Statistical Significance

Is it not enough ? So why some journals enforce requirements like practical significance?

by

Dr. Mohammad H. Omar
Department of Mathematical Sciences

Dec 19, 2006

Presented at Statistic Research (STAR) colloquium,
King Fahd University of Petroleum & Minerals,
Dhahran, Saudi Arabia.

SEMINAR OUTLINE

- *Mean differences*
- *What is statistical significant difference?*
- *How do we show statistical significance?*
- *Is statistical significance enough*
  to meet some journal publication requirements?
- *APA statements*
  for journals following the APA submission format
- *List of journals requiring effect sizes*
- *What is practical significance?*
- *Jacob Cohen's practical significance index*
- *Ranges of importance*
- *Is it easy to calculate these indices?*
- *An Example of practical significance index in journal articles*

Mean differences

- Often time in research, we want to compare treatment groups.
- We want to compare a control group with another group taking an experimental treatment, usually we call this a treatment group.
- Let's say we found that the sample mean for

  Treatment group > control group
- Often we express this as mean difference

  Treatment – Control > 0

Is this difference good enough?

Mean Difference

Is this difference good enough?

- Even if we find a difference, we don't know if there is a real mean difference in the population.
  - That is, we don't even have an idea about how likely is it that the population means are different.
  - Note that we never know the population.
- We don't know how likely until we take our random sampling chance error into account.
  - Thus, we need to do some statistical work like hypothesis testing for mean difference.

What is statistically significant mean difference?

Sample mean difference is

difference between

the sample means

Statistically significant mean difference is

large enough mean difference that cannot be due to chance error alone

How do we show statistical significance?

- We have to take into consideration the sampling error for these mean differences
- Also, take the standard error of the sample mean differences into account
- Often, these techniques are covered in statistics textbooks and courses under the topic of hypothesis testing

Hypothesis Tests for Two Population Means

Two Population Means, Independent Samples

Lower tail test:	Upper tail test:	Two-tailed test:
$H_0: \mu_1 \geq \mu_2$ $H_A: \mu_1 < \mu_2$ i.e., $H_0: \mu_1 - \mu_2 = 0$ $H_A: \mu_1 - \mu_2 < 0$	$H_0: \mu_1 \leq \mu_2$ $H_A: \mu_1 > \mu_2$ i.e., $H_0: \mu_1 - \mu_2 = 0$ $H_A: \mu_1 - \mu_2 > 0$	$H_0: \mu_1 = \mu_2$ $H_A: \mu_1 \neq \mu_2$ i.e., $H_0: \mu_1 - \mu_2 = 0$ $H_A: \mu_1 - \mu_2 \neq 0$

Hypothesis tests for $\mu_1 - \mu_2$

Population means, independent samples

σ_1 and σ_2 known	Use a **z** test statistic
σ_1 and σ_2 unknown, n_1 and $n_2 \geq 30$	Use s to estimate unknown σ, approximate with a **z** test statistic
σ_1 and σ_2 unknown, n_1 or $n_2 < 30$	Use s to estimate unknown σ, use a **t** test statistic and pooled standard deviation

Pooled s_p t Test: Example

Is there a difference in dividend yield between stocks listed on the NYSE & NASDAQ? A financial analyst for a brokerage firm collected the following data:

	NYSE	NASDAQ
Number	21	25
Sample mean	3.27	2.53
Sample std dev	1.30	1.16

Assuming equal variances, is there a difference in average yield (α = 0.05)?

Solution

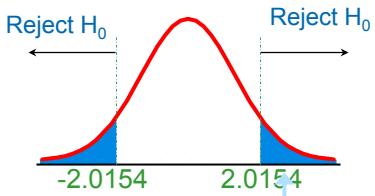
H_0: $\mu_1 - \mu_2 = 0$ i.e. ($\mu_1 = \mu_2$)

H_A: $\mu_1 - \mu_2 \neq 0$ i.e. ($\mu_1 \neq \mu_2$)

$\alpha = 0.05$

df = 21 + 25 - 2 = 44

Critical Values: t = ± 2.0154



2.040

Test Statistic:

$$t = \frac{3.27 - 2.53}{1.2256\sqrt{\frac{1}{21} + \frac{1}{25}}} = 2.040$$

Reject H_0 at α = 0.05

There is statistical evidence of a difference in means.

Calculating the Test Statistic

The test statistic is:

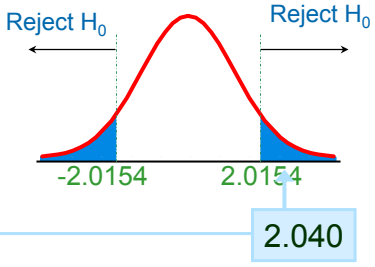
$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(3.27 - 2.53) - 0}{1.2256\sqrt{\frac{1}{21} + \frac{1}{25}}} = 2.040$$

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(21-1)1.30^2 + (25-1)1.16^2}{21 + 25 - 2}}$$

$$= 1.2256$$

Solution

H_0: $\mu_1 - \mu_2 = 0$ i.e. ($\mu_1 = \mu_2$)
H_A: $\mu_1 - \mu_2 \neq 0$ i.e. ($\mu_1 \neq \mu_2$)
$\alpha = 0.05$
df = 21 + 25 - 2 = 44
Critical Values: t = ± 2.0154



Test Statistic:

$$t = \frac{3.27 - 2.53}{1.2256\sqrt{\frac{1}{21} + \frac{1}{25}}} = 2.040$$

Reject H_0 at $\alpha = 0.05$

There is statistical evidence of a difference in means.

