

# Understanding Statistics 6

**Elliot Abt**

*Department of Dentistry, Illinois Masonic Medical Center, Chicago, Illinois, USA*

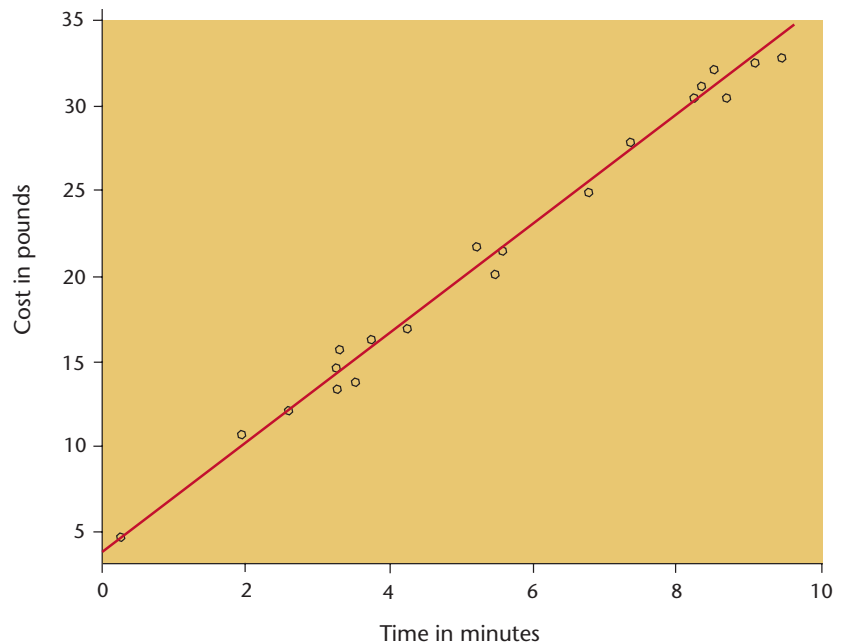
We have previously described three general domains of statistics: differences between groups; association between groups; and time-to-event (survival) data. This article will describe the statistics commonly used in these last two domains: associations between groups; and survival data.

## Associations between groups

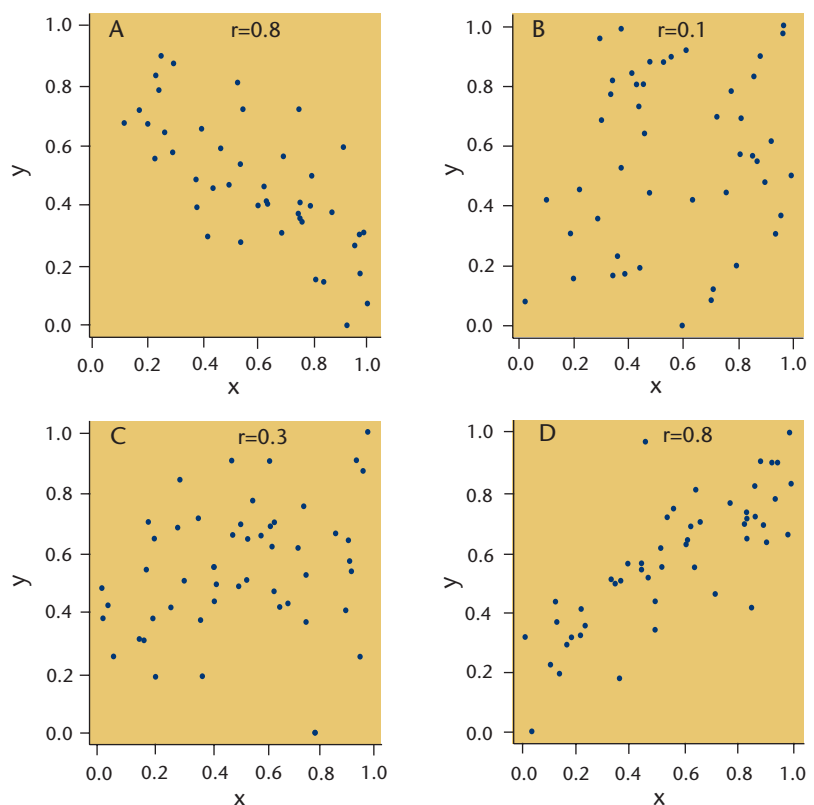
Statistical analysis in this domain generally deals with the correlation and regression. Although these two terms are used synonymously, they refer to slightly different concepts. Correlation is the strength of the association between groups, whereas regression is the nature of this association. For example, if we wanted to know if there is an association between the number of minutes on our cell phone and our phone bill, we could graph this (quantitative) data (Figure. 1).

