Journal of Mathematical Sciences 3 (2016) 32-41
ISSN 2372-5214

Betty Jones & Sisters Publishing
http://www.bettyjonespub.com

Concept Map for learning Applied Statistics

Osama Ajami Rashwan
Mathematics Department, Ajman University of Science & technology, Fujairah 2202, UAE

Abstract: In this paper a challenge in studying applied statistics has been treated. To teach students studying statistics as a subsidiary course and using it in their work, a need for a map for the method of teaching and the instructor's guide interrelation. This can be eased by using a concept map with the most available tools, Normal Calculator and Microsoft Excel to create a collaboration between using the technology tools and teaching the concepts.

Keywords: concept map, applied statistics, hypothesis test, microsoft excel, data analysis

1. Introduction

A concept map given in [1], using the most available tools, the Calculator and the Microsoft Excel has been modified in this work to be below (see Fig. 1).

Following [2], the goals of statistics educational reforms are to change attitudes towards statistics, and to improve the teaching and learning of statistics. To achieve these goals, a large number of research studies conducted from various perspectives, which divided into three categories:

1) Teaching and learning methods;

2) Using technology in statistics education;

3) The evaluation of the teaching and learning methods suggested by researchers.

Some of the aims of this work is to let students gain understanding of ideas such as the following:

1) The idea of variability of data and summary statistics.

2) The normal distribution is a useful model though it is seldom perfect fits.

3) The usefulness of sample characteristics (and inference made using these measures) depends critically on how sampling conducted.

4) A correlation between two variables does not imply cause and effect.

5) Statistical conclusions should not blindly accepted [3].

Statistics formulas will not be written because this is not the issue unless to clarify the concept.

The starting point is to distinguish between data's type:

Quantitative Data: A data that is expressed numerically. Any mathematical manipulation carried out on them will have meaning, i.e: height, length, volume, etc.

Qualitative Data: A data that identify an item non numerically, e.g., marital status, car color, occupation.

Below are some of the basic concepts:

Independent variable (Factor): It is an explanatory variable manipulated by the experimenter. Each factor has two or more levels (groups or treatments) which are aspects of the factor. Combinations of factor levels (treatments).

Table 1 below shows independent variables, factors, levels, and treatments for a hypothetical experiment.

Table 1 Independent variables, factors, levels, and treatments for a hypothetical experiment

		Vitamin C		
		0 mg	250 mg	500 mg
Vitamin E	0 mg	Treatment 1	Treatment 2	Treatment 3
	400 mg	Treatment 4	Treatment 5	Treatment 6

Corresponding author: Osama Ajami Rashwan, Ph.D., research fields: group theory. Email: oarashwan@gmail.com.

Dependent variable: In the hypothetical experiment above, the researcher is looking at the effect of vitamins on health. The dependent variable in this experiment would be some measure of health (annual doctor bills, number of colds caught in a year, number of days hospitalized, etc.).

There are two types of statistics in following sections.

