Block 12: Cross-tabulation and hypothesis testing

(Activity solutions can be found at the end of the document.)

Once data have been collected and prepared for analysis, basic analysis should be conducted. Typical forms include a **frequency distribution**, **cross-tabulation** and **hypothesis testing**. Sometimes market research projects do not go beyond basic data analysis. However, such analysis guides multivariate analysis.

Learning Objectives

- Discuss data analysis associated with frequencies, including measures of location, measures of dispersion and measures of shape
- Explain data analysis associated with cross-tabulations and the associated statistics: chi-squared, phi coefficient, contingency coefficient, Cramer's VV and lambda coefficient
- Describe data analysis associated with parametric hypothesis testing for one sample, two independent samples and paired samples.

Reading List

Malhotra, N.K., D. Nunan and D.F. Birks. Marketing Research: An Applied Approach. (Pearson, 2017) 5th edition [ISBN 9781292103129] Chapter 20 (excluding non-parametric tests).

12.1 Cross-tabulation and hypothesis testing

For each section of *Cross-tabulation and hypothesis testing*, use the LSE ELearning resources to test your knowledge with the Key terms and concepts flip cards.

A general procedure for hypothesis testing

Step 1: Formulate H0 and H1

A **null hypothesis** is a statement of the status quo, one of 'no difference' or 'no effect'. If the null hypothesis is not rejected, no changes will be made. An **alternative hypoth**esis is one in which some difference or effect is expected. Deciding on the alternative hypothesis will lead to changes in opinions or actions. The null hypothesis refers to a specified value of the population parameter (such as μ, σ or π), not a sample statistic (such as $\bar{X}$, S or P).

A null hypothesis may be rejected, but it *can never be accepted* based on a single test. In classical hypothesis testing, there is no way to determine whether the null hypothesis is true. In market research, the null hypothesis is formulated in such a way that its rejection leads to the desired conclusion.

The alternative hypothesis represents the conclusion for which evidence is sought. For example, for the proportion of internet shoppers, we might test:

$$H_0 : \pi = 0.40 \quad \text{vs.} \quad H_1 : \pi > 0.40$$

The test of the null hypothesis is a **one-tailed test**, because the alternative hypothesis is expressed directionally. If this is not the case, then a **two-tailed test** would be required, and the hypotheses would be expressed as:

$$H_0 : \pi = 0.40 \quad \text{vs.} \quad H_1 : \pi \neq 0.40$$

Step 2: Select an appropriate statistical technique

The **test statistic** measures how 'close' the sample has come to the null hypothesis. The test statistic often *follows a well-known distribution*, such as the standard normal, *t*, or chi-squared distribution.

In our example, the *z* statistic, which follows the standard normal distribution, would be appropriate:

$$Z = \frac{P - \pi}{\sigma\rho} \sim N(0,1)$$

where:

$$\sigma\rho = \sqrt{\frac{\pi(1-\pi)}{n}}$$

Step 3: Choose the level of significance, α

A **Type I error** occurs when the sample results lead to the rejection of the null hypothesis when it is, in fact, true. The probability of a Type I error, α, is also called the **significance level**.

A **Type II error** occurs when, based on the sample results, the null hypothesis is not rejected when it is, in fact, false. The probability of a Type II error is denoted by β. Unlike α, which is specified by the researcher, the magnitude of β depends on the actual value of the population parameter.

The **power of a statistical test** is the probability, 1−β, of rejecting the null hypothesis when it is false and should be rejected.

Although β is unknown, it is related to α. An extremely small value of α (such as α = 0.001) will result in intolerably high β errors. So it is necessary to *balance the two types of error*. For a test of H_0 : π=0.40 vs. H_1: π = 0.45, the probabilities of Type I, α, and Type II, β, errors are shown in Figure 20.4 of the textbook.

Step 4: Collect data and calculate test statistic

The required data are collected and the value of the test statistic is computed. In our example, the value of the sample proportion is:

$$P = \frac{17}{30} = 0.5667$$

The value of σ_P can be determined as follows:

$$\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.40 \times 0.60}{30}} = 0.0894$$

The test statistic value can be calculated as follows:

$$z = \frac{p - \pi}{\sigma\rho} = \frac{0.5667 - 0.40}{0.0894} = 1.86$$

Step 5: Determine the p-value or critical value

Using standard normal tables, the probability of obtaining a *z*-value, or **p-value**, of 1.86 (or more extreme) can be calculated. The area between $-\infty$ and 1.88 is 0.9686. Therefore, the area to the right of $z = 1.86$ is $1 - 0.9686 = 0.0314$.

Alternatively, the *critical value* of *z*, which will give an area to the right side of the critical value of 0.05, is between 1.64 and 1.65 and equals 1.645. Note that in determining the critical value of the test statistic, the area to the right of the critical value is either α or α/2. It is α for a one-tailed test and α/2 for a two-tailed test.

Step 6: Compare the p-value/critical value, reject or do not reject H0

If the probability associated with the calculated or observed value of the test statistic, TS_{cal}, is less than the significance level, α, the null hypothesis is rejected.

In our example, the probability associated with the calculated or observed value of the test statistic is 0.0314. This is the probability of getting a sample proportion of 0.5667 (or more extreme) when $\pi = 0.40$. This is less than the significance level of 0.05, hence the null hypothesis is rejected.

