12 What You Need To Know About Statistics

Katrina D. Maxwell

Abstract: How do you measure the value of data? Not by the amount you have, but by what you can learn from it. Statistics provides a way to extract valuable information from your data. It is a science concerned with the collection, classification, and interpretation of data according to well-defined procedures. For a manager, however, statistics is simply one of many diverse techniques that may improve decision-making.

The purpose of this chapter is to develop a deeper understanding of the statistical methods used to analyse software project data. The methods used to analyse software project data come from the branch of statistics known as multivariate statistical analysis. These methods investigate relationships between two or more variables. However, before we delve into detailed explanations of chi-square tests, correlation analysis, regression analysis, and analysis of variance, you need to understand some basic concepts.

Keywords: Statistical concepts, Regression, Correlation, Distribution, sampling.

12.1 Describing Individual Variables

In this section, you will learn how to categorise and meaningfully summarise data concerning individual variables.

12.1.1 Types of Variables

All data is not created equal. Information can be collected using different scales. This has an impact on what method you can use to analyse the data. There are four main types of scales: nominal, ordinal, interval, and ratio.

Nominal scales – Variables such as business sector, application type, and application language are nominal-scale variables. These variables differ in kind only. They have no numerical sense. There is no meaningful order. For example, let’s say that a business sector has four categories: bank,
insurance, retail, and manufacturing. Even if we label these with numbers instead of names in our database (say 101, 102, 103, and 104), the values of the numbers are meaningless. Manufacturing will never be “higher” than bank, just different.

**Ordinal scales** – The values of an ordinal-scale variable can be ranked in order. The 10 risk factors discussed in Chapter 5 of my book “Applied Statistics for Software Managers” are ordinal-scale variables. It is correct to say that Level 5 is riskier than Level 4, and Level 4 is riskier than Level 3, and so on; however, equal differences between ordinal values do not necessarily have equal quantitative meaning. For example, even though there is an equal one-level difference between 3 and 4, and 4 and 5, Level 4 may be 50% more risky than Level 3, and Level 5 may be 100% more risky than Level 4.

**Interval scales** – The values of an interval-scale variable can be ranked in order. In addition, equal distances between scale values have equal meaning. However, the ratios of interval-scale values have no meaning. This is because an interval scale has an arbitrary zero point. A start date variable is an example of an interval-scale variable. The year 1993 compared to the year 1992 only has meaning with respect to the arbitrary origin of 0 based on the supposed year of the birth of Christ. We know that 1993 is one year more than 1992, and that 1991 is one year less than 1992. Dividing 1993 by 1992 makes no sense. For example, we could decide to make 1900 year zero and count from there. In this case, 1993 would simply become 93 and 1992 would become 92 in our new scale. Although in both cases there is a one-year difference, the ratio 1993/1992 does not equal the ratio 93/92.

Another example of an interval scale is a Likert-type scale. Factors are rated on a scale of equal-appearing intervals, such as very low, low, average, high, and very high, and are assigned numerical values of 1, 2, 3, 4, and 5, respectively. However, in real life, it is virtually impossible to construct verbal scales of exactly equal intervals. It is more realistic to recognise that these scales are approximately of equal intervals. Thus, a Likert scale is really somewhere between an ordinal scale and a true interval scale.

**Ratio scales** – Variables such as effort, application size, and duration are measured using a ratio scale. Ratio-scale variables can be ranked in order, equal distances between scale values have equal meaning, and the ratios of ratio-scale values make sense. For example, it is correct to say that an application that required 10 months to develop took twice as long as an application that took 5 months. Another ratio scale is a percentage scale. For example, the percentage of COBOL used in an application is also a ratio-type variable.