Inferential Statistics

t-test, Chi Square, Variance ratio and Analysis of Variance

t-test

Comparing 2 samples of correlated/related scores

E.g., Experiment to compare time to conduct a task using two different interfaces

Participant	Interface 1	Interface 2
Participant 01	5.0 s	4.7 s
Participant 02	6.1 s	5.5 s
. . .	. . .	. . .
Participant 10	8.9 s	7.2 s
Participant 11	4.8 s	3.5 s
Participant 12	7.3 s	6.1 s

Correlated/related design: Same participants take part in the different experimental conditions

Can see by eye that conducting the task using interface 2 appears to take less time than when using interface 1, and this would be reflected in a difference between the mean times spent conducting the task using interface 1 compared to interface 2. but how can we be sure of this?

$$t = \frac{\text{Particular sample mean} - \text{population mean}}{\text{standard error of sample means}}$$

Degrees of freedom = *N*-1

Typical reporting style: e.g., "$t(49) = 2.96$, $p < 0.05$, two-tailed
T-test value was 2.96, degrees of freedom was 49, and the difference between the two groups is statistically significant at the 0.05 or 5% level of probability (using a two-tailed significance level).

t-test

Comparing 2 samples of unrelated/uncorrelated scores

E.g., Response times using a novel interface for two groups (UG students and PhD students)

UG	PhD
2.3 s	4.7 s
4.1 s	4.6 s
. . .	. . .
5.9 s	5.2 s
4.8 s	6.1 s
2.3 s	3.9 s

Uncorrelated/unrelated design: Different participants take part in the different experimental conditions.

Unrelated *t*-test tells you whether the 2 means are statistically significant or not.

Unrelated *t*-test combines variation in the 2 sets of scores to estimate standard error.

$$t = \frac{\text{Sample 1 mean} - \text{Sample 2 mean}}{\text{standard error of differences between sample means}}$$

Degrees of freedom = N-2

Typical reporting style is same as for the related *t*-test

Corrections

An assumption of the t-test is that it requires equal variances in the 2 samples (assumption of homogeneity of variance).

Can perform a check to see if both variances are the same (Levene's test). Standard statistical packages will conduct the Levene's test.

If the levene's test is significant (variances significantly different) another test will be conducted (for equal variances not assumed), this test will have a slightly lower degrees of freedom value.

Chi Square (X^2)

Differences between samples of frequency data

Category	Sample 1	Sample 2	Sample 3
Category 1	27	21	5
Category 2	19	20	19
Category 3	9	17	65

Chi-square is used with category data in the form of frequency counts

E.g. Samples could be previous experience of working in robotics: no experience (category 1), medium experience (category 2), a lot of experience (category 3).

Categories could be type of interface chosen: tactile (sample 1), tactile-visual (sample 2), visual alone (sample 3).

Chi Square (X^2)

Differences between samples of frequency data

Category	Sample 1	Sample 2	Sample 3	Row frequencies
Category 1	27	21	5	**53**
Category 2	19	20	19	**58**
Category 3	9	17	65	**91**
Column frequencies	**55**	**58**	**89**	**Overall frequencies = 202**

E.g. Samples could be previous experience of working in robotics: no experience (category 1), medium experience (category 2), a lot of experience (category 3).

Categories could be type of interface chosen: tactile (sample 1), tactile-visual (sample 2), visual alone (sample 3).

In null-hypothesis-defined population, expect 53 out of every 202 to prefer category 1, 58 out of every 202 to prefer category 2, and 91 out of every 202 to prefer category 3.

On left are *Observed frequencies*

Expected frequency of each cell needs to be calculated.

E.g., for Sample1, column frequency = 55. If null hypothesis is true expect 53 out of every 202 to prefer category 1. So expected frequency for preferring category 1 in sample 1 is:
55 x 53/202 = 14.43
And continue for all the cells.

Chi Square (X^2)

Differences between samples of frequency data

For each cell calculate observed frequency and expected frequency.

Chi-square statistic:

Differences between observed and expected frequencies

Greater the difference between observed and expected frequencies – less likely null hypothesis is true.

