Estimating Correlations from Scatterplots

Joachim Meyer and David Shinar

Human Factors: The Journal of the Human Factors and Ergonomics Society 1992 34: 335

DOI: 10.1177/001872089203400307

HUMAN FACTORS, 1992, 34(3), 335–349

Estimating Correlations from Scatterplots

JOACHIM MEYER *and* DAVID SHINAR,[1] *Ben-Gurion University of the Negev, Beer-Sheva, Israel*

Previous attempts to establish the function relating intuitive estimates of correlations from scatterplots to accepted statistical measures have led to unsatisfying results. In this study two experiments dealt with the effects of the statistical training of the viewer and various characteristics of the display on estimates. Statistical knowledge was related to higher estimates of correlations and the use of a wider range of values, but people with and without statistical knowledge were equally affected by the type of dispersion of the point cloud, the mere display of the regression line, and the slope of the regression line. Results indicate that estimates of correlations from scatterplots are partly based on perceptual processes that are influenced by visual properties of the display and are unrelated to the cognitive structures created by formal statistical training.

INTRODUCTION

Inferences about correlations between variables are a major component of most decision-making processes. Although formal statistical methods are sometimes used, most people are "intuitive statisticians" who assess various aspects of the phenomena they perceive and draw conclusions that seem appropriate to them.

Nisbett and Ross (1980) identified two different ways in which people evaluate correlations. One is *theory driven*, whereby people apply their preconceptions about certain phenomena and draw conclusions that are often based on false, stereotyped beliefs unsubstantiated by the data (e.g., Chapman, 1967; Chapman and Chapman, 1967, 1969). The results are usually gross overestimates of correlation between variables. A second way in which people assess correlations is *data driven*. With this method, subjects have no a priori beliefs or knowledge about the correlations between the variables of interest and, when presented with covariate data, typically underestimate correlations, compared with the statistical measures used to indicate the strength of the correlation (e.g., Cleveland, Diaconis, and McGill, 1982; Jennings, Amabile, and Ross, 1982). Although most recent research on estimates of correlation has focused on theory-driven estimates, it is important to understand the principles of data-driven estimates of correlation, considering the increasing use of computer graphic displays.

The bivariate distribution of two interval-scaled variables is usually plotted in two-dimensional space, thereby creating a scatterplot. The most common way of forming an impression about the possible correlation between two variables is by inspecting such a

[1] Requests for reprints should be sent to David Shinar, Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, P.O.B. 653, Beer-Sheva, Israel.

scatterplot. Wainer and Thissen (1979) pointed out that when the assumption of bivariate normality is violated, inspection of a scatterplot might yield better insights into the relations between variables than would use of the standard product-moment correlation. Anscombe (1973) demonstrated how different dispersions of data points result in the same correlation coefficient and how merely viewing the scatterplot allows one to form a deeper understanding of the relation between the variables.

As a visual display, the scatterplot affects the viewer's perception in various ways that lead to different estimates of the strength of the association between the two variables. Bobko and Karren (1979) have shown how perceptions that have been influenced by the display and cognitive processes involved in estimating correlations from them might bias legal decisions. Understanding these biasing mechanisms is important because public data are often presented in the form of scatterplots to decision makers and the general public. With the increasing use of powerful graphic software packages for personal and business computers, and with the incorporation of scatterplots in many decision support systems, the cognitive aspects of the intuitive assessment of correlations from scatterplots have gained further importance.

The present study aims to clarify the impact of various statistical and perceptual properties of scatterplots on estimates of correlation. Some of these characteristics can easily be manipulated by the graphic designer who draws the scatterplot and, thereby, manipulates the impressions generated by the displayed data. Furthermore, we investigated whether people with formal statistical training are less biased by the graphical features of a display than are people with little or no statistical knowledge.

Previous studies (Bobko and Karren, 1979; Cleveland et al., 1982; Collyer, Stanley, and Bowater, 1990; Lauer and Post, 1989) have shown that subjects consistently underestimate correlations relative to the statistical coefficient of r. This finding was indirectly anticipated in an early study by Pollack (1960), who found that discrimination between different levels of correlation is easier when the correlations are high. That is, given a constant difference in r^2, two scatterplots look more alike when the correlations are low than when they are high. In only one study were no systematic underestimations found (Wainer and Thissen, 1979); that study allowed subjects to view prototypes of scatterplots with different correlations throughout the experiment. The subjects were expected to estimate the correlations in stimulus scatterplots by comparing them with the prototypes. Of the other studies, one (Strahan and Hansen, 1978) claimed that estimates approach values of r^2, whereas the rest showed that subjects generated values that fell even below r^2. Jennings et al. (1982) suggested that the *coefficient of alienation*—defined as $r' = 1 - \sqrt{(1 - r^2)}$—may be an appropriate function for relating estimates and correlations. Later studies (Cleveland et al., 1982; Cleveland, Harris, and McGill, 1983) have shown that this function, though closer to fitting the data than r or r^2, still falls short of adequately describing the relation between correlation estimates and objective measures.

