Statistics refresher seminar series

How much data should I collect?

13-Jun-2019

Kim Colyvas

statsupport@newcastle.edu.au

Our website, search for StatSS on university web site
How To Analyse My Data
3- 5 July 2019

Outlines
• Exploratory data analysis and visualising data
• Formulating research questions
• Data types and related statistical tests
• How to interpret statistical results

♦ Explanation of common statistical tests
♦ Workbook with worked examples then hands on practice
♦ Use statistical software to create output (SPSS)
♦ SPSS software guide provided
♦ Focus on understanding, concepts and interpretation of results

Instructors
Nic Croce, Fran Baker

Statistical Support Service
Notes for all seminars can be downloaded from the **Courses, Seminars and Workshops** section at


Easier however is to type **StatSS** into the university web site’s search box. Our site is the first result in the list – choose the heading **Courses, seminars and workshops heading.**
Intent of this session

• What information is needed to determine sample size for a study.
• How this information is used.
• Interpreting the results of a sample size determination.
• Understanding effect sizes.
• Not how to do sample size and power calculation (but will get some idea).
Variable types

• **Numeric** – values that “mean numbers”
  – *Continuous*: temperature, weight, speed, distance
  – *Discrete*: #defects, result of die toss, product count

• **Categorical** – values based on categories
  – *Nominal*
    gender – male/female  colour - blue/green/yellow
  – *Ordinal*
    Grades - FF, P, C, D, HD,
    Temperature - Low, Medium, High
<table>
<thead>
<tr>
<th>Response</th>
<th>Explanatory</th>
<th>Specific question(s)</th>
<th>Displays</th>
<th>Statistical method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categorical</td>
<td>Categorical</td>
<td>How do proportions in response depend on the levels of the explanatory variable?</td>
<td>Tables</td>
<td>Chi-squared statistic</td>
</tr>
<tr>
<td>Categorical</td>
<td>Numeric (Continuous)</td>
<td>How does the proportion in response depend on the explanatory variable?</td>
<td>Tables (X groups)</td>
<td>Logistic regression</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Correlation (for a binary response only)</td>
</tr>
<tr>
<td>Numeric (Continuous)</td>
<td>Categorical</td>
<td>How does mean level in response change with the levels of the explanatory variable? If so how does it vary?</td>
<td>Box plots Mean plots CI plots</td>
<td>t test (2 groups)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ANOVA</td>
</tr>
<tr>
<td></td>
<td>Numeric (Continuous)</td>
<td>How does mean level of response change with the explanatory variable</td>
<td>Scatter plots</td>
<td>Correlation Regression</td>
</tr>
<tr>
<td>Numeric (Continuous)</td>
<td>Numeric (Continuous)</td>
<td>How does mean level in response change with the levels of the explanatory variable</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Todays focus will be on the 2 research questions in red**

- Kappa (2x2 - Agreement)
- McNemar’s test (2x2 - bias in agreement)
- Paired t test
- Repeated measures
- ANOVA
Differences between 2 groups

Purpose: Test differences between 2 treatments, genders etc

- **Outcome is categorical**
  
  *Increase awareness of service following training intervention from 35% to 54%.*

- **Outcome is numeric**
  
  *Improvement in pain index after treatment was 17.3.*
Statistical significance & Practical significance

• **Statistical significance**
  A statistical test is carried out and we find the difference is significant based say on a p value.

• **Practical significance**
  Whether the difference is meaningful within our field of study.

• It is easy with large sample sizes to obtain statistically significant differences that are not meaningful.

