfieldcollege.edu)
Dr. Kimberly Kostelis (kostelisk@ccsu.edu)

## Missing Data

- What situations lead to "missing data"?
- How is missing data represented in an SPSS data file?
- May not be random
- Need to examine data sets to see if patterns exist with missing data
- Using Explore and Descriptive Statistics options in SPSS can assist in reviewing the data
- Code missing data – examine frequencies and cases

13

## SPSS Output

Statistics

| | | class | sex | cigs per day | hours exercise per week | height | weight | age | sys1 | dia1 | hr1 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| N | Valid | 65 | 65 | 64 | 65 | 65 | 65 | 64 | 64 | 64 | 65 |
| | Missing | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 |
| Mean | | 1.89 | 1.62 | 1.23 | 1.83 | 66.91 | 145.74 | 19.50 | 125.25 | 74.14 | 74.45 |
| Median | | 2.00 | 2.00 | 1.00 | 2.00 | 67.00 | 140.00 | 19.00 | 121.50 | 74.00 | 75.00 |
| Mode | | 1 | 2 | 1 | 1* | 62 | 115* | 19 | 114 | 69* | 70 |
| Std. Deviation | | 1.091 | .550 | .684 | .894 | 4.072 | 27.428 | 1.285 | 19.869 | 11.717 | 8.639 |
| Skewness | | 1.112 | .103 | 3.349 | 1.022 | .397 | .609 | .834 | .945 | .899 | -.022 |
| Std. Error of Skewness | | .297 | .297 | .299 | .297 | .297 | .297 | .299 | .299 | .299 | .297 |
| Kurtosis | | .510 | -.899 | 11.019 | .478 | -.533 | -.060 | .286 | .887 | .590 | -.499 |
| Std. Error of Kurtosis | | .586 | .586 | .590 | .586 | .586 | .586 | .590 | .590 | .590 | .586 |
| Range | | 4 | 2 | 3 | 3 | 18 | 127 | 5 | 94 | 52 | 37 |
| Minimum | | 1 | 1 | 1 | 1 | 60 | 103 | 18 | 93 | 51 | 56 |
| Maximum | | 5 | 3 | 4 | 4 | 78 | 230 | 23 | 187 | 103 | 93 |

a. Multiple modes exist. The smallest value is shown

14

## Handling Missing data

- Deleting cases that have caused problems
  - Not a bad alternative if only a few cases have missing values
- Missing values may be concentrated to a few variables
  - Entire variable may be dropped depending on its importance
- Estimate missing values
  - Prior knowledge for a replacement value
  - Calculation of the means using available data for the variables with missing values
  - Regression approach – several IVs are used to develop an equation that can be used to predict the value of the DV

15 . . . Regardless –repeat analysis with missing cases

## Listwise versus Pairwise Deletion

- With SPSS listwise or pairwise deletion is available to remove missing data points
  - **Listwise –** data for the participant is ignored for all calculations – can result in a smaller N; but calculations are made on the same set of participants
  - **Pairwise –** analyses will be made using data from all participants who had non-missing values for the particular pair of variables – preserves the maximum possible N for the computation, the N will vary across computations

16

## Impossible Scores

- You should also review your data for impossible scores
  - Is the score within the range of the data?
    - Examples:
      - Age of 18-30 year olds and there is a "62" for a case
      - Were the numbers reversed? Need to follow up with original data
  - Frequencies procedure can help you to determine impossible scores

17

## SPSS Output

Statistics

| | | class | sex | cigs per day | hours exercise per week | height | weight | age | sys1 | dia1 | hr1 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| N | Valid | 65 | 65 | 64 | 65 | 65 | 65 | 64 | 64 | 64 | 65 |
| | Missing | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 |
| Mean | | 1.89 | 1.62 | 1.23 | 1.83 | 66.91 | 145.74 | 19.50 | 125.25 | 74.14 | 74.45 |
| Median | | 2.00 | 2.00 | 1.00 | 2.00 | 67.00 | 140.00 | 19.00 | 121.50 | 74.00 | 75.00 |
| Mode | | 1 | 2 | 1 | 1* | 62 | 115* | 19 | 114 | 69* | 70 |
| Std. Deviation | | 1.091 | .550 | .684 | .894 | 4.072 | 27.428 | 1.285 | 19.869 | 11.717 | 8.639 |
| Skewness | | 1.112 | .103 | 3.349 | 1.022 | .397 | .609 | .834 | .945 | .899 | -.022 |
| Std. Error of Skewness | | .297 | .297 | .299 | .297 | .297 | .297 | .299 | .299 | .299 | .297 |
| Kurtosis | | .510 | -.899 | 11.019 | .478 | -.533 | -.060 | .286 | .887 | .590 | -.499 |
| Std. Error of Kurtosis | | .586 | .586 | .590 | .586 | .586 | .586 | .590 | .590 | .590 | .586 |
| Range | | 4 | 2 | 3 | 3 | 18 | 127 | 5 | 94 | 52 | 37 |
| Minimum | | 1 | 1 | 1 | 1 | 60 | 103 | 18 | 93 | 51 | 56 |
| Maximum | | 5 | 3 | 4 | 4 | 78 | 230 | 23 | 187 | 103 | 93 |