A number of factors affect intuitive estimates of correlations. One is the dispersion of the data point cloud. In general, correlational analyses assume bivariate normal distributions, which in a scatterplot form oval-shaped point clouds. Often this assumption does not hold for empirical data. In actual data sets one frequently encounters outliers—that is, data points that fall outside the boundary of the point cloud, defined by the majority of points. One disadvantage of r as a statistical measure is its vulnerability to such outliers, which violate the assumption of bivariate normality and which considerably diminish the value of the correlation coefficient. People's subjective estimates are influenced far less by outliers than are the accepted statistical measures (Bobko and Karren, 1979).

The dispersion of the data point cloud is determined by characteristics of the data set itself and is only one factor that affects estimates of correlations from scatterplots. Other factors are the result of characteristics of the display (i.e., the way the scatterplot is drawn). One such characteristic is the density or size of the point cloud in the display frame. The size can easily be manipulated by changing the scales of the axes. Cleveland et al. (1982) showed that smaller, denser point clouds lead to higher estimates of correlation than do more dispersed ones.

Another characteristic of the display we examined is the influence of drawing a regression line through the data. In one study (Mosteller, Siegel, Trapido, and Youtz, 1981) subjects chose slopes close to the slope of the first principal component of the data, with lines passing close to the centroids. Recently Collyer (1988; Collyer et al., 1990) required subjects to position the best-fitting regression line and to estimate the correlation coefficient. Collyer concluded that these judgments involve independent cognitive processes. Although people seem to have some approximately correct intuitions about the placement of the regression line, no study has dealt with whether or not the mere presentation of a regression line in a point cloud affects estimates of correlation.

Still another variable might be relevant to the intuitive estimation of correlations. Cleveland et al. (1982) tested subjects with various degrees of statistical training: faculty at statistics and mathematics departments, practicing statisticians, and students in statistics courses. However, they did not analyze the potential differences in correlation estimates among these groups. Bobko and Karren (1979) showed that psychologists active in evaluation and measurement—and therefore relatively knowledgeable in statistics—are also likely to be influenced by characteristics of the displays. Lewandowsky and Spence (1989) asked subjects to decide which of two strata in a scatterplot had the higher apparent correlation. Experts (faculty and senior graduate students) responded more slowly but more accurately than did novices (undergraduate students). Lane, Anderson, and Kellam (1985) tested the influence of three statistical components of the Pearson correlation (slope, error variance, and variance of *X*) on undergraduate students' and Ph.D. and Ed.D. professionals' estimates of correlations. Although both groups were more strongly influenced by the error variance, some differences between the two groups appeared (which were not statistically analyzed). These studies imply that statistical training may be relevant to intuitive estimates of correlations, but this issue has not yet been studied directly.

EXPERIMENT 1

The first experiment tested estimates of correlations by people with extensive statistical training (university faculty) and people with limited knowledge in statistics (undergraduate students). Three different levels of correlation, from moderate to high, were presented. We expected to observe the previously noted effect of underestimation relative to accepted statistical measures.

We were also interested in the effect of different types of dispersion of the data point clouds on intuitive estimates of correlations. The product-moment coefficient (*r*) assumes bivariate normal distributions, which generally cause oval dispersions of the data point cloud. Two additional types of point clouds were tested: one had a uniform dispersion of points around the regression line, whereas the other had a trumpet-shaped dispersion, in which the variance between the data points increased at higher levels of x. The latter dispersion is similar to that produced by the presence of outliers, so estimates of correlation should be higher for this type of dispersion than for the other types. The trumpet shape was selected in order to retain the linearity of the least-square prediction line, which might be lost if outliers were positioned randomly in space.

Finally, we tested the influence of superimposed regression lines on correlation estimates. We hypothesized that the line would act as a "perceptual center of gravity" that would reduce the effects of the dispersion of the data point cloud and thus increase the estimated correlation.

Method

Subjects. Two groups of subjects participated in this experiment. One group consisted of 19 first- and second-year undergraduate students in the Department of Behavioral Sciences who had completed only a basic introductory course in statistics. The second group consisted of 10 faculty members from the Departments of Behavioral Sciences, Education, and Industrial Engineering and Management, all of whom were involved in quantitative research and had experience teaching statistics and correlational analysis.