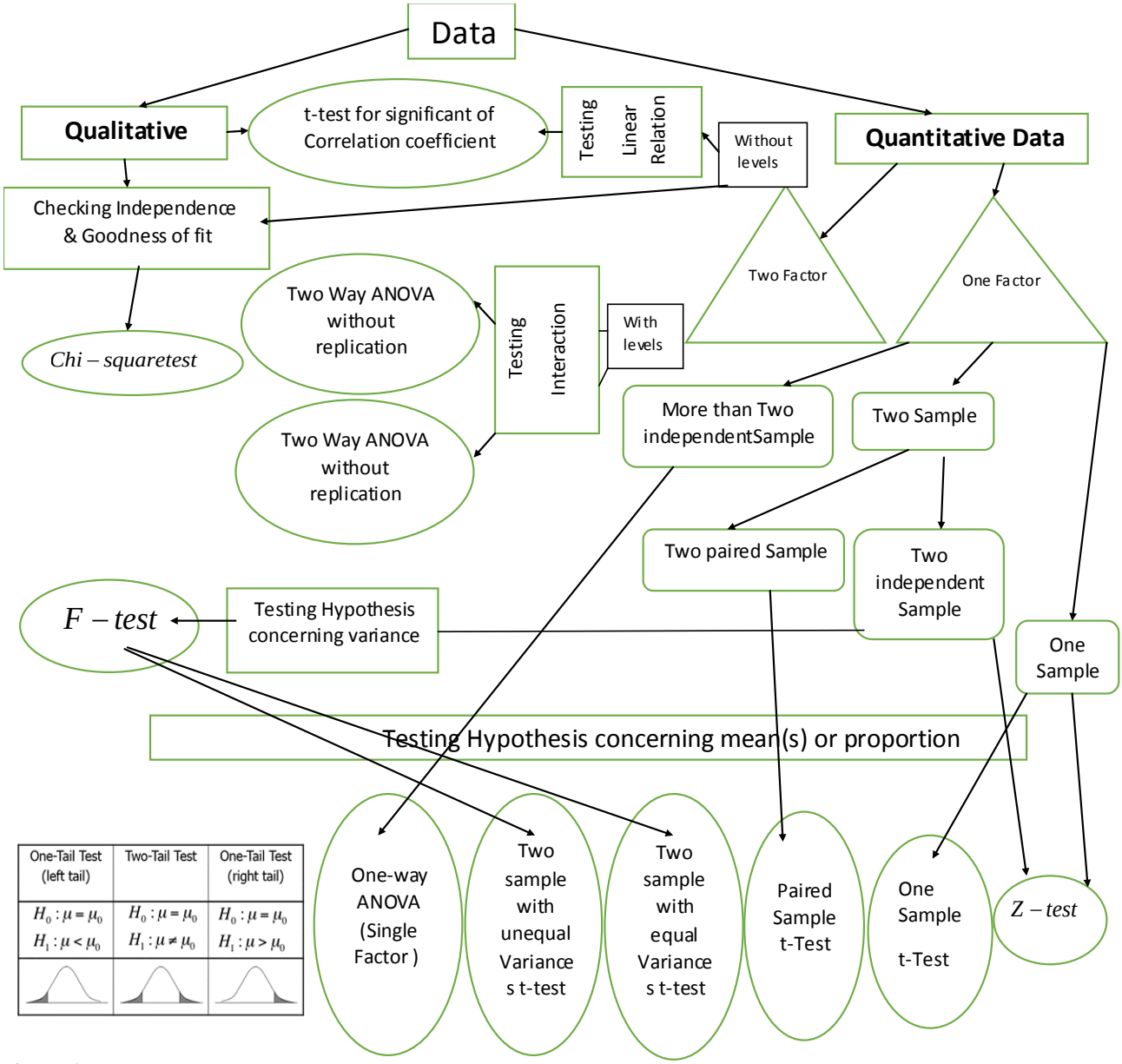


Fig 1 Concept map.

2. Descriptive Statistics

2.1 Describing Central Tendency

It is a value that indicates where the middle of the data set is located) such as:

1) Mean, μ, is the average or expected value;

2) Median, M_d, is the middle point of the ordered measurements;

3) Mode, M_o, is the most frequent value;

4) The *p*th percentile of a data set is a value such that at least *p* percent of the items take on this value or less and at least (100 - *p*) percent of the items take on this value or more.

2.2 Measures of Variation

It is a measure of the amount that the data values vary among themselves) such as:

1) The range is the largest minus the smallest measurement;

2) The variance is the average of the sum of the square of the deviations from the mean;

3) The standard deviation is the square root of the variance.

3. Inferential Statistics

Applies to generalizations made about the group studied such as, Biostatistics (The mathematics of collection, organization and interpretation of numeric data having to do with living organisms).

4. Normal Populations

If a population has mean μ and standard deviation σ and described by a bill-shape curve (see Fig. 2)

The standard normal distribution is a normal distribution with a zero mean and a standard deviation is one (see Fig. 3).

Normal distributions, has mean μ and standard deviation σ, can be transformed to standard normal distributions that is called *z*-scores, by the formula: $Z = \frac{X_i - \mu}{\sigma}$.

4.1 Statistical Testing (Hypothesis Testing):

A statistical hypothesis is an assumption about a population parameter. This assumption may or may not be true. Hypothesis testing refers to the formal procedures used by statisticians to accept or reject statistical hypotheses.

The two hypotheses are the null hypothesis and the other the alternative or research hypothesis. The usual notation is:

H_0: — *the "null" hypothesis*

H_1: — *the "alternative" or "research" hypothesis*

The null hypothesis will always state that the parameter equals the value specified in the alternative hypothesis.