Chi-square (X^2) = $\sum \frac{(O-E)^2}{E}$

O = observed frequency,
E = expected frequency

Variance Ratio Test

The variance ratio test (the F-ratio test) assesses whether the variances of 2different samples are significantly different from one another- i.e., tests whether the spread of scores for the 2 samples is significantly different.

F = $\frac{\text{One variance estimate (a)}}{\text{Another variance estimate (b)}}$

F-ratio is a one-tailed test. The 5% or 0.05 criterion level applies to the upper (right-hand) tail of the distribution.

Larger F-ratio, more likely that (a) variance estimate is significantly larger than the (b) variance estimate.

F-ratio often used as part of other statistical techniques such as analysis of variance (ANOVA).

ANOVA

One-way uncorrelated (unrelated) ANOVA

Samples of scores unrelated.

Scores are dependent variable, groups are independent variable.

ANOVA estimates variance in the population due to the cell means (between variance) and the variance in the population due to random (or error) processes (within variance). These are compared using the F-ratio.

Error is variation which is outside of the researcher’s control.

If the ANOVA test is significant -> overall some of the means differ from each other.

ANOVA

One-way uncorrelated (unrelated) ANOVA

Group 1	Group 2	Group 3
5	5	10
12	4	12
6	5	6
10	6	7

E.g. Three groups of people with different haptic devices: Group 1 with haptic device 1, Group 2 with haptic device 2, Group 3 with haptic device 3 provide an estimate for the distance of a virtual object (cm).

Factor: haptic device
with 3 levels (3 types of device)

Any given measure/score can have a true part and an error part, e.g., 15 (obtained score) = 12 (true component) + 3 (error component)

$$F = \frac{\text{variance estimate (of true scores)}}{\text{variance estimate of error scores}}$$

Example of a typical output of a one-way unrelated ANOVA:

Source variation	Sum of squares	Degrees of freedom	Mean square	*F*-ratio
Between groups	68.222	2	34.111	10.6
Within groups	19.334	6	3.222	
Total	**87.556**	**8**	**10.944**	

Variance due to “true” scores

“Error” variation

Variance estimate

ANOVA

Correlated (repeated measures) ANOVA

Comparing 2 or more related samples of means. E.g., same group of participants is assessed 3 times on a measure.

Scores are dependent variable, the different occasions the measure is taken constitutes the independent variable.

Since individuals are measured more than once, can obtain a separate assessment of the variation in the data due to individual differences.

Amount of error variance is lower in related designs. What remains of the error is called the residual.

Significant value of F-ratio shows that the means in the conditions differ from each other overall.

ANOVA

Correlated (repeated measures) ANOVA

Case	Treatment 1	Treatment 2
Person 1	9	12
Person 2	7	10
Person 3	5	6
Person 4	2	7

E.g. Four participants take part in 2 experiments (treatments) to complete an audio task (time taken (s) to identify a musical note)
In treatment 1 they wear headphones, in treatment 2 they do not.
Factor: headphone
with 2 levels (with and without)

Any given measure/score can have a true part and an error part, e.g.,
15 (obtained score) = 12 (true component) + 3 (error component)

$$F = \frac{\text{between−treatments variance estimate}}{\text{error (residual) variance estimate}}$$

Example of a typical output of a one-way unrelated ANOVA:

Source variation	**Sum of squares**	**Degrees of freedom**	**Mean square**	***F*-ratio**
Between treatments	40.000	2	20.000	$\frac{20.000}{3.917}$ = 5.11
Between people	8.667	4	2.167	
Error (i.e., residual)	31.333	8	3.917	
Total	**80.000**	**14**		

As in the experiment

Individual differences

Variance estimate

Sample Article

Below is a sample of the Materials and Methods section taken from a published article: Cassarino, M., Maisto, M., Esposito, Y., Guerrero, D., Chan, J.S. and Setti, A., 2019. Testing attention restoration in a virtual reality driving simulator. *Frontiers in psychology*, *10*, p.250.