Apparatus and design. Each subject received a booklet containing 54 different scatterplots, each with 21 data points, displaying the relation between two numerical variables (for examples of the stimuli see Figure 1). The variance on the x-axis had seven discrete levels labeled consecutively from 100 to 700. At each level were three data points. A computer program generated the data points so that a predetermined correlation (accurate at ±0.01) and a predetermined shape of the data point cloud were obtained. Two of the three points were generated randomly, and the third point was computed so that the average of the three data points for a given value of x would fall on the linear regression line. We thereby ensured that the best possible correlation coefficient would be the linear one and that no nonlinear relations were present.

The scatterplots that served as experimental stimuli had one of three levels of correlation, r^2 = 0.40, 0.65, and 0.90, and their data point clouds were of one of three types of dispersion: uniform around the regression line; oval, with decreasing variance toward the lower and higher levels of x; and trumpet shaped, with increasing variance at the higher levels of x. For each of the nine possible configurations, six different plots were created. Each subject saw half of the plots with the regression line drawn in. The sets with the regression line varied randomly between subjects, so that half of the subjects saw a certain set of data with the regression line and the other half of the subjects saw the same set of data without a line. Order of displays was also randomized within and between subjects.

Procedure. Members of the academic staff were approached individually and asked to estimate the correlations in their free time, whereas students participated in the experiment in the classroom. Instructions for filling out the booklet appeared on the cover. We gave a short explanation of the scatterplots, stating that they represented the relation between two variables, such as investment in advertising and sales; no specific context was given to any scatterplot. The axes were unlabeled, and subjects were not encouraged to consider the displayed data as showing the relation between any two specific variables. Thus we hoped to obtain purely data-driven estimates unaffected by subjects' preconceptions (Nisbett and Ross, 1980). The subjects marked their correlation estimate on a scale printed under the scatterplot, ranging from 0 to 100, with tick marks for every 10 units.

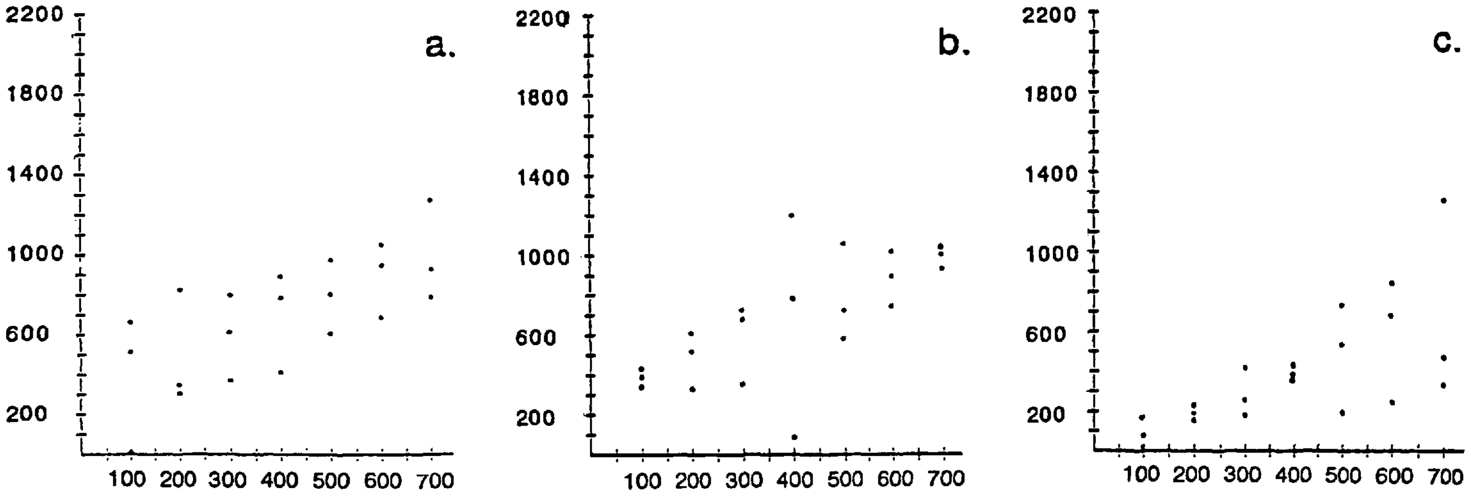

Figure 1. *Examples of scatterplots used in Experiment 1, all with $r^2 = 0.65$; (a) uniform, (b) oval, and (c) trumpet-shaped dispersions of data point clouds.*

No time limit was set, and subjects required approximately 15 min to complete the experiment.