a. Multiple modes exist. The smallest value is shown

18 Impossible Scores!!!

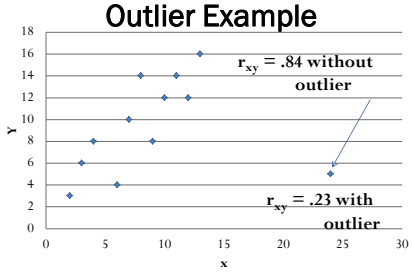
Dr. Tracey Matthews (tmatthews@springfieldcollege.edu)
Dr. Kimberly Kostelis (kostelisk@ccsu.edu)

## Outliers

- Unusual or extreme scores
  - Data entry errors
  - Person is not a member of the population for which sample is intended
  - Person is simply different from remainder of the sample

19

## Outlier Example




| x | y |
|---|---|
| 2 | 3 |
| 3 | 6 |
| 4 | 8 |
| 6 | 4 |
| 7 | 10 |
| 8 | 14 |
| 9 | 8 |
| 10 | 12 |
| 11 | 14 |
| 12 | 12 |
| 13 | 16 |
| 24 | 5 |

20

| Group 1 | Group 2 | Group 3 |
|---|---|---|
| 15 | 17 | 6 |
| 18 | 22 | 9 |
| 12 | 15 | 12 |
| 12 | 12 | 11 |
| 9 | 20 | 11 |
| 10 | 14 | 8 |
| 12 | 15 | 13 |
| 20 | 20 | 30 |
| | 21 | 7 |

**Descriptives with Outlier**

score

| | N | Mean | Std. Deviation | Std. Error | 95% Confidence Interval for Mean | | Minimum | Maximum |
|---|---|---|---|---|---|---|---|---|
| | | | | | Lower Bound | Upper Bound | | |
| 1 | 8 | 13.5000 | 3.85450 | 1.36277 | 10.2776 | 16.7224 | 9.00 | 20.00 |
| 2 | 9 | 17.3333 | 3.53553 | 1.17851 | 14.6157 | 20.0510 | 12.00 | 22.00 |
| 3 | 9 | 11.8889 | 7.10409 | 2.36803 | 6.3867 | 17.4111 | 6.00 | 30.00 |
| Total | 26 | 14.2692 | 5.50315 | 1.07928 | 12.0465 | 16.4920 | 6.00 | 30.00 |

**Descriptives without Outlier**

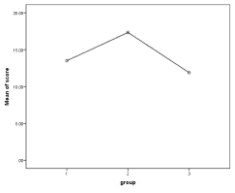
score2

| | N | Mean | Std. Deviation | Std. Error | 95% Confidence Interval for Mean | | Minimum | Maximum |
|---|---|---|---|---|---|---|---|---|
| | | | | | Lower Bound | Upper Bound | | |
| 1.00 | 8 | 13.5000 | 3.85450 | 1.36277 | 10.2776 | 16.7224 | 9.00 | 20.00 |
| 2.00 | 9 | 17.3333 | 3.53553 | 1.17851 | 14.6157 | 20.0510 | 12.00 | 22.00 |
| 3.00 | 8 | 9.6250 | 2.50357 | .88515 | 7.5320 | 11.7180 | 6.00 | 13.00 |
| Total | 25 | 13.6400 | 4.56326 | .91265 | 11.7564 | 15.5236 | 6.00 | 22.00 |

