Inf1B: Data and Analysis Experimental Design and Statistics

Contents

1. Statistical tests for Confirmatory Experiments
2. Summative Evaluation of Standup
3. Writing up Experiments and Empirical Studies
4. References

1. Statistical tests for Confirmatory Experiments

Maths teaching: Classroom v. ITS

Experimental Hypothesis: ***teaching Maths using an ITS increases Maths performance***
Measure: ***performance on Higher Exam***

Compare distributions of:

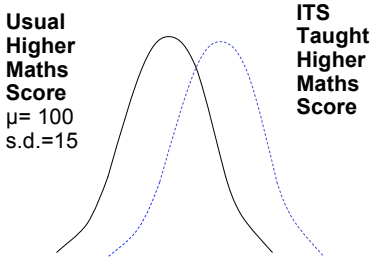
1. Usual Higher Exam Score Distribution
2. ITS taught students Score Distribution

Assume:

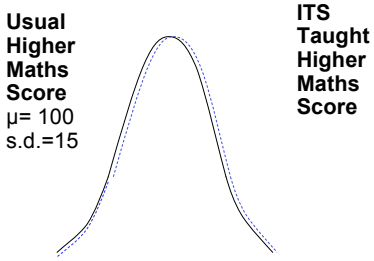
- Higher Exam Score Normally distributed
- Test is designed to ensure normal distribution

Null Hypothesis: ***ITS taught score distribution will be the same as the normal score distribution***

Score Distributions



Null Hypothesis



Example Student: Peter

Peter: Taught by ITS
Does his test score come from the normal distribution?
1. Test him = 120
2. Find position in the distribution:
3. Find Z score = x - μ(mean)/σ (standard deviation)
= 120-100/15 = 1.33
4. Look this up in tables: 1.33 > probability of 0.0918

So 9.18% of the usual-score population score higher than Peter - not very convincing….

What is he scored 145 instead?
Z = 3, p = 0.0013 of the score higher…
So, 0.13% (1 in 769) would score higher
-> so rare that it seems likely this comes from a different population
-> Reject the null hypothesis

Rejecting the Null Hypothesis

What do we set as a cut-off for this?
- Reject if the probability is less than or equal to 5% or 0.05
- Write as **"p < 0.05" or "significant at p = 0.05"**

Means that:
when the score from the unknown distribution could only arise from the known one
(i.e have the same distribution)
with a less then 5 from 100 chance or less,
we reject the null hypothesis,
and say the score is from a different distribution.
p of 0.05 = significance level

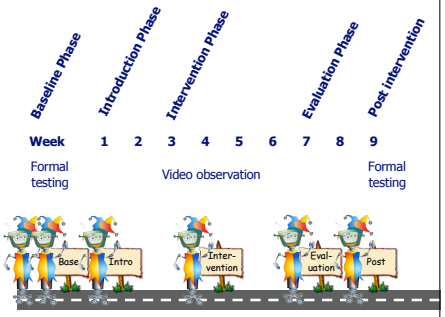
Significance Level of 0.05

This means that there is a greater than 95% chance we are correct in accepting a different distribution
[in only 5% of extreme cases would this score arise by chance]
To be more certain, we may take the 0.01 level (1% chance)
[99% confident in claim of differences]

All statistical tests follow this basic logic:
- Research hypothesis predicts a difference in distribution
- Null hypothesis predicts no difference

2. Summative Evaluation of Standup

Research Methodology



Evaluation Instruments

CELF Clinical Evaluation of Language Fundamentals (Semel, Wiig, Secord, 1995)
- **CELF Linguistic concepts** (participants are asked to point to…: "the blue line", "the line that is not yellow"; participants must point to a stop sign if they think they cannot do what they are asked to do.)
- **CELF Sentence structure** (e.g. show me…: "The girl is not climbing", "The dog that is wearing a collar is eating a bone")
- **CELF Oral directions** (e.g. point to…: "The black circle", "The last white triangle and the first black square")
- **CELF Word classes** (participants choose two related items from a set of four, e.g. "**girl boy** car table", "slow **nurse doctor** rain")

PIPA Preschool and primary inventory of phonological awareness (Frederickson, Frith and Reason, 1997)

Evaluation Instruments: The KMT

Keyword Manipulation Task (O'Mara, 2005): standardised across 57 children, including language impaired children; 5 – 12 years.