The best way to determine whether a statistical hypothesis is true would be to examine the entire population. Since that is often impractical, researchers typically, examine a random sample from the population. If the sample data are not consistent with the statistical hypothesis, the hypothesis will be rejected. There are two possible errors. A type I error occurs when we reject a true null hypothesis. A type II error occurs when we do not reject a false null hypothesis. Table 2 explains when a test is one tail or two tails.

The smaller the *p*-value, the more statistical evidence exists to support the alternative hypothesis. What is meant by the value of α:

1) If $0 < \alpha < 0.01$, there is an overwhelming evidence (highly significant);

2) If $0.01 < \alpha < 0.05$, there is a strong evidence (significant);

3) If $0.05 < \alpha < 0.1$, there is a weak evidence (not significant);

4) If $\alpha > 0.10$, there is no evidence (not significant).

Unlike the stat-tables, Microsoft Excel has more choices for the value of α.

4.2 Degrees of Freedom

Is the number of scores in our sample that are free to vary and it is equal to the data size minus 1.

Now, have sample(s), so

1) Determine the appropriate test;

2) Establish the level of significance: α;

3) Formulate the statistical hypothesis;

4) Calculate the test statistic;

5) Determine the degree of freedom (not for *Z*-test);

6) Compare computed test statistic against a tabled (critical) value;

7) It is easier to compare the assumed level of significant α, with the *p*-value in the Excel result. If the resulting *p*-value greater than the assumed level of significant α, there is no difference, accept the null

hypothesis. Contrarily, if the resulting p-value less than assumed level of significant α, there is difference, reject the null hypothesis, and say that our alternative hypothesis is statistically significant.

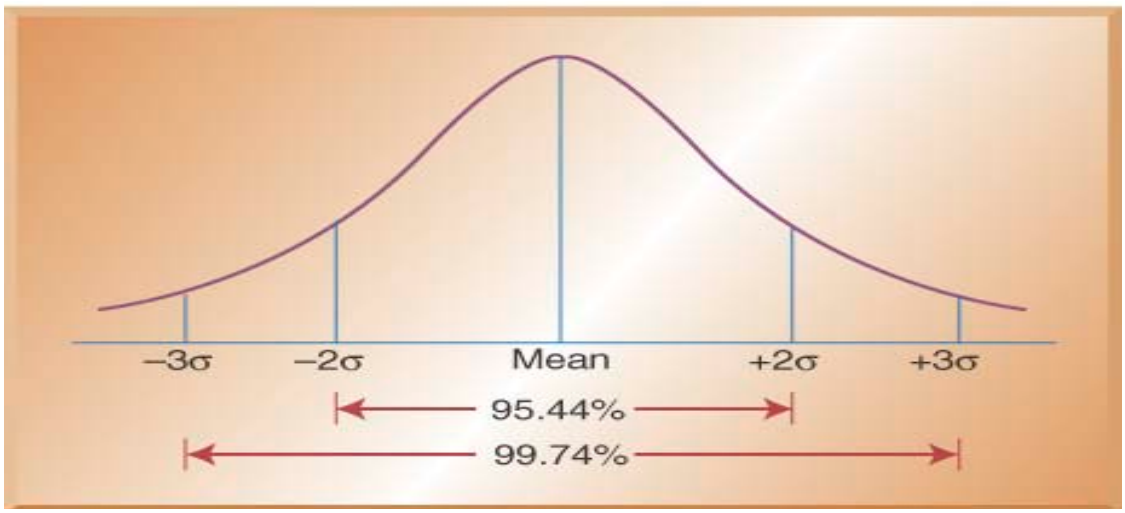


Fig 2 Normal Distribution with mean μ and standard deviation σ.

