CRELLA PG Forum
19 Feb 2014

Introduction of statistical analyses for language testing/learning research (Part 1)

Dr. Sathena Chan

Why statistics?

Statistics provide a set of tools to help us systematically **collect**, **organise**, **analyse**, **interpret** and **present** data.

Aim

- To introduce the basic concept of statistics and
  - Null hypothesis
  - Descriptive and Inferential statistics
  - Variables
  - Measurement scales
- To introduce the common statistical analyses performed in language testing / learning research
  - Testing differences
  - Correlation
  - Rater reliability and item analysis (Part 2)
  - Factor analysis (Part 2)
  - Multiple regressions (Part 2)

Research Hypothesis

- **Null Hypothesis**

e.g. *There is **no** difference in test takers' performance between paper-based and computerized-based writing tests.*

e.g. *There is **no** relationship between test takers' listening ability and their performance on a speaking test.*

- **Alternative Hypothesis**

e.g. *There is a difference in test takers' performance between paper-based and computerized-based writing tests.*

e.g. *There is a relationship between test takers' listening ability and their performance on a speaking test.*

Variables

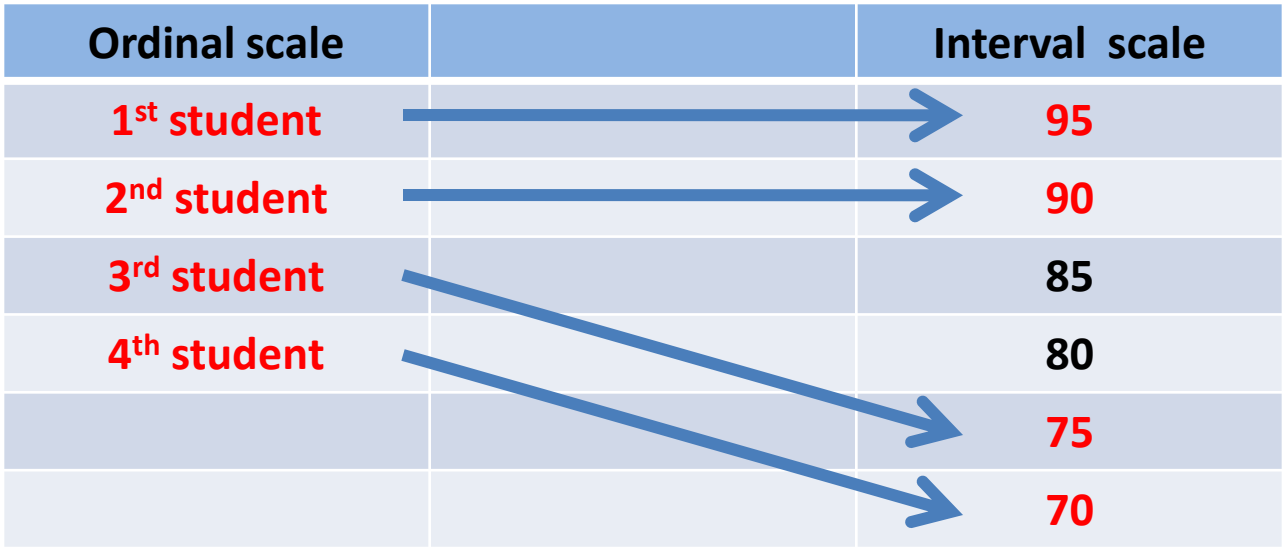
- A **variable** can have different value. It helps us to measure the 'constructs' which we aim to investigate in the research.

  Two-variable design (Bivariate)

  *There is **no** difference in test takers' performance between paper-based and computerized-based writing tests.*
  - dependent variable: test tasks' performance
  - independent variable: test delivery mode (pp vs cb)

Measurement scales

- **Nominal scale:** a scale with mutually exclusive groups, e.g. test delivery mode, native language, academic discipline, etc.
- **Ordinal scale:** a scale with classification and rank, e.g. ranking of students in terms of test scores
- **Interval scale:** a scale in which differences between points on the scale are equal



Types of statistics

- **Descriptive statistics**: to quantitatively **describe** the main features of a set data

- **Inferential statistics**: to draw **conclusions** (make inferences) from data on the basis of **probability** theory