Alternatively (and equivalently), if the calculated value of the test statistic is greater than the critical value of the test statistic, TS_{cr}, the null hypothesis is rejected. The calculated value of the test statistic z = 1.86 lies in the rejection region, beyond the value of 1.645. Again, the same conclusion to reject the null hypothesis is reached.

Note that the two ways of testing the null hypothesis are equivalent, but *mathematically opposite in the direction of comparison*. If the *p*-value of TS_{cal}< significance level, α, then reject H_0, but if $TS_{cal} > TS_{cr}$ then reject H_0.

Step 7: Draw market research conclusion

The conclusion reached by hypothesis testing must be expressed in terms of the market research problem.

In our example, we conclude that there is evidence that the proportion of internet users who shop via the internet is significantly greater than 0.40. Hence the recommendation to the department store would be to introduce the new internet shopping service.

Cross-tabulation

While a frequency distribution describes one variable at a time, a cross-tabulation describes two or more variables simultaneously. Cross-tabulation results in tables which reflect the *joint distribution of two or more variables* with a limited number of categories or distinct values, for example:

Internet usage	**Gender**		**Row total**
	Male	**Female**	
Light	5	10	15
Heavy	10	5	15
Column total	15	15	

Since two variables have been cross-classified, percentages could be computed either *column-wise*, based on column totals, or *row-wise*, based on row totals. The general rule is to compute the percentages in the *direction of the independent variable*, across the dependent variable.

Internet usage	**Gender**	
	Male	**Female**
Light	33.3%	66.7%
Heavy	66.7%	33.3%
Column total	100%	100%

Figure 20.7 of the textbook shows the possible outcomes after introducing a third variable in cross-tabulation.

Refining an initial relationship:

Purchase of 'designer' clothes	**Marital status**	
	Married	**Unmarried**
High	31%	52%
Low	69%	48%

Column	100%	100%
Number of respondents	700	300

52% of unmarried participants fell in the high-purchase category, as opposed to 31% of the married participants. Before concluding that unmarried participants purchase more designer clothing than those who are married, a *third variable*, the buyer's gender, is now introduced into the analysis.

Purchase of 'designer' clothes	**Gender**			
	Male marital status		**Female marital status**	
	Married	**Unmarried**	**Married**	**Unmarried**
High	35%	40%	25%	60%
Low	65%	60%	75%	40%
Column	100%	100%	100%	100%
Number of respondents	400	120	300	180

In the case of females, 60% of those unmarried fall into the high-purchase category, compared with 25% of those who are married. On the other hand, the percentages are much closer for males, with 40% of those unmarried and 35% of those married falling into the high purchase category. Hence the introduction of gender (the third variable) has *refined the relationship* between marital status and purchase of designer clothing (the original variables). Unmarried participants are more likely to fall into the high-purchase category than married participants, and this effect is much more pronounced for females than for males.

Spurious initial relationship:

Own expensive car	**Education**	
	Degree	**No degree**
Yes	32%	21%
No	68%	79%
Column	100%	100%
Number of respondents	250	750

This shows that 32% of those with university degrees own an expensive car, compared with 21% of those without university degrees. Realising that *income may also be a factor*, the researcher decides to re-examine the relationship between education and ownership of expensive cars in light of income level.

Own expensive car	**Income**			
	Low-income education		**High-income education**	
	Degree	**No degree**	**Degree**	**No degree**
Yes	20%	20%	40%	40%
No	80%	80%	60%	60%
Column	100%	100%	100%	100%
Number of respondents	100	700	150	50

The percentages of those with and without university degrees who own expensive cars are the same for both income groups. When the data for the high-income and low-income groups are examined separately, the association between education and ownership of expensive cars disappears, indicating that the initial relationship observed between these two variables was *spurious*.

Revealing a suppressed association:

Desire to travel abroad	**Age**	
	Under 45	**45 or older**
Yes	50%	50%
No	50%	50%
Column	100%	100%

Number of respondents	500	500

The table shows *no association* between desire to travel abroad and age. However, when gender is introduced as the third variable, the following table is obtained.

Desire to travel abroad	**Gender**			
	Male age		**Female age**	
	Under 45	**45 or older**	**Under 45**	**45 or older**
Yes	60%	40%	35%	65%
No	40%	60%	65%	35%
Column	100%	100%	100%	100%
Number of respondents	300	300	200	200

Among men, 60% of those under 45 indicated a desire to travel abroad, compared with 40% of those 45 or older. The pattern was reversed for women, where 35% of those under 45 indicated a desire to travel abroad compared with 65% of those 45 or older. Since the association between desire to travel abroad and age runs in the opposite direction for males and females, *the relationship between these two variables is masked* when the data are aggregated across gender. However, when the effect of gender is controlled, the *suppressed association* between desire to travel abroad and age is revealed for the separate categories of males and females.

No change in the initial relationship:

Eat frequently in fast-food restaurants	**Family size**	
	Small	**Large**
Yes	65%	65%

No	35%	35%
Column	100%	100%
Number of respondents	500	500

Consider the cross-tabulation of family size and the tendency to eat out frequently in fast-food restaurants - no association is observed.

