Statistics for those who hate statistics

Dr. Ned Kock

Full reference: Kock, N. (2004), Statistics for Those who Hate Statistics, Appendix of *Business Process Improvement through E-Collaboration: Knowledge Sharing through the Use of Virtual Groups*, Idea Group Publishing, Hershey, PA.

The importance of statistical tests

When we analyze quantitative evidence, or numbers, that describe a particular situation or phenomenon, we often need to generate coefficients based on specific statistical tests to reach reasonable conclusions. Visually inspecting a table full of numbers, for example, can be quite confusing, and the related conclusions may be deceiving. This is one of the reasons why statistical tests are important. The more quantitative evidence we have to analyze, the more difficult it is to inspect it visually, and so the more important those statistical tests become.

For example, we may want to know whether a particular variable, such as the degree of e-collaboration technology use by a business process improvement group, has any effect on the duration (or lifetime) of the group, measured in days. One way of testing that is to analyze the duration of several groups, some of them conducted using e-collaboration technology support, and others conducted without any e-collaboration technology support.

By simply comparing group duration averages (also known as "means", in statistics lingo) for each condition (i.e., with and without technology support), we may find that e-collaboration technology-supported groups have, on average, a duration in days that is, say, 13 percent higher than the groups conducted without any e-collaboration technology support.

In the situation above, the following question arises. Is the 13 percent difference large enough to be significant? If the answer is "yes", and other circumstances (e.g., group size, cultural background of the participants) were the same regarding the two group conditions (i.e., with and without technology support), then we can conclude that the use of e-collaboration technology had a significant impact on the duration of the business process improvement groups. The answer to this type of question, which is quite important in behavioral research in general, is one of the most important outcomes of statistical tests.

Statistical tests are widely used in areas other than behavioral research on the impact of technologies on people. For example, similar types of questions to the one above are whether a particular medical drug has any significant effect on individuals suffering from a certain disease, and whether a difference in the number of votes for two competing candidates in a pre-election poll is significant enough to warrant optimism in the camp of the candidate with the higher number of votes.

Three main types of statistical tests of significance used in previous chapters of this book are comparison of means, correlation, and distribution trend tests. Comparison of means tests are aimed at establishing whether the differences between the means, or averages, of two or more conditions differ significantly from each other (as illustrated through the example above). Correlation tests aim to establish whether two variables, e.g., degree of e-collaboration technology use and likelihood of success of a business process improvement group, vary together in a significant way. Distribution trend tests aim to establish whether an observed distribution trend, e.g., the distribution of user perceptions about an e-collaboration tool's impact on group outcome quality, is significant enough to allow for the conclusion that it is caused by a particular variable, e.g., e-collaboration tool support. Each of these tests is discussed in more detail below.

Comparing means from different conditions

Let us assume that we facilitated 20 business process improvement groups. Half of those groups (i.e., 10 groups) used an e-collaboration system to communicate, whereas the other half communicated face-to-face. Let us also assume that the outcomes of those business process improvement groups, that is, the business process redesigns generated by them, were scored in terms of quality. The scores ranged from 1 (very poor quality) to 7 (very high quality).

Table A.1 shows the scores obtained for each of the business process improvement groups. A simple visual inspection of Table A.1 suggests that the e-collaboration technology-supported groups seemed to have had generally higher scores than the face-to-face groups, but a simple visual inspection is not usually enough for us to establish with certainty how much better the e-collaboration technology-supported groups did on average, and whether that difference is statistically significant.

	Group outcome quality (scores from 1 to 7)	
Groups	**Face-to-face**	**E-collaboration**
1 and 2	4	7
3 and 4	5	6
5 and 6	3	7
7 and 8	2	5
9 and 10	5	6
11 and 12	1	2
13 and 14	4	5
15 and 16	5	7
17 and 18	3	5
19 and 20	2	6

Table A.1: Outcome quality scores for 20 business process improvement groups

To find out how much better the e-collaboration technology-supported groups did on average we can calculate the mean (or average) scores obtained for both face-to-face and e-collaboration technology-supported groups. Those means are 3.4 and 5.6, respectively, which suggest that the scores for e-collaboration technology-supported groups were on average about 65 percent higher than the scores obtained for face-to-face groups.

When we calculate means we also usually calculate standard deviations (noted as "SD" in Table A.2), which in the example in question are a measure of how much variation there is in the scores for each condition (i.e., face-to-face and e-collaboration). The standard deviations obtained for face-to-face and e-collaboration technology-supported groups are 1.43 and 1.51, respectively, which suggests that the degree of variability between the two conditions is similar – this is usually considered a good thing, from a statistical analysis perspective. The standard deviations can also tell us much more, but a discussion about that would be somewhat technical and beyond the scope of this appendix (whose goal is not to induce readers to hate statistics even more than they already may do).

