

Addis College

M.Sc : Construction Technology & Management

Course: Research Methods and Statistical Application

Lecture 6: Research Design Cont'd from Lecture 5

By: Dagnachew A. (PhD)

August 13, 2020

1

Contents of the presentation

- Lecture notes
- Individual assignment

Statistical methods

Some common statistical methods used in data analysis are :

Non-parametric tests

- sign test
- rank-sum test
- chi-square test
- goodness of fit

Parametric tests

- *t-test*
- ANOVA (analysis of variance)
- regression and correlation
- time series
- index numbers.

- Parametric and non-parametric tests are available for testing hypotheses related to differences in the metric and non-metric scales.
- In the parametric case, the *t-test is used to* examine hypotheses related to the population means.
- In the non-parametric case, popular tests include the chi-square, and binomial test.

Non-parametric tests

- Non-parametric tests are distribution free and so are more flexible in application.

1. Sign test

- The sign test examines paired data using positive (+) and negative (–) signs.

Example

The sign test

A sample of **Engineers** are asked to rate the performance of two types of roofing tiles, A and B; the scorings are: excellent = 5, to very poor = 1, on a five-point Likert response format.

Architect No.	1	2	3	4	5	6	7	8	9	10	11	12
(1) Type A	5	4	5	2	1	2	3	2	5	5	3	1
(2) Type B	3	4	2	1	3	3	4	5	3	4	1	5
Sign [(1) – (2)]	+	0	+	+	–	–	–	–	+	+	+	–

No. +ve (p)	6
No. –ve (q)	5
No. 0	1
Total	<u>12</u>

$H_0: p = 0.5$ (there is no difference between tiles A and B)

$H_A: p \neq 0.5$ (there is a difference between tiles A and B)

excluding zeros:

$$n = 11$$

$$\bar{p} = 6/11 \text{ (proportion of 'successes')}$$

$$\bar{q} = 5/11 \text{ (proportion of 'failures')}$$

for 'no difference', $pH_0 = qH_0 = 0.5$.

(continued)

Standard error of the proportion:

$$\sigma_{\bar{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.5)(0.5)}{11}} \\ = \underline{\underline{0.151}}$$

As $H_a: p \neq 0.5$ (i.e. concerned with larger or smaller), a two-tailed test is required.

At 0.05 level of significance and as np and $nq < 5$, the normal distribution approximates to the binomial, the z-value for 0.475 (i.e. 0.5 minus $\frac{1}{2} \times 0.05$) of the area under one tail of the normal curve is 1.96, then:

$$pH_0 + 1.96\sigma_{\bar{p}} = 0.5 + (1.96)(0.151)$$

and

$$pH_0 + 1.96\sigma_{\bar{p}} = 0.5 - (1.96)(0.151)$$

So, the range of acceptance is:

$$0.204 \rightarrow 0.796$$

The sample proportion,

$$\bar{p} \left(= \frac{6}{11} \right) = 0.545$$

As $0.204 < 0.545 < 0.796$, there is *no difference* in the **Engineers'** perceptions of the tiles (H_0 is accepted).

Chi-square (χ^2) test

- The *chi-square test* is used to compare observed and expected frequencies of a *variable* that has three or more categories, to test whether more than two population proportions can be considered to be equal.
- Generally, the χ^2 distribution should *not be used if any* cell contains an expected frequency of less than 5.

Example

- The number of male and female workers is noted over three construction sites. The researcher wishes to investigate possible sex discrimination between the construction sites and so wishes to know if the data provide any evidence.

Solution

H_0 : there is no difference in the proportion of female workers employed on each site.

	Site A	Site B	Site C	Total
Male	52	48	60	160
Female	13	15	12	40
Total	65	63	72	200

Normalising the data (rounding to whole numbers):

	Site A	Site B	Site C	Total
% Male	33	30	37	100
% Female	33	38	29	100
% Total	33	31	36	100

Hence, the question is, given a sample of 200 workers on the construction sites, is it reasonable for the female workers to be distributed $A = 13$; $B = 15$; $C = 12$, if there is no sexual discrimination between those sites?