Stimulus: *How can you tell there has been an elephant in your fridge?*
Footprints in the butter.

Keyword Alternates:
Mouse. Giraffe. Cat. Rabbit.

Stimulus: *What do you get when you cross a car and a sandwich?*
A traffic-jam.

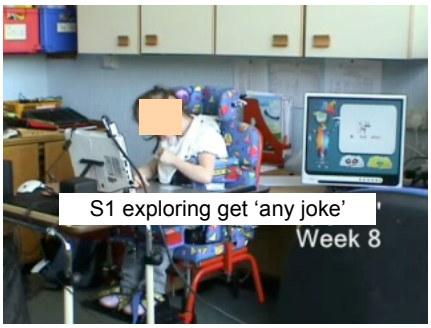
Keyword Alternates:
Bicycle. Plane. Train. Truck.

Participants

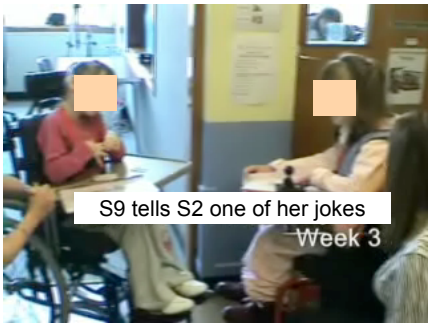
Level	Participant	Communication	Head switch	Direct access
Early Primary	S1, female; age: 8y4m	Dynavox DV4 user: PCS	✓	
Middle Primary	S2, female; age: 10y10m	Intelligible speech: poor articulation		✓
	S3, female; age: 10y9m	Communication book: gross fist & eye gaze	✓	
	S4, male; age: 10y3m	Communication Board: PCS, TechSpeak		✓
	S5, male; age: 10y3m	Clear speech		✓
Senior Primary	S6, male; age: 11y3m	Dynavox DV4 user: PCS	✓	
	S7, male; age: 12y9m	Speech: poor intelligibility uses PCS	✓	
	S8, male; age: 11y10m	Dynavox DV4 user: PCS		✓
	S9, female; age: 11y3m	Intelligible speech		✓

For all participants: Aetiology: Cerebral Palsy
Mobility: Wheelchair
Literacy: Emerging and assisted

STANDUP in use 1



STANDUP in use 2



Preliminary Results: Pre/Post Testing

		CELF Word Classes (out of 27)		PIPA Rhyme (out of 12)	
		Pre-test	Post-test	Pre-test	Post-test
Early Primary	S1, female; age: 8y4m	19	25	10	11
Middle Primary	S2, female; age: 10y10m	11	18	3	3
	S3, female; age: 10y9m	23	26	11	11
	S4, male; age: 10y3m	0	2	10	9
	S5, male; age: 10y3m	17	26	11	11
Senior Primary	S6, male; age: 11y3m	1	4	1	8
	S7, male; age: 12y9m	17	24	12	11
	S8, male; age: 11y10m	9	8	5	3
	S9, female; age: 11y3m	12	13	10	11

CELF WC: choose 2 related items from set of 4, e.g. "girl boy car table"
PIPA Rhyme: Phonological awareness

Statistical Comparison: T-test

The t-test assesses whether the means of two groups are statistically different from each other, assuming that paired differences are independent and normally distributed.
Given two paired sets Xi and Yi of n measured values:

$t = (\text{meanX} - \text{meanY}) \times \text{sqrt}\left[(n(n-1)) \; / \; \sum((X'i - Y'i)^2)\right]$