• This session is concerned with designing studies to find **practically significant** effects, i.e. important clinically, biologically, environmentally, socially etc.
My Study – most important variables

1) **Response Type:**

2) **Explanatory Type:**
   Numbers of levels for each if categorical

3) **Dependent/Independent:**

4) **Practical significance**
   What size change is important?
   - Previous experience or research
   - Don’t know, use Cohen’s effect size (see later)

5) **How large is the variability?**
   Prior information, prior research, guess,
   Don’t know, use Cohen’s effect size (see later)
Break in lecture
for ~10 mins
for class exercise
Sample size estimation

• We will not be using formulae.
• Free software is available.
• The demonstrations in this talk use the **Power and Sample Size** program.
• See first reference on last slide for link to download.
• Also G*Power 3 is an alternative free program.
Sample Size and Power analysis

• Key idea is to know before a study what is the chance you will discover something.
• Sample size is a key driver of this.
• Can save wasted effort and disappointment if proper planning is carried out prior to the study.
Statistical Concepts for sample size

- **Type I Error** – alpha (typically $\alpha = 0.05$)
  Chance that an effect will be declared as real when in reality there is no effect.

- **Type II Error** – beta (typically $\beta=0.20$)
  Chance a real effect **WILL NOT** be detected.

- **Power** $(1-\beta)$ (typically 0.8)
  Chance a real effect **WILL** be detected.
  Power of 0.8 means we have an 8 in 10 chance of detecting the effect.
  2 in 10 chance of NOT detecting the effect.\textsuperscript{13}
Categorical outcome – difference of 2 proportions

• Difference between 2 groups, propose that Control Group = 35%, Treatment group = 54%

Sample size (n) required in each group ~ 115, ie total N = 230
What is n for 35% vs 70%?
What is n for 35% vs 18%?
Program setup
First screen for previous slide then click Graphs button

Control group
Treatment group
=1 means equal size groups
Program setup - second screen

This is used to create the graphical output two slides before
Sensitivity analysis

• Very useful not to do just a single sample size calculation.

• Try a range of options to get a feel for how they might impact the effectiveness of your planned study.

• How would the power of your study be affected by different sample sizes? For example loss to follow-up.

• How would sample size change for other sizes of practical significance?
Varying sample size

• As the size of the **difference** changes the effect on the power of the test is shown below.

• Or for a constant sample size the power changes
Numeric outcome difference between 2 means

- Mean difference between 2 independent groups
  103.6 Low group, 96.4 in High group, Diff = 7.1

Sample size (n) required in each group ~ 65, ie total N = 130
What is n for difference = 4?
Effect on sample size if variability (σ) was larger?
Effect size (ratio of signal/noise)

- See Cohen references

\[ d = \frac{(mean_1 - mean_2)}{SD} \]

Change in means. What is the **practical/clinical significance** of this?

Variability within each group controls (or limits) the ability to detect a difference.
Sample sizes for means

For a single (1) sample compared to a reference value.
For two (2) samples, between two groups for example.
(Howell 2002)

<table>
<thead>
<tr>
<th>Number of samples</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small effect</td>
<td>d = 0.2</td>
<td>196</td>
</tr>
<tr>
<td>Medium effect</td>
<td>d = 0.5</td>
<td>32</td>
</tr>
<tr>
<td>Large effect</td>
<td>d = 0.8</td>
<td>13</td>
</tr>
</tbody>
</table>

$\alpha = 0.05$, power = 0.8
Other effect size measures

Table 1
ES Indexes and Their Values for Small, Medium, and Large Effects

<table>
<thead>
<tr>
<th>Test</th>
<th>ES index</th>
<th>Effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Small</td>
</tr>
<tr>
<td>1. $m_A$ vs. $m_B$ for independent means</td>
<td>$d = \frac{m_A - m_B}{\sigma}$</td>
<td>.20</td>
</tr>
<tr>
<td>2. Significance of product-moment $r$</td>
<td>$r$</td>
<td>.10</td>
</tr>
<tr>
<td>3. $r_A$ vs. $r_B$ for independent $r$s</td>
<td>$q = z_A - z_B$ where $z$ = Fisher’s $z$</td>
<td>.10</td>
</tr>
<tr>
<td>4. $P = .5$ and the sign test</td>
<td>$g = P - .50$</td>
<td>.05</td>
</tr>
<tr>
<td>5. $P_A$ vs. $P_B$ for independent proportions</td>
<td>$h = \phi_A - \phi_B$ where $\phi$ = arcsine transformation</td>
<td>.20</td>
</tr>
<tr>
<td>6. Chi-square for goodness of fit and contingency</td>
<td>$w = \sqrt{\sum_{i=1}^{k} \frac{(P_{1i} - P_{0i})^2}{P_{0i}}}$</td>
<td>.10</td>
</tr>
<tr>
<td>7. One-way analysis of variance</td>
<td>$f = \frac{\sigma_m}{\sigma}$</td>
<td>.10</td>
</tr>
<tr>
<td>8. Multiple and multiple partial correlation</td>
<td>$f^2 = \frac{R^2}{1 - R^2}$</td>
<td>.02</td>
</tr>
</tbody>
</table>

