

SPECIAL NOTES FOR NEW USERS OF EZZE CORRE (CORRELATION)

WARRANTY

The author is not liable for errors contained herein or for incidental or consequential damages in connection with the furnishing, performance or use of this material. **All technical application software is inherently complex and users are cautioned to verify the results.**

TERMINOLOGY

In the EZZE CORRE program there is some terminology used which may not be familiar to you. Definitions and significance of calculated parameters can be found in the references listed below or the USER GUIDE.

TECHNOLOGY

EZZE CORRE is a data analysis utility which includes the following:

(i) HYPOTHESIS (Null, Student - t, Chi Square) TESTS OF DATA

The purpose of hypothesis testing is to test the viability of a hypothesis in the light of experimental data. Depending on the data, the hypothesis either will or will not be rejected as a viable possibility. Which test, when and how to use it is discussed in the references below and the tutorial.

(ii) CORRELATION (Pearson) COEFFICIENT

The coefficient calculates the strength of the correlation between up to 5 measured variables with which one may predict future results or outcomes based on regression analysis of sets of measured variables.

(iii) REGRESSION MODELLING

The program uses Least Squares criteria for "Goodness-of-fit to develop a model and evaluate the model's validity. Regression models are powerful tools for predicting a score based on some other score. With the model EZZE CORRE estimates exact scores, called point estimates and intervals of scores(interval estimates) for up to 5 measured variables (10 relationships).

(iv) PROBLEM TEMPLATES

Included are sample problems which may be used as a templates in your investigation:
The user is cautioned to verify the results using his knowledge of the data source and good engineering judgement.

REFERENCES


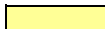

BOOKS OR MANUALS (FREE)

<http://www.itl.nist.gov/div898/handbook/index.htm>
AND
DataMyte Handbook, A practical guide to computerized data collection for Statistical Process Control.
<http://www.ab.com/events/pressrel/9603/960319.html>

INTERNET SITES

www.isixsigma.com	good reference & forum
http://yeivier.20m.com/statistics/MSA/MSA.html	Excellent and complete article covers theory that you need
http://www.aiqusa.com/index.asp	organization + excellent reference material

NOTE: THE PROGRAM DOES NOT REQUIRE THE STATISTICS "TOOLPAK" ADD-IN IT REQUIRES ONLY THE STANDARD EXCEL FUNCTIONS, EZZE CORRE DOES THE "MATH"

- 2 **DATA ENTRY CELLS ARE HIGHLIGHTED** 
- CALCULATED DATA IS HIGHLIGHTED** 
- USER INPUT & MANUAL DATA ENTRY HIGHLIGHTED** 
- 3 THIS PROGRAM HAS FOUR KEYPADS.
- | | |
|-----------|-------------------------------------|
| KEYPAD | CORRELATION ANALYSIS AND ESTIMATION |
| TKEYPAD | STUDENT - t TESTS |
| CHIKEYPAD | CHI SQUARE TESTS |
| SHEET1 | NULL HYPOTHESIS TESTS |

THE PROGRAM IS COMPLEX DUE TO THE NUMBER OF TESTS EACH KEYPAD CAN PERFORM. THIS REQUIRES INTERACTION WITH THE PROGRAM TO ENSURE THE INPUTS YOU REQUIRE ARE DONE PROPERLY BY CLEARLY DEFINING YOUR PROBLEM AND PROVIDING THE INFORMATION REQUIRED BY THE TEST.

SHEETS THIS PROGRAM

READMEFIRST	THIS PAGE (CONTAINS "USER GUIDE")
KEYPAD	DATA ENTRY AND PROGRAM KEYPAD CORRELATION
Sheet1	DATA ENTRY AND PROGRAM KEYPAD NULL HYPOTHESIS
TKEYPAD	DATA ENTRY AND PROGRAM KEYPAD T HYPOTHESIS
CHIKEYPAD	DATA ENTRY AND PROGRAM KEYPAD CHI SQUARE
CORRELATIONS	WORKSHEET CORRELATION AND REGRESSION
NULLWKST	NULL TEST PROGRAM SHEET
TESTDATA	HYPOTHESIS TEST EXAMPLES
OTHERTESTS	CHI & STUDENT t WORKSHEET
DATA COR	SAMPLE CORRELATION DATA
TREPORT	STUDENTS- t HYPOTHESIS REPORT
CHIREPORT	CHI SQUARE HYPOTHESIS REPORT
CORREP	CORRELATION REGRESSION REPORT
QUICKCHART	GRAPHICAL REGRESSION ANALYSIS OF SCATTER PLOT

USER'S GUIDE

Hypothesis testing often confuses people but it is the keystone of most statistical applications. Every acceptance sampling test, designed experiment, and control chart* is a statistical hypothesis test.