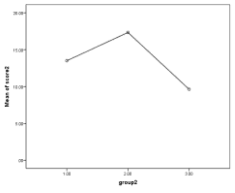
21

## Impact of Outliers

With outlier – no differences



Without outlier - differences



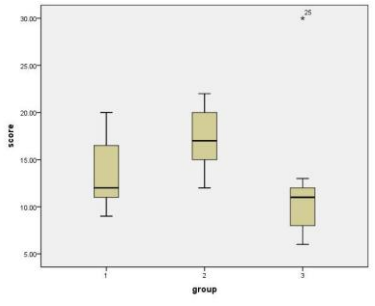
22

## Decisions about Outliers

- Standardize values to determine outliers
  - ± 3.0 (Approximately 99% of scores should lie within 3 standard deviation units)
- If outliers are due to data entry or instrumentation error, drop and redo analysis
- If outlier is not due to these types of errors, outlier should not be dropped
- Suggestion – run analyses with and without outliers
- Outliers – not necessarily bad – but may provide reasons for further study

23

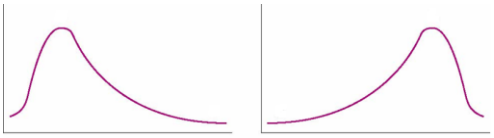
## SPSS Boxplot



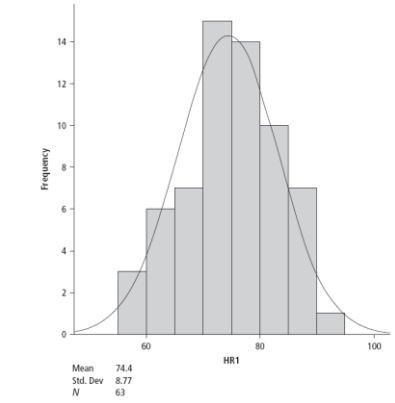

24

## Distribution of Scores: Skewness

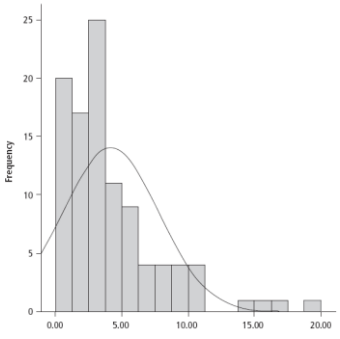
- Positive skewness:
  - Distribution of scores at the lower end of the scale and few scores at the upper end
- Negative skewness:
  - Distribution of scores at the higher end of the scale and few scores at the lower end



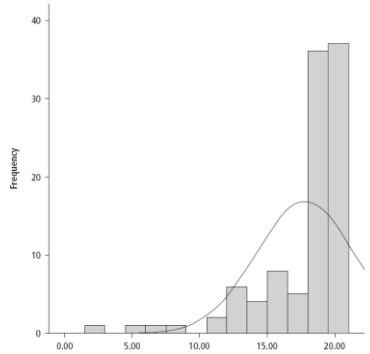
25




26




27



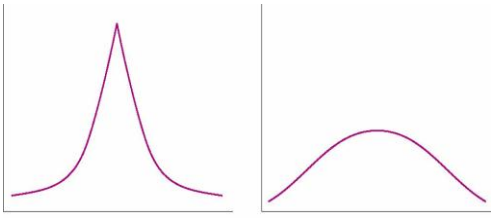

28

## Distribution of Scores: Kurtosis

- Level of peaked ness
- Mesokurtic – normal, bell-shaped
- Leptokurtic = peaked
  - Homogeneous scores within distribution
- Platykurtic = flat
  - Heterogeneous scores within distribution

29

## Kurtosis

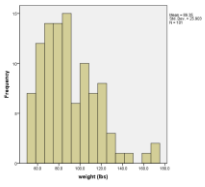


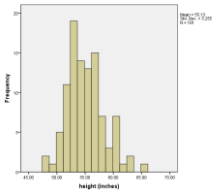
Leptokurtic Platykurtic

Dr. Tracey Matthews (tmatthews@springfieldcollege.edu)
Dr. Kimberly Kostelis (kostelisk@ccsu.edu)

**Statistics**

| | | weight (lbs) | height (inches) |
|---|---|---|---|
| N | Valid | 101 | 101 |
| | Missing | 0 | 0 |
| Mean | | 89.351 | 55.1287 |
| Std. Deviation | | 25.9032 | 3.25473 |
| Skewness | | .929 | .618 |
| Std. Error of Skewness | | .240 | .240 |
| Kurtosis | | .721 | .653 |
| Std. Error of Kurtosis | | .476 | .476 |
| Minimum | | 53.0 | 48.00 |
| Maximum | | 172.0 | 65.75 |



