Hypotheses, Effect Size, Significance and Power
OUTLINE

- **Hypotheses**
  + Defining Hypotheses
  + Core logic of Hypotheses Testing
  + One- vs. Two-tailed Hypotheses
  + Controversies in Null Hypothesis Significance Testing

- **Effect Size**
  + Definition of effect size
  + Measures of effect size
  + Defining ‘small’, ‘medium’ and ‘large’ effect sizes
OUTLINE

✶ Significance
  + Defining significance
  + What does statistical significance really tell us?
  + Use of significance in practice
  + Common Pitfalls
  + Signal-to-noise ratio conceptualisation of significance
  + Alternatives to significance testing

✶ Power
  + What is the power of a test
  + Type II error and statistical power
  + Using a priori power calculations to determine sample size
  + Components of a statistical power analysis
HYPOTHESIS TESTING
HYPOTHESIS TESTING (1)

DEFINING A HYPOTHESIS

- A hypothesis is a specific conjecture about a property of a population of interest.
- It should be interesting and deserve testing.
- The null hypothesis (H₀) is a specific baseline statement and generally takes the form of “no effect” or “no difference.”
- The alternative (or research; H₁) hypothesis is the denial of the null hypothesis.
HYPOTHESIS TESTING (2)

What makes a good hypothesis?

(1) A hypothesis should be specific enough to be falsifiable (otherwise the hypothesis cannot be tested successfully)

(2) A hypothesis should be a conjecture about a population (parameter) and not a sample (statistic)

(3) A valid hypothesis is not based on the sample to be used to test the hypothesis
HYPOTHESIS TESTING (3)

Example:

Distribution of when babies begin to walk

[Graph showing age (months) and Z scores]

Population $M = 14$
Population $SD = 3$
THE CORE LOGIC OF HYPOTHESIS TESTING

We determine whether or not our manipulation had an effect by first of rejecting the notion that it did not have an effect (i.e. $H_0$)
HYPOTHESIS TESTING (5)

The hypothesis testing process
(1) Restate the question as a research & null hypotheses
(2) Determine the characteristics of the comparison distribution
(3) Determine the cut-off sample score on the comparison distribution at which the null hypothesis should be rejected
(4) Determine your sample score on the distribution
(5) Decide whether or not to reject the null hypothesis
HYPOTHESIS TESTING (6)

1. State the null & alternative hypotheses
2. Determine the test size (significance level)
3. Compute the test statistics & p-value or construct the confidence interval
4. Reject or do no reject the null hypothesis
5. Draw conclusion & interpret
# HYPOTHESIS TESTING (6)

Three Approaches to Hypothesis Testing

<table>
<thead>
<tr>
<th>Step</th>
<th>Test Statistic Approach</th>
<th>p-value approach</th>
<th>Confidence Interval Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>State $H_0$ and $H_1$</td>
<td>State $H_0$ and $H_1$</td>
<td>State $H_0$ and $H_1$</td>
</tr>
<tr>
<td>2</td>
<td>Determine test size $\alpha$ and find the critical value</td>
<td>Determine test size $\alpha$</td>
<td>Determine test size $\alpha$ or $1 - \alpha$, and a hypothesised value</td>
</tr>
<tr>
<td>3</td>
<td>Compute test statistic</td>
<td>Compute a test statistic and its $p$ value</td>
<td>Construct the $(1 - \alpha)$ 100% confidence interval</td>
</tr>
<tr>
<td>4</td>
<td>Reject $H_0$ if $TS &gt; CV$</td>
<td>Reject $H_0$ if $p &lt; \alpha$</td>
<td>Reject $H_0$ if a hypothesised value does not exist in CI</td>
</tr>
<tr>
<td>5</td>
<td>Substantive interpretation</td>
<td>Substantive interpretation</td>
<td>Substantive interpretation</td>
</tr>
</tbody>
</table>
HYPOTHESIS TESTING (7)

Directional (one-tailed) vs. Non-directional (two-tailed hypotheses

+ One-tailed: $H_0: \mu \geq 0$ or $H_0: \mu \leq 0$
+ Two-tailed: $H^0: \mu = 0$
Controversies in Null Hypothesis Significance Testing (NHST)

Rozeboom (1997): NHST is “surely the most boneheadedly, misguided procedure ever institutionalized in the rote training of science students”
HYPOTHESIS TESTING (9)

Bayes Theorem

\[ p(\text{Null} | \text{Data}) = \frac{p(\text{Data} | \text{Null}) \cdot p(\text{Null})}{p(\text{Data})} \]
EFFECT SIZE
EFFECT SIZE (1)

The measure of the strength of the relationship between two variables in a statistical population, or a sample based estimate of that quantity.
Effect size can refer to a statistic calculated from a sample of data OR a parameter of a hypothetical statistical population.