	Face-to-face	E-collaboration
Mean	3.40	5.60
SD	1.43	1.51

T coefficient	*3.349*
P	*0.004*

Table A.2: Descriptive and T test statistics

Table A.2 also allows us to establish whether the 65 percent difference between the mean scores obtained for face-to-face and e-collaboration technology-supported groups is significant from a statistical standpoint. For that, we can use a variety of comparison of means tests, of which one of the most common is the T test. A typical T test will generate two coefficients, a T coefficient and a P coefficient, which are 3.349 and .004, respectively, in the example discussed here.

Even though the test is called a T test, it is the P coefficient that really matters most, because that coefficient is the probability that the 65 percent difference between the mean scores obtained for face-to-face and e-collaboration technology-supported groups is due to chance. In our example, the P coefficient is .004, which means that the probability that the difference between the mean scores is due to chance is .4 percent (less than half of 1 percent). In most statistical tests, a chance probability below 5 percent is considered very low, and an indication that the effect under

consideration is statistically significant. Therefore, in our example, the .4 percent probability allows us to safely say that the 65 percent difference between the mean scores obtained for face-to-face and e-collaboration technology-supported groups is NOT due to chance. Or, in other words, we can safely say that the use of the e-collaboration system had a significant and positive effect on the quality of the outcomes generated by the business process improvement groups.

I know that the above may sound like a convoluted and complicated way of stating the obvious. This is, incidentally, why many people hate statistics. But it is important to stress that the method behind the procedure discussed above has been very carefully developed, and is widely used in a variety of areas. One good example is the pharmaceutical industry. To prove that a vaccine is effective against a certain disease, the developer of the vaccine has to test it in a group of individuals, who are usually paid to voluntarily participate in experiments involving the administration of the vaccine (other ethical considerations may exist – e.g., no health risks are involved in either getting or not getting the vaccine, under the test circumstances).

One particularly convincing type of test would involve a group of individuals taking the vaccine, while another group (of about the same number) would take a placebo (i.e., a drug containing no active ingredients). If a comparison of means test, such as the T test, indicates that there is a statistically significant difference in the average resistance to the disease in favor of those who were administered the vaccine (when compared with those who were administered the placebo), than the pharmaceutical company that developed the vaccine strikes gold.

There are many types of comparison of means tests, and there are many statistical software packages that allow one to run those tests. Examples of other fairly widely used comparison of means tests are the one-way ANOVA and Mann-Whitney U tests. The widely used T test, which was illustrated above, can be run on many commercial spreadsheet packages, including MS Excel – which was used to generate the statistical coefficients above.

Checking for correlations between variables

There is another way of testing the statistical significance of the impact of the e-collaboration technology support on the quality of the outcomes generate by business process improvement groups, which was tested in the previous section through a T test. Namely, we can calculate the correlation between two variables – the degree to which e-collaboration technology support was available and the group outcome quality scores. The former variable, the degree to which e-collaboration technology support was available, can have basically two values – 1, for no support (face-to-face groups), and 2, for some support (e-collaboration technology-supported groups). The group outcome quality scores are the same as in the previous section.

Table A.3 shows the scores for the two variables that are tested for correlation using one of the most common correlation tests, the Pearson correlation test. The results of the Pearson correlation test are shown at the bottom of Table A.3. They are the *r* coefficient, and the related P coefficient.

Group	E-collaboration support	Outcome quality
1	1	4
2	1	5
3	1	3
4	1	2
5	1	5
6	1	1
7	1	4
8	1	5
9	1	3
10	1	2
11	2	7
12	2	6
13	2	7
14	2	5
15	2	6
16	2	2
17	2	5
18	2	7
19	2	5
20	2	6

r coefficient	*0.620*
P	*0.004*

Table A.3: Testing an effect through a Pearson correlation test

For most statistical analysis purposes, an *r* coefficient that is generated through a Pearson correlation test, and that is higher than .6, is generally seen as an indication of a strong correlation between two variables. This is consistent with the low P coefficient of .004 obtained (the same as in the T test employed in the previous section), which suggests that the strong correlation suggested by the Pearson correlation test has a .4 percent probability of being due to chance (which, again, is substantially lower than the 5 percent threshold used to draw conclusions from most statistical tests).