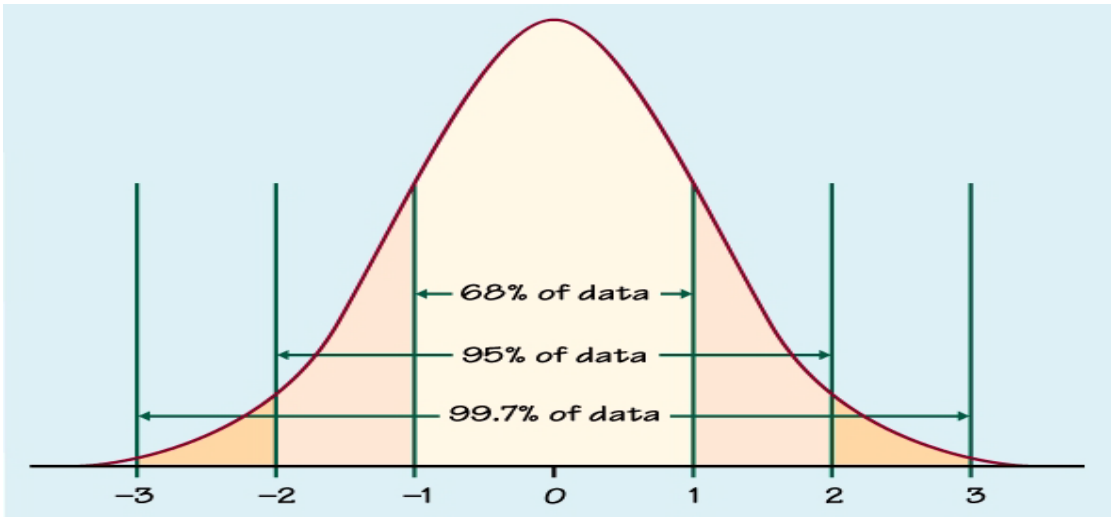
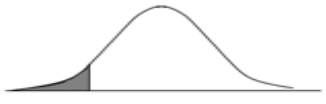
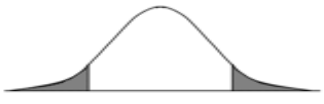
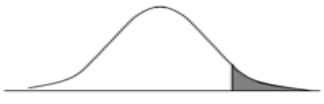


Fig 3 Normal Distribution with mean μ=0 and standard deviation σ = 1.

Table 2 Right tail, Left tail, and Two-tail.

One-Tail Test (left tail)	Two-Tail Test	One-Tail Test (right tail)
$H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0$	$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$
		

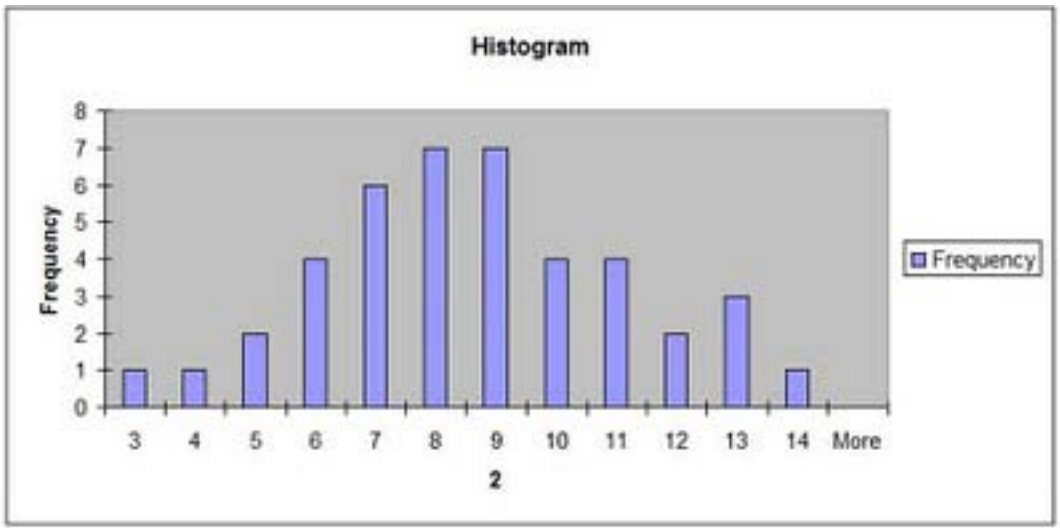


Fig. 4 Histogram.

5. Method and Contents

Using the normal calculator as well as Microsoft Excel to manage the applied statistics calculations, avoiding memorizing the formula and the long theoretical method of calculation, letting some of students discussing their graduate projects that have statistical ideas to know how to choose the right concepts and method. Before using the Microsoft Excel, we have some general notes:

1) Avoid shading the title of the (sample) column or row when the used excel function window does not have the label choice. Such as ANOVA—two ways (Factors) with replication does not have the label choice but ANOVA—two way (Factors) without replication has and so does ANOVA one way (Single Factors);

2) The Microsoft Excel result does not give an answer for your statistical questions or analysis. Therefore, the concepts are very important beside this tool.