All hypothesis tests have unavoidable, but quantifiable, risks of making the wrong conclusion. Statistical tests always involve Type I (producer's or alpha) and Type II (consumer's or beta) risks. The Type I risk is the chance of deciding that a significant effect is present when it isn't. The Type II risk is the chance of not detecting a significant effect when one exists.

Null and Alternate Hypothesis

A null hypothesis is a statistical hypothesis that is tested for possible rejection under the assumption that it is true (usually that observations are the result of chance). The concept was introduced by R. A. Fisher.

"Accepting the null hypothesis" is like acquitting a defendant. It does NOT prove that the null hypothesis is true, or that the defendant is innocent. It means there is a reasonable doubt about the defendant's guilt. In statistical testing, the significance level, Type I risk, or alpha risk is the "reasonable doubt." It is the chance of wrongly rejecting the null hypothesis when it is true. In acceptance sampling, it is the producer's risk, or risk of wrongly rejecting a lot that meets requirements.

The alternate hypothesis is that the process change or treatment has an effect, or something is wrong with the process. The Type II risk is the chance of accepting the null hypothesis when it is false. The "consumer's risk" is the Type II risk for an acceptance sampling plan. It is the chance of passing a lot that does not meet the requirements. If the Type I risk is the chance of crying wolf, the Type II risk is the chance of not seeing a real wolf. The following table explains hypothesis testing and risks.

Hypothesis Testing

Hypothesis testing is the use of statistics to determine the probability that a given hypothesis is true. The usual process of hypothesis testing consists of four steps.

- 1 Go to Sheet1 (NULL KEYPAD)
 - 2 DEFINE YOUR PROBLEM - PROVIDE SAMPLE DATA IN THE BLANKS
 - 3 CLICK THE NULL ACTION BUTTON TO PROCESS YOUR HYPOTHESIS
 - 4 CHECK YOUR COMPLETED NULL REPORT
- ITS THAT EASY - THE PROGRAM AUTOMATICALLY SELECTS THE EQUATIONS SUITABLE TO YOUR PROBLEM AND DOES THE CALCULATIONS !!
- REMEMBER TO CHECK TESTDATA - IT HAS A SET OF SAMPLE PROBLEMS WHICH MAY BE HELPFUL IN SOLVING OR DEFINING YOUR PROBLEM

Other Hypothesis Testing.

STUDENTS T

IT IS USED TO ESTABLISH CONFIDENCE LIMITS AND TEST THE HYPOTHESIS WHEN THE POPULATION VARIANCE IS NOT KNOWN AND THE SAMPLE SIZE IS SMALL (<30)

THE "T" TEST IS BASED ON THE ASSUMPTION THAT THE SAMPLES COME FROM A NORMALLY DISTRIBUTED POPULATION. ALSO THE TEST REQUIRES AN ASSUMPTION ABOUT THE TYPE OF POPULATION OR PARAMETERS AND IS CONSEQUENTLY KNOWN AS A PARAMETRIC TEST.

EZZE CORRE RUNS THREE STUDENT - T TESTS FROM "TKEYPAD".
THE INSTRUCTIONS FOR THE 3 TESTS LISTED ON THE KEYPAD ARE AS FOLLOWS:

- (i) SELECT THE TEST TO BE RUN (<30 samples)
 1. Difference between two samples dependent samples or Matched Paired Observations
 2. Comparison of two samples Means Independent samples
 3. Comparison of Sample and Population means
- (ii) ENTER YOUR DATA INTO DATA TABLE BELOW PER TEST
- (iii) ENTER CRITICAL t VALUE CORRESPONDING TO THE ALPHA OF YOUR TEST
- (iv) ENTER HYPOTHESIS + PROBLEM TITLE/DESCRIPTION
- (v) CLICK BUTTON CORRESPONDING TO TEST YOU ARE RUNNING
- (vi) CHECK YOUR DATA AND THE RESULTS OF YOUR T TEST
- (vii) PREPARE AND PRINT YOUR REPORT