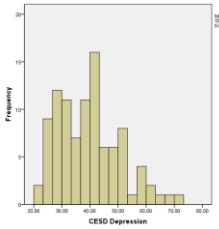




31

## Another Example – Depression - Total

**Descriptive Statistics**

| | N | Range | Minimum | Maximum | Mean | Std. Deviation | Skewness | | Kurtosis | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Statistic | Statistic | Statistic | Statistic | Statistic | Statistic | Statistic | Std. Error | Statistic | Std. Error |
| CESD Depression | 98 | 52.00 | 21.00 | 73.00 | 39.4517 | 11.17662 | .675 | .244 | .086 | .483 |
| Valid N (listwise) | 98 | | | | | | | | | |




32

## Comparing Groups

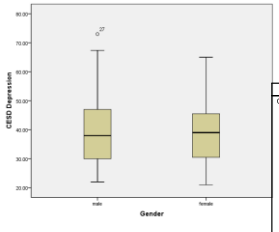
**Case Processing Summary**

| Gender | | Cases | | | | | |
|---|---|---|---|---|---|---|---|
| | | Valid | | Missing | | Total | |
| | | N | Percent | N | Percent | N | Percent |
| CESD Depression | male | 22 | 100.0% | 0 | .0% | 22 | 100.0% |
| | female | 76 | 100.0% | 0 | .0% | 76 | 100.0% |

**Descriptives**

| Gender | | | | Statistic | Std. Error |
|---|---|---|---|---|---|
| CESD Depression | male | Mean | | 39.9187 | 2.84706 |
| | | 95% Confidence Interval for Mean | Lower Bound | 33.9979 | |
| | | | Upper Bound | 45.8394 | |
| | | 5% Trimmed Mean | | 39.0896 | |
| | | Median | | 38.0000 | |
| | | Variance | | 178.327 | |
| | | Std. Deviation | | 13.35390 | |
| | | Minimum | | 22.00 | |
| | | Maximum | | 73.00 | |
| | | Range | | 51.00 | |
| | | Interquartile Range | | 17.75 | |
| | | Skewness | | 1.054 | .491 |
| | | Kurtosis | | .811 | .953 |
| | female | Mean | | 39.3165 | 1.21149 |
| | | 95% Confidence Interval for Mean | Lower Bound | 36.9031 | |
| | | | Upper Bound | 41.7299 | |
| | | 5% Trimmed Mean | | 38.9131 | |
| | | Median | | 39.0000 | |
| | | Variance | | 111.545 | |
| | | Std. Deviation | | 10.56150 | |
| | | Minimum | | 21.00 | |
| | | Maximum | | 65.00 | |
| | | Range | | 44.00 | |
| | | Interquartile Range | | 15.50 | |
| | | Skewness | | .462 | .276 |
| | | Kurtosis | | -.454 | .545 |

33



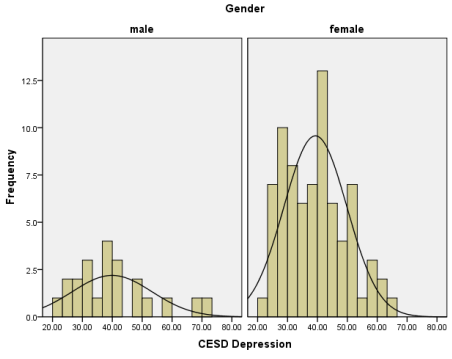

**Extreme Values**

| Gender | | | Case Number | Value |
|---|---|---|---|---|
| CESD Depression | male | Highest 1 | 27 | 73.00 |
| | | 2 | 30 | 67.37 |
| | | 3 | 6 | 58.00 |
| | | 4 | 3 | 52.00 |
| | | 5 | 98 | 48.00 |
| | | Lowest 1 | 75 | 22.00 |
| | | 2 | 21 | 25.00 |
| | | 3 | 58 | 26.00 |
| | | 4 | 87 | 28.00 |
| | | 5 | 78 | 28.00 |
| | female | Highest 1 | 38 | 65.00 |
| | | 2 | 22 | 63.00 |
| | | 3 | 68 | 63.00 |
| | | 4 | 50 | 58.00 |
| | | 5 | 77 | 58.00[a] |
| | | Lowest 1 | 83 | 21.00 |
| | | 2 | 97 | 24.00 |
| | | 3 | 31 | 24.00 |
| | | 4 | 81 | 25.00 |
| | | 5 | 79 | 25.00[b] |

a. Only a partial list of cases with the value 58.00 are shown in the table of upper extremes.

b. Only a partial list of cases with the value 25.00 are shown in the table of lower extremes.