Eat frequently in fast-food restaurants	**Income**			
	Low-income family size		**High-income family size**	
	Small	**Large**	**Small**	**Large**
Yes	65%	65%	65%	65%
No	35%	35%	35%	35%
Column	100%	100%	100%	100%
Number of respondents	250	250	250	250

Income is introduced as a third variable in the analysis. Again, no association is observed. This is an example where there is *no change in the initial relationship*.

Activity 12.1

Define a spurious association.

Activity 12.2

What is meant by a suppressed association? How is it revealed?

Chi-squared test of association

To determine whether a *systematic association* exists, the probability of obtaining a chi-squared value as large or larger than the one calculated from the cross-tabulation is estimated. An important characteristic of the chi-squared statistic is the number of *degrees of freedom* (df) associated with it, that is:

$$df = (r-1) \times (c-1)$$

where *r* is the number of rows in the cross-tabulation, and *c* is the number of columns.

The null hypothesis, H_0, of no association between the two variables will be rejected only when the calculated value of the test statistic is greater than the critical value of the chi-squared distribution with the appropriate degrees of freedom.

The **chi-squared statistic**, $\chi 2$, is used to test the statistical significance of the observed association in a cross-tabulation. The *expected frequency* for each cell can be calculated by using the simple formula:

$$f_e = \frac{row\ total\ \times column\ total}{n}$$

Recall:

Internet usage	**Gender**		**Row total**
	Male	**Female**	
Light	5	10	15
Heavy	10	5	15
Column total	15	15	

For the data in the previous table, the expected frequencies for the cells, going from left to right, and from top to bottom, are:

$$\frac{15 \times 15}{30} = 7.5, \quad \frac{15 \times 15}{30} = 7.5 \quad \frac{15 \times 15}{30} = 7.5, \quad \frac{15 \times 15}{30} = 7.5$$

The test statistic value is calculated as:

$$\sum \frac{(f_o - f_e)}{f_e} = \frac{(5-7.5)^2}{7.5} + \cdots + \frac{(5-7.5)^2}{7.5} = 0.833 + \cdots + 0.833 = 3.333.$$

The chi-squared distribution is a skewed distribution whose shape depends solely on the number of degrees of freedom. As the number of degrees of freedom increases, the chi-squared distribution becomes more symmetrical. Statistical tables typically contain upper-tail areas of the chi-squared distribution for different degrees of freedom. For 1 degree of freedom, the probability of exceeding a chi-squared value of 3.841 is 0.05. For the example, there is $(2-1)\times(2-1)=1$ degree of freedom. The calculated chi-squared statistic had a value of 3.333. Since this is less than the critical value of 3.841, the *null hypothesis of no association cannot be rejected*, indicating that the association is not statistically significant at the 5% significance level. Figure 20.8 of the textbook provides a graphical representation.

Cross-tabulation statistics

The **phi coefficient**, ϕ, is used as a measure of the *strength* of association in the special case of a table with two rows and two columns (i.e. a 2×2 table). The phi coefficient is *proportional to the square root of the chi-squared statistic*:

$$\varphi \sqrt{\frac{x^2}{n}}$$

It takes the value of 0 when there is no association, which would be indicated by a chi-squared value of 0 as well. When the variables are perfectly associated, ϕ assumes the value of 1 and all the observations fall just on the main or minor diagonal.

While the phi coefficient is specific to a 2×2 table, the **contingency coefficient**, C, can be used to assess the strength of association in a *table of any size*:

$$C = \sqrt{\frac{x^2}{x^2 + n}}$$

The contingency coefficient varies between 0 and 1. The maximum value of the contingency coefficient depends on the size of the table (the number of rows and the number of columns). For this reason, it should be used *only to compare tables of the same size*.

Cramer's V is a modified version of the phi correlation coefficient, ϕ, and is used in tables larger than 2×2. Cramer's *V* is:

$$V = \sqrt{\frac{\varphi^2}{\min(r-1, c-1)}} = \sqrt{\frac{X^2/n}{min(r-1, c-1)}}\ .$$

Asymmetric lambda measures the *percentage improvement in predicting* the value of the dependent variable, given the value of the independent variable. Lambda also varies between 0 and 1. A value of 0 means no improvement in prediction. A value of 1 indicates that the prediction can be made without error. This happens when each independent variable category is associated with a single category of the dependent variable. Asymmetric lambda is computed for each of the variables (treating it as the dependent variable).

A **symmetric lambda** is also computed, which is a kind of average of the two asymmetric values. The symmetric lambda does not make an assumption about which variable is dependent. It measures the overall improvement when prediction is done in both directions.

Other statistics like **tau** b, **tau** c, and **gamma** are available to measure the association between two ordinal-level variables - both tau b and tau c adjust for ties. Tau b is the most appropriate with *square tables* in which the number of rows and the number of columns are equal. Its value varies between −1 and +1. For a rectangular table in which the number of rows is different than the number of columns, tau c should be used. Gamma does not make an adjustment for either ties or table size. It also varies between −1 and +1 and generally has a higher numerical value than tau b or tau c.

Cross-tabulation in practice

Test the null hypothesis that there is no association between the variables using the chi-squared statistic. If you fail to reject the null hypothesis, then there is no relationship.

If H_0 is rejected, then determine the strength of the association using an appropriate statistic (phi coefficient, contingency coefficient, Cramer's *V*, lambda coefficient, or other statistics).