We would likely find that as we spent more time on our cell phone, our phone bill would increase. If for every additional minute our bill increased by exactly the same amount, we would have perfect positive correlation, with a correlation coefficient of 1, and all the data points would lie on the diagonal line. What is more likely is that we will not have perfect correlation, and the data points will be scattered on either side of the line. **The closer the points are to the line, the stronger the correlation, and the higher the correlation coefficient.** The correlation coefficient takes values between -1 and 1. A value of -1 means perfect negative correlation (ie every increase in x leads to an exact decrease in y) with 0 being the null value.

Again, regression describes the nature of the relationship between variables. In this case, for every change in x, how much of an increase (or decrease) in y do we experience? Simple linear regression involves finding the best straight line to fit the relationship between two variables, and is essentially the slope of the line. In our example, for every



**Figure 1 Telephone call costs**



**Figure 2 Regression analysis plots**

## TOOLBOX

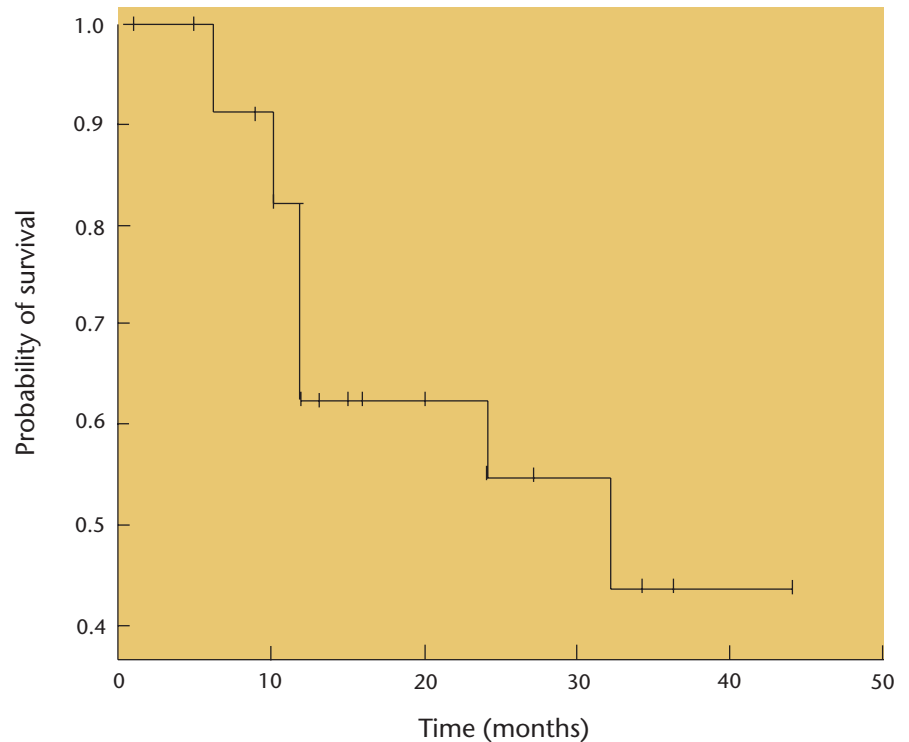
additional increment of time, the regression would describe how much of an increase will there be in our phone bill....ie the steepness (slope) of the line. For example, an additional one minute call could result in a small increase of 20 pence or a large increase of 2 pounds. The null value = 0 which is a horizontal line. That is, for every change in x, y does not change.