Descriptive statistics

1. Measures of **central tendency**
   - Mean
   - Median
   - Mode
   - Frequency

Table 7: IELTS mean band scores

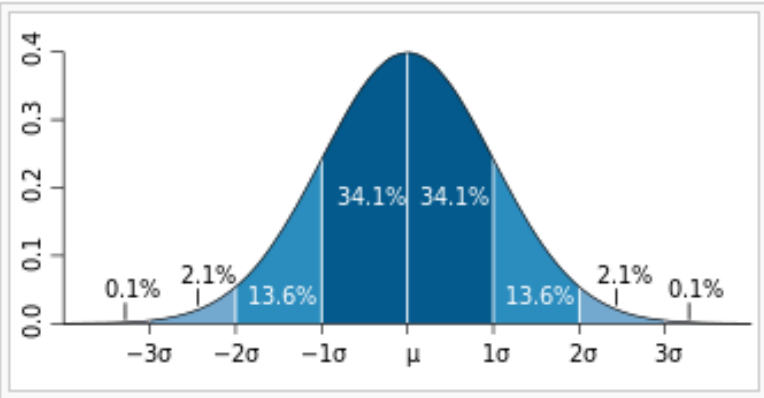
	IELTS (Overall)
Current study Participants (2011)	5.93
All Chinese 2010*	5.6
All Taiwanese 2010*	5.8

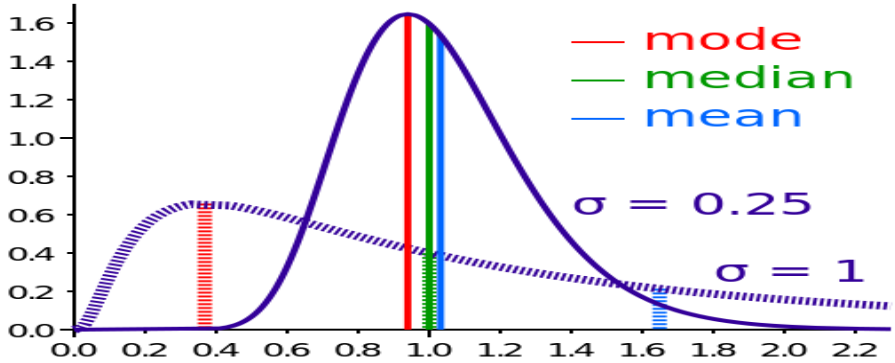
Table 6: Participants' IELTS scores

	IELTS (Overall)		
Bands	Freq.	Per cent	Total Per cent
4.5	0	0	
5	2	1.1	79.5%
5.5	67	39.2	
6.0	67	39.2	
6.5	25	14.6	19.3%
7.0	8	4.7	
7.5	1	0.6	1.2%
8	1	0.6	
8.5	0	0	
Total	171	100	

Descriptive statistics

2. Measures of **dispersion**
   - Standard deviation
   - Minimum/Maximum
   - **Distribution** (normal distribution* by K-S test)
   - Kurtosis/Skewness: descriptors of the shape of a probability distribution