If H_0 is rejected, interpret the pattern of the relationship by computing the percentages in the direction of the independent variable, across the dependent variable.

If the variables are treated as ordinal rather than nominal, use tau *b*, tau *c*, or gamma as the test statistic. If H_0 is rejected, then determine the strength of the association using the magnitude, and the direction of the relationship using the sign of the test statistic.

Activity 12.3

Discuss the reasons for the frequent use of cross-tabulations. What are some of the limitations?

Hypothesis testing procedures

Parametric tests assume that the variables of interest are measured on at least an interval scale. Non-parametric tests (non-examinable) assume that the variables are measured on a nominal or ordinal scale.

These tests can be further classified based on whether one, two or more samples are involved.

Independent samples refers to samples which are drawn randomly from different populations. For the purpose of analysis, data pertaining to different groups of participants (such as males and females) are generally treated as independent samples.

Paired samples refers to sample which relate to the same group of participants.

Figure 20.9 of the textbook provides a classification of hypothesis testing procedures (a reminder that non-parametric tests are non-examinable).

Activity 12.4

Present a classification of hypothesis testing procedures.

Parametric tests

The t **statistic** assumes that the variable is *normally distributed* and the mean is known (or assumed to be known) and the *population variance is estimated from the sample*.

Assume that the random variable XX is normally distributed, with mean μ and unknown population variance $\sigma 2$ (which is estimated by the sample variance s^2). Hence $T = \bar{X} - \mu / S_{\bar{X}}$ is t-distributed with n−1 degrees of freedom.

The *t* distribution is similar to the (standard) normal distribution in appearance. Both distributions are bell-shaped and symmetric. As the number of degrees of freedom increases, the tt distribution approaches the standard normal distribution.

Formulate the null hypothesis, H_0, and the alternative hypothesis, H_1. Select the appropriate formula for the *t* statistic. Select a significance level, α, for testing H_0 - typically, the 5% significance level is selected. Take one or two random samples and compute the mean and standard deviation for each

sample. Calculate the *t* statistic assuming H_0 is true. Calculate the degrees of freedom and estimate the probability of getting a more extreme value of the test statistic (alternatively, calculate the critical value of the *t* statistic).

If the probability computed is smaller than the significance level selected, reject H_0. If the probability is larger, do not reject H_0. (Alternatively, if the value of the calculated *t* statistic is larger than the critical value, reject H_0. If the calculated value is smaller than the critical value, do not reject H_0.) Failure to reject H_0 does not necessarily imply that H_0 is true. It only means that the true state is not significantly different than that assumed by H_0. Express the conclusion reached by the *t* test in terms of the market research problem.

Consider the following frequency distribution of internet familiarity:

Value label	**Value**	**Frequency**
Very unfamiliar	1	0
	2	2
	3	6
	4	6
	5	3
	6	8
Very familiar	7	4
Missing	9	1
Total		30

Suppose we wanted to test the hypothesis that the mean familiarity rating exceeds 4.0, the neutral value on a seven-point Likert scale. A significance level of α=0.05 is selected. We test $H_0 : \mu = 4.0$ vs. $H_1 : \mu > 4.0$, and obtain:

$$t = \frac{\bar{x} - \mu}{s_{\bar{X}}} = \frac{4.724 - 4.0}{0.293} = 2.471$$

where:

$$s_{\bar{X}} = \frac{1.579}{\sqrt{29}} = 0.293$$

The degrees of freedom for the *t* statistic to test the hypothesis about one mean are $n - 1$, i.e. in this case 29−1=28. From statistical tables or a computer, the probability of getting a more extreme value than 2.471 is less than 0.05. Alternatively, the critical *t*-value for 28 degrees of freedom and a significance level of 0.05 is 1.7011, which is less than the calculated value. Hence the null hypothesis is rejected, and so the familiarity level does exceed 4.0.

Note that *if the population standard deviation was assumed to be known* as 1.5, rather than estimated from the sample, a zz test would be appropriate. In this case, the value of the zz statistic would be:

$$Z = \frac{\bar{x} - \mu}{\sigma_{\bar{X}}} = \frac{4.724 - 4.0}{0.279} = 2.595$$

where:

$$\sigma_{\bar{X}} = \frac{1.5}{\sqrt{29}} = 0.279$$

From statistical tables or a computer, the probability of getting a more extreme value of *z* than 2.595 is less than 0.05. Alternatively, the critical *z*-value for a one-tailed test and a significance level of 0.05 is 1.645, which is less than the calculated value. Therefore, the null hypothesis is rejected, reaching the same conclusion arrived at earlier using the *t* test.

The procedure for testing a null hypothesis with respect to a proportion was illustrated earlier.