CHI SQUARE

THERE ARE TIMES WHEN IT IS IMPOSSIBLE TO MAKE ASSUMPTIONS ABOUT THE SAMPLES POPULATION DISTRIBUTION HENCE THE NEED FOR NON PARAMETRIC TESTS LIKE CHI SQUARE

CHI SQUARE(χ^2) CAN BE USED TO EVALUATE A RELATIONSHIP BETWEEN TWO NOMINAL OR ORDINAL VARIABLES. IT IS A MEASURE OF THE DIVERGENCE OF OBSERVED AND EXPECTED FREQUENCIES. IF THERE IS NO DIFFERENCE THEN $\chi^2 = 0$, A DIFFERENCE WILL RESULT IN A VALUE >0 . χ^2 ENABLES US TO DETERMINE IF THE DIVERGENCE BETWEEN THEORY AND FACT IS SIGNIFICANT OR NOT.

KEY POINT

IF THE VALUE OF χ^2 IS VERY SMALL COMPARED TO ITS TABLE VALUE THEN THE FIT IS VERY GOOD.
IF THE VALUE OF χ^2 IS VERY LARGE COMPARED TO ITS TABLE VALUE THEN THE DIVERGENCE BETWEEN EXPECTED AND OBSERVED FREQUENCES IS VERY BIG AND THE FIT IS POOR

TO RUN A CHI TEST GO TO 'CHIKEYPAD' AND FOLLOW THE INSTRUCTIONS (REF COPY BELOW)

- 1 ENTER THE HYPOTHESIS & PROBLEM DESCRIPTION
- 2 ENTER DATA INTO 'CHI DATA ENTRY'
**NOTE MULTIPARAMETER(READING) IS LIMITED TO
A 4X6 MATRIX**
- 3 WHEN DATA ENTRY IS COMPLETE NOTE USER ENTRY
PORT BELOW FOR DEGREES OF FREEDOM ENTER CRITICAL CHI VALUE
- 4 CLICK THE APPROPRIATE CHI TEST KEY
FINISHED - CHECK REPORT
FINISHED - PRINT REPORT

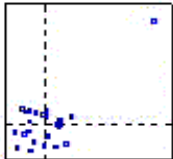
CORRELATIONS - ESTIMATION

Correlation analysis measures the relationship between two items, for example, a security's price and an indicator. The resulting value (called the "correlation coefficient") shows if changes in one item (e.g., an indicator) will result in changes in the other item (e.g., the security's price).

When comparing the correlation between two items, one item is called the "dependent" variable and the other the "independent" variable. The goal is to see if a change in the independent variable will result in a change in the dependent variable. This information helps you predict the "value" of a dependent variable from a given value of the independent variable

Correlations tend to be positive. Pick any two variables at random and they'll almost certainly be positively correlated, if they're correlated at all--height and weight; saturated fat in the diet and cholesterol levels; amount of fertilizer and crop yield; education and income. Negative correlations tend to be rare--automobile weight and fuel economy; number of cigarettes smoked and child's birth weight.

A note of caution what do you do if your scatter plot has an orphan value (outlier) as per the following



The correlation is 0 within the bulk of the data in the lower left-hand corner. The outlier in the upper right hand corner increases both means and makes the data lie predominantly in quadrants I and III. Check with the source of the data to see if the outlier might be in error. Errors like these often occur when a decimal point in both measurements is accidentally shifted to the right. Even if there is no explanation for the outlier, it should be set aside and the correlation coefficient or the remaining data should be calculated. The report must include a statement of the outlier's existence. It would be misleading to report the correlation based on all of the data because it wouldn't represent the behavior of the bulk of the data.

Correlation coefficients are appropriate only when data are obtained by drawing a random sample from a larger population. However, sometimes correlation coefficients are mistakenly calculated when the values one of the variables--X, say--are determined or constrained in advance by the investigator. In such cases, the message or the outlier may be real, namely, that over the full range of values, the two variables tend to increase and decrease together. It's poor study design to have the answer determined by a single observation and it places the analyst in an uncomfortable position. It demands that we assume the association is roughly linear over the entire range and that the variability in Y will be no different for large X from what it is for small X. Unfortunately, once the study is performed, there isn't much that can be done about it. The outcome hinges on a single observation.