(continued)

Expected female workers (rounded):

$$\text{Site A : } \frac{65}{200} \times 40 = 13$$

$$\text{Site B : } \frac{63}{200} \times 40 = 13$$

$$\text{Site C : } \frac{72}{200} \times 40 = 14$$

Expected male workers:

$$\text{Site A : } \frac{65}{200} \times 40 = 52$$

$$\text{Site B : } \frac{63}{200} \times 160 = 50$$

$$\text{Site C : } \frac{72}{200} \times 160 = 58$$

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

where:

f_o = observed frequency

f_e = expected frequency

(continued)

so

$$\begin{aligned}\chi^2 &= \left[\frac{(52 - 52)^2}{52} + \frac{(48 - 50)^2}{50} + \frac{(60 - 58)^2}{60} \right. \\ &\quad \left. + \frac{(13 - 13)^2}{13} + \frac{(15 - 13)^2}{13} + \frac{(12 - 14)^2}{14} \right] \\ &= [0 + 0.08 + 0.067 + 0.308 + 0.286] \\ &= \underline{\underline{0.741}}\end{aligned}$$

The number of degrees of freedom of a χ^2 distribution is:

$$(\text{no. of rows} - 1)(\text{no. of columns} - 1)$$

Hence, in the example, the degrees of freedom are:

$$(2 - 1)(3 - 1) = 2$$

The tabulated value of χ^2 with 2 degrees of freedom and $\alpha = 0.5$ is 5.991. α = level of significance.

Thus, as the calculated χ^2 is less than the tabulated value, the null hypothesis cannot be rejected; there appears to be no sexual discrimination in worker employment between the construction sites.

Parametric test: *t*-test

The *t*-test is used to determine if the mean of a sample is similar to the mean of the population.

$$t = \frac{\bar{x} - \mu}{\hat{\sigma}_{\bar{x}}}$$

where:

$\hat{\sigma}_{\bar{x}}$ = estimated standard error of the mean

Degrees of freedom applicable are $(n - 1)$

t_{tab} = value taken from tabulated *t*-distribution curve

If $t_{\text{calc}} < t_{\text{tab}}$, the mean of the sample is not significantly different from the mean of the population at the selected level of confidence (significance). Usually, tests start by using the 95% level ($p = 0.05$), if the test at that level is passed, a higher level may be tested or *vice versa*. The 95% confidence level indicates that, although the data support the conclusion with 95% probability, there is a 5% chance that the conclusion is wrong.

The test may be used to examine the means of two samples:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\text{standard error of the difference in means}}$$

where:

$\bar{x}_1$ = mean of sample 1

$\bar{x}_2$ = mean of sample 2

$\bar{x}_1 - \bar{x}_2$ = difference between the means.

Parametric test: *Analysis of variance (ANOVA)*

- ANOVA is based on a statistical method called the *F-test*, which is the ratio of the variance among conditions (between-groups variance) to the variance within conditions (within-groups, or error variance).
- The larger the between-groups variance relative to the within-groups variance, the larger the calculated value of *F* and the more likely it is that the differences among the condition means reflect true effects of the independent variable rather than error variance.
- ANOVA and *t-test* are developed in the context of experimental research to test the differences among the means of the experimental groups, but they are also widely used to analyze data from non-experimental studies.
- In non-experimental studies, participants are not randomly assigned to groups (as in a true behavioral experiment) but are categorized into naturally occurring groups (e.g. architects and engineers); then a *t-test* or ANOVA is used to analyze the differences among the means of these groups.

$$\begin{array}{l} H_O: \mu_1 = \mu_2 = \dots = \mu_n \\ H_A: \mu_1 \neq \mu_2 \neq \dots \neq \mu_n \end{array}$$

