STATS at George Mason University

checking out the facts and figures behind the news

What does it mean for a result to be “statistically significant”?

How can we tell whether two events happen at the same time by chance, or for a reason? A headache cured by an aspirin might have gone away without the aspirin. The fact that a sequence of five coin flips turned up five heads may or may not indicate a biased coin. When are the observations we make – such as that Republicans go to church more than Democrats, and men earn more money than women in similar jobs – due to chance, and when are they truly correlated events, with an underlying reason?

Measuring the likelihood that an event occurs by chance is the idea behind “statistical significance.” If there is, at most, a **five percent chance** of two events would happen together by coincidence, we may legitimately infer that there is a reason that the events occurred together. Such results are called statistically significant, and the events are considered correlated. If there is more than a five percent chance of occurring randomly, the possibility that the events occurred together just by luck is too high to dismiss, and we conclude nothing. The five percent line is arbitrary, but has become standard in the field of biomedical research; statistical significance is the golden measuring stick for evaluating data. Why five percent and not ten? More on that below.

Statistical significance is extremely important. Suppose we want to test the effectiveness of a medicine to reduce the likelihood of a heart attack. We design a [controlled study](http://www.stats.org/in_depth/faq/Controlled%20vs%20Observational%20vs%20Case.doc) of two groups of people. Group A takes the medicine, and Group B takes a placebo. Suppose that Group A has a much lower rate of heart attacks than Group B. Is this due to chance, or the medicine?

If the rate of Group A heart attacks is just slightly lower than that of Group B, then we are more likely to believe that the medicine didn’t cause the effect, since any two groups of people are likely to have small differences due to random fluctuations. Similarly, if there are a small number of people in the study, we believe that chance plays a larger role. The formula for determining statistical significance therefore depends not only on the actual rates of heart attack in these two groups, but also the number of people in each group. The p-value is the likelihood that the observed relationship between medicine and heart attack rate occurred by chance.

Suppose the p-value for the study is .04. This means that there is a four percent (.04 x 100) chance that Group A would have as low a rate (or lower) as it did in the study just by chance. Since p <.05, the result is considered statistically significant and researchers are justified in concluding that the drug is [*correlated*](http://www.stats.org/in_depth/faq/Causation%20Versus%20Correlation.doc) with a reduced heart attack rate. If p had been .1, then there would be a ten percent (.1 x 100) chance that Group A has a lower rate due to chance. We would have less confidence that the medicine played a role, and the result would be inconclusive. It does not mean that the drug does not help, but that it has not been proven to help.

The fact that statistical significance is achieved when there is as much as five percent chance that the observation is due to chance is controversial. For some, a five percent chance that an observation was due to chance is very high, and for others it’s very low. For every twenty studies published claiming an association between events at p =.05, one of them is flawed. For some, this makes biomedical research untrustworthy. For others, the fact that a result with p =.1 is not considered reliable means that important correlations are not being reported to the public, with possibly hard consequences. There are cases when scientists hold research to higher (or lower) standards of demonstrating statistical significance, and certainly stronger or weaker correlations are remarked upon in the literature. However, no matter how you conduct the research there is always a small possibility that you observed an association when one isn’t really there. For the sake of having a standard of some kind, scientists have agreed on p =.05.

A result that is statistically significant has more weight in the scientific community than one that is not. There is nonetheless a (small) possibility that the result is just due to chance, which is why scientists keep their eyes open for other studies that might discredit the first. Similarly, even if a conjectured correlation has not been demonstrated to be statistically significant, there may still be a good chance that the association really exists, which is why more tests are often called for. Statistical significance should not, however, be dismissed as lacking certainty — we can be 95 percent certain that a statistically significant result is true. This is why statistical significance is the stamp of approval by the biomedical community.