THE MORAL - ALWAYS LOOK AT THE SCATTERPLOTS!

The correlation coefficient is a numerical summary and, as such, it can be reported as a measure of association for any batch of numbers, no matter how they are obtained. Like any other statistic, its proper interpretation hinges on the sampling scheme used to generate the data.

The correlation coefficient is most appropriate when both measurements are made from a simple random sample from some population. The sample correlation then estimates a corresponding quantity in the population. If the data does not constitute a simple random sample from some population, it is not clear how to interpret the correlation coefficient. This distortion most commonly occurs in practice when the range of one of the variables has been restricted

Another source of misleading correlation coefficients is *the ecological fallacy*. It occurs when correlations based on grouped data are incorrectly assumed to hold for individuals.

Imagine investigating the relationship between food consumption and cancer risk. One way to begin such an investigation would be to look at data on the country level and construct a plot of overall cancer risk against per capita daily caloric intake. A scatterplot might show that cancer increases with food consumption. But it is people, not countries, who get cancer. It could very well be that within countries those who eat more are less likely to develop cancer. On the country level, per capita food intake may just be an indicator of overall wealth and industrialization.

The ecological fallacy is in studying countries when one should be studying people.

WHY SHOULD ONE DO MULTIPLE CORRELATIONS IN EZZE CORRE?

Correlation is not causation. The observed correlation between two variables might be due to the action of a third, unobserved variable. Yule (1926) gave an example of high positive correlation between yearly number of suicides and membership in the Church of England due not to cause and effect, but to other variables that also varied over time. (Can you suggest some?) Mosteller and Tukey (1977, p. 318) give an example of aiming errors made during bomber flights in Europe. Bombing accuracy had a high positive correlation with amount of fighter opposition, that is, the more enemy fighters sent up to distract and shoot down the bombers, the more accurate the bombing run! The reason being that lack of fighter opposition meant lots of cloud cover obscuring bombers from the fighters and the target from the bombers, hence, low accuracy.

NOW ON TO CALCULATING CORRELATION COEFFICIENTS!!

PEARSON

The *Pearson Product-Moment Correlation Coefficient* (r), or correlation coefficient for short is a *measure of the degree of linear relationship between two variables*, usually labeled X and Y. While in regression the emphasis is on predicting one variable from the other, in correlation the emphasis is on the degree to which a linear model may describe the relationship between two variables. In regression the interest is directional, one variable is predicted and the other is the predictor; in correlation the interest is non-directional, the relationship is the critical aspect.

The computation of the correlation coefficient is most easily accomplished with the aid of a statistical calculator. The value of r was found on a statistical calculator during the estimation of regression parameters in the last chapter. Although definitional formulas will be given later in this chapter, the reader is encouraged to review the procedure to obtain the correlation coefficient on the calculator at this time.

The correlation coefficient may take on any value between plus and minus one.

$$-1.00 \leq r \leq +1.00$$

The sign of the correlation coefficient (+, -) defines the direction of the relationship, either positive or negative. A positive correlation coefficient means that as the value of one variable increases, the value of the other variable increases; as one decreases the other decreases. A negative correlation coefficient indicates that as one variable increases, the other decreases, and vice-versa.

REGRESSION ANALYSIS - LEAST SQUARES

Regression models are used to predict one variable from one or more other variables. Regression models provide the scientist with a powerful tool, allowing predictions about past, present, or future events to be made with information about past or present events. The scientist employs these models either because it is less expensive in terms of time and/or money to collect the information to make the predictions than to collect the information about the event itself, or, more likely, because the event to be predicted will occur in some future time. Before describing the details of the modeling process, however, some examples of the use of regression models will be presented.

In order to construct a regression model, both the information which is going to be used to make the prediction and the information which is to be predicted must be obtained from a sample of objects or individuals. The relationship between the two pieces of information is then modeled with a linear transformation. Then in the future, only the first information is necessary, and the regression model is used to transform this information into the predicted. In other words, it is necessary to have information on both variables before the model can be constructed.

The situation using the regression model is analogous to that of the interviewers, except instead of using interviewers, predictions are made by performing a linear transformation of the predictor variable. Rather than interviewers in the above example, the predicted value would be obtained by a linear transformation of the score. The prediction takes the form

$$Y' = a + bX$$

where a and b are parameters in the regression model.