Effect size can refer to a standardised measures of effect (e.g. Cohen’s $d$) or to an unstandardised measure (e.g. the raw difference between group means).

Reporting effect size is considered good statistical practice.
(1) Pearson r correlation

Used as an effect size when paired quantitative data are available (e.g. birth weight & longevity).

Can also be used when the data are binary

Will vary in magnitude from -1 to 1, with -1 indicating a perfect negative linear relationship and 1 indicating a perfect positive linear relationship; 0 indicates no relationship

Cohen’s guidelines for the social sciences:

\[ r = 0.1 - 0.23 \rightarrow \text{small effect size} \]
\[ r = 0.24 - 0.36 \rightarrow \text{medium effect size} \]
\[ r > 0.37 \rightarrow \text{large effect size} \]
EFFECT SIZE (4)

Effect sizes based on means:

\[ \theta = \frac{\mu_1 - \mu_2}{\sigma} \]

Cohen’s \( d \): the difference between two means divided by the standard deviation for the data

\[ d = \frac{\bar{x}_1 - \bar{x}_2}{s} \]
EFFECT SIZE (5)

Glass’s $\Delta$

$$\Delta = \frac{\bar{X}_1 - \bar{X}_2}{s_2}$$

Hedges’ $g$

$$g = \frac{\bar{X}_1 - \bar{X}_2}{s^*}$$
EFFECT SIZE (6)

Distribution of effect sizes based on means

- If data has Gaussian distribution a scaled Hedge’s g follows a noncentral t distribution, with $n_1 + n_2 - 2$ degrees of freedom
- Likewise, Glass’ $\Delta$ is distributed with $n_2 - 1$ d.f.
- From the distribution it is possible to compute the expectation & variance of effect sizes
- For large samples approximations for the variance are used, e.g.

$$\theta^2(g^*) = \left( \frac{n_1 + n_2}{n_1 n_2} \right) + \left( \frac{(g^*)^2}{2(n_1 + n_2)} \right)$$
EFFECT SIZE (7)

Cohen’s $f^2$

An appropriate effect size measure to use in the context of an F-test for ANOVA or multiple regression

**Multiple Regression**

\[ f^2 = \frac{R^2}{1 - R^2} \]

**Hierarchical Multiple Regression**

\[ f^2 = \frac{R_{AB}^2 - R_A^2}{1 - R_{AB}^2} \]

**Factorial Analysis of Variance**

\[ f_{\text{EFFECT}} = \sqrt{\left(\frac{df_{\text{EFFECT}}}{N}\right)(F_{\text{EFFECT}} - 1)} \]

For a balanced design, the corresponding population parameter of $f^2$ is:

\[ SS \left( \mu_1, \mu_2, \ldots, \mu_K \right) \]

\[ \frac{K \times \sigma^2}{K} \]
**EFFECT SIZE (8)**

\( \phi \), Cramer’s \( \phi \), or Cramer’s \( V \)

The best measure of association for chi-square test.
Related to the point biserial correlation co-efficient and Cohen’s \( d \).
Estimates the extent of the relationship between two variables (2 x 2).
Cramer’s \( \psi \) can be used with variables with more than two levels.

\[
\phi = \sqrt{\frac{\chi^2}{N}}
\]

Cramer’s \( \phi \)

\[
\phi_c = \sqrt{\frac{\chi^2}{N(k-1)}}
\]
EFFECT SIZE (9)

Odds Ratio
- Appropriate when both variables are binary
- Effect size is computed by noting the relative odds of one condition/group/treatment vs. the other
- Favoured for case-control and retrospective studies

Relative Risk (or risk ratio)
- The risk (probability) of an event relative to some independent variable.
- Differs from the odds ratio in that it compares probabilities rather than odds.
- Effect size is computed in the same way as in the odds ratio, but using probabilities instead.
- Favoured for randomised controlled trials and cohort studies.
Defining Effect Size

- ‘Small’, ‘medium’, or ‘large’ effect size (a.k.a ‘T-shirt’ effect sizes) designation depends on substantial context and operational definition
- Cohen’s conventional criterions are near ubiquitous across fields
- Power analysis or sample size estimation require an assumed population parameter of effect size and many adopt Cohen’s standard as default
- Small, medium, and large are relative not only to each other but to the area of science, the specific content and research method being employed.
STATISTICAL SIGNIFICANCE
Statistical significance ≠ “significance”

Significance level (critical p-value) = the amount of evidence required to accept that an event is unlikely to have arisen by chance

Critical p-value = \(\alpha\) = the probability of deciding to reject \(H_0\)

When the obtained or calculated p-value is < \(\alpha\), then we reject the null hypothesis.

