Introduction to Statistical Principles

Part II

LTC Renee Cole, PhD, RDN, LDN
So far …

Classification of Stats
Scales of Measurement
Descriptive Statistics
Correlation Coefficients

Now let’s continue …
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
Probability

Basis for all inferential statistics
– Where inferences are made about the population based only on data from samples

Symbol is $p$ or $p$-value

“Sig.” or significance level is synonym in some contexts
$p$ and $\alpha$

$p$ is probability of committing a Type I error. (obtained from statistical test)

$\alpha$ is researcher’s maximum level of tolerance for committing a Type I error. (set by researcher in advance)

- Risk tolerance = willing to accept risk of Type I error up to level of alpha but no more
- Alpha is usually set to 0.05
- Means there is 5% probability of Type I error
Properties of a Normal Distribution

- Symmetrical
- Unimodal
- Points of Inflection at μ & σ ± 1
- Tails that approach the horizontal axis
- Considered parametric
Skewed Distributions

Mean & Median are not at same point on curve
Skewness: (+)-skew (tail on higher end); (-)-skew (tail on lower end)
Not normally distributed
No longer parametric

Positively Skewed
(+ skewness coefficient)

Negatively Skewed
(- skewness coefficient)
Four Parametric Assumptions

The sample is randomly drawn from the target population

Data are normally distributed

Homogeneity of variance
  – SD of group 1 $\cong$ SD of group 2

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)
  – Ordinal intervals "might be" equal

• Accounted for by “robust” nature of inferential tests
Homogeniety

Normal distribution
Variance about the means are not equal
No longer parametric when comparing 2 groups
Sampling Error: Sampling Distribution of Means

Central Limit Theorem
“The sampling distribution of means will approach the normal curve as ‘n’ increases”

This concept allows researchers to “use sampling distributions and probabilities associated with a normal curve to predict population characteristics for any distribution”

Portney & Watkins (2009) p. 408, 410
When two groups have different mean scores, there are two interpretations based upon probability:

The difference represents sampling error ($p > \alpha$)

Two samples from the same population

The difference represents "real" (population) difference ($p \leq \alpha$)

Two samples from 2 different populations
Hypothesis Terms

Statistical hypothesis = Null hypothesis
- Ho; this is always the hypothesis tested statistically
- Ho: $\mu_1 = \mu_2$ OR Ho: $\mu_1 - \mu_2 = 0$

Alternative hyp ($H_1$ or $H_a$):
- $H_a$: $\mu_1 \neq \mu_2$ OR $H_1$: $\mu_1 - \mu_2 \neq 0$
- AKA “research hypothesis”; it is what the researcher expects to find

Directional (1-tailed) OR non-directional (2-tailed)
Hypothesis Testing: Definitions

**Null Hypothesis (H₀)**
- In context of comparing 2 sample means with a t-test:
  \[ H₀ : \overline{X}_1 = \overline{X}_2 \]
  States that the population means (not sample means) are equal

No difference between mean APFT scores of Ranger vs. Medical Unit

**Alternate Hypothesis (Hₐ)**
- Directional: \[ Hₐ : \overline{X}_1 > \overline{X}_2 \] Ranger scores are higher
- Non-directional: \[ Hₐ : \overline{X}_1 \neq \overline{X}_2 \] Different, but not sure which unit will be higher

**Research Hypothesis**
- Can be \( H₀ \) or \( Hₐ \)
Decision Rule: Traditional Hypothesis Testing

If $p \leq \alpha$: reject the null hyp.
- We accept the risk of being wrong

If $p > \alpha$: accept the null hyp.
- Risk of Type I error is greater than our threshold for risk tolerance

Example:
- $\alpha = 0.05$
- Mean APFT scores of Rangers is 285 ± 40 pts and the Medics score is 260 ± 40 pts... significant difference...