34

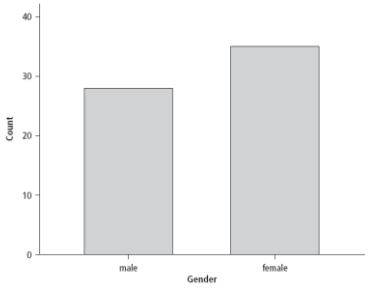



35

## Data Screening for Two Categorical Variables

When you have two categorical variables, what graphs or statistics do you need to obtain to characterize your sample and help you plan appropriate analyses?

36




37

GENDER * SMOKE Crosstabulation

| | | | SMOKE | | | |
|---|---|---|---|---|---|---|
| | | | non smoker | light smoker | heavy smoker | Total |
| GENDER | male | Count | 22 | 4 | 1 | 27 |
| | | Expected Count | 21.8 | 3.9 | 1.3 | 27.0 |
| | | % within GENDER | 81.5% | 14.8% | 3.7% | 100.0% |
| | female | Count | 28 | 5 | 2 | 35 |
| | | Expected Count | 28.2 | 5.1 | 1.7 | 35.0 |
| | | % within GENDER | 80.0% | 14.3% | 5.7% | 100.0% |
| Total | | Count | 50 | 9 | 3 | 62 |
| | | Expected Count | 50.0 | 9.0 | 3.0 | 62.0 |
| | | % within GENDER | 80.6% | 14.5% | 4.8% | 100.0% |

Figure 4.32 ♦ Cross-Tabulation of Gender by Smoking Status

NOTE: Expected cell frequencies less than 10 in three cells.

38

GENDER * smoke2 Crosstabulation

| | | | smoke2 | | |
|---|---|---|---|---|---|
| | | | 1 | 2 | Total |
| GENDER | male | Count | 22 | 5 | 27 |
| | | Expected Count | 21.8 | 5.2 | 27.0 |
| | | % within GENDER | 81.5% | 18.5% | 100.0% |
| | female | Count | 28 | 7 | 35 |
| | | Expected Count | 28.2 | 6.8 | 35.0 |
| | | % within GENDER | 80.0% | 20.0% | 100.0% |
| Total | | Count | 50 | 12 | 62 |
| | | Expected Count | 50.0 | 12.0 | 62.0 |
| | | % within GENDER | 80.6% | 19.4% | 100.0% |

Figure 4.36 ♦ Crosstabs Using the Recoded Smoking Variable (SMOKE2)

39

## Data Transformations

- Mathematical procedures that can be used to modify variables that violate the statistical assumptions of normality, linearity, and homoscedasticity
- What extent of the basic assumption has been violated?
  - Robustness: relative insensitivity of a statistical test to violations of the underlying inferential statistics
- Use data transformations through the compute procedure in SPSS

40

## Data Transformations


- Caution
  - While data transformations can significantly improve the precision of a multivariate analysis
  - Transformation can also provide data interpretation problems
  - Difficulty interpreting – transformed data vs raw data
  - Skewed distributions: taking the log or square root of scores can help
  - Nonlinear transformations – base 10 log of X

41

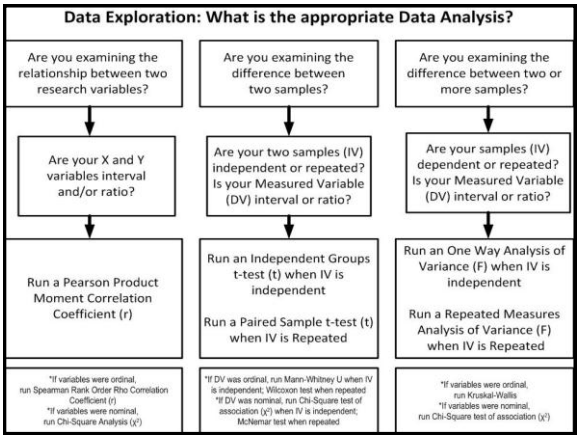
## Reporting about data screening

Researchers should describe data screening procedures and explain what steps were taken to identify and correct problems

For example: if some scores were deleted or recoded, report how many scores were changed and the rationale for making the changes.

