Mark Van Selst

San Jose State University

Cozby & Bates: Methods in Behavioral Research

Chapter 13: Understanding Research Results (Statistical Inference)

Summer 2014

Statistics

Descriptive Statistics: describe the sample

Inferential Statistics: describe the likely characteristics of the Population (the one that you are trying to make INFERENCES about)

The California State University
WORKING FOR CALIFORNIA

Hypothesis Testing

Null Hypothesis

- H_0: $Mean_{Treatment} = Mean_{Control}$

Research Hypothesis (a.k.a. alternate hypothesis)

- H_1: $Mean_{Treatment} \neq Mean_{Control}$

The logic of statistical hypothesis testing is that if we test to see if we can reject the “no difference” hypothesis. We reject H_0 where there is sufficient evidence that there is likely to be a “real” difference.

The “significance” level reflects the likelihood (***PROBABILITY***) that the obtained difference is a product of chance (random error, etc.) alone rather than reflecting a “real” difference in the populations that the samples represent.

The California State University
WORKING FOR CALIFORNIA

Probability

- When do we reject H_0?
- The probability (*p*-value) reflects the likelihood that the obtained results are a product of chance only
- We TYPICALLY set the alpha level (maximum obtained p-value to call a result statistically significant rather than due to chance) at $\alpha = .05$ (1/20).

t-tests

The “**student’s t” (t-test**) compares two means.

t = between group difference / within group variability

$t = (M_1 – M_2) / \sqrt{(sd_1/N_1 + sd_2/N_2)}$

Degrees of Freedom refer to the number of scores that are free to vary once the means are known (e.g., for a t-test $df = N_1 + N_2$ -2); df will determine the critical values that t must obtain in order to reject Ho.

One vs. Two-Tailed Tests

- H_1 Can be a directional hypothesis
  - (e.g., men will be taller than women)
- When H_1 is directional, a one-tailed test may be appropriate
  - (e.g., since if you ever find that women are taller than men you will fail to reject the null hypothesis)
- You can compute a one-tailed test by halving the "p-value" reported for the (two-tailed) statistical test.

ANOVA (Analysis of Variance)

- The statistic associated with ANOVA is “F”
- F values reflect explained variance / error variance
- $F = MS_{TMT} / MS_{error}$
- $F = MS_{Between} / MS_{within}$
- Any time *F<1*, the result is not statically significant
- *F* has 2 degrees of freedom terms: one for the treatment, one for the error term
  - The treatment yields “systematic variance” (aka between groups variance)
  - The error term reflects the “random error” that occurs within groups (aka within-group variance)
  - $F(2,35) = 23.33, p=.002$

Effect Size

Effect size is an (important!) general term that refers to the strength of association between variables. In its most generic sense, the effect size indicates how much of the variability in one variable is attributable to its relation with particular other variables. Pearson's r is a measure of the effect size (recall: $-1.0 < r < 1.0$).

Small:	**/r/ <= 0 to .20**
Medium:	**/r/ <= .20 to .30**
Large:	**/r/ <= .30 to .40+**

r^2 is "variance accounted for"; that is how much (%) of the variation across Variable A is associated with variation in Variable B. The technical term is that r^2 is the "*coefficient of determination*".

Effect Size

Correlation coefficient ($-1<r<1$) gives a measure of effect size (r2 is variance accounted for)

- $r = \sqrt{(t^2/(t^2+df))}$

Cohen's d: expresses effect size in terms of units of standard deviation (i.e., an effect of .75 is ¾ of a standard deviation difference between groups… this is a lot!)

- $d = (M_1 – M_2) / \sqrt{((sd_1^2 + sd_2^2)/2)}$

Note: r is limited to $-1<r<1$; There is no maximum value for d.