Descriptive statistics

Figure 5: Participants' performance on GEPT Writing Task 1

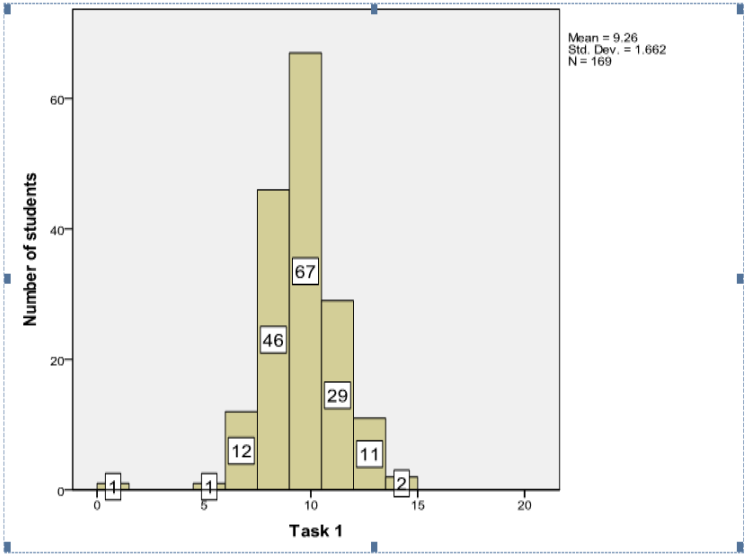


Figure 2: GEPT advanced writing test overall band

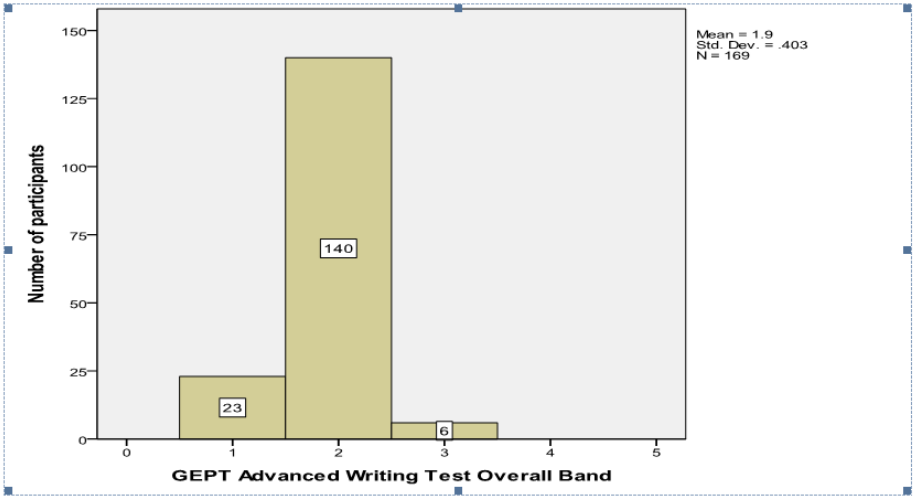
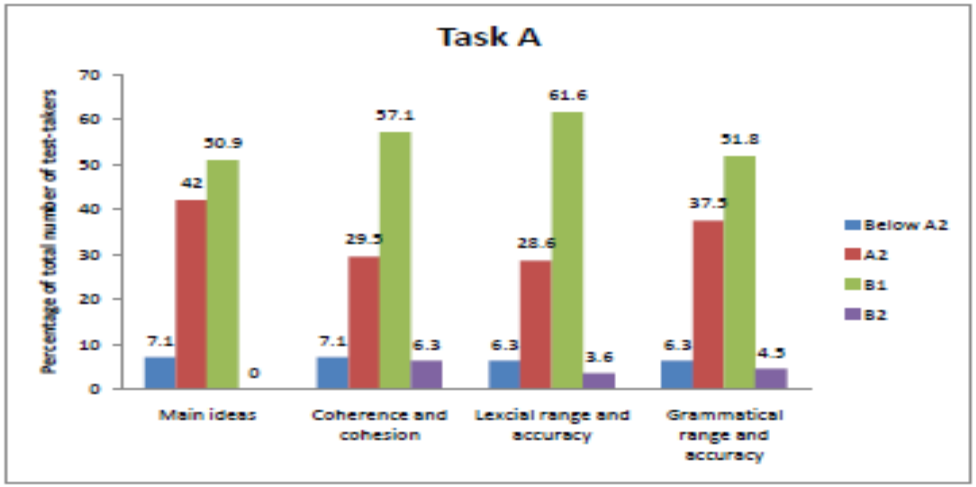


Figure 1 Task A



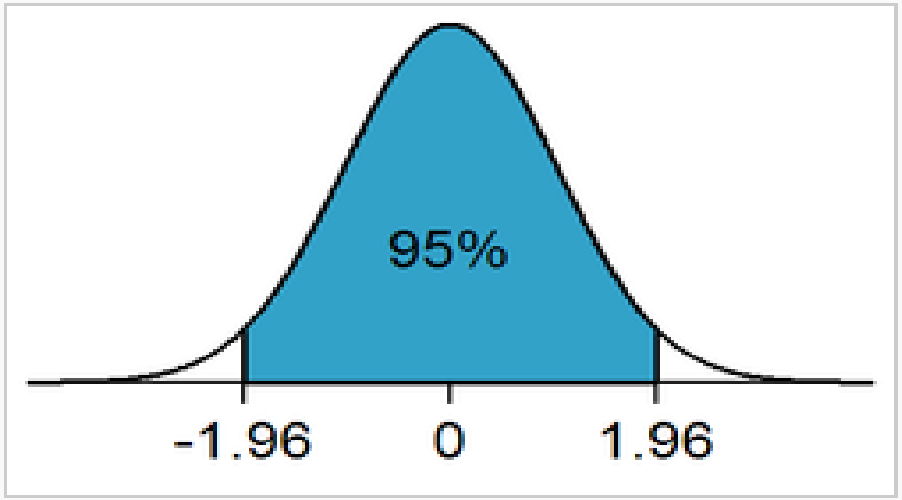
Questions to ask

Try not to jump into influential statistics analysis. Asking ourselves the following questions would help us to interpret the data.

- Are the averages (means) different in the way expected from the hypothesis?
- Is the difference between the means intuitively large, given the length of the score scale?
- Are the spreads of scores of the two groups markedly different?
- Are the standard deviations large in relation to the scale, indicating massive general disagreement?
- Given the picture from the graphs and descriptive statistics, do you think the sample difference is big and clear enough to suggest a general difference?

Inferential statistics

- **Statistical significance**: to show the that a result is **not** due to just **chance** alone
- If a ***p*-value** was found to be less than 0.05, the result would be considered statistically significant.