The method assumes that each sample is drawn from a normal population; each sample has the same variance.

$$F = \frac{\text{between groups estimated variance}}{\text{within groups estimated variance}}$$

Sample variance:

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

Variance among samples means:

$$s_{\bar{x}}^2 = \frac{\sum (\bar{x} - \bar{\bar{x}})^2}{k - 1}$$

where:

$\bar{\bar{x}}$ = the ground mean (i.e. the arithmetic mean of all the values of all the samples)

k = the number of samples

As the standard error of the mean, $\sigma_{\bar{x}}$ is $\sigma/\sqrt{n}$, Levin and Rubin (1990: p. 439) show that the first estimate of the population variance, the between-groups variance, is:

$$\hat{\sigma}^2 = \frac{\sum n_j(\bar{x}_j - \bar{\bar{x}})^2}{k - 1}$$

n_j = number of items in sample j .

The within-group variance:

$$\text{sample variance, } S^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

The second estimate of the population variance, the within-group variance, is:

$$\hat{\sigma}^2 = \sum \left(\frac{n_j - 1}{n_T - k} \right) S_j^2$$

where:

$$n_T = \sum n_j$$

As $F \rightarrow 1$, the likelihood that H_0 is valid increases; as the value of F increases, the likelihood of H_0 being valid decreases.

Degrees of freedom in the numerator: $(k - 1)$

Degrees of freedom in the denominator: $(n_T - k)$

Using tables of the F -distributions and the appropriate degrees of freedom; if $F_{\text{calc}} < F_{\text{tab}}$, the null hypothesis should not be rejected.

Note: You can get the F-table from any statistics books

- To be valid, the *F-test can be applied to large samples only, $n \geq 100$* (Yeomans 1968: p. 101).
- When research designs involve more than two conditions (therefore involving more than two means), it is better to analyze the data using **ANOVA** rather than many *t-tests* in order to reduce the chances of Type I error.
- ANOVA analyzes differences between all condition means simultaneously. Rather than testing the difference between each pair of means as a *t-test* does, *ANOVA determines whether any of a set of means differs from another using a single statistical test, regardless of how many group means are involved in the test.*

Example

Independent groups

Condition	Groups	Mean
A	1	μ_1
B	2	μ_2
C	3	μ_3
D	4	μ_4
E	5	μ_5

Instead of conducting 10 *t*-tests (i.e. groups 1 vs 2, 1 vs 3, 1 vs 4 and so on) among all pairs of five means (each *t*-test with 5% chance of Type I error), ANOVA performs a single, simultaneous test on all condition means with 0.05 chance of Type I error.

- In factorial designs, the between-groups variance can be broken down further into other components to test for the presence of different main effects and interactions, that is when the research design involves more than one independent variable, ANOVA can test whether the between-group variance is related to each of the independent variables, as well as whether the between-group variance is produced by interactions among the variables.
- In a 2×2 factorial design involving two variables, α and β , the total variance would come from:
 - (1) the error variance,
 - (2) the main effect of α ,
 - (3) the main effect of β and
 - (4) the $\alpha \times \beta$ interaction (combined effect of the two variables).

- *F-values are calculated for the main effect of α , β and $\alpha \times \beta$ interaction. Each of these F-values is then compared to the critical value of F.*
- *When an F-test is statistically significant, that is when the calculated value of F exceeds the critical value, at least one of the group means differs from one of the others.*
- *If the calculated F-value is less than the critical value, the null hypothesis is not rejected and so conclude that the condition means do not differ.*

Example

Factorial design

Variables	α_1	α_2
β_1		
β_2		

The two conditions of variable β are absence and presence of anti-frost additive and the two conditions of variable α are ambient temperature levels of +4 and -4 °C. Concrete cube strength is then tested while holding other variables constant, for example, water at 4 °C.