In summary, the degree to which e-collaboration technology support was available seems to have strongly and positively affected business process improvement group outcome quality, in the example discussed above. If the Pearson correlation coefficient had been negative, then we could conclude the opposite, that is, that the degree to which e-collaboration

technology support was available strongly and *negatively* affected group outcome quality.

Generally speaking, two variables are highly correlated when an x-y graph (i.e., a bi-dimensional graph) plotting their values looks like a line; in such a graph, the x values would be those of one of the variables and the y values would be those of the other variable. The more similar to a line the graph is, the higher is the correlation between the variables. Conversely, the less similar to a line the graph is, that is, the more dispersed the x-y intersection points are, the lower is the correlation between the variables.

Figure A.1 illustrates the above relationship. The graph at the top (Figure A.1.a) plots the relationship of two highly correlated variables, whose Pearson correlation coefficient is .98. The maximum Pearson correlation coefficient possible is 1, which would be obtained if the relationship between two variables was completely linear, which would in turn make the graph look like a perfect line. The graph at the bottom (Figure A.1.b) plots the relationship of two variables whose correlation is low, with a Pearson correlation coefficient of only .08.

Figure A.1.a: High correlation (Pearson r = .98)

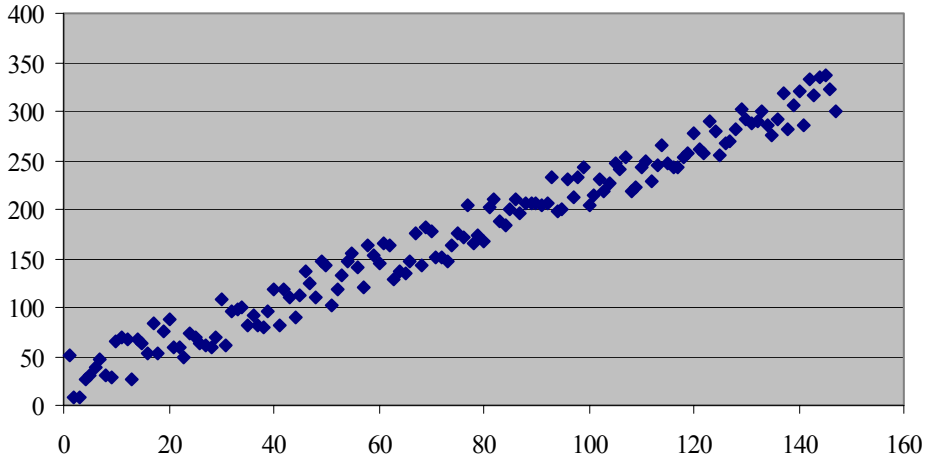

Figure A.1.b: Low correlation (Pearson r = .08)

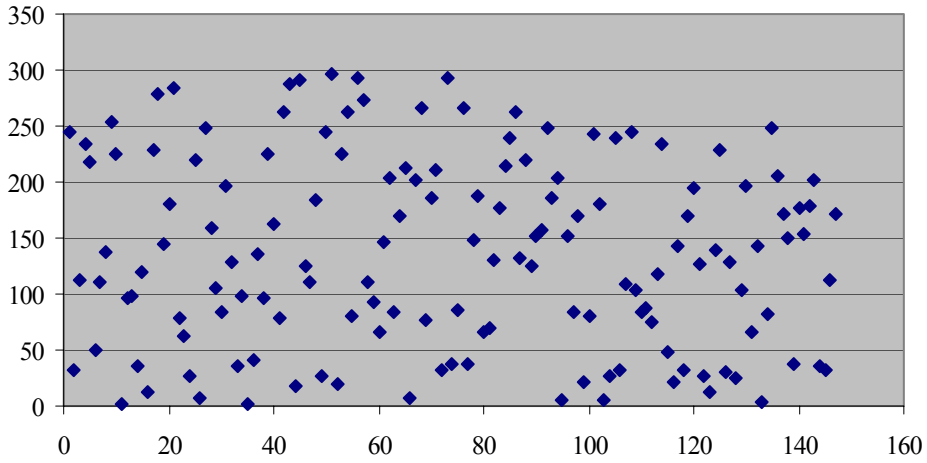

Figure A.1: Examples of high and low correlations between variables

There are several different types of correlation tests, although not as many as there are comparison of means tests, and, as with other statistical tests, there are many statistical software packages that allow one to run those different correlation tests. The widely used Pearson correlation test, which was illustrated above, can also be run on many commercial spreadsheet packages, including MS Excel – which was used to generate the statistical coefficients and the illustrative graphs above.