3) Sometimes we have to find out how to compute some concepts involved in the process of testing or function such as Chi-square test.

4) Remember first, to activate the data analysis icon in your excel.

6. Checking the Normality of Quantitative Data

The larger the sample size, the better the judgment of normality. The mean is the best measure of central tendency if the distribution is normal, i.e.,

Most scores "bunched up" in middle;

Extreme scores are less frequent, therefore less probable.

We stress on checking normality because it is an assumption for most of statistics tests.

6.1 Method 1

If the Histogram of any data is symmetrical bill-shape similar to the normal distribution curve, now we are ready to create a Histogram with Excel:

Click Data→Data Analysis→Histogram

In the coming dialogue box, highlight the input data and bin range data. Hit the OK button. If the result looks like this (see Fig. 4).

Compare this to the normal distribution curve.

6.2 Method 2

When Mena = Median = Mode, for any data, it is normally distributed. It is easy get them by accessing the descriptive Statistics Excel:

Click Data→Data Analysis→Descriptive Statistics.

This method fails when our sample has more than one mode (it is possible).

6.3 Method 3

If the data Skewness (has value between "-1" and "+1") equals zero, the histogram is symmetric about the mean. As in Method 2, it is easy to get the data Skewness by accessing the descriptive Statistics Excel.

6.4 Method 4

The kurtosis characterizes the relative peakedness or flatness of a distribution compared to the normal distribution. The kurtosis of a normal distribution is three. It is in the descriptive data too.

6.5 Method 5

If the relation between the original data X_i and its *z*-scores is linear, then the data is normally distributed. To check that, first, you order the data ascendingly without repetition and find the corresponding *z*-scores using the function "STANDARDIZE" which needs the mean and the stander deviation beside the value that need to be standardized. Then we use the plotting option scattered with straight line and marks.

Before giving some types of statistical tests, recall that: It is easier to compare the assumed level of significant α, with the p-value in the Excel result. If the resulting *p*-value greater than the assumed level of significant *α*, there is no difference, accept the null hypothesis. Contrarily, if the resulting *p*-value less than assumed level of significant *α*, there is difference, reject the null hypothesis, and say that our alternative hypothesis is statistically significant.

7. Comparing Mean(s)

We have the following three cases:

7.1 One Sample Test

To compare the population mean with a value x, the null hypothesis, is $H_0 : \mu_0 = x$.

For one sample, no need to input the α-value and there is no functioning to the Data Analysis of Microsoft Excel, so we use the function and its result is the *p*-value.

1) *Z*-test: Used if the standard deviation, sigma of the sample population is given or the sample size is big (more than 30) in that case take the standard deviation of thesample (array) instead of the unknown sigma.

Click: FORMULA→Insert Function→Z.TEST, then shading the array (sample), insert the compared meanx, and the standard deviation of the population, where the Excel command is "= ZTEST (array, x, sigma)" and click OK. The Excel result is the p-value P (Z < a).

Note that Excel has only a right tail z-test, so for the left tail we subtract the value from 1. In the case of two-tails z-test, p-value= (1/2) (1-the result).

2) *T*-test: It is used the standard deviation of population is unknown or your sample size is small (less than 30).

The *T*-test function is not for one sample, but it is for two samples t-test.

Click FORMULA→Insert Function→T.TEST

Here is the trick, you have to input two arrays one of them is your sample data and the other is similar array size but with the same values as the one you want to compare with (up to the null hypothesis). In addition, you choose 3 in the box of the type (you consider they are unequal variance samples), and the Excel command is "= T.TEST (array 1, array 2, tails, type)". Choose 1 for one tail or 2 for two tails. Then by clicking OK, The result is the *p*-value *P* (t < a).

7.2 Two Samples Test

To compare the population means μ_1 and μ_2, the null hypothesis is $H_0 : \mu_1 = \mu_2$ which means there is no difference between the means.

1) Z-test: When the variances of the two populations are given or the sample sizes are big (more than 30) in that case take the variances of the two samples (array) instead of the unknown two population variances.

Click Data→ Data Analysis→Z.TEST: Two Sample for Means.