Where $X'i = (Xi - \text{meanX})$ $Y'i = (Yi\_ - \text{meanY})$

Statistical Comparison: T-test Performance on Pipa Test

Pre-intervention:
Mean = 12.1 Standard Deviation = 7.87
Post-intervention:
Mean = 16.2 Standard Deviation = 9.76
Difference:
Mean = -4.11 Standard Deviation = 3.30

The results of a paired t-test
t= -3.74 degrees of freedom = 8

The probability of this result, assuming the null hypothesis, is 0.006
So cannot assume the null hypothesis

Statistical Comparison: T-test Performance on Celf Test

Pre-intervention:
Mean = 8.11 Standard Deviation = 4.01
Post-intervention:
Mean = 8.67 Standard Deviation = 3.39
Difference:
Mean = -0.556 Standard Deviation = 2.60

The results of a paired t-test
t=-0.640 degrees of freedom = 8

The probability of this result, assuming the null hypothesis, is 0.540
So no reason not to accept the null hypothesis

Preliminary Results: Feedback

Unexpected Outcomes impact on school curriculum
Questionnaires with parent, teachers and Classroom assistants (not significant issues raised but all positive)
Semi-structured interviews with SLTs

Participant Feedback using Talking Mats

Bad OK Good

Good:
Jester character
Way screen changes
Way of telling jokes

OK
Jokes
Scanning

Bad
Voice

Participant Feedback using Talking Mats

Bad OK Good

Good:
Jester character

OK
Touchscreen

OK/Bad
Way screen changes
Way of telling jokes
Voice

Bad
Jokes

STANDUP: some initial conclusions

Interfaces CAN be designed which provide children with CCN with successful access to complex underlying technology

Using STANDUP:
- the generative capabilities allows opportunity for natural language development, cf DA choosing punchline first
- the generative capabilities allows novel explorative learning, cf NI searching subjects

All children benefited
- enhanced desire to communicate
- knock on effect on other AAC usage
- illustrated children's abilities and potential of AAC

Illustrated use of technology within a wider environment

STANDUP: some initial conclusions

Issues with interface design
- scanning
- voice output
- improved appropriateness of vocabulary

The telling of the joke is important - what is the impact of STANDUP:
- on interactive conversation
- on joke comprehension and vocabulary acquisition

Do we want better jokes? (yes)

Use with speaking children with language impairment and other user groups

3. Writing up Experiments and Empirical Studies

Writing-up empirical studies 1

Abstract: Short summary of the problem, the results and the conclusion.

Introduction: What is the problem? What related work have other people done?
[Should go from general statement of the problem to a succinct and testable statement of the hypothesis].

Method:
Participants: state number, background and any other relevant details of participants

Materials: exactly what test materials, teaching materials, etc. were used, giving examples

Procedure: clear and detailed description of what happened at each stage in the experiment
[Someone reading should be able to duplicate it. Should also clearly indicate what data was collected and how.]

Writing-up empirical studies 2

Results:
Give actual data, or a summary of it.
Provide an analysis of data, using statistical tests where/if appropriate.
Use tables and graphs to display data clearly.
*[Interpretation of results does **not** go here, but in discussion section].*

Discussion:
Interpretation of results; restating of hypothesis and the implications of results; discussion of methodological problems such as weaknesses in design, unanticipated difficulties, confounding variables, etc.
Wider implications of the work should also be considered here, and perhaps further studies suggested.

Conclusion:
Statement of overall conclusion of the study.

References

Cohen, P. (1995) *Empirical Methods for Artificial Intelligence*, MIT Press, 1995.

Corbett, A.T. and Anderson, J.R., (1990) The Effect of Feedback Control on Learning to Program with the Lisp Tutor, *Proceedings of the 12th Annual Conference of the Cognitive Science Society*, LEA, New Jersey, 1990

Dix, A., Finlay, J., Abowd, R. and Beale, R. (2004) *Human-Computer Interaction.* Prentice Hall

Preece, J., Rogers, Y., Sharp, H., Benyon, D. Holland, S. and Carey, T. (1994). *Human-Computer Interaction*. Addison-Wesley