Because the two parameters of the regression model, a and b, can take on any real value, there are an infinite number of possible models, analogous to having an infinite number of possible interviewers. The goal of regression is to select the parameters of the model so that the least-squares criterion is met, or, in other words, to minimize the sum of the squared deviations. The procedure which transforms the scale of X to the scale of Y, such that both have the same mean and standard deviation will not work in this case, because of the prediction goal.

The result of these calculations is a regression model of the form:

$$b = \frac{N \sum XY - \sum X \sum Y}{N \sum X^2 - (\sum X)^2}$$

Solving for the a parameter is somewhat easier.

$$a = \bar{Y} - b\bar{X}$$

notation is used to mean the standard deviation of Y given the value of X is known. The standard error of estimate is defined by the formula

$$s_{YX} = \sqrt{\frac{N-1}{N-2} s_y^2 (1 - r^2)}$$

where:
 s_y^2 is the variance of Y
 r^2 is the correlation coefficient squared

WITH "a" AND "b" KNOWN YOU CAN ESTIMATE 'Y' FOR A "GIVEN" X. THIS A POINT ESTIMATE!

TO GET AN INTERVAL(RANGE) ESTIMATE WE USE 'THE STANDARD ERROR OF ESTIMATE' AND 'CRITICAL VALUES @ DIFFERENT CONFIDENCE LEVELS' USED IN OUR NULL HYPOTHESIS TESTS

$$\text{INTERVAL ESTIMATE} = \text{ESTIMATED 'Y'} \pm (\text{CV}) * S(\text{YX})$$

$$\text{WHERE CV} = (1.96@95\%, 2.58@99\%, 3@99.73\%)$$

LET'S BACKTRACK A BIT AND CALCULATE THE 'SLOPE' AND 'INTERCEPT' OF THE X AND Y DATA USING EXCEL

VOILA YOU GET THE SAME "a" AND "b" VALUES AS THE FORMULAE ABOVE - THIS CONFIRMS OUR MATH IS CORRECT AND THE VALIDITY OF THE FIVE VARIABLE TABLE, QUICK ESTIMATOR AND CHART ESTIMATOR IN THIS PROGRAM(SEE BELOW)

IN THE 5 VARIABLE TABLE ARE LEAST SQUARES(FIRST ESTIMATE), REGRESSION MODEL(MATH), AND LINEAR(EXCEL FUNCTIONS) 'Y' ESTIMATES. AS EXPECTED THE LEAST SQUARE (FIRST ESTIMATE) IS/CAN BE VASTLY DIFFERENT. USE THE LINEAR VALUE IN YOUR CALCULATIONS BECAUSE IT IS THE BEST FIT LINEAR REGRESSION FOR THE DATA SET.

ANOTHER "STARTLING" EXCEL FACT IS YOU CAN DO A SCATTER PLOT OF YOUR DATA GENERATE A CHART THEN USING EXCEL MAGIC YOU CAN DO A GRAPHICAL ESTIMATE OF 'Y' (QUICKCHART) THE CHART VALUE = THE TABLE VALUE. THE MAGIC IS THE TRENDLINE FUNCTION IN EXCEL WHICH DRAWS A BEST FIT STRAIGHT LINE (ONE OF SEVERAL OPTIONS) THE LINEAR TREND LINE USES THE LINEAR EQUATION DEVELOPED ABOVE WHICH SHOULD AND DOES PROVIDE THE SAME ESTIMATED VALUE AS THE NUMERICAL METHOD. NOW HOW TO USE THE KEYPAD

- 2 ENTER YOUR CONFIDENCE INTERVAL
- 3 CLICK THE CORRELATION BUTTON AND YOUR CORRELATIONS AND REGRESSIONS ARE COMPLETE
- 4 IF YOU WANT TO DO ONLY ONE REGRESSION ESTIMATE GO TO A80 FILL IN THE DATA AND YOUR REGRESSION IS COMPLETE
- 5 GO TO CORREP CHECK AND FORMAT YOUR REPORT - PRINT IT

bb646@yahoo.ca

USER'S SUPPORT

R. Cuthbert - Hamilton, ON Canada

AM I AN EXPERT YET?

The more I learn the less I know - AN EXPERT IN LESS NOT MORE