Click Ok after inserting the two samples. Also The population of the first sample variance, and the

population of the second sample variance (if known or use the sample 1 variance and sample 2 variance instead). Note that you tick labels if you insert the label cell with our data. Finally input the α-value. The result is including the one tail and two tails.

2) *t*-test: It is used when the variances of the two populations are unknown or the two samples sizes are small (less than 30). It is important to know that there are three different types of two samples *t*-test and the right one fit to the case of study chosen:

(a) Paired-two samples: tests the relationship between 2 linked samples, e.g. means obtained in two conditions by a single group of participants; (b) Independent two-samples: (equal variance); and (c) Independent two-samples: (unequal variance).

Click Data→ Data Analysis and choose the suitable *t*-test type. Click Ok after inserting the two samples and the *α*-value (Alpha).The result including the one tail and two tails.

Note that, to decide using (b) or (c), first use the *F*-test (to compare the two populations' variances) as in the following section.

To compare the population means of more than two samples use:

7.3 One-way Anova (Single Factor)

It is to test the null hypothesis $H_0 : \mu_1 = \mu_2 = ... = \mu_c$, which means there is no difference among the means. Click Data→ Data Analysis, and choose: Anova: single factor. Then click Ok, to insert, in the coming window, the samples (in columns or rows and tick your choice) and the α-value. The output given as two tables, the first provides descriptive statistics and the second the actual ANOVA table. In the result, compare the *p*-value with the assumed α to accept or reject the null hypothesis.

8. Comparing Variances (F-test)

Testing that there is no difference between the variances σ_1^2, σ_2^2 of the two populations $H_0 : \sigma_1^2 = \sigma_2^2$.

Click Data→Data Analysis, choose *F*-test Two Sample for Variance and click Ok. In the coming window, insert your samples (in columns or rows and tick your choice) and the assumed α. Excel calculates the correct *F* value, which is the ratio of the first sample and the second one. Be sure that the variance of the first sample is higher than the variance of the second sample. If not swap them. In the result, compare if the *F* stat value is less than the *F* critical, which is only for one tail, therefore the variances of the two populations are equal; otherwise, the variances of the two populations are unequal. In addition, we may compare the p-value with the assumed α with the p-value *P* ($F \leq f$).

8.1 Testing the Interaction (Two-way ANOVA)

Mainly either test for the main effect of one of the two factors. It is an extension to One-way ANOVA because there are two independent variables (nominal factors), hence the name two-way ANOVA is given. Each nominal factor has two or more levels, and the degrees of freedom for each factor is one less than the number of levels.

The assumptions needed to use the two-way ANOVA are:

All samples drawn from normally distributed populations;

All samples drawn independently from each other;

All populations have common variance;

Within each sample, the observations sampled randomly of each other.

A two way ANOVA usually done with replication (more than one observation for each combination of the nominal variable level) and called a two-way ANOVA with replication. Nevertheless, if there is only one observation for each for each factor level called a two-way ANOVA without replication but this is less informative (we can't test the interaction term).

Let us know how to do each type of two-way ANOVA:

1) A two-way ANOVA with replication:

Unlike any other statistical tests, there are three sets of hypothesis H_0:

(a) The population means of the first factor are equal. This is like the one-way ANOVA for the row factor,

(b) Similar to the one-way ANOVA for the column factor, the population means of the second factor are equal, and

(c) There is interaction between the two factors or not; that means whether the effect of one factor depends on the other factor. This is similar to performing a test for independence with contingency tables, as we see later in the section "Test of Independence and Goodness Fit".

Click Data→Data Analysis. Use Anova: Two factor with replication.

Click Ok to input your data in the coming window, and the α-level or use the default (5%).

The output given as two tables: the first provides descriptive statistics and the second the actual ANOVA table. In the actual ANOVA table:

If the *p*-value for columns (rows) is larger than the chosen alpha level, suggesting that any effect we saw could well have been due to chance;

If the *p*-values for columns (rows) is smaller than the chosen alpha level, suggesting that any effect we saw could not well have been due to chance;

If the *p*-values for interaction is larger than your chosen alpha level, suggesting that there are no significant differences in the interaction between the two factors;

Finally if the *p*-values for columns (rows) is smaller than the chosen alpha level, suggesting that conclude there are significant differences in the interaction between the two factors.

2) A two-way ANOVA without replication:

Here it is required to assume that there is no interaction because there are no replications. So there are two sets of hypothesis:

The population means of the first factor are equal. This is like the one-way ANOVA for the row factor;

Similar to the one-way ANOVA for the column factor, the population means of the second factor are equal.