There are different tests for association, which depend on the type of data. Thus, for categorical data, a  $X^2$  test of association or a Kappa statistic is commonly used for inter-rater agreement. For ordinal data, either a Spearman's Rank or Kendall's Tau test would be used. Perhaps the most familiar test is the Pearson's Product Moment Correlation Coefficient, or simply, a Pearson's r. The Pearson's r is used for association between continuous (quantitative) variables, but it has often been misused due to assumptions that have not been followed, such as:

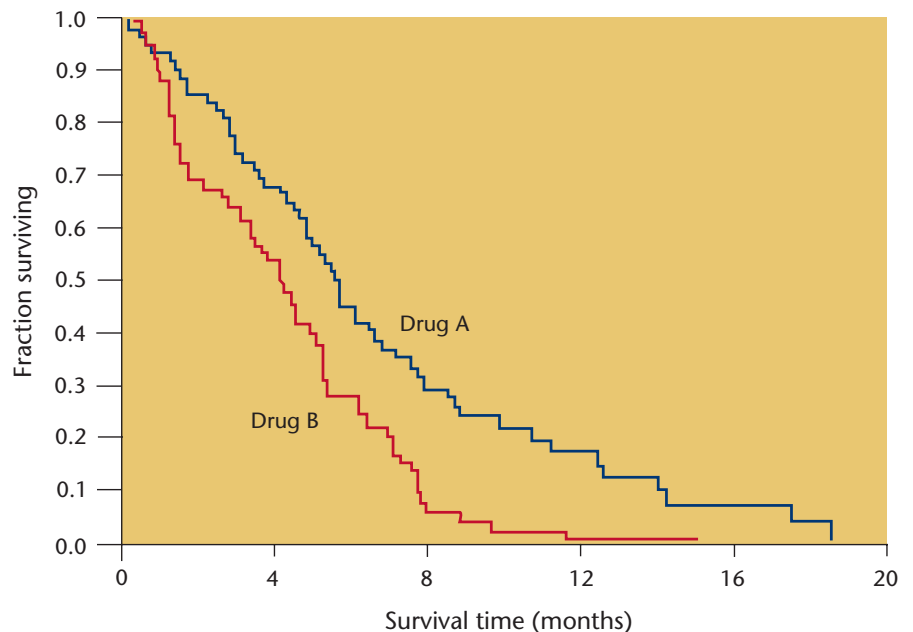
- The data must be continuous, otherwise a non-parametric equivalent test should be run. If not, it can overestimate the strength of the relationship between variables.
- The data must be independent, one should not be forced to vary with the other, otherwise a paired test should be done.
- The data must be normally distributed, which should be a linear pattern when graphed. If non-linear, a non-parametric test should be done.
- P-values for a Pearson's r are quite often below 0.05 with a decent sample size, (around 30 per group) so it is the magnitude of the r that is important. Values below 0.8 do not show a great deal of strength between variables, even with very low p-values.

Most notably, even if all assumptions have been met, a Pearson's r of say, 0.9 should never infer that A caused B or vice-versa (again, think periodontitis and grey hair). Randomised trials, in theory, are the only study design that generally shows causality as, apart from the intervention, we are keeping all other variables equal. If observational studies are examining association versus causality, the Bradford-Hill criteria should be kept in mind.

$R^2$  is the coefficient of determination: what percent of the variability of y is determined by input variable x. If  $r=0.9$   $R^2 = 0.81$



**Figure 3 Kaplan-Meier Plot**



**Figure 4 Kaplan-Meier Plot comparing survival rates on two drug treatments**

meaning 81% of variability of y explained by x, 19% remains unexplained. If  $r=0.7$   $R^2 = 0.49$ ; thus over half (51%) of the variability of y remains unexplained. This is why correlation coefficients less than 0.8 don't really show much strength between variables as there is a great deal of scatter about the regression line (Figure. 2).

We can also define other types of regression:  
Multiple regression - what is the nature

of the relationship between two or more input variables and one continuous output variable....ie how does smoking status, age, pregnancy status and weight affect diabetes expressed as serum glucose levels?

Logistic regression – what is the nature of the relationship between two or more input variables and one dichotomous output variable...ie how does smoking status, age, pregnancy status and weight affect

diabetes expressed as a binary outcome (diabetic/non-diabetic)?