  - If $p = 0.048$? $p < 0.05$; Yes, reject $H_0$, truly different
  - If $p = 0.052$? $p > 0.05$; No, accept $H_0$, different by chance
Hypothesis Testing: Truth vs. Outcome

$H_0 = \text{No difference in mean APFT scores of Ranger vs. Medical Units}$

<table>
<thead>
<tr>
<th>Decision</th>
<th>$H_0$ is true</th>
<th>$H_0$ is false</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reject $H_0$</td>
<td>Type I Error ($p$)</td>
<td>$\square$</td>
</tr>
<tr>
<td>Accept $H_0$</td>
<td>$\square$</td>
<td>Type II Error ($\beta$)</td>
</tr>
</tbody>
</table>

Ranger APFT = Medic APFT (Truth)
Ranger APFT $\neq$ Medic APFT (Ranger APFT $\neq$ Medic APFT)
Type I and Type II Errors

Type I error \((\alpha)\) *(can only happen when \(p \leq \alpha\))*
- “Rejecting” the \(H_0\) when \(H_0\) is true
- Said another way…Finding a significant difference when none exists in the target population
- Rangers have a higher mean APFT scores than medics when in reality they do not!

Type II error \((\beta)\) *(can only happen when \(p > \alpha\))*
- “Accepting” the \(H_0\) when \(H_0\) is false
- Said another way…Failing to find a significant difference when there is one in the target population
- No difference between Ranger & Medics APFT score when in reality one really exists!
- Occurs when your sample size is too small; you don’t have enough power to find a true difference
When two groups have different mean scores, there are two interpretations:

- The difference represents sampling error \((p > \alpha)\)

  ![Diagram showing sampling error]

- The difference represents "real" (population) difference \((p \leq \alpha)\)

  ![Diagram showing population difference]
Standard Error of the Mean

SEM (SE$\bar{x}$ or SS$\bar{x}$)

- Formula: $s\bar{x} = \frac{s}{\sqrt{n}}$

Practical use: 95% CI value of population mean

We can be 95% confident that the population mean lies somewhere within ~2 SEMs on either side of the sample mean

5% CI = $\bar{X} \pm \sim 2s\bar{x}$
Not the same thing!

Statistical significance determined by \( p \) (in relation to \( \alpha \))

Clinical significance is determined by effect size (in relation to MCID)

- Important concept when looking at effect size and clinical relevance of research findings

- “…the smallest change score associated with a patient’s perception of a change in health status”*

- Depends on attributes of outcome scale, clinical context, clinical judgment
What are Types of Inferential Stats Tests?

- Independent t-test
- 1-ANOVA
- RM-ANOVA
- MANOVA
- Regression
- Paired t-test
- Chi-square
Assumptions for t-Tests
(Basic Assumptions for Parametric Tests)

Statistical method to decide whether an observed difference in sample scores represents a “real” difference in the population

Two basic types:

- Independent (unpaired) t-test

*Portney & Watkins (2009) p. 434*
Samples are randomly drawn from populations

Normal distributions
  – Test with Kolmogorov-Smirnov test
  – In SPSS: Analyze, Nonparametric Tests, 1-Sample KS

Homogeneity of variance
  – Test with Levene’s test, Bartlett’s test, etc.
  – In SPSS: Levene’s test output automatic with t-tests

Data from ratio or interval scales (continuous data)
One-tailed vs. Two-tailed Tests

**One-tailed t-test**

**Directional hypothesis**

“Ranger APFT score is **better than** Medic APFT score”

---

**Two-tailed t-test**

**Non-directional hypothesis**

“Ranger APFT score is **different from** Medic APFT score”

---

**FIGURE 18.6** Standard normal distribution of z-scores showing critical values for a one-tailed test (non-directional) at α₂ = .05 and the calculated z-score for IQ.

**FIGURE 18.7** Standard normal distribution of z-scores showing critical values for a one-tailed test (directional) at α₁ = .05 and the calculated z-score for IQ.
Independent t-Test

Two independent groups

Parametric assumptions met

Check homogeneity of variance

- Levene’s test ($H_0: s^2_1 = s^2_2$)

Equal & unequal variance options

- Practical focus is on $p$-value
  - If $p \leq \alpha$, reject $H_0$
  - If $p > \alpha$, accept $H_0$

- Choose 1-tailed or 2-tailed results
  - Default in SPSS is 2-tailed
  - Divide $p_{(2\text{-tailed})}$ by 2 to get $p_{(1\text{-tailed})}$
Independent t-Test

Objective: Cross Sectional Observational Study

Hypothesis: Assess difference in APFT scores between 2 unit APFT programs. One unit has a team sports-based APFT program whereas the other uses standard PRT.