Confidence Interval

- We often want to provide the reader with some “real” idea of the stability of the data we are reporting. This is often done by reporting the “Confidence Interval” of our parameter estimates.
- In most practical research, the standard deviation for the population of interest is not known. In this case, the standard deviation *σ* is replaced by the estimated standard deviation (*SE)*, also known as the *standard error.*
- *The standard error is an estimate for the true value of the standard deviation, the distribution of the sample mean (M) is no longer normal with mean μ and standard deviation σ/(√n). Instead, the sample mean follows the t distribution with mean μ and standard deviation SE/(√n).*
- *The t distribution is described by its degrees of freedom. For a sample of size n, the t distribution will have n-1 degrees of freedom. The notation for a t distribution with k degrees of freedom is t(k) where k = n-1. As the sample size n increases, the t distribution becomes closer to the normal distribution as the standard error approaches the true standard deviation σ with a large enough n.*
- *For a population with unknown mean and unknown standard deviation, a confidence interval for the population mean, based on a sample of size of n, is* $M \pm t^{crit}\, SE/(\sqrt{n})$*, where* t^{crit} *is the upper (1-C)/2 critical value for the t distribution with n-1 degrees of freedom,* $t^{crit}(n-1)$*.*
- Typically C = .95 (i.e., the 95% confidence interval)

Type I and Type II errors

	POPULATION		
DECISION		H_0 True	H_0 False
	Reject H_0	**TYPE I ERROR (α)**	**Correct Decision (1-β)**
	Fail to Reject H_0	**Correct Decision (1- α)**	**Type II ERROR (β)**

POWER = 1- β … more powerful experiments are less likely to fail to find effects if they exist

Statistical Significance

- “practical significance”
  - very large N → very powerful experiments
  - Very powerful experiments can pick up very small effects.
  - Very small effects may or may not be “meaningful” in the real world / application.
- “statistical significance”
  - Depends on the alpha level chosen by the experimenter (minimum required p-value)
  - Typically set at .05
  - Familywise error rate increases with increases in the “family” of statistical tests run (more tests→ greater chance of Type I error)

Choosing the Appropriate Inferential Statistic

- Nonparametric tests: for nominal or ordinal Dependent Variables
  - Chi-square (“goodness of fit” and/or “test of independence)
- Parametric tests: for (approximately) interval or ratio data
  - Relationships: correlation
  - Two means: t-test
  - 2+ means: oneway ANOVA
  - Multiple IVs: Factorial ANOVA, multiple regression, …

Choosing the Appropriate Inferential Statistic

Five issues must be considered when choosing statistical tests.

- Scale of measurement
- Number of samples/groups
- Nature of the relationship between groups
- Number of variables
- Assumptions of statistical tests

Choosing the Appropriate Inferential Statistic

IV	DV	Statistic
Nominal (Men/Women)	Nominal (Vegitarian? Y/N)	Chi-Square (χ^2)
Nominal (Men/Women)	Interval (or Ratio) (e.g., gpa)	t-test (t)
Nominal (BAC: .00/.05/.08)	Interval (or Ratio) (test score)	One-Way ANOVA (F)
2+ Nominal	Interval (or Ratio)	Factorial ANOVA
Interval/Ratio (Beck Depression)	Interval (or Ratio) (sick days taken)	Pearson Correlation (r)
2+ Interval/Ratio	2+ Interval/Ratio	Multiple Regression

POWER

POWER = 1- β … more powerful experiments are less likely to fail to find effects (if they exist)

Effect Size: *r*	Power = .80	Power = .90
.10	n=786	n=1052
.20	n=200	n=266
.30	n=88	n=116
.40	n=52	n=68
.5	n=28	n=36

Chapter 13 Terminology

- Alpha level
- ANOVA
- Confidence Interval
- Chi-Square Test
- Degrees of Freedom (df)
- Error Variance
- Inferential versus descriptive statistics
- Null Hypothesis (H_0)
- Alternate (Research) Hypothesis (H_1)
- Power
- Probability
- Sampling Distribution
- Statistical Significance
- Systematic Variance
- Cohen's d
- Type I, II error
- t-test
- Experimental Replication

CSU The California State University

www.calstate.edu

www.sjsu.edu/psych