**Example 1**

You have just read a study examining the association between salivary *Streptococcus mutans* counts and caries prevalence in 1,800 4-5 year old children. The authors provide a Pearson's *r* of 0.31 and state that *p*<0.001. They also state that the data are skewed.

With strong evidence against the null hypothesis (*p*<0.001), should we believe a strong relationship exists between these two variables?

Answer: no. It is the magnitude of the correlation coefficient that is relevant, not the *P*-value. With an *r* value of 0.31, there is little correlation between variables. Additionally, with skewed data, a non-parametric test of association, rather than the parametric Pearson's correlation test should have been done.

**Time-to-event (survival) data**

Time-to-event data expands on the concepts of risk ratios and odds ratios for binary data. Binary data let us know the proportion of implant failures in the experimental versus the control group. However, we, along with our patients may want to know more than simply, 'did the dental prosthesis survive'? We may want to know how long did it survive, and what was the average time to failure. Survival data can provide these answers.

If a patient experiences an event of interest (ie death/prosthesis failure) during the study, their survival time is said to be **exact**. If the event is not observed, ie patient survives, patient dies from some unrelated cause, or withdraws from the study, the observation is said to be **censored**.

If we have complete follow-up for each patient (no censored data) we can estimate the survival rate as:

$$S = \frac{\text{number of subjects surviving beyond time } t}{\text{number of individuals recruited in the study}}$$

In other words, we would treat survival as if it were binary data.

If we have censored observations, a different method of calculation is required as the above method would be ignoring a significant part of the data (especially people who survived) and tends to underestimate survival.

The **Kaplan-Meier** estimate gives us the cumulative probability of survival accounting for censored observations (Figure. 3). The probability usually (but not always) starts at the value 1.0 and falls in steps as each event occurs. The survivor function will not reach the minimum value of 0.0 if there are subjects who remain alive (or have not experienced the event) at the end of the study period. Kaplan-Meier estimates will also give us the mean and median survival time. The median is a better estimate as lifetime distributions tend to be positively skewed (time-to-event data are generally considered to be non-parametric, that is, not assumed to be normally distributed).

Similar to a chi-square test, the **Logrank** test is the most common method for comparing treatment groups that allows for censored observations. It compares the observed number of events in each group to expected values assuming identical survival profiles.

The hazard function is the instantaneous failure rate, or the probability of an event happening at a particular point in time among those at risk. The ratio of two hazard rates is known as the **hazard ratio** (HR) and quantifies the difference between survival patterns in two groups. In the absence of censored observations, the hazard ratio equals the relative risk.

The Logrank test examines the hypothesis that the hazard ratio = 1.0 (null value),

and if the 95% confidence interval of the HR contains the null value, this corresponds to *P* > 0.05. The HR assumes that the relative risk of death between two groups remains constant. In Figure 4, it appears that patients on the drug A had a better survival rate than patients on drug B, and we might like to quantify this difference. In other words, is this difference statistically significant or clinically relevant?

**Example 2**

Two thousand one hundred dental implants in 575 patients were evaluated for risk factors for implant failure. Using a Cox regression model, the authors report a HR of 2.9 (95%CI 1.6-5.3) for current tobacco use.

- How do we interpret this result in plain English?
- Is this statistically significant?
- Is this clinically relevant?

Answer:

At any time, roughly three times as many smokers are experiencing implant failure compared with non-smokers.

It is statistically significant, as the null value of '1' is not contained within the 95% confidence interval.

To examine clinical relevance, we first must know the baseline failure rate for dental implants. If we take this rate to be 5%, then a three-fold risk of failure is roughly 15%. This may not be clinically relevant to the majority of practitioners. However, we must now look at the upper limit of the 95% CI, which is 5.3. This represents over a 25% risk of failure, which may be clinically relevant to practitioners. Thus, we might say that this study is not clinically relevant, but it is also indecisive, as a clinically relevant effect of smoking cannot be ruled out.

*Evidence-Based Dentistry* (2012) 13, 29-31. doi:10.1038/sj.ebd.6400849.