Introduction to Statistical Principles

Part I

LTC Renee Cole, PhD, RDN, LDN
Purpose

The goal of this training session is to gain (or refresh) basic knowledge of statistical principles.

Why do we care?

• Identify statistical tests needed to:
  – Effectively answer your research question
  – Assess data collected
  – Improve ability to critically appraise the quality of evidence-based literature

Note: Sessions present common statistical methods used. It is not meant to be all inclusive.
References

References that may helpful:


Acknowledgements

- Dr. Steven Allison, PT, retired Air Force Colonel
- Dr. John Childs, PT, Air Force Lt Col
- Dr. Shane Koppenhaver, PT, Army LTC

Some slide materials were extracted from their presentations for use in this 3-Part series
Objectives

• Describe variable scales of measurements
• Describe the principles involved in statistical inference & hypothesis testing
• Determine the appropriate statistical tests:
  – Descriptive analysis
  – Correlation analysis
  – T-tests analysis
  – ANOVA & Repeated Measures models
1st and Foremost...

• Methods & statistical test must be able to address the research question
• The type of statistical methods will depend upon:
  • Research design
  • Variables (independent vs. dependent)
  • Scales of measurement
  • Normality & homogeneity of data
• Always compute descriptive statistics to paint the picture of your sample population / setting
Classifications of Statistics

- **Descriptive** – *What are the characteristics of?*
- **Comparative** – *What’s the difference between?*
- **Correlation** – *What’s the relationship (association) between?*
- **Regression** – *What predicts?*
- **Reliability** – *How reproducible are scores, a technique or device?*

- **Parametric (inferential)** - *With underlying assumptions*
- **Nonparametric** – *Does not meet assumptions*

- **Univariate** – *Using a single dependent variable*
- **Multivariate** – *Simultaneous use of multiple dependent variables*
Making Choices

- If you already know all the statistical tests and how they are used, it’s “easy” …
- For the rest of us:
  - Broad categories help to clarify our thoughts
  - Algorithms help to pick the right test (see attachments)
  - Remember you will get better with practice
  - Statistical programs do most of the work
Variable Scales of Measurement

- **Nominal** ("naming") variables
- **Ordinal** ("ordered") variables
- **Interval** (equal intervals) variables
- **Ratio** (equal intervals, true zero) variables
Variable Scale of Measurement

Why Bother Identifying the Variable Scale of Measurement?

• Required to choose the right statistical test
• Helps to determine if parametric or non-parametric version of statistical test
Four Parametric Assumptions

1. The sample is randomly drawn from the target population
2. Data are normally distributed
3. Homogeneity of variance
   – SD of group 1 $\approx$ SD of group 2
4. Data are on interval/ratio scales (continuous)
   - May be justifiable to violate assumptions
     - Random population sampling is complex and costly (random sampling $\neq$ random group assignment)
     - Accounted for by “robust" nature of inferential tests
Nominal Variables

- Dichotomous or categorical
- Discrete and mutually exclusive categories
- To analyze must assign a # for each category / response for identification only (dummy coding)

**Examples:**
- **Sex** (1=Male, 2=Female) or **Deployed** (1=Yes, 2=No)
- **Race** (1=Caucasian, 2=Black/AA, 3=Asian, etc.)
- **Medical Condition** (1=DM2, 2=HTN, 3=CVD, etc.)

**Math:** descriptive counts only (frequencies or percent)

- No quantitative value
  - Mean value will not make sense (i.e. Sex mean value = 1.3)
  - Frequency (n= 200 Men); percent (70% men)
Ordinal Variables

• Rank-ordered categories

• Relationship for adjacent categories (> or <)
  – Zero < trace < poor < fair < good < normal
  – “Unequal intervals”

• Dummy codes for variables required for analysis

• Examples:
  – Rank (1=E1-E4, 2=E5-E9, 3=W01-CW5, 4=O1-O3, 5=O4-O6, etc.)
  – Satisfaction (1=Unsatisfied, 2=Neither Unsat or Sat, 3=Satisfied)
  – BMI Category (1=Underweight, 2=Normal weight, 3=Overweight, 4=Obese)

• Math: descriptive counts only (frequencies)

• Note: Survey items are typically ordinal, but when a set of survey items are treated as a score and totaled, the scale is often considered continuous
Interval Variables