In the case of means for two independent samples, the hypotheses take the following form

$$H_0 : \mu 1 = \mu 2 \text{ vs } H_1 : \mu 1 \neq \mu 2$$

The two populations are sampled, and the means and variances computed based on samples of size *n*1 and *n*2, respectively. If both populations are found to have the same variance, a *pooled variance estimate* is computed from the two sample variances as follows:

$$S_p^2 = \frac{\overline{\sum_{i=1}^{n1}(X_{il} - X_1)^2 + \sum_{i=l}^{n_2}(X_{i2} - X_2)^2}}{n_1 + n_2 - 2} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

The standard error in the test statistic can be estimated as:

$$S_{\bar{X}_1 - \bar{X}_2} = \sqrt{S_p^2 \left(\frac{1}{n^1} + \frac{1}{n^2}\right)}$$

The appropriate value of *t* can be calculated as:

$$T = \frac{X_1 - X_2 - (\mu_1 - \mu_2)}{S_{\bar{X}_1 - \bar{X}_2}} \sim t_{n_1 + n_2 - 2}$$

The degrees of freedom in this case are *n*1 + *n*2 − 2

An *F* test of sample variances may be performed if it is not known whether the two populations have equal variances. In this case, the hypotheses are:

$$H_0 : \sigma_1^2 = \sigma_2^2 \qquad \text{vs} \qquad H_1 : \sigma_1^2 \neq \sigma_2^2$$

The *F* statistic is computed from the sample variances as follows:

$$F = \frac{S_1^2}{S_2^2} \sim F_{n_{1-1},\ n_{2-1}}$$

When testing for a difference between two population proportions, the null and alternative hypotheses are:

$$H_0 : \pi_1 = \pi_2 \qquad \text{vs} \qquad H_1 : \pi_1 \neq \pi_2$$

A *z* test is used as in testing the proportion for one sample. However, in this case the test statistic is given by:

$$Z = \frac{P_1 - P_2}{S_{P_{1-P_2}}} \sim N(0,1)$$

In the test statistic, the *numerator is the difference between the proportions* in the two samples, P_1 and P_2. The *denominator is the standard error* of the difference in the two proportions and is given by:

$$S_{P_{1-P_2}} = \sqrt{P(1-P)\left(\frac{1}{n^1} + \frac{1}{n^2}\right)}$$

where:

$$P = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2}$$

For paired samples, the difference in these cases is examined by a paired samples *t* test. To compute *t* for paired samples, the *paired difference variable*, denoted by *D*, is formed and its mean and standard deviation are calculated. Next, the *t* statistic is computed. Its degrees of freedom are *n*−1, where *n* is the number of paired data values. We test:

$$H_0 : \mu_D = 0 \qquad \text{vs.} \qquad H_1 : \mu_D \neq 0$$

using the test statistic:

$$T = \frac{D - \mu_D}{S_{\bar{D}}} \sim t_{n-1}$$

Also:

$$D = \frac{1}{n}\sum_{i=l}^{n} D_i \qquad \text{and} \qquad S_D = \sqrt{\frac{1}{n-1}\sum_{i=l}^{n}(D_i - \overline{D})^2}$$

and:

$$S_{\overline{D}} = \frac{S_D}{\sqrt{n}}.$$

In the internet usage example, a paired *t* test could be used to determine if the participants differed in their attitudes toward the internet and attitudes toward technology.

Activity 12.5

Describe the general procedure for conducting a t test.

Activity 12.6

What is the major difference between parametric and non-parametric tests?

Discussion forum and case studies

To access the solutions to these questions and case study, click here to access the printable Word document or click here to go to LSE's Elearning resources.

Exercises on the block's topics

1. In each of the following situations, indicate the statistical analysis you would conduct and the appropriate test or test statistic which should be used.

    a. Consumer preferences for The Body Shop shampoo were obtained on an 11-point Likert scale. The same consumers were then shown a commercial about The Body Shop. After the commercial, preferences for The Body Shop were again measured. Has the commercial been successful in inducing a change in preferences?
    b. Does the preference for The Body Shop shampoo follow a normal distribution?
    c. Participants in a survey of 1,000 households were classified as heavy, medium, light and non-users of ice cream. They were also classified as being in high-, medium-, or low-income categories. Is the consumption of ice cream related to income level?
    d. In a survey of 2,000 households, participants were asked to rank 10 supermarkets, including Lidl, in order of preference. The sample was divided into small and large households based on a median split of the household size. Does preference for shopping in Lidl vary by household size?

2. The current advertising campaign for Red Bull would be changed if less than 30% of consumers like it.

    a. Formulate the null and alternative hypotheses.
    b. Discuss the Type I and Type II errors which could occur in hypothesis testing.

c. Which statistical test would you use? Why?
d. A random sample of 300 consumers was surveyed and 84 participants indicated that they liked the campaign. Should the campaign be changed? Justify your answer.

3. An electrical goods chain is having a New Year sale of refrigerators. The numbers of refrigerators sold during this sale at a sample of 10 stores were:

80 110 0 40 70 80 100 50 80 30.

a. Is there evidence that on average more than 50 refrigerators per store were sold during this sale? Use α=0.05
b. What assumption is necessary to perform this test?

4. In a survey pre-test, data were obtained from 45 participants on Benetton clothes. These data are given in the file *Benetton.sav*, which gives the usage, gender, awareness, attitude, preference, intention and loyalty toward Benetton of a sample of Benetton users. Usage was coded as 1, 2 or 3, representing light, medium or heavy users, respectively. Gender was coded as 1 for females and 2 for males. Awareness, attitude, preference, intention and loyalty were measured on a 7-point Likert-type scale (1 = Very unfavourable, 7 = Very favourable). Note that five participants have missing values which are denoted by 9. (An Excel version of the dataset is *Benetton.xlsx*.)

Analyse the Benetton data to answer the following questions. In each case, formulate the null and alternative hypotheses and conduct the appropriate statistical test(s).

a. For each of the following variables calculate appropriate descriptive statistics: awareness, attitude, preference, intention and loyalty toward Benetton.