Assessing the significance of distribution trends

Let us now consider a different type of research question, which was asked and answered several times in previous chapters of this book. The question is the following: How can we establish whether a trend in a distribution of perceptions regarding a single variable is due to chance? In this case we don't have two different variables to correlate, or two different conditions to compare, which prevents us from using comparison of means or correlation tests.

For example, let us assume that 100 individuals routinely conduct business process improvement group discussions by interacting in smaller groups through face-to-face meetings. Those 100 individuals then participate in e-collaboration technology-supported business process improvement group discussions in groups of similar size. Following that, they are asked whether the e-collaboration technology support decreased, had no effect, or increased the quality of the outcomes generated by their business process improvement groups.

The answers provided by the 100 individuals are distributed according to Figure A.2. As it can be seen, there seems to be an underlying trend toward perceiving e-collaboration technology support as having increased business process improvement group outcome quality. More individuals (namely 50) perceived group outcome quality as having been increased by e-collaboration technology support than those who perceived outcome quality as not having been affected (30 individuals), or having been decreased (20 individuals). One of the ways to test the significance of that trend is to run a Chi-squared test comparing the observed distribution with a distribution in which there was no clear trend, that is, a distribution in which the same number of individuals perceived e-collaboration technology support as having increased, had no effect, and decreased group outcome quality. That number is 100 divided by 3, or approximately 33 individuals. The results of the Chi-squared test are shown at the bottom of Figure A.2 (Chi-squared = 13.9, P = .0009), and suggest that the probability that the observed distribution trend is due to chance is only .09 percent (much lower than the 5 percent threshold used to draw conclusions from most statistical tests).

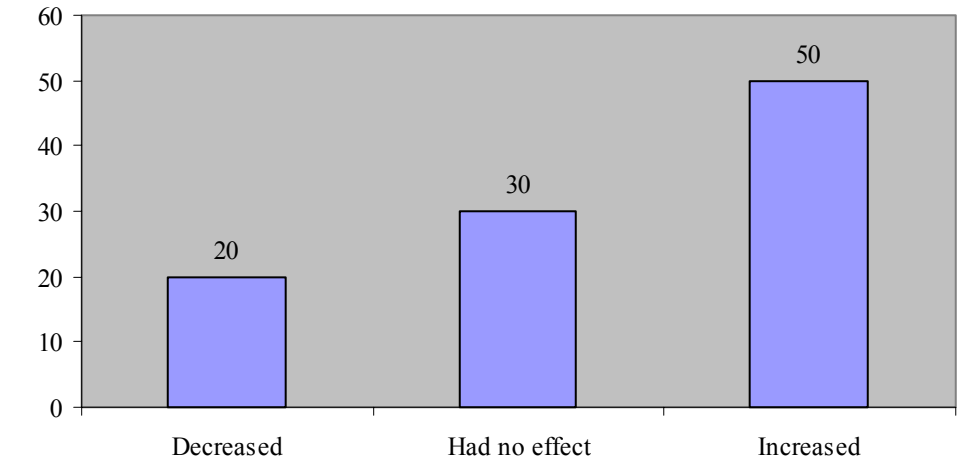

Figure A.2: Distribution of answers suggesting a strong trend
(Chi-squared = 13.9, P = .0009)

The conclusion that can be drawn based on the discussion above, and on the distribution trend suggested by Figure A.2, is that e-collaboration technology support strongly and positively affected business process improvement group outcome quality.

If the trend was not as skewed toward the "increased" perception, the Chi-squared test would not have been as conclusive. For example, the distribution of answers shown in Figure A.3 would still suggest a perception trend that is obviously leaning toward a general perception that group outcome quality was increased by e-collaboration technology support. However, the trend is weaker than that shown in Figure A.2, which is indicated by the Chi-squared test results at the bottom of Figure A.3. Those results suggest that the probability that the observed trend is due to chance is 17 percent, which is too high when compared with the 5 percent level suggested by statisticians as the upper limit used to draw conclusions from most statistical tests. In other words, a distribution of perceptions like the one in Figure A.3 would not allow us to conclude with much certainty that e-collaboration technology support positively affected business process improvement group outcome quality.