- Rank-ordered
- Equal intervals
- No true zero, but considered ‘continuous’
- Examples:
  - Years (B.C. or A.D.) or Temp (°F, C or K) or Time
- Math: add or subtract (no ratios)
  - No dummy codes necessary!
- Often displayed as Mean ± Std Deviation (SD)
Ratio Variables

- Interval scale with true zero
- Examples:
  - Steps/day, Calories, Age, or Height
- Math: all operations, including ratios
  - No dummy codes necessary!
- Often displayed as Mean ± Std Deviation (SD)

For SPSS Interval & Ratio variables are treated the same = continuous
Let's Practice…

How can you manipulate these variables to change the scale?

- Education
- Calories
- Passed a Test
- Behavioral Risk Score
- Military Rank
- Weight Status
- Age

Can convert from continuous to nominal or ordinal, but cannot go in the opposite direction

Education Possibilities

- Continuous
  - Total # of years of education
- Ordinal
  - 1= High School Diploma
  - 2= Associates Degree
  - 3= Bachelor’s Degree
  - 4= Masters / Grad Degree
- Nominal
  - 1= Received BS diploma
  - 2= Did Not receive BS diploma
How to answer your Research Question?

Determine if:

- Descriptive study only
- Correlation / Relationship (no causation)
  - Pearson’s r, Spearman’s rho, Kendall’s-tau, Phi-coefficient
- Statistical Inference (infer to larger population)
  - T-Test, Chi-Square, ANOVA
  - Linear regression, Multivariate regression

Describe Population  Find Relationships  Cause & Effect
Descriptive Analysis

- Describes your research sample and/or situation
- Typically many independent variables (IV)
  - Not manipulating – just observing
- What do we typically want to know?
  - Demographics of Population (Sex, Age, Education, etc.)
  - Trying to paint a picture of the situation
    - Amputee weight status through healing process
    - Bariatric patients compliance with their meal plan
    - Dieting habits of Soldier trying to meet ABCP standards
Descriptive Analysis

How are outcomes displayed?

• Nominal / Ordinal
  – Frequencies (%)
  – Mode (most frequent #) or Median (mid-pt of data)
  – Presented in tables (%), histograms or pie charts

• Interval / Ratio (Continuous data)
  – Mean ± standard deviation (SD)
  – Presented in tables, figures, graphs
### Examples...

**Table 1. Participant Descriptive Demographics Dichotomized by Body Mass Index**

<table>
<thead>
<tr>
<th>Continuous Data</th>
<th>Nominal / Ordinal Data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong>&lt;sup&gt;*&lt;/sup&gt;</td>
<td><strong>Normal BMI</strong>&lt;br&gt;n=105</td>
</tr>
<tr>
<td>Mean SD</td>
<td>Mean SD</td>
</tr>
<tr>
<td>Age</td>
<td>30.1 8.6</td>
</tr>
<tr>
<td>BMI</td>
<td>30.0 4.2</td>
</tr>
<tr>
<td>Normal BMI status</td>
<td>n %</td>
</tr>
<tr>
<td>168 56.9</td>
<td></td>
</tr>
<tr>
<td>Sex (Male)&lt;sup&gt;*&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>209.0 70.8</td>
<td>53 50.5</td>
</tr>
<tr>
<td>Education</td>
<td></td>
</tr>
<tr>
<td>High School</td>
<td>50 16.9</td>
</tr>
<tr>
<td>College Credits up to BS</td>
<td>208 35.3</td>
</tr>
<tr>
<td>Graduate Credits</td>
<td>37 12.5</td>
</tr>
<tr>
<td>Rank</td>
<td></td>
</tr>
<tr>
<td>Enlisted</td>
<td>211.0 35.8</td>
</tr>
<tr>
<td>Officer</td>
<td>84.0 9.5</td>
</tr>
<tr>
<td>Branch (Army)</td>
<td>263.0 90.7</td>
</tr>
<tr>
<td>Ethnicity/Race</td>
<td></td>
</tr>
<tr>
<td>Black/AA</td>
<td>42.0 14.3</td>
</tr>
<tr>
<td>White/Caucasian</td>
<td>166.0 56.7</td>
</tr>
<tr>
<td>Hispanic</td>
<td>49.0 16.7</td>
</tr>
<tr>
<td>Asian</td>
<td>16.0 5.5</td>
</tr>
<tr>
<td>Other</td>
<td>20.0 6.8</td>
</tr>
</tbody>
</table>