Parametric test: *Analysis of variance (ANOVA)*

- Usually, regression & correlation are considered together in expressing a relationship between two variables: one or more known values, realizations, of the independent variable; & the other unknown, the dependent variable. To keep research clear, it is advisable, at least in the beginning, to consider variables in pairs – one independent & the other dependent.
- Regression & correlation statistics establish only any *relationship between the* realized values of the variables that occur; they *do not establish causality, that is* the province of theory, evidence & logical reasoning, in the light of the statistics.
- Conventionally, the independent variable is plotted on the *x-axis* & the *hypothesized* dependent variable on the *y-axis*. *Simple or linear regression considers straight line* hypothesized relationships only. Initially, a scatter plot of the raw data should be examined to indicate any likely relationship b/n the variables & shape of relationship.

- The standard form equation for a straight line is:

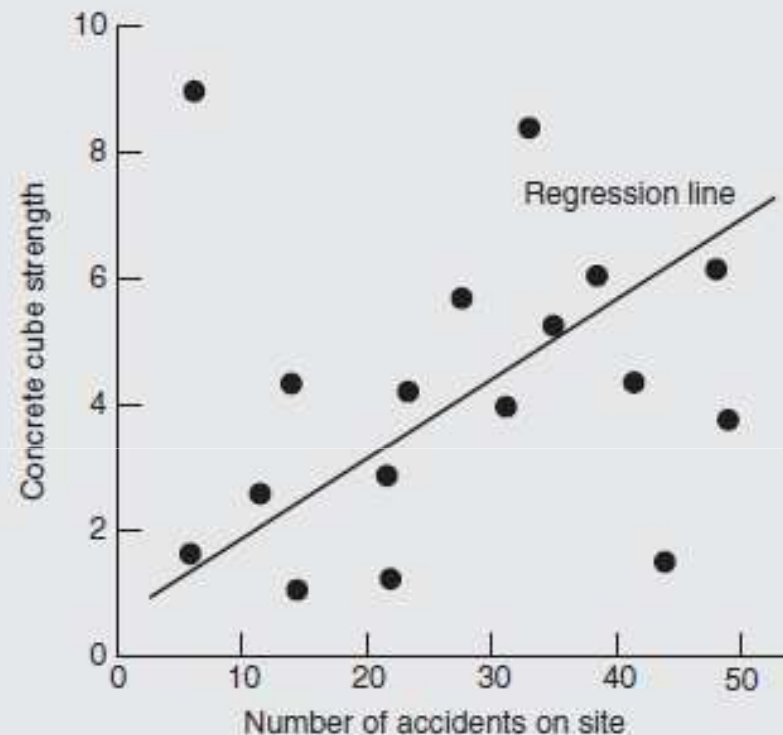
$$y = a + bx$$

- where '*a*' is the intercept of the line on the y-axis & '*b*' is the slope of the line.

- If we have given at least two data points on a scatter plot (a graph of the associated values of *x* and *y*), a regression line can be drawn – like a regression line shown next slide:

Example

Consider the following scatter plot:



Examination of the scatter plot and regression line leads to the conclusion that there may be a purely coincidental relationship; it is unlikely that there is any causal relationship, in either direction, between strengths achieved in concrete test cubes and the number of accidents on construction sites.