Question: Do the mean APFT scores differ between the 2 programs?

Primary outcome is the APFT Score (points).

Compared to the standard PRT Program:

- Sports team APFT unit has a significantly higher APFT score:
  - Mean APFT score = 230 pts
  - APFT score range = 180-280 pts (± 50 points)

Compared to the Team Sports Program:

- Unit exercise programs have similar APFT results:
  - Mean APFT score = 210 pts
  - APFT score range = 160-260 pts (points)

Independent t-Test: p = 0.132; p = 0.004; the Sports team APFT unit is significantly higher.
Paired t-Test

Used to compare one group measured twice
- pre- vs. post intervention

Use with matched groups
- twins, age-matched, etc.

Used when parametric assumptions met

Choose one-tailed or two-tailed results
- SPSS default: 2-tailed for “Sig.” level (synonym for $p$-value)
- Simply divide $p_{(2\text{-tailed})}$ by 2 to get $p_{(1\text{-taile})}$

Practical focus is on $p$-value
- If $p \leq \alpha$, reject $H_0$
- If $p > \alpha$, accept $H_0$
Paired t-Test

Objective: Double-Blind Randomized Controlled Cross-Over Trial

Hypothesis: Assess difference in serum Vit B12 values when two different Vit B12 supplements (1000 mcg) are taken by patients with low Vit B12 serum levels (200 pg/ml). Patients are randomly assigned to Supplement A (or B) for 8 weeks with a 4 week washout period before the start of Supplement B (or A).

Question: Is Vit B12 Suppl A (soft gel) better than Suppl B (tablet) to increase serum B-12? Primary outcome is the Vit B12 serum value.

$p = 0.065$; Both B-12 supplements provide similar results

$p = 0.015$; You can conclude that one B-12 supplement is better than the other

Vitamin B12
- B12 serum = 200 ± 40 pg/ml
- B12 serum = 650 ± 40 pg/ml

Before B-12 serum = 200 ± 40 pg/ml
After B-12 serum = 550 ± 40 pg/ml
Analysis of Variance (ANOVA): Comparison of 3 or more groups

Are observed differences in whole set of means greater than would be expected by chance alone?

F statistic
– Test statistic for the ANOVA
– The larger the F, the more likely it is significant

p-value: similar interpretation as with t-test

Independent groups assessment or repeated times

Parametric assumptions:
– random sample
– Normal distribution
– homogeneity of variance
Decision Rule:
Hypothesis Testing with ANOVA

$H_0$: no difference among means

If $p \leq \alpha$: reject $H_0$
If $p > \alpha$: retain $H_0$

Alpha is usually set to 0.05
- 5% probability of Type I error
- Statement of risk tolerance (willing to accept risk of Type I error up to level of alpha but no more)
One-way ANOVA: Independent samples

Used with a single between-subjects factor
- synonym: non-repeated measures IV

Which diet results in more weight lost in 1 month (Mediterranean, Zone, Atkins, Paleo Diets)?
- n=11 per each of 4 groups
- Random assignment to group
- Complete pre & post-test
- What is DV? (Assume normal distribution of DV)
- Number of levels of IV?

To determine where the difference lies (or where multiple differences lie), we must do multiple comparison tests (posthoc)
One-Way ANOVA

**Objective:** Randomized Controlled Trial

**Hypothesis:** Assess effectiveness of four different weight loss diets (Mediterranean, Zone, Atkins, Paleo Diets).

**Question:** Is one diet better than the other at facilitating weight loss?

**Primary outcome is change in weight (lbs).**

To analyze in SPSS: Compare Means and click on One-Way ANOVA

$p=0.023$ obtained within the SPSS ANOVA output table, but are they all significantly different from each other? Which one is better?

We can’t tell by just looking at the graph.