<sup>*</sup> p<0.05 between BMI weight categories

BMI = Body Mass Index; SD = Standard Deviation; BS = Bachelors Degree; AA = African American
• Remember…
• You always want to describe your population
• Plan to run descriptive analysis regardless of the research design
• Even if RCT study you plan for…
  – Sample demographic descriptive analysis
  PLUS
  – T-test inferential statistics and/or correlations
• Descriptive analysis should be included in the ‘Methods Section’ of protocol or manuscript
How to answer your Research Question?

Determine if:

- Descriptive study only
- Correlation / Relationship *(no causation)*
  - Pearson’s r, Spearman’s rho, Kendall’s-tau, Phi-coefficient
- Statistical Inference *(infer to larger population)*
  - T-Test, Chi-Square, ANOVA
  - Linear regression, Multivariate regression

Describe Population

Find Relationships

Cause & Effect
Correlation Coefficients

• Assessing a relationship between variables
  • Do they vary together?
  • How strongly are they associated?
  • What is the nature of the relationship?

• Correlation vs. Causation
  • Not the same
  • Both may be caused by another factor
  • Causation statements come from inferential statistics (t-test, chi-square, analysis of variance, etc.)
Correlation Coefficients

• Vary between zero and ± 1.00
• Closer to |1.00|, higher strength of relationship
• Sign indicates direction
  • + increase together
  • - as one ↑ the other ↓
• Tighter grouping means higher coefficient
• Visualize with scatterplots

Portney & Watkins (2009) p. 524
Is it a good correlation?

Look at “r”

- $r = 0.00$ to $0.25 =$ little or no relationship
- $r = 0.26$ to $0.50 =$ fair relationship
- $r = 0.51$ to $0.75 =$ moderate to good
- $r = 0.75$ to $1.00 =$ good to excellent

Always create a scatterplot to observe values

- One outlier can influence the results
- Scatter points might not be uniform
Scatterplots

Current vs. Time

\[ \text{Current} = 0.0032 \times t + 0.44, \ r = 0.52 \]

Body Mass (kg)

HHD Force (newtons)

\[ r = 0.31 \]
Correlation Coefficient

- Covariance: as one changes, the other also changes
- Does **NOT** assess differences
- May **not have** statistically significant difference, but excellent correlation coefficient
- May **have** significant difference but weak correlation coefficient
- Does not assess agreement (ICCs do)
Significance of coefficient ($p$-value)

- Null hypothesis: the correlation coefficient between variable x and variable y is not significantly different from zero.
  - $H_0$: $r = 0$

- Very sensitive to sample size
  - trivial coefficients ($r = 0.1$ to $0.2$) are often statistically significant if sample is large enough
Design: Cross Sectional Assessment of BMI vs Intuitive Eating Scale Score (IES)

Purpose: Determine if there is a correlation between BMI status and intuitive eating scores.

Question: Do normal BMI Soldiers exhibit more intuitive in eating habits than overweight BMI Soldiers?

Primary outcome (dependent variable) is IES Score:
• 21-item questionnaire on a 5-point Likert Scale
• Point range from 21-105; the higher the score, the more intuitive

Grouping variable (independent variable):
BMI – dichotomized – normal (18-24.9 kg/m²) & overweight (≥25 kg/m²)

What type of demographic data should I collect?
Do you think there will be other associations with weight status and/or Intuitive Eating Score?
## Correlation matrix

<table>
<thead>
<tr>
<th></th>
<th>BMI</th>
<th>IES Score</th>
<th>Gender</th>
<th>Trying to Lose Wt</th>
<th>Meet 5 days/wk for 30 min</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BMI</strong></td>
<td>Correlation Coefficient</td>
<td>1.000</td>
<td>-0.555</td>
<td>-0.241**</td>
<td>-0.395**</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td></td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td><strong>IES Score</strong></td>
<td>Correlation Coefficient</td>
<td>-</td>
<td>1.000</td>
<td>-0.030</td>
<td>0.222**</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td></td>
<td>0.612</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td>Correlation Coefficient</td>
<td>-</td>
<td>-</td>
<td>1.000</td>
<td>-0.190**</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td></td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td><strong>Trying to Lose Wt</strong></td>
<td>Correlation Coefficient</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td></td>
<td></td>
<td></td>
<td>1.000</td>
</tr>
<tr>
<td><strong>Meet 5 days/wk for 30 min</strong></td>
<td>Correlation Coefficient</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* = Correlation is significant at the 0.05 level (2-tailed).
** = Correlation is significant at the 0.01 level (2-tailed).