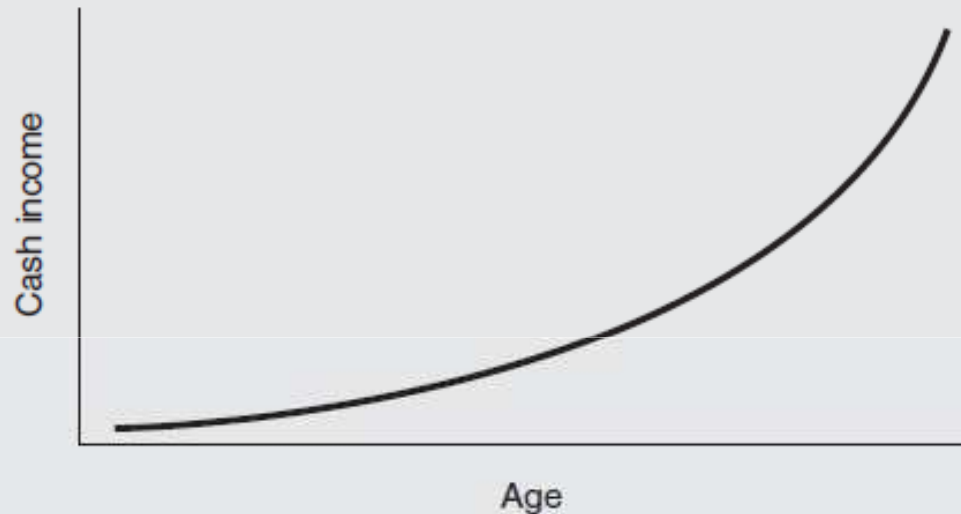
Note, however, that there is a positive relationship when the regression line has a positive slope, upwards from left to right.

- Regression assumes that the scatter of data points around the line of best fit is 'random', otherwise called *homoscedastic*.
- *If there is a pattern to the scatter of the data points* about the line, which shows the scatter to be different at different points, *heteroscedasticity* is present and so regression is questionable.
- Strictly, homoscedasticity is where the error has constant variance; heteroscedasticity is where the variance of the error changes along the length of the regression line (*Pindyck & Rubinfeld 1981: p. 49*).

- Not all scatter plots suggest that straight lines are the best fit, so curve fitting may be appropriate (*refer Figure next slide*). Fortunately, as for a straight line, there are standard forms of equation for various types of curves – usually determined by the approximate shapes of the curves, such as slope changes and any turning points.
- Use of computer packages is helpful for both linear and non-linear regression.
- The shape of the line and its nature should be detectable, along with any close alternatives, from observation of the scatter diagram and, in particular, from the nature of the relationship suggested by any underpinning theory. *Theory is important!*

Example

In the long term, the relationship between a person's age and his/her cash income is likely to be of the form shown in the following section, whilst that person is of working age and in continuous employment.



The line is an increasing geometric curve as income increases by an approximately constant percentage each year, so that the cash sum received per annum increases progressively. Economic theory has been used to determine the nature of the relationship. Collection of data and application of statistics would allow analytic checking and quantification of the relationship. Fortunately, most statistics packages, via simple commands, manipulations of data and setting of criteria, allow the programme to determine the line of best fit.

- The *coefficient of correlation* ,*r*, identifies the degree and nature of the relationship between the two variables, from a perfect positive relationship (+1) to a perfect negative relationship (−1), that is

$$-1 \leq r \leq +1$$

- *r = +1* means that an increase in variable *x* is matched an equiproportional increase in *y*.
- If *r = 0*, there is no relationship; changes in the variables are quite independent of each other; they are random.
- However, it is common to wish to know how much of the change in the values of a dependent variable is caused, given the logic of the relationship, by a change in the values of the independent variable. The statistic required is the *coefficient of determination*, *r*².
- If *r*² = 0.81 (i.e. *r* = 0.9), 81% of the changes in *y* is caused (explained) by the changes in *x*.

The coefficient of determination can be calculated using the following method:

A sample of data has variation about its own mean, which, in terms of least squares error, is:

$$\sum (y - \bar{y})^2$$

Similarly, the variation of the data about the regression line is:

$$\sum (y - \hat{y})^2$$

The sample coefficient of determination is:

$$r^2 = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}$$

The coefficient of determination measures the strength of a *linear* relationship between two variables.

Levin and Rubin (1990: p. 510) note a ‘short-cut method’ for calculation of r^2 , and most statistics packages can calculate r^2 directly:

$$r^2 = \frac{a \sum y + b \sum xy - n\bar{y}^2}{\sum y^2 - n\bar{y}^2}$$

If the true regression line for the population is given by $y = a + bx$ and the line estimated from the sample is $\hat{y} = a + bx$, the standard error of the regression coefficient can be used to test null hypotheses – for example that the slope of the regression line is unchanged from what has been found in the past (the ‘proportionate’ relationship between y and x). This is analogous to the use of standard deviation.