We need to do post-hoc analysis.
Error Associated with Multiple Tests

ANOVA (Omnibus test): $\alpha = 0.05$; reduces type one error

Post Hoc Test:
- Mediterranean vs Zone: $\alpha = 0.05$
- Mediterranean vs Atkins: $\alpha = 0.05$
- Mediterranean vs Paleo: $\alpha = 0.05$
- Zone vs Atkins: $\alpha = 0.05$
- Zone vs Paleo: $\alpha = 0.05$
- Atkins vs Paleo: $\alpha = 0.05$
- Family-wise type 1 error rate $\approx$ add all 0.05 + omnibus $\alpha = 0.35$

Need to balance between Type 1 & 2 error
- Bonferroni divides $\alpha$ by # of statistical tests
  - i.e. $\alpha/7 = 0.05/7$; thus $\alpha = 0.007$ for each post-hoc test
**Multiple Comparison Tests**

synonym: *post-hoc tests*

Post-Hoc tests
- Newman-Keuls (S-N-K in SPSS)
  - Higher statistical power: protects against Type II error
- Scheffe’s
  - Lower statistical power: protects against Type I error
- Tukey’s HSD
  - Intermediate level of statistical power

For Repeated Measures ANOVA & MANOVA
- Sidak – more powerful but less conservative
- Bonferroni – most commonly used

Typically found under “Post Hoc” option within the SPSS test view
Two-way ANOVA

One DV again & Two IV – between subjects
Participants complete both interventions:
- 2 IVs: Diet and Exercise
- DV: Weight Loss

What is the effect of Diet (Low Carb Diet vs. Low Fat Diet) on weight when tested with aerobic activity (30 min/d vs. 60 min/d)?

n=60 (random assignment to groups and assume normal distribution of DV)

How many levels for each IV?

<table>
<thead>
<tr>
<th></th>
<th>Low Carb (A₁)</th>
<th>Low Fat (A₂)</th>
</tr>
</thead>
<tbody>
<tr>
<td>30 min PA (B₁)</td>
<td>Mean (A₁B₁)</td>
<td>Mean (A₂B₁)</td>
</tr>
<tr>
<td>60 min PA (B₂)</td>
<td>Mean (A₁B₂)</td>
<td>Mean (A₂B₂)</td>
</tr>
</tbody>
</table>
Interaction Effects

Main Effect
- Effect of diet on weight when ignoring amount of aerobic activity?
- Effect of aerobic activity on weight when ignoring diet?

Interaction Effect
- Interaction between diet & aerobic activity
- Effect on weight?

[Graph showing weight change with different diet and aerobic activity levels]
**Two-Way ANOVA**

**Context:** Randomized Controlled Trial (2 x 2 ANOVA)

**Purpose:** Assess effectiveness of weight loss diet (Low Carb vs. Low Fat) and exercise amount (30 min/d vs. 60 min/d aerobic activity).

**Question:** Is one diet and exercise prescription better than the other for weight loss? Primary outcome is change in weight (lbs).

To analyze in SPSS: General Linear Model and choose Univariate

\[ p = 0.044 \], but are they all significantly different from each other?

Again, we need to do post-hoc analysis.

What if we had 3 levels of measurement for one of our IV?
- Low Fat, Low CHO, High PRO or
- 30 min, 45 min, 60 min PA
- 2 x 3 ANOVA
Repeated Measures ANOVA

This is the principal statistical analysis in many randomized controlled trial (RCT)

Used for a **within-subjects** factor
- synonym: repeated-measures IV
- Repeated measures IV: usually “time”
- i.e. baseline, mid-treatment, end-of-treatment, follow up

One **between-subjects** factor
- Non-repeated IV: usually “group”
- i.e. placebo group, treatment group

One **dependent variable**
- i.e. Weight Loss…or…changing biochemical parameter…or…

More statistical power than a 1-way ANOVA for a non-repeated factor

Usual parametric assumptions plus one more:
- Sphericity (Mauchly’s Test) – if fail, read *p*-value Greenhouse-Geisser or
**Objectives:**
Prospective study repeated measures RCT (0, 2 months, 4 months, 6 months)

**Hypothesis:**
Assess effectiveness of a Diabetes Telehealth program compared to Standard program (Face-to-Face)

**Question:**
Which program will yield a greater improvement in blood sugar?

**Primary outcome:**
Change in fasting blood glucose (FBG)

**Methodology:**
- Recruit participants and randomize into one of two conditions
- Baseline measurements
- Tele-health intervention
  - 2 month post intervention assessment
  - 4 month post intervention assessment
  - 6-month follow-up
- In-person group sessions
  - 2 month post intervention assessment
  - 4 month post intervention assessment
  - 6-month follow-up