The smaller the value of \(\alpha\) the greater confidence is in the estimation of statistical significance but the risk of Type II error is also greater.
Selection of $\alpha$ inevitably involves a compromise between power & significance.

Statistical significance is often expressed as $1 - \alpha$

Simply, the significance level is defined as the probability that a decision to reject $H_0$ will be made when it is in fact true (Type I error)

The probability of rejecting $H_0$ is no more than the stated probability
Common Pitfalls

- A statistically significant result is of practical significance
- There is a tendency for multiple comparisons to yield spurious significant differences
  + In such case p-value is adjusted in order to control for the FAMILYWISE ERROR RATE or the FALSE DISCOVERY RATE
    \[
    \text{FAMILYWISE ERROR} = 1 - (0.95)^n \quad n = \text{number of tests carried out on the data}
    \]
- Frequentist analyses of p-values are considered by some to overstate statistical significance
- No statistically significant difference ≠ No difference
STATISTICAL SIGNIFICANCE (4)

Signal-to-noise conceptualisation of significance

\[
\text{confidence} = \left( \frac{\text{signal}}{\text{noise}} \right) \times \sqrt{\text{sample size}}
\]

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Parameter Increases</th>
<th>Parameter Decreases</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOISE</td>
<td>Confidence decreases</td>
<td>Confidence increases</td>
</tr>
<tr>
<td>SIGNAL</td>
<td>Confidence increases</td>
<td>Confidence decreases</td>
</tr>
<tr>
<td>SAMPLE SIZE</td>
<td>Confidence increases</td>
<td>Confidence decreases</td>
</tr>
</tbody>
</table>
STATISTICAL SIGNIFICANCE (5)

Alternatives to significance testing

Calculation of effect sizes

The empirical investigation of the replicability of results

  e.g. actual replication; cross-validation; jack-knife; bootstrap
What is the *power* of a test?

- The probability that it will correctly reject a false null hypothesis (Greene 2000)
- The ability of a test to detect an effect, if that effect actually exists (High 2000)
- The probability that the analysis of a true effect will result in the conclusion that the phenomenon exists (Cohen 1988)
Type II error and statistical power

- Type II error ➔ failure to reject the null hypothesis when it is false (wrongly accepting your null hypothesis/ rejecting your research hypothesis)
- As power increases the chances of making a Type II error decrease
- The probability of a Type II error ➔ false negative rate (\( \beta \))
- Therefore: \( \text{POWER} = 1 - \beta \)
POWER (2B)
POWER (3)

Statistical power explore the relationship between four components (Mazen, Hemmasai, Lewis 1985)

(1) Standardized effect size: (a) effect size & (2) variation
(2) Sample size (N)
(3) Test size (significance level $\alpha$)
(4) Power of the test ($1 - \beta$)
Components of a statistical power analysis

- Standardised Effect Size
- Test Size
- Sample Size
- Power

- Positive (+) relationship between Standardised Effect Size and Test Size
- Positive (+) relationship between Test Size and Sample Size
- Positive (+) relationship between Sample Size and Power
- Positive (+) relationship between Standardised Effect Size and Sample Size
- Positive (+) relationship between Test Size and Power
- Positive (+) relationship between Sample Size and Power
- Positive (+) relationship between Standardised Effect Size and Power
- Positive (+) relationship between Test Size and Power
POWER (5)

- *A priori* vs. *post hoc* power calculations
  - *A priori* = prospective: sample size determinations
  - *Post hoc* = retrospective: observed power

- Controversies in *post hoc* power calculations
  - Useful or useless?
  - When to use retrospective power calculations?
    - Evaluation of the ability of a study to determine whether a pre-specified target has been met
    - Making comparisons between different studies

- All retrospective power calculations should be accompanied by a sensitivity analysis
  - *How can the variation in the output of a mathematical model be apportioned to different sources of variation in the input of a model?*

- Retrospective power calculations is no substitute for the proper planning of research.
Power calculation tools:

http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/PowerSampleSize
  (software download)

http://www.danielsoper.com/statcalc/
  (number of free online statistical calculators; a priori & post hoc calculators for multiple and hierarchical regression and t-test)

http://statspages.org
  (links to hundreds of pages that perform statistical calculations; including power, sample size and experimental design calculators for one- & two-sample tests, survey design, ANOVAs, multiple regression and others)