Video walkthrough of activity 4a.

b. Conduct a cross-tabulation of usage with gender. Interpret the results.
Use **Analyze** >> **Descriptive Statistics** >> **Crosstabs**…….

Video walkthrough of activity 4b.

c. Does the average awareness of Benetton exceed 3.0? Hint: conduct a one-sample tt test. To perform the test, **Analyze** >> **Compare Means** >> **One-Sample T Test**…….

Video walkthrough of activity 4c.

d. Do males and females differ in their average awareness of Benetton? Their average attitude toward Benetton? Their average loyalty for Benetton? Hint: conduct an independent samples *t* test. Use **Analyze** >> **Compare Means** >> **Independent-Samples T Test**…….

Video walkthrough of activity 4d.

e. Do the participants in the pre-test have a higher average level of awareness than loyalty? Hint: conduct a paired-samples *t* test. Use **Analyze** >> **Compare Means** >> **Paired-Samples T Test**…….

Video walkthrough of activity 4e.

f. Does awareness of Benetton follow a normal distribution?

Use **Analyze** >> **Nonparametric Tests** >> **Legacy Dialogs** >> **1-Sample K-S** to perform the Kolmogorov-Smirnov test of normality. Add the required variable to the 'Test Variable List' and check 'Normal' under 'Test Distribution'.

The Kolmogorov-Smirnov test tests the null hypothesis that the data are normally distributed, i.e. $N(\bar{x}, s^2)$, against the alternative hypothesis that the data do not follow a $N(\bar{x}, s^2)$, distribution. (If the data have a normal distribution, the most likely (maximum likelihood) estimates of μ and σ^2 would be $\bar{x}$ and s^2 respectively.)

Video walkthrough of activity 4f.

g. Is the distribution of preference for Benetton a normal distribution?

Video walkthrough of activity 4g.

Discussion forum, exercises and discussion points

Discussion points

1. 'Why waste time doing basic forms of data analysis? Why not just go straight to performing multivariate analyses, whose outputs from most software packages will include basic analyses?'
2. Why do managers find cross-tabulations so appealing? What would it take to make managers more appreciative of statistical analyses which go beyond cross-tabulations?

Learning outcomes checklist

Use this to assess your own understanding of the chapter. You can always go back and amend the checklist when it comes to revision!

- Discuss data analysis associated with frequencies, including measures of location, measures of dispersion and measures of shape
- Explain data analysis associated with cross-tabulations and the associated statistics: chi-squared, phi coefficient, contingency coefficient, Cramer's V and lambda coefficient
- Describe data analysis associated with parametric hypothesis testing for one sample, two independent samples and paired samples.

Block 12: Cross-tabulation and hypothesis testing

Solution to Exercise 12.1

Two variable cross-tabulations may show an association between the variables. However, the introduction of a third variable in the cross-tabulation might reveal that there is no association between the two variables despite the observed initial association. Such a relationship is called a spurious association.

Solution to Exercise 12.2

Sometimes the cross-tabulation of two variables may indicate little association. However, after introducing a third variable, the cross-tabulation might reveal some third association between the two variables although no association was initially observed. This is known as a suppressed association.

Solution to Exercise 12.3

Cross-tabulations are popular for the following reasons.

- *Ease of comprehension* – i.e. cross-tabulation analysis and results can be easily interpreted and understood by managers who have little statistical knowledge.
- *Versatility* – i.e. a series of cross-tabulations may provide greater insights into a complex phenomenon than a single multivariate analysis.
- *Clarity* – i.e. the clarity of interpretations provides a stronger link between the research results and managerial action.
- *Simplicity* – i.e. cross-tabulation analysis is simple to conduct and appealing to the less sophisticated researcher. Cross-tabulation, though meant for describing the joint distribution of two or more variables, is seldom used in computations involving more than three variables. This is because the interpretation becomes quite complex. Also, since the number of cells increases multiplicatively, maintaining an adequate number of participants in each cell becomes problematic. Consequently, the statistics computed could be unreliable. Besides, since only two or three variables are tabulated at a time, a cross-tabulation is not a very efficient way of examining the relationships when there are several variables.

Solution to Exercise 12.4

Based on the measurement scale of the variables involved, hypothesis testing procedures can be classified into two categories: parametric and non-parametric.

Parametric tests:

- These tests assume that the variables of interest are measured on at least an interval scale.
- The most popular test is a tt test used to examine hypotheses about means.
- The tt test could be conducted on the mean of one sample or two samples of observations.

In the latter case, the samples could be independent or paired. An independent samples tt test is used when the samples are independent and a paired tt test is used when the samples are paired.

Non-parametric tests (non-examinable):

These tests assume that the variables are measured on a nominal or ordinal scale. Popular tests based on observations drawn from one sample include:

- Kolmogorov-Smirnov test
- chi-squared test
- runs test
- binomial test.

In the case of two independent samples, the popular tests are:

- Mann-Whitney test
- median test
- Kolmogorov-Smirnov test.

If the two samples are paired, the popular tests are:

- sign test
- Wilcoxon signed-rank test.

Parametric and non-parametric tests are also available for conducting hypotheses relating to more than two samples.