**Analysis in SPSS:**
- General Model and choose Repeated Measures

**Graph:**
Change in Fasting Blood Sugar

Graph showing comparison between F2F and Telehealth groups over time (0, 2, 4, 6 months) with decreasing FBG values.
MANOVA Designs

Often called “mixed-model” ANOVA
Principal statistical analysis in randomized controlled trial (RCT)

One **within-subjects** factor
- Repeated measures IV: usually “time”
- i.e. baseline, mid-treatment, end-of-treatment, follow up

One **between-subjects** factor
- Non-repeated IV: usually “group”
- i.e. placebo group, treatment group

More than 1 DV (unlike the standard RM ANOVA)
- Have multiple outcome variables
- Weight loss, ↓ in LDL, ↑ Intuitive Eating Score
Nonparametric Tests
Nonparametric tests

Violates Parametric Assumptions
- Sampling – Convenience or Small Samples
- Normality – Skewed or Bimodal
- Homogeneity – significant variance between groups
- Scale of Measurement – Nominal or Ordinal data

Less sensitive and based upon
- Comparison of ranks of scores
- Comparisons of counts (yes/no) or “signs” of scores

Less powerful compared to parametric
- Need n=100 for non-parametric compared to n=70 in parametric sample
Chi-Square

Chi-square is the non-parametric version of t-test

Types of Analysis:

1. Goodness of Fit – observed frequencies different than expected frequencies
2. Test of Independence (association) – are frequencies of 1 variable associated with another variable?

To analyze in SPSS:
- Descriptive Statistics and choose Crosstabs
- Under statistics tabs, choose Chi-Square
Chi-Square Contingency Table
(2 Groups & 2-Level Variable)

<table>
<thead>
<tr>
<th>Group * Gender Crosstabulation</th>
<th>Gender</th>
<th>Count</th>
<th>% within Group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male</td>
<td>Female</td>
<td></td>
</tr>
<tr>
<td>Group A</td>
<td>42</td>
<td>34</td>
<td>55.3%</td>
</tr>
<tr>
<td>% within Group</td>
<td></td>
<td></td>
<td>100.0%</td>
</tr>
<tr>
<td>Group B</td>
<td>33</td>
<td>40</td>
<td>45.2%</td>
</tr>
<tr>
<td>% within Group</td>
<td></td>
<td></td>
<td>100.0%</td>
</tr>
<tr>
<td>Count</td>
<td>75</td>
<td>74</td>
<td>100.0%</td>
</tr>
<tr>
<td>% within Group</td>
<td>50.3%</td>
<td>49.7%</td>
<td></td>
</tr>
</tbody>
</table>

Is the between-group gender proportion difference statistically significant?

\[ p = 0.220 \ (p > \alpha) \]

No statistically significant difference in proportion of gender between the 2 groups.

Fisher’s Exact Test (2 x 2 tables only) for small samples with expected frequencies less than 5 (not needed here).
Chi-Square Contingency Table (3 Groups & 2-Level Variable)

Is the between-group proportion difference statistically significant?

\[ p = 0.022 \ (p < \alpha) \]

Statistically significant difference in proportion of patients regaining weight among the 3 groups
# How to Run Nonparametric T-Test & ANOVA in SPSS

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Parametric Test</th>
<th>Nonparametric Test</th>
<th>How to run in SPSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independent Groups</td>
<td>Unpaired t-test</td>
<td>Mann-Whitney U test</td>
<td>Analyze, Non-parametric tests, 2-Independent Samples, Mann-Whitney U</td>
</tr>
<tr>
<td>Paired t-test</td>
<td>Wilcoxon Signed-Ranks test</td>
<td></td>
<td>Analyze, Nonparametric Tests, 2 Related Samples</td>
</tr>
<tr>
<td>One-Way ANOVA</td>
<td>Friedman Two-Way ANOVA by Ranks</td>
<td></td>
<td>Analyze, Non-parametric tests, K Related Samples</td>
</tr>
</tbody>
</table>
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
Thank You!

For additional questions contact: renee.e.cole.mil@mail.mil; 508-233-5808 office