Note. ES = population effect size.

From Cohen, J., A power primer, Psychological Bulletin, 1992, 112(1), 155-159
Numeric variables for best power

• If you can collect data on a variable in the numeric form rather than as categorical you will have greater power.

• There might be other considerations that require a categorical form, but if possible use the numeric.
Numeric variables for best power (2)

- **Apparently** easier interpretation of results is the wrong reason for making categories.
- The results that follow illustrate the extent of the loss in power if continuous variables are converted to categorical.
• Can treat both variables as numeric
• Y as numeric, X as categorical
• Both X and Y as categorical

Numeric vs categorical variables

\[ r = -0.50 \]

Graph showing the distribution of data points for numeric vs categorical variables with an r-value of -0.50.
Numeric vs categorical variables (2)

- Converting the numeric variables to categorical variables leads to the following tables.

<table>
<thead>
<tr>
<th>Effect size</th>
<th>r</th>
<th>X=Low</th>
<th>X=High</th>
<th>Diff</th>
<th>SD(Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large</td>
<td>-0.5</td>
<td>106.0</td>
<td>94.0</td>
<td>12.0</td>
<td>13.8</td>
</tr>
<tr>
<td>Medium</td>
<td>-0.3</td>
<td>103.6</td>
<td>96.4</td>
<td>7.1</td>
<td>14.6</td>
</tr>
<tr>
<td>Small</td>
<td>-0.1</td>
<td>101.2</td>
<td>98.8</td>
<td>2.3</td>
<td>14.95</td>
</tr>
</tbody>
</table>

**Percentages for Y=High**

<table>
<thead>
<tr>
<th>Effect size</th>
<th>r</th>
<th>X=Low</th>
<th>X=High</th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large</td>
<td>-0.5</td>
<td>61.3%</td>
<td>28.1%</td>
<td>33.2%</td>
</tr>
<tr>
<td>Medium</td>
<td>-0.3</td>
<td>54.3%</td>
<td>35.1%</td>
<td>19.2%</td>
</tr>
<tr>
<td>Small</td>
<td>-0.1</td>
<td>47.7%</td>
<td>41.7%</td>
<td>6.1%</td>
</tr>
</tbody>
</table>

This detail is provided for the interested reader and can be skipped.

Both categorical
## Numeric vs categorical variables (3)

<table>
<thead>
<tr>
<th>Effect size</th>
<th>r</th>
<th>Sample size (N) total of both groups</th>
<th>Both Numeric</th>
<th>Y - numeric X Categorical</th>
<th>Both Categorical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large</td>
<td>-0.50</td>
<td>30</td>
<td>44</td>
<td>68</td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>-0.30</td>
<td>85</td>
<td>134</td>
<td>208</td>
<td></td>
</tr>
<tr>
<td>Small</td>
<td>-0.10</td>
<td>784</td>
<td>1328</td>
<td>2154</td>
<td></td>
</tr>
</tbody>
</table>

Statistical test correlation coef.  t-test  chi-squared

- Converting one numeric variable to categorical sample size increases range from 45% to 70% (Large to small effect size)
- Converting both numeric variables to categorical sample size increases range from 125% to 175%.