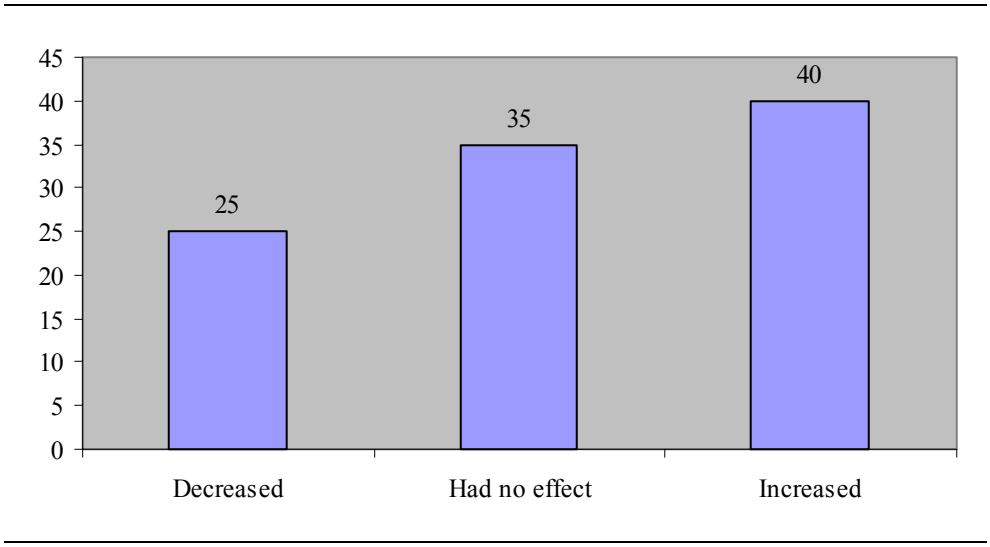

Figure A.3: Distribution of answers suggesting a weak trend
(Chi-squared = 3.5, P = .17)

There are a few different types of tests that can be used to analyze distribution trends. As with the other statistical tests discussed in this chapter and used in previous chapters of this book, there are many statistical software packages that allow one to run those distribution trend significance tests. The relatively widely used Chi-squared test, which was illustrated above, can also be run on many commercial spreadsheet packages, including MS Excel. That package was used to generate the Chi-squared and P coefficients above, as well as the graphs used to illustrate the distribution trends used in the Chi-squared tests.

Summary and concluding remarks

In previous chapters of this book the results of several statistical tests were discussed. The three main types of tests employed were comparison of means, correlation, and distribution trend tests. A typical statistical test will generate two main coefficients, a test coefficient, which is usually named after the test (e.g., T coefficient, for a T test), and a P coefficient. It is usually the P coefficient that matters most, because it indicates the probability of chance of the test, and whether the yielded result is

significant or not. In most statistical tests, a P below 5 percent (i.e., below .05) is considered very low, and an indication that the effect under consideration is statistically significant.

Comparison of means tests are aimed at establishing whether the difference between the means, or averages, of two or more conditions differ significantly from each other. One of the most common comparison of means tests is the T test. There are many types of comparison of means test, and there are many statistical software packages that allow one to run those tests. Examples of other fairly widely used comparison of means tests are the Mann-Whitney U and the one-way ANOVA tests.

Correlation tests aim to establish whether two variables, e.g., degree of e-collaboration technology use and likelihood of success of a business process improvement group, vary together in a significant way. One of the most common correlation tests is the Pearson correlation test, which generates an *r* coefficient and a P coefficient. If an *r* coefficient is higher than .6, this is generally seen as an indication of a strong correlation between the two variables under consideration. When two variables are highly correlated an x-y graph plotting their values will look like a line. The more similar to a line the graph is, then the higher is the correlation between the variables; the less similar to a line the graph is, the lower is the correlation between the variables.

Distribution trend tests aim to establish whether an observed distribution trend, e.g., the distribution of user perceptions about an e-collaboration tool's impact on group outcome quality, is significant enough to allow for the conclusion that it is caused by a particular variable, e.g., e-collaboration tool support. One of the ways to test the significance of a distribution trend is to run a Chi-squared test comparing the observed distribution with a distribution in which there was no clear trend.

There are many statistical software packages that allow one to run the several statistical tests discussed in this chapter. One such statistical analysis package is SPSS, which is commercialized by a company of the same name. The T, Pearson correlation, and Chi-squared tests, which

have been discussed in this chapter, can also be in large part run on many commercial spreadsheet packages, including MS Excel.

Most statistical tests, including the ones discussed in this chapter, are best run when what is known as the “sample size” is relatively large, otherwise they loose their power. For example, let us assume that we want to test the statistical significance of the impact of the e-collaboration technology support on the quality of the outcomes generated by business process improvement groups. It will be better to run a correlation test based on evidence collected from 50 groups, than on evidence collected from only 10 groups. That is, the former will yield more reliable results than the latter. Conversely, the larger the sample size, the less strong the underlying effect needs to be to yield a statistically significant result, which means that with very large samples, even weak effects will be statistically significant.