1. Testing differences

Two variables (comparing two means)

- **t-test**: to determine if two sets of data are significantly different from each other
  - Independent samples (e.g. test results of Group A and Group B)
  - Paired Samples (e.g. Group A's results on Test A and Test B)

Table 3.12 Comparisons of the proficiency of the participants who did Test Task A and Test Task B

	Participants who did only Test Task A (n=79)		Participants who did only Test Task B (n=59)		
	Mean	Std Dev	Mean	Std Dev	Independent samples t-test
IELTS Reading	5.91	0.481	5.73	0.601	t(275)=1.901, p=0.060 (n.s.)
IELTS Writing	5.59	0.534	5.58	0.513	t(270)=1.177, p=0.860 (n.s.)

- **Mann-Whitney U test:** is used on data with non-normal distribution (the non-parametric version of t-test)

1. Testing differences

Multiple variables (comparing different means)

- **ANOVA** (ANalysis Of VAriance between groups) *(MANOVA is the non-parametric option)*
- Example: Do the use of different reading strategies, namely ***guessing***, ***dictionary*** and ***glossary,*** lead to different levels of performance on a reading test?
- Prerequisite tests: *normality test* and *homogeneity of variance*

ANOVA

Readingscores

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	10.533	2	5.267	4.938	.027
Within Groups	12.800	12	1.067		
Total	23.333	14			

The result shows that there was a significant difference in the mean scores of the three groups of students. However, the Sig value does not tell us which groups' mean scores are different. We have to perform **post hoc tests** to find out where the differences are.

Multiple Comparisons

Reading scores

Tukey HSD

(I) Conditon	(J) Conditon	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
guessing	dictionary	.60000	.65320	.639	-1.1426	2.3426
	glossary	2.00000*	.65320	.025	.2574	3.7426
dictionary	guessing	-.60000	.65320	.639	-2.3426	1.1426
	glossary	1.40000	.65320	.123	-.3426	3.1426
glossary	guessing	-2.00000*	.65320	.025	-3.7426	-.2574
	dictionary	-1.40000	.65320	.123	-3.1426	.3426

*. The mean difference is significant at the 0.05 level.

2. Correlation

- **Pearson correlation coefficient** (**Pearson's *r***): to measure the linear correlation between two variables
  - **Pearson's *r is*** a value between +1 and −1

    (**1** is total positive correlation, **0** is no correlation, and **−1** is total negative correlation)
- **Spearman correlation:** for non parametric data

2. Correlation

- For example: Is there any relationship between test takers' performance on a writing test and their real-life writing performance?

		Mean real-life score
Test Task A total scores (n=160)	Pearson Correlation	.306**
	Sig. (2-tailed)	.000

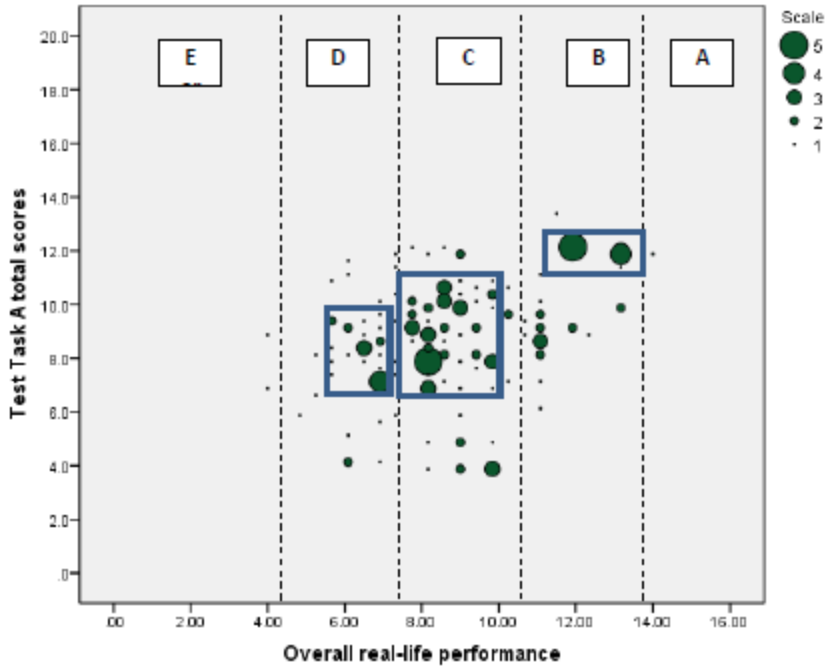


Figure 6.6 Relationships between Test Task A and real-life performance

Useful books to read

- Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.
- Mayers, A. (2013). *Introduction to statistics and SPSS in psychology*. Harlow: Pearson.
- Tabachnick, B. G., & Fiedell, L. S. (2001). *Using multivariate statistics (4th edition)*. MA: Allyn & Bacon.

Thank you!