Effect size labels, small, medium and large using Cohen’s criteria – earlier slides

r is Pearson correlation coefficient
Relationship between 3 variables

1 Response
2 Explanatory
Case study: Proposed design
Pre-Post/Control-Treatment

- This is a common high quality study design
- Some literature data available for variability but only for no intervention case

Re effect of Gastric banding on bone growth

SD’s from Table 2 at each time period range from 0.12 to 0.14
Pre, post data are dependent
Control, Treatment are independent

- **Response:** Serum concentration of enzyme
- **Treatment:** Fortified dairy products.
- **Control:** Normal dairy products.

- What is a clinically significant improvement?
- Do better than control group by at least 20%.
- Assume control does not worsen.
- Treatment reduced by 20%.
- Need to know correlation between pre and post scores, or SD of differences.
Scenario 1

- 20% reduction for treatment compared to control

<table>
<thead>
<tr>
<th></th>
<th>Pre</th>
<th>Post</th>
<th>(Pre-Post)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>0.16</td>
<td>0.16</td>
<td>0</td>
</tr>
<tr>
<td>Treatment</td>
<td>0.16</td>
<td>0.128</td>
<td>0.032</td>
</tr>
</tbody>
</table>

Difference (Treatment - Control) 0.032

- SD = 0.12

- Correlation between pre & post scores = 0.50
  (guess, conservative, not available from paper)

- n=220 in each group, too much work!
Scenario 2

• Fixed sample size, n=50, all that can be handled

<table>
<thead>
<tr>
<th></th>
<th>Pre</th>
<th>Post</th>
<th>(Pre-Post)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>0.16</td>
<td>0.16</td>
<td>0</td>
</tr>
<tr>
<td>Treatment</td>
<td>0.16</td>
<td>0.085</td>
<td>0.075</td>
</tr>
<tr>
<td>Difference (Treatment - Control)</td>
<td></td>
<td></td>
<td>0.075</td>
</tr>
</tbody>
</table>

• SD = 0.14
  correlation between pre & post scores = 0.50

• Difference of .075/.16 =
  Only 47% reduction is achievable.

• Both cases were with power = 0.80, type 1 error = .05
Result of sample size analysis

• Desired practical significance to detect 20% change requires more resources than we can afford (n=220 in each group, total N=440).

• Alternative based on resources we can afford (n=50 in each group, total N=100). Can only detect a large change 47%.

• What do you do? Only you and your supervisor can answer that question.

• But at least you know you have an issue to solve!
Which test should I use?

<table>
<thead>
<tr>
<th>Test</th>
<th>Response</th>
<th>Explanatory</th>
<th>This lecture</th>
<th>P &amp;S software</th>
</tr>
</thead>
<tbody>
<tr>
<td>m&lt;sub&gt;a&lt;/sub&gt; vs. m&lt;sub&gt;b&lt;/sub&gt; for independent samples</td>
<td>Continuous</td>
<td>Categorical 2 levels</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Correlation (r) = 0</td>
<td>Continuous</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Independent r&lt;sub&gt;a&lt;/sub&gt; vs. r&lt;sub&gt;b&lt;/sub&gt;</td>
<td>Continuous</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sign test (P = 0.5)</td>
<td>Categorical</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Independent P&lt;sub&gt;a&lt;/sub&gt; vs. P&lt;sub&gt;b&lt;/sub&gt;</td>
<td>2 categories</td>
<td>2 categories</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>χ&lt;sup&gt;2&lt;/sup&gt; test</td>
<td>2 or more categories</td>
<td>2 or more categories</td>
<td></td>
<td></td>
</tr>
<tr>
<td>One-way ANOVA (Between Subjects)</td>
<td>Continuous</td>
<td>3 or more categories</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regression</td>
<td>Continuous</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>

Pre, Post measures with Treatment & Control groups ✓
Sample Size and Power Calculation

References

- Power and Sample Size program, Dupont WD and Plummer WD: PS power and sample size program available for free on the Internet. Controlled Clin Trials, 1997;18:274
  http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/PowerSampleSize
- G*Power 3 – free on the Internet – wider range of calculations than Power and Sample size, most suited to social sciences, strong Psychology basis
  http://www.gpower.hhu.de
  (Continuous variables only)
  (Overview of a range of methods – full text available on-line as PDF)
  Confidence intervals – the better direction.