It is a good idea to look at distributions of scores after you delete errors, remove outliers, or apply transformations to ensure that the remedies had the intended effect.

42




## Basic Assumptions for t-Test for Independent Groups

**Assumptions:**

1. Initially 2 sample groups come from same population (randomness)
2. Population is normally distributed
3. Two groups are representative samples; that is, they have approximately equal variances (Homogeneity of variance)

## Basic Assumptions for t-Test for Dependent Groups

**Assumptions:**

1. Paired differences are random sample from a normal population
2. Equal variances assumption is unnecessary, since you will be working with one group

## Basic Assumptions for One Way ANOVA

**Assumptions:**

1. Samples are randomly drawn from a normally distributed population
2. Variances of samples are approximately equal

## Basic Assumptions for Repeated Measures ANOVA

**Assumptions:**

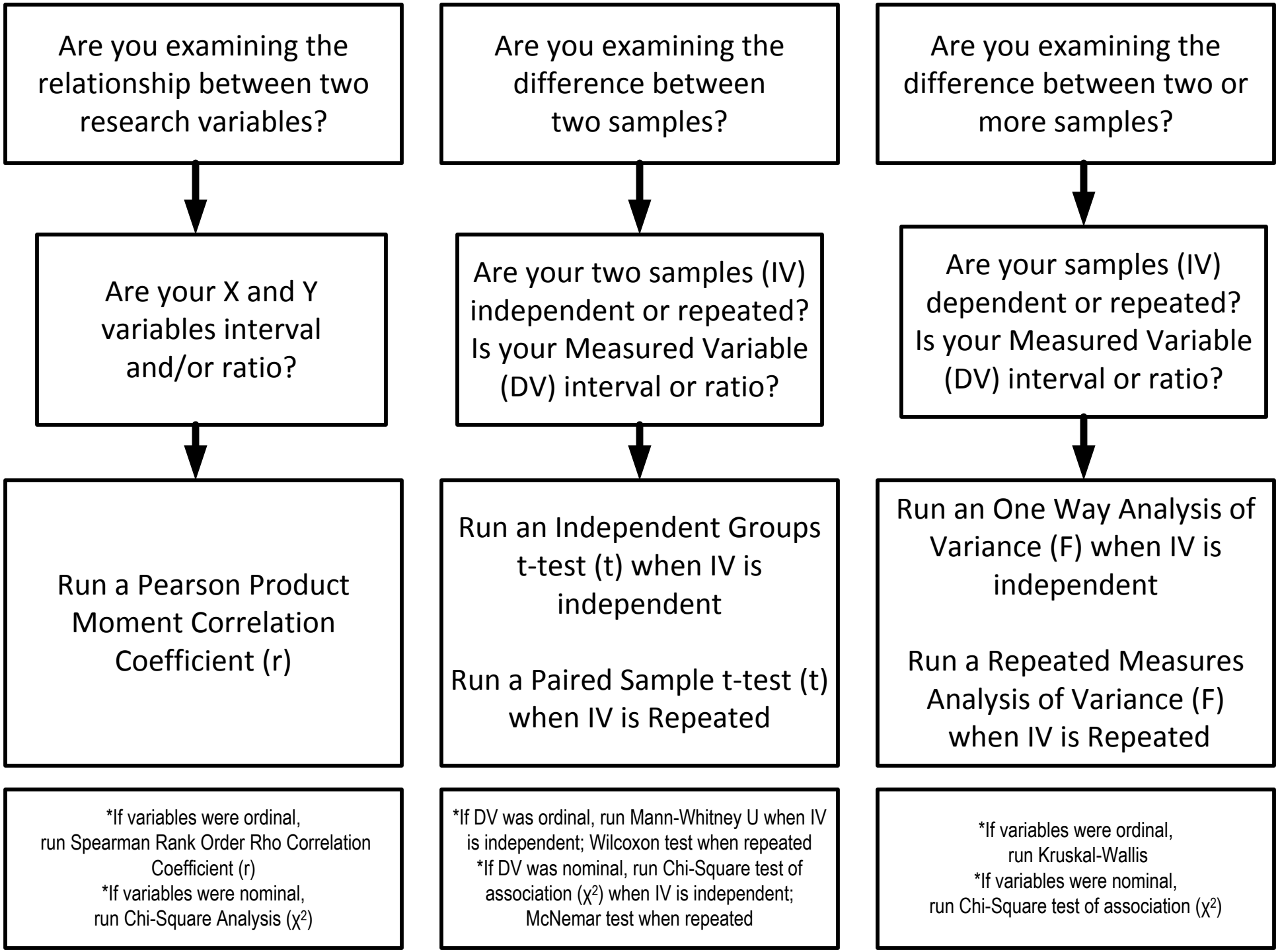
1. Samples are randomly selected from a normal population
2. Variances for each measurement are approximately equal