IES = Intuitive Eating Score (ranges from 21-105 pts); Gender = 1 Male, 2 Female; Trying to Lose Weight = 1 Yes, 2 No; Meet 5 d/wk (30 min) physical activity = 1 No, 2 Yes

**What does this mean??**

Let’s draw out simple line graphs
Sometimes it helps to visualize...

Negative association...
- BMI increases as IES Score decreases
- Normal BMI tends to be more intuitive in eating characteristics
- Moderate to good relationship

$r = -0.555 \ (p < 0.001)$

Negative association...
- BMI decreases as dummy code for Gender increases
- How is Gender scored?
  - 1=M and 2=F
- Men are more likely to have a higher BMI than women
- Little to no relationship

$r = -0.241 \ (p < 0.001)$

Positive association...
- As dummy code of Lose Wt increases, so does IES Score
- How is Trying to Lose Wt scored?
  - 1=Yes, 2=No
- Participants trying to lose weight tend to be less intuitive in eating characteristics
- Little to no relationship

$r = 0.222 \ (p < 0.001)$
Is it a good correlation? 
Look at “r²”

r² = “coefficient of determination”

- The square of the correlation coefficient is called the “coefficient of determination”
- Value is conceptually helpful
- Interpreted as: “the percent of variance in x that is explained (or accounted for) by y”
- So, consider correlation of BMI & IES:
  - r = -0.555 means r² = 0.308
  - ~31% of the variance in IES is explained by BMI
Correlation Coefficients

• If correlation analysis is used, it should be noted in the ‘methods section’ of protocol or manuscript
• Typically note the type of correlational analysis, which is based upon scale of measurement
• Statistical programs do not always use same theoretical nomenclature as a text book
## Which Test Do I Use?

### Table: Type of Data

<table>
<thead>
<tr>
<th>Type of Data</th>
<th>Ratio</th>
<th>Interval</th>
<th>Ordinal</th>
<th>Nominal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ratio</td>
<td><strong>Pearson r</strong></td>
<td><strong>Pearson r</strong></td>
<td><strong>Spearman rho</strong></td>
<td><strong>Point biserial</strong></td>
</tr>
<tr>
<td>Interval</td>
<td></td>
<td><strong>Pearson r</strong></td>
<td><strong>Spearman rho</strong></td>
<td><strong>Point biserial</strong></td>
</tr>
<tr>
<td>Ordinal</td>
<td></td>
<td></td>
<td><strong>Spearman rho or Kendall’s tau</strong></td>
<td><strong>Rank biserial</strong></td>
</tr>
<tr>
<td>Nominal</td>
<td></td>
<td></td>
<td></td>
<td><strong>phi coefficient</strong></td>
</tr>
</tbody>
</table>

- **Pearson r** – Both variables are continuous (ratio or interval) (same in SPSS)
- **Spearman rho** – One variable is ordinal & the other is continuous or ordinal (same in SPSS) *(also used for continuous data instead of Pearson r if the data is not normally distributed)* *(Kendall’s tau assess small discrepancies; more accurate w/ small sample)*
- **Point biserial** – One variable is nominal & the other is continuous (Pearson in SPSS)
- **Rank biserial** – One variable is nominal & the other is ordinal (Kendalls-tau in SPSS)
- **Phi-coefficient** - Both variables are nominal (if discrete and dichotomous; 2x2 contingency table) *(Crosstab in SPSS)*
How to answer your Research Question?

Determine if:

- Descriptive study only
- Correlation / Relationship (no causation)
  - Pearson’s r, Spearman’s rho, Kendall’s-tau, Phi-coefficient
- Statistical Inference (infer to larger population)
  - T-Test, Chi-Square, ANOVA
  - Prediction, Linear regression, multivariate regression

Describe Population

Find Relationships

Cause & Effect
Continue on with Part 2

Inferential Statistics