As b denotes the slope of the population’s regression line, S_b denotes the standard error of the regression coefficient of b .

$$S_b = \frac{S_e}{\sqrt{\sum x^2 - n\bar{x}^2}}$$

$$\left(S_e \text{ is the standard error of estimate; } = \sqrt{\frac{\sum y^2 - a \sum y - b \sum xy}{n - 2}} \right)$$

Using the t -distribution with $n - 2$ degrees of freedom, the limits for the acceptance region in this instance, showing that b is unchanged, are:

Upper: $b + t(s_b)$

Lower: $b - t(s_b)$

Confidence intervals can be calculated in a similar way.

- For some types of data, such as opinion surveys, which have employed Likert scale/response formats or something similar, the data are not suitable for analysis by regression & correlation due to the nature of the scales employed.

Example

Data on satisfaction with projects procured by various approaches have been collected from clients and contractors using a five-point Likert response format (1-totally satisfied, 5-totally dissatisfied). The data are summarised in the following section. All cell figures are numbers of respondents scoring per cell.

Procurement Method	Client Satisfaction					Contractor Satisfaction				
	1	2	3	4	5	1	2	3	4	5
Traditional	2	4	7	6	3	1	5	10	0	4
Construction management	3	5	1	8	2	5	5	5	3	2
Management contracting	4	6	2	8	0	6	6	5	2	1
Project management	2	4	6	2	6	3	7	3	4	3
Design and build	5	2	6	3	4	6	6	5	1	2
Design and management	2	3	7	4	4	5	5	3	5	2
BOOT ^a	1	6	7	3	3	6	6	4	2	2

^aThe procurement method: Build, Own, Operate, Transfer.

Likert scale

A method used to measure attitudes, which involves respondents indicating their degree of agreement or disagreement with a series of statements. Scores are summed to give a composite measure of attitudes.

(Continued)

Means and standard deviations are not appropriate. Rank correlations should be used as only the rankings can be compared. Rankings of satisfaction are obtained from examining the scorings in the totally satisfied column (column 1) – only in the event of tied scores are the next (and then subsequent) columns considered to determine the ranks.

	Clients	Contractors	<i>D</i>	<i>D</i> ²
	(Difference)			
Traditional	4	7	3	9
Construction management	3	4	1	1
Management contracting	2	1	1	1
Project management	5	6	1	1
Design and build	1	2	1	1
Design and management	6	5	1	1
BOOT	7	3	4	16
				<u>30</u>

The approach is adopted because respondents interpret the scorings differently in terms of both levels of the score categories (1–5 and associated word (semantic) descriptors, if any) and the intervals between the score categories. The latter

(continued)

differences may be accentuated if the data are collected only against semantic descriptors.

The coefficient of correlation between the ranks is a measure of the association between two variables, which is determined from the ranks of observations of the variables. It is calculated using Spearman's coefficient of rank correlation, ρ :

$$\begin{aligned}\rho &= 1 - \frac{6 \sum D^2}{n(n^2 - 1)} = \frac{6 \times 30}{7(7^2 - 1)} \\ &= 1 - \frac{180}{7 \times 48} = 1 - 0.54 = \underline{0.46}\end{aligned}$$

Individual Assignment 5 (10%)

- What are the methods researchers commonly use to collect data?
(2%)
- What is the difference between purposive & random sampling?
(2%)
- Describe the parametric and non-parametric tests where researchers use during conducting a scientific research & identify the advantages & disadvantages of each of the tests (6%)

Note:

Copying at the level of post graduate is strictly forbidden. If any two or more papers are found to be similar in any aspect, they will be disqualified.

Thank you

Your ? Through email