## Basic Assumptions for Pearson Product-Moment Correlation Coefficient

1. **Interval/Ratio** data of both variables
2. **Normal distribution**
   - Homogeneity of Variance - variation in scores for both X and Y scores must be similar, this is known as **heteroscedasticity**. Assumed unless either distribution is skewed
3. **Linear Relationship** - association between X and Y is linear. Relationship has to form a straight line. **Curvilinear relationships** (in which an increase in X is accompanied by an increase in Y up to a point, and is then accompanied by a decrease in Y) should not be assessed by Pearson r

Dr. Tracey Matthews (tmatthews@springfieldcollege.edu)
Dr. Kimberly Kostelis (kostelisk@ccsu.edu)

# Data Exploration: What is the appropriate Data Analysis?



# SPSS Commands for Data Exploration and Analysis

**Explore Missing Data**:

Analyze
- Descriptive Statistics
  - Frequencies
    - Move all variables that you want to explore into variable(s) box
  - Statistics
    - You can select from different descriptive data, such as mean, standard deviation, skewness, kurtosis
  - Charts
    - You can select graphs, such as a histogram to view your data

OK

**Explore Group Differences**:

Analyze
- Compare Means
  - Move DV to Dependent List
  - Move IV to Independent List
- Options
- Move over: mean, sd, n, kurtosis, skewness & SE of kurtosis and skewness

OK

**Explore Univariate Normality:**

Analyze
- Descriptive Statistics
  - Explore
    - Move IV to Factor List
    - Move DV to Dependent List
  - Statistics
    - Make sure descriptive and outliers are checked
  - Plots
    - Check histograms
  - Continue
- OK

Dr. Tracey Matthews (tmatthews@springfieldcollege.edu)
Dr. Kimberly Kostelis (kostelisk@ccsu.edu)

**Remove Impossible/Extreme Scores**

Data

Select Cases

Use a logical "if" statement to assist in excluding data

IF → sex ≅ 3

Continue

Select – If condition is satisfied

Output – Select filter out unselected cases

OK

## INDEPENDENT GROUPS t-TEST ANALYSIS

Analyze

Compare Means

Independent Sample T Test

Move DV into Test Variable box

Move IV into Grouping Variable box

DEFINE GROUPS

In Group 1 box enter in 1 (or code utilized)

In Group 2 box enter in 2 (or code utilized)

Continue

OK

## REPEATED MEASURES t-RATIO ANALYSIS

Analyze

Compare Means

Paired-Samples t Test

Highlight both conditions and move to Paired Variable Box

OK

Dr. Tracey Matthews (tmatthews@springfieldcollege.edu)
Dr. Kimberly Kostelis (kostelisk@ccsu.edu)

## SPSS: ONE-WAY INDEPENDENT GROUPS ANALYSIS OF VARIANCE

Analyze

Compare Means

One-Way ANOVA

Your DV → Dependent Variable Box
Your IV → Factor Box

Options

X Descriptives
X Homogeneity of Variance
X Means Plot

Continue

OK

## SPSS: REPEATED MEASURES ANALYSIS OF VARIANCE

Analyze

General Linear Model

Repeated Measures

Change Within-Subject Factor name (Factor 1) to the name of your repeated variable (i.e. Treatment)

Number of Levels - enter appropriate number of levels

ADD

Define

Highlight and Move TX1 --> (1) (Be careful here
Highlight and Move TX2 --> (2) to put these names
Highlight and Move TX3 --> (3) in logical order)

Plots

Highlight repeated measures factor name
and move it over to Horizontal Axis

ADD

Continue

Options

Highlight repeated measures factor name and
move it to the Display Means Box on Right

Display

X Descriptives
X Estimates of Effect Size
X Parameter Estimates

Continue

OK

Dr. Tracey Matthews (tmatthews@springfieldcollege.edu)
Dr. Kimberly Kostelis (kostelisk@ccsu.edu)