Solution to Exercise 12.5

The *t* test is used to test hypotheses about population means. The test is based on the (Student's) *t* statistic. It is assumed that the random variable *X* is normally distributed with mean, μ, and unknown population variance, σ^2, which is estimated by the sample variance, S^2. It is known that the standard error of the sample mean $\bar{X}$ is estimated as $S_{\bar{X}} = \frac{S}{\sqrt{n}}$. $T = (\bar{X} - \mu)/S_{\bar{X}}$ IS *t*-distributed with *n*−1 degrees of freedom.

The procedure for hypothesis testing using the *t* statistic is given below.

- Formulate the null hypothesis, H_0, and the alternative hypothesis, H_1.
- Select the appropriate formula for the *t* statistic.
- Select a significance level, α, for testing H_0. Typically the 5% significance level is selected.
- Take one or two random samples and compute the mean and standard deviation for each sample.
- Calculate the *t* statistic assuming that H_0, is true.
- Calculate the degrees of freedom and determine the probability of getting a more extreme value of the statistic under H_0, i.e. the *p*-value.
- If the *p*-value of the test statistic is smaller than the significance level, reject H_0. This means that the true state is significantly different from that assumed by H_0.

- Express the conclusion in terms of the market research problem.

Solution to Exercise 12.6

The difference lies in the measurement scale of the variables involved. Parametric tests assume that the variables of interest are measured on at least an interval scale (metric data). Non-parametric tests assume that the variables are measured on a nominal or ordinal scale (non-metric data).

Discussion forum and case studies

To access the solutions to these questions and case study, click here to access the printable Word document or click here to go to LSE's Elearning resources.

Solutions to exercises on the block's topics

Exercise 1:

a. Measure the significance of the change in the mean score of the scale, $\bar{X}\bar{X}$, using a tt test for paired samples.

b. Use a non-parametric Kolmogorov-Smirnov one-sample test in which KKequals the largest absolute difference between the observed results and the normal distribution.

c. Conduct a cross-tabulation focusing on Cramer's VV statistic.

d. Use a non-parametric hypothesis test for two independent samples to test if the means of the two populations are equal. The Mann-Whitney, median or Kolmogorov-Smirnov test can also be used.

Exercise 2:

a. We test:

$$H_0 : \pi = 0.30 \qquad \text{vs.} \qquad H_1 : \pi < 0.30.$$

Type I errors occur when H_0 is rejected when it is in f

b. A z statistic because we assume that the population standard deviation is known (see (d)).

c. We have n=300, p=84/300=0.28 and:

$$\sigma_P \sqrt{\frac{0.30 \times 0.70}{300}} = 0.027.$$

Therefore:

$$z = 0.28 - \frac{0.30}{0.027} = -0.7407$$

This z value corresponds to a p-value of $P\,(Z \leq -0.7407) = 0.23$, which is not significant. Therefore, we do not reject H_0, i.e. do not change the current advertising campaign.

Exercise 3:

a. Use a one-sample *t* test to test:

$H_0: \mu = 50 \quad$ vs $\quad H_1: \mu > 50.$

We have $\bar{x} = 64, \quad s = 33.73, \quad n = 10$ and so:

$$t = \frac{64 - 50}{33.73/\sqrt{10}} = 1.313$$

The pp-value is P(T ≥ 1.313)=0.11, where T~t_9, so we do not reject H_0 and hence we do not have significant evidence that on average more than 50 refrigerators per store were sold during the sale.

b. It is assumed that the sales of refrigerators is normally distributed and that the mean is known to be 50.

Exercise 4:

a. Some appropriate descriptive statistics are:

	Awareness	**Attitude**	**Preference**	**Attitude**	**Loyalty**
Mean	4.18	4.07	4.23	4.05	3.95
Median	4.00	4.00	4.00	4.00	4.00
Mode	6.00	3.00	4.00	3.00	5.00
Standard deviation	1.88	1.91	1.54	1.71	1.68
Variance	3.55	3.65	2.37	2.95	2.84
Skewness	-0.25	0.05	0.12	0.01	-0.02
Kurtosis	-1.09	-1.06	-0.60	-0.70	-1.19

b. We test H_0: No association between usage and gender vs. H_1: There is an association between usage and gender. Cross-tabulation results are shown next.

			Usage			**Total**
			Light user	**Medium user**	**Heavy user**	
Gender	**1**	**Count**	14	5	5	24
		Expected count	10.1	5.3	8.5	24.0
	2	**Count**	5	5	11	21

		Expected count	8.9	4.7	7.5	21.0
Total		**Count**	19	10	16	45
		Expected count	19.0	10.0	16.0	45.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	6.341[a]	2	.042
Likelihood Ratio	6.545	2	.038
Linear-by-Linear	6.182	1	.013
Association			
N of Valid Cases	45		

[a] 1 cells (16.7%) have expected count less than 5. The minimum expected count is 4.67.

The Pearson Chi-Square test statistic value is 6.341, with a p-value of 0.042. We reject H_0 at the 5%, but not 1%, significance level and conclude that males are heavier users of Benetton relative to females, who tend to be light users (having compared the counts and expected counts).