Click Data→ Data Analysis. Use Anova: Two factor with replication.

Click Ok to input your data in the coming window, and the α-level.

The output given as two tables: the first provides descriptive statistics and the second the actual ANOVA table. In the actual ANOVA table: Similar to the two way ANOVA, if we find a significant difference in the means for one of the main effects, we should not know whether that difference was consistent for different values of the other main effect.

8.2 Regression Analysis

To predict the value of one variable (the dependent variable), on the basis of the other variables (the independent variables), the Linear regression equation and the correlation coefficient are easy to get using the normal calculator as well as the excel function is "CORREL".

There are two type of the correlation coefficient, one for the quantitative data called Pearson Correlation Coefficient and the other for qualitative data called Spearman rank order Correlation Coefficient. Note that, if we replace the quantitative data by their order rank, we may use Spearman rank order Correlation Coefficient. Pearson correlation coefficients measure only linear relationships. Spearman correlation coefficients measure only monotonic relationships. Therefore, a meaningful relationship can exist even if the correlation coefficients are zero.

1) If the correlation coefficient is close to +1 that means you have a strong positive relationship;

2) If the correlation coefficient is close to "-1" that means you have a very strong negative relationship;

3) If the correlation coefficient is close to 0 that means you have no correlation.

t-test for significant of correlation: to test the null hypothesis: $H_0: \rho = 0$, i.e., there is no correlation

between the two variables in the population, against the alternative hypothesis:

$H_1: \rho \neq 0$, i.e., there is a significant correlation between the two variables. There is no such test in Excel, so we compare from the *t*-table the critical values $\frac{t_a}{2}, n-2$ with the stat one $t = r\sqrt{\frac{n-2}{1-r^2}}$, where r is the correlation coefficient of the couple variable sample and n is the number of couples.

If $t_{\frac{\alpha}{2},n-2} > t = r\sqrt{\frac{n-2}{1-r^2}}$, this means that there is no correlation between the two variables.

If $t_{\frac{\alpha}{2},n-2} < t = r\sqrt{\frac{n-2}{1-r^2}}$, this means that there is a significant correlation between the two variables in the population.

Now, let us complete the study of association and goodness fit:

Chi-square test: to determine how well observed values of the two variable match with values expected by a theoretical model.

H_0: The two variables are independent

H_1: The two variables are associated

There is no function included in the Data Analysis of Excel, so we use the function of the test. Here we cannot go for the Excel function CHISQ.TEST unless we get a table of the correspondences expected frequencies using

$F_e = F_r F_c / N$, where F_r is the row frequency, F_c is the column frequency and N is the sum of F_r or F_c (the result are equal).

Click FORMULA→ Insert Function CHISQ. TEST

After shading the actual observations, and the expected observations. Note that Excel result of the Chi-square test is not the Chi-square value, but it is the p-value. Then we compare the result with the assumed α-value to reject or accept the null-hypothesis.

9. Conclusion

1) The statistics concepts should be understood not memorizing their formulas;

2) The most available tools such as the normal calculators and Microsoft Excel is better to use in teaching statistics;

3) One must know what the case of study is;

4) It is important to know how to interpret the results when using a software for statistical analysis;

5) Statistics taught in labs is better to ease the calculations and concentrate in the concepts;

6) Theoretical statistics exam involves questions about the concepts and their meaning;

7) Discussing graduate projects, that have statistical ideas to know how to choose the right concepts and method to apply, is worth, more than teaching theoretical examples;

8) Statistical background and statistical software skill are necessary in solving statistical problems in different areas. Applied Statistics course description should split into three parts. The first is teaching the statistics concept and its link to the students' field. The second part is how to use the tools to solve the statistical problems. The third includes some applications (see Fig. 5)

Fig 5

References

[1] A. Gosofsky, Instructor's Guide to Using Research Methods and Statistics Concept Maps, Office of Teaching Resources Psychology, 2011.

[2] S. Tishkovskaya, G.A. Lancaster, Statistical Education in the 21st Century: A review of challenges, teaching innovations and strategies for reform, Journal of Statistics Education 20 (2) (2012).

[3] J.M. Wicherts, M. Bakker, D. Molenaar, Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results, PLOS ONE, 2011.