Is statistical significance enough to meet some journal publication requirements?

- Some journals say **'YES'**
- Some journals say **'NO'**
- Those journals that say **'NO'** want extra information on practical significance of findings
  - One group of journals adopted the American Psychological Association **(APA)** policy.
  - Sawyer, A.G. and Ball, A. D. (1981). Statistical Power and Effect Size in Marketing Research. Journal of Marketing Research, vol XVIII, 275-290.

APA Task force on statistical significance

- Members:
  - Robert Rosenthal, Robert Abelson, & Jacob Cohen
  - Leona Aiken, Mark Appelbaum, Gwyneth Boodoo, David A. Kenny, Helena Kramer, Donald Rubin, Bruce Thompson, Howard Wainer, & Leland Wilkinson.
- Senior Advisors:
  - Lee Cronbach, Paul Meehl, Frederick Mosteller, & John Tukey
- After 2years of meetings came up with APA statements on effect sizes (practical significance)

APA statements
On Hypothesis Testing

- It is hard to imagine a situation in which a dichotomous accept-reject decision is better than reporting an actual *p*-value or, better still, a confidence interval.
- Never use the unfortunate expression "accept the null hypothesis."
- Always provide some effect-size when reporting a *p*-value.
- Cohen (1994) has written on this subject in this journal. All psychologists would benefit from reading his insightful article.

Source: Wilkinson, L. (1999). Statistical Methods in Psychology Journals: Guidelines and Explanations. *American Psychologists*, 54(8), 594-604

APA statements
On Effect Sizes

continued

- Always present effect sizes for primary outcomes.
- If the units of measurement are meaningful on a practical level (e.g., number of cigarettes smoked per day),
  - then we usually prefer an unstandardized measure (regression coefficient or mean difference) to a standardized measure (r or d).
- It helps to add brief comments that place these effects sizes in a practical and theoretical context.

Source: Wilkinson, L. (1999). Statistical Methods in Psychology Journals: Guidelines and Explanations. *American Psychologists*, 54(8), 594-604

APA statements

On Effect Sizes

continued

- APA's (1994) publication manual included an important new "encouragement" (p.18) to report effect sizes.
- Unfortunately, empirical studies of various journals indicate that the effect size of this encouragement has been negligible (Keselman et al., 1998; Kirk, 1996; Thompson & Snyder, 1998).
- We must stress again that reporting and interpreting effect sizes in the context of previously reported effects is essential to good research.
- It enables readers to evaluate the stability results across samples, designs and analyses.
- Reporting effect sizes also informs power analyses and meta-analyses needed in future research.

Source: Wilkinson, L. (1999). Statistical Methods in Psychology Journals: Guidelines and Explanations. *American Psychologists*, 54(8), 594-604

APA statements

On Effect Sizes

continued

- Fleiss (1994), Kirk (1996), Rosenthal (1994), and Snyder and Lawson (1993) have summarized various measures of effects sizes used in psychological research.
- Consult these articles for information on computing them.
- For a simple, general purpose display of the practical meaning of an effect size, see Rosenthal and Rubin (1982).
- Consult Rosenthal and Rubin (1994) for information on the use of "counternull intervals" for effect sizes, as alternatives to confidence intervals.

Source: Wilkinson, L. (1999). Statistical Methods in Psychology Journals: Guidelines and Explanations. *American Psychologists*, 54(8), 594-604

Marketing Research

"Statistical Power and effect size are not considered sufficiently by marketing researchers. The authors discuss how better attention to these two factors can improve the planning, execution, and reporting of marketing and consumer research. Suggestions are offered about how to increase effect size and improve statistical power."

Source:

Sawyer, A.G. and Ball, A. D. (1981). Statistical Power and Effect Size in Marketing Research. *Journal of Marketing Research*, vol XVIII, 275-290.

24 Applied Research Journals now requiring *effect size* reporting

Educational and Psychological Measurement
Educational Technology Research and Development
Journal of Educational Psychology (APA)
Journal of Educational and Psychological Consultation
Journal of Experimental Education
Measurement and Evaluation in Counseling and Development
The Professional Educator
Reading and Writing
Contemporary Educational Psychology
Research in the Schools
Early Childhood Research Quarterly
Health Psychology (APA)
Career Development Quarterly
Exceptional Children
Journal of Agricultural Education
Journal of Applied Psychology
Journal of Community Psychology
Journal of Consulting & Clinical Psychology
J. of Counseling and Development (ACA)
Journal of Experimental Psychology: Applied
Journal of Learning Disabilities
Language Learning
J. of Early Intervention
J. of Personality Assessment

Source: http://www.coe.tamu.edu/~bthompson/index.htm

What is practical significance?