Note the output warns that one of the expected counts is less than 5 (4.67 for medium usage for males). Technically, for a chi-squared test to be valid, *all* expected counts should be at least 5 to ensure that the test statistic is (approximately) chi-squared distributed. For our purposes, we can treat 4.67 as close to 5 and hence continue to perform the test.

c. We test H_0: μ = 3 vs. H_1: μ > 3. Mean awareness = 4.18 and the test statistic value is:

$$t = \frac{4.18 - 3}{1.883/\sqrt{44}} = \frac{4.18 - 3}{0.284} = 4.162$$

Note SPSS excluded the missing value (in observation 29) and so used n=44n=44 values. We have n−1=44−1=43 degrees of freedom. The two-sided pp-value is 0.000. However, we are performing a one-sided test and so we need to divide this pp-value by 2 (which is still 0.000!). Hence the test is highly significant and so we conclude that there is strong evidence that the mean awareness for Benetton exceeds 3.00.

Also, note the 95% confidence interval for the mean difference (between $\bar{x}$ = 4.18 $\bar{x}$ = 4.18 and the hypothesised value of 3) is (0.61,1.75)(0.61,1.75) which excludes 0.

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean

Awareness	44	4.18	1.883	.284

One-Sample Test

	Test Value = 3					
	t	dt	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Awareness	4.162	43	.000	1.182	.61	1.75

d. SPSS output is:

Group Statistics

	Gender	**N**	**Mean**	**Std. Deviation**	**Std. Error Mean**
Awareness	**1**	23	3.57	1.903	.397
	2	21	4.86	1.652	.360
Attitude	**1**	24	3.58	1.998	.408
	2	20	4.65	1.663	.372
Loyalty	**1**	23	4.17	1.696	.354
	2	21	3.71	1.678	.366

Independent Samples Test

	Levene's Test for Equality of Variances		t-test for Equality of Means					
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference

									Lower	Upper
Awareness	Equal variances assumed	1.249	.270	-2.394	42	.021	-1.292	.540	-2.381	-.203
	Equal variances not assumed			-2.410	41.903	.020	-1.292	.536	-2.374	-.210
Attitude	Equal variances assumed	.395	.533	-1.900	42	.064	-1.067	.561	-2.200	.066
	Equal variances not assumed			-1.933	42.000	.060	-1.067	.552	-2.181	.047
Loyalty	Equal variances assumed	.014	.905	.902	42	.372	.460	.509	-.568	1.487
	Equal variances not assumed			.903	41.719	.372	.460	.509	-.568	1.487

Assuming equal variances throughout, there is only a significant difference between the genders in terms of awareness (the means are 4.86 and 3.57) with the pp-value of the test being 0.021. The other variables (attitude and loyalty) are insignificant.

e. The difference between awareness (mean = 4.21) and loyalty (mean = 3.98) is not significant. t=0.62, df = 42 and the *p*-value = 0.538 (two-tailed test).

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Awareness	4.21	43	1.897	.289
	Loyalty	3.98	43	1.697	.259

Paired Samples Correlations

		N	Correlation	Sig.
Part 1	Awareness & Loyalty	43	.068	.664

Paired Samples Test

	Paired Differences					t	df	Sig.(2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
Pair 1 Awareness - Loyalty	0.233	2.458	0.375	-0.524	0.989	0.621	4 2	0.538

f. We test the null hypothesis that the variable is normally distributed $\mathrm{N}(4.18, (1.883)^2$, against the alternative hypothesis that it is not. The null hypothesis cannot be rejected as the *p*-value is 0.268 which exceeds conventional significance levels. (The test statistic value is the Kolmogorov-Smirnov *Z* value, which here is 1.002.)

One-Sample Kolmogorov-Smirnov Test

		Awareness
N		44
Normal Parameters [a,b]	Mean	4.18
	Std. Deviation	1.883
Most Extreme Differences	Absolute	.151
	Positive	.104
	Negative	-.151
Kolmogorov-Smirnov Z		1.002
Asymp. Sig. (2-tailed)		.268

a. Test distribution is Normal

b. Calculated from data

g. Similarly, the distribution for preference is normal $N(4.23, (1.538)^2$ as we do not reject the null hypothesis because the *p*-value is 0.146. (The test statistic value is the Kolmogorov-Smirnov *Z* value, which here is 1.143.)

One-Sample Kolmogorov-Smirnov Test

		Preference
N		44
Normal Parameters [a,b]	Mean	4.23
	Std. Deviation	1.538
Most Extreme Differences	Absolute	.172
	Positive	.172
	Negative	-.123
Kolmogorov-Smirnov Z		1.143
Asymp. Sig. (2-tailed)		.146

a. Test distribution is Normal

b. Calculated from data

Commentary and Summary of Discussion points

1. Each data analysis technique has its own purpose. Although sophisticated multivariate techniques are powerful, they usually have stringent assumptions, which must be met, may be difficult to interpret and are quite costly. Basic data analysis can often provide a high degree of insight relatively quickly and at low cost, and can suggest which sophisticated techniques might be most helpful in further analyses.

2. One of the fundamental concepts of marketing is market segmentation, i.e. catering for differences in target markets. All decision-makers appreciate this concept and the cross-tabulation represents differences between groups in the most simple manner. Cross-tabulation can be used effectively to help the researcher and decision-makers gain an understanding of the relationships within the data. If decision-makers appreciate the limitations of cross-tabulations, they may be drawn into overcoming these by engaging with more sophisticated techniques.