Compare understanding pre-and post lecture:

Statistical Analyses:

"Participants' performance at the SART was analyzed in terms of d-prime (d': a measure of signal detection sensitivity, calculated as the standardized difference (z-scores) between the proportion of correct responses on non-lures minus the proportion of incorrect responses on lures), overall mean accuracy (proportion of correct responses on lures and non-lures), mean accuracy on non-lures (pressing the bar), accuracy on lures (not pressing the bar when number three appears), reaction times (in milliseconds) of correct responses (related to pressing the bar in the presence of a non-lure), and inverse efficiency, a measure of speed-accuracy trade-off calculated as the ratio of reaction times over accuracy on non-lures (Bruyer and Brysbaert, 2011). Comparisons between the two exposure groups in terms of gender were conducted using Chi-square test and potential differences in age and driving experience were investigated via an independent samples t-test."

Sample Article

Below is a sample of the Materials and Methods section taken from a published article: Cassarino, M., Maisto, M., Esposito, Y., Guerrero, D., Chan, J.S. and Setti, A., 2019. Testing attention restoration in a virtual reality driving simulator. *Frontiers in psychology*, *10*, p.250.

Compare understanding pre-and post lecture:

Statistical Analyses:

"A 2 × 2 mixed-design ANOVA was conducted with Environment (rural vs. urban) as the between-subjects factor, and SART (pre- vs. post-drive) as the within-subjects factor to investigate effects of environmental exposure on changes in attentional performance pre- and post-drive. Post hoc comparisons were conducted via t-test statistics. Comparisons between exposure groups in terms of driving behavior were assessed via independent t-test. In addition, potential effects of driving on attention were tested through a 2 (SART session) × 2 (environmental exposure) × 2 (driving vs. passenger condition) ANOVA with Driving (driver or passenger) and Environment (urban vs. rural) as the between-subject factors, and SART (pre- vs. post-drive) as the within-subjects factor. We conducted a test of normality on the ANOVA unstandardized residuals as well as the Levene's test of homogeneity; for measures that did not appear to meet the assumptions of normality, we conducted the analyses using non-parametric tests and found no differences in results."

Sample Article

Below is a sample of the Materials and Methods section taken from a published article: Cassarino, M., Maisto, M., Esposito, Y., Guerrero, D., Chan, J.S. and Setti, A., 2019. Testing attention restoration in a virtual reality driving simulator. *Frontiers in psychology*, *10*, p.250.

Compare understanding pre-and post lecture:

Results:

"Environmental Exposure Effects on Attention: The two exposure groups (n = 19 in each group) did not differ significantly in terms of gender (χ21= 0.11, p = 0.74), age (t36 = -0.42, p = 0.67) or driving experience (t36 = 0.16, p = 0.87).

The 2 × 2 mixed-design ANOVA indicated no significant interaction between environmental exposure and SART pre- and post-drive for any of the measures of interest.

There was a main effect of environmental exposure for the measure of d' ($F_{1,36}$ = 4.18, p = 0.048, μ^2 = 0.11), with participants in the rural exposure group (M = 1.26, SD = 1.07) showing overall higher sensitivity (i.e., better performance) than the urban exposure group (M = 0.62, SD = 0.84). There was also a main effect of environmental exposure for the measure of accuracy on lures ($F_{1,36}$ = 4.61, p = 0.04, μ^2 = 0.11), with participants in the rural group (M = 0.64, SD = 0.25) being overall more accurate than those in the urban group (M = 0.48, SD = 0.21). In both cases, however, the size of the effect was small."

Summary

Descriptive and inferential statistics

Frequency distributions
Variations in the normal curve
Cumulative frequency
Percentiles
Measures of central tendency
Variability
Z-scores
Correlation coefficients
Hypothesis testing
T-tests
Chi square
Analysis of variance

Resources

Essential:

Research Methods and Statistics by Bernard C. Beins and Maureen A. McCarthy, Part III.

Research Methods in Human-Computer Interaction by Jonathan Lazar, Jinjuan Heidi Feng, Harry Hochheiser, 2017 2nd Edition. Elsevier. Chapter 4.

Supplementary:

Research Methods for Human-Computer Interaction by Paul Cairns and Anna L. Cox, 2016. Chapter 6.

Introduction to Statistics in Psychology, D. Howitt and D. Cramer, 4th Edition. Pages 88-133 and 134-152, 159-219.