Shows **how practically important** is the findings found in a study.

Illustration.

- **Hubble telescope – NASA scientists in late 1980s thought that their Hubble telescope is calibrated within specifications. So, their telescope focus error was thought to be small.**
- However, a small difference in calibration was important enough to hinder them from detecting new galaxies and making new discoveries.
- When they corrected this calibration error, new discoveries were easily made.

Practical significance index

- Jacob Cohen (1988,1994) introduced some effect size indices to address practical significance
  - For mean differences (2 means)

$$d = \frac{\overline{x}_1 - \overline{x}_2}{\hat{\sigma}}$$

    - similar to mean of Z scores with 0 as center
  - For mean differences (3 or more means, ANOVA setting)

$$\eta^2 = \frac{SS_B}{SS_T}$$

  - For correlation/regression setting

$$r \text{ and } r^2 = \frac{SS_R}{SS_T}$$

- Related to statistical power

Effect sizes in research manuscripts: selecting, calculating, reporting, and interpreting

Ranges of importance for d

Mean differences (2 means) for psychological, educational, & behavioral constructs (in absolute values)

0.2	small
0.5	medium
0.8	large

Source: Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed). Lawrence Erlbaum Associates: New Jersey, pp25-27.

Is it easy to calculate these indices?

- Should be easy.
  - Mean difference (2 means) effect size
    - Need to take this mean difference and divide by pooled standard deviation (or std dev of control group).
  - Mean differences (3 or more means, ANOVA)
    - Need to take sum of squares between means and divide with sum of squares total (without group)
  - Correlation & Regression analyses
    - Use r &
    - r^2=% variation explained by regression.

Reconsider Example: Pooled s_p t Test

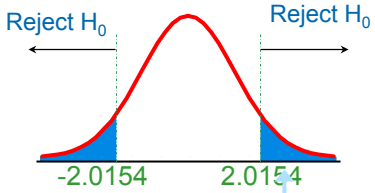
Is there a difference in dividend yield between stocks listed on the NYSE & NASDAQ? A financial analyst for a brokerage firm collected the following data:

	NYSE	NASDAQ
Number	21	25
Sample mean	3.27	2.53
Sample std dev	1.30	1.16

Assuming equal variances, is there a difference in average yield (α = 0.05)?

Results were statistically significant. But, what is the size of the effect?

H_0: $\mu_1 - \mu_2 = 0$ i.e. ($\mu_1 = \mu_2$)
H_A: $\mu_1 - \mu_2 \neq 0$ i.e. ($\mu_1 \neq \mu_2$)
α = 0.05
df = 21 + 25 - 2 = 44
Critical Values: t = ± 2.0154



2.040

Test Statistic:

$$t = \frac{3.27 - 2.53}{1.2256\sqrt{\frac{1}{21} + \frac{1}{25}}} = 2.040$$

Reject H_0 at α = 0.05
There is statistical evidence of a difference in means.

d =(3.27-2.53)/1.2256 = 0.604
Medium effect size

An Example of practical significance index in APA journal articles

- ***Custer, M., Omar, M. H. & Pomplun, M.* (2006).** Vertical Scaling With the Rasch Model Utilizing Default and Tight Convergence Settings With WINSTEPS and BILOG–MG. Applied Measurement in Education 19(2), 133-149.
  - Item Response Theory (IRT) Vertical Scaling Paper
    - Grades (K to 10) estimation of student ability mean and variances
    - Simulated abilities were done mimicking properties of the California Achievement Test (CAT) Norms
      - Simulation was done with SAS 8.0 at Riverside Publishing Company
      - with Normal ability vectors and Skewed ability distribution vectors (skewness as defined by the CAT ability distributions at different grades)
      - with Item parameter matrix following the difficulties on the CAT Norms
  - Rasch model
  - Estimation of item parameter and person ability parameter vectors can be achieved using two well-known IRT estimation techniques

An Example of practical significance index in APA journal articles *Continued.*

- ***Custer, M., Omar, M. H. & Pomplun, M.* (2006).** Vertical Scaling With the Rasch Model Utilizing Default and Tight Convergence Settings With WINSTEPS and BILOG–MG. Applied Measurement in Education 19(2), 133-149.
  - Item Response Theory (IRT) Vertical Scaling Paper
  - Rasch model
  - Estimation of item parameter and person ability parameter vectors can be achieved using two well-known IRT estimation techniques
    - Marginal Maximum Likelihood Estimation (MMLE)
    - Unconditional Maximum Likelihood Estimation (UCON) – utilizing person raw scores as sufficient statistics
  - No closed form solution exists for item and person ability parameters
    - Instead, numerical analyses and numerical integration algorithms are used to estimate these parameters
    - So, computer softwares are used to estimate these parameters
  - Results from WINSTEP (UnConditional MLE) and BILOG-MG (Marginal MLE) were compared
- Usual simulation results were tabulated and presented in the paper
- In addition, publication required **effect sizes** to show any practical importance of differences. Thus, these were also calculated in the paper.


Thank You
Thank You
THANK YOU