Results and Discussion

The correlation estimates were analyzed in a four-way analysis of variance (ANOVA), in which the group (faculty vs. students) was a between-subjects variable and correlation level ($r^2 = 0.4, 0.65$, and 0.9), dispersion of the data point cloud (uniform, oval, and trumpet-shaped), and presence or absence of the regression line were within-subjects variables. Dependent variables were means of the estimates across the three repetitions of each of the combinations of the within-subjects variables. Mean estimates and standard deviations for the various conditions in the experiment appear in Table 1.

All four main effects, as well as three two-way interactions, were significant, whereas none of the higher-order interactions approached significance. As expected, the most significant effect was that of the correlation level, $F(2,54) = 112.68$, $p < 0.0001$, $MS_e = 0.045$. Subjects were obviously able to distinguish among the three levels of correlations. Their mean estimates for the squared correlation levels of 0.40, 0.65, and 0.90 were 0.34, 0.48, and 0.67, respectively. These estimates fell considerably below the true values of r^2 and therefore also below the corresponding values of r, which were 0.63, 0.81, and 0.95.

The faculty gave significantly higher estimates than did students, r(est) = 0.57 vs. 0.45, $F(1,27) = 8.71$, $p < 0.01$, $MS_e = 0.197$, and the magnitude of the difference was larger for higher correlation levels, as seen in Figure 2 and in the significant interaction of Group × Correlation Level, $F(2,54) = 3.95$, $p < 0.03$, $MS_e = 0.045$. It appears that people with extensive statistical training use a wider range of estimates and that their estimates are closer to the accepted statistical measures. Nevertheless, they too underestimate the correlation relative to the objective values of r^2.

Dispersion of the data point clouds had also a significant effect on the estimates, $F(2,54) = 15.52$, $p < 0.001$, $MS_e = 0.012$. Across all levels of correlation, estimates for the trumpet-shaped cloud were the highest, r(est) = 0.53,

TABLE 1

Experts' and Novices' Mean Estimates (and Standard Deviations) in Experiment 1 as a Function of Correlation Level, Type of Dispersion, and Display of the Regression Line

	Without Regression Line			*With Regression Line*		
Correlation Level (r^2)	*0.4*	*0.65*	*0.9*	*0.4*	*0.65*	*0.9*
Experts						
Uniform dispersion	0.28 (0.101)	0.52 (0.107)	0.76 (0.078)	0.37 (0.113)	0.60 (0.126)	0.85 (0.074)
Oval dispersion	0.28 (0.113)	0.52 (0.137)	0.75 (0.105)	0.39 (0.131)	0.56 (0.089)	0.77 (0.082)
Trumpet-shaped dispersion	0.48 (0.151)	0.57 (0.100)	0.79 (0.089)	0.43 (0.140)	0.59 (0.093)	0.83 (0.076)
Novices						
Uniform dispersion	0.24 (0.194)	0.39 (0.168)	0.57 (0.123)	0.35 (0.195)	0.45 (0.161)	0.65 (0.185)
Oval dispersion	0.25 (0.194)	0.39 (0.156)	0.52 (0.184)	0.35 (0.185)	0.45 (0.155)	0.61 (0.182)
Trumpet-shaped dispersion	0.34 (0.181)	0.42 (0.158)	0.61 (0.180)	0.37 (0.187)	0.48 (0.184)	0.69 (0.144)

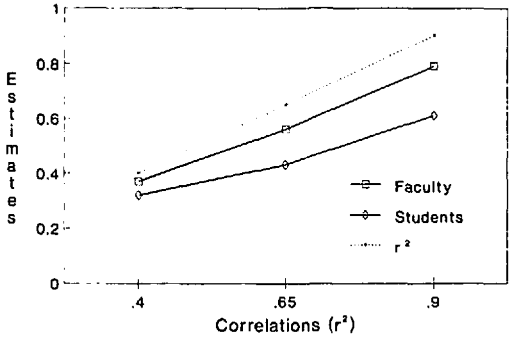

Figure 2. *Correlation estimates as a function of objective correlation level (r^2) and statistical training (university faculty vs. undergraduate students).*

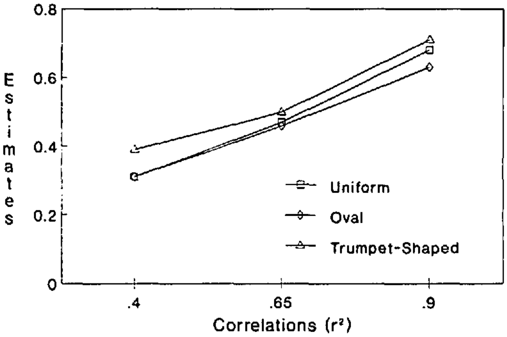

Figure 3. *Correlation estimates as a function of objective correlation level (r^2) and type of dispersion of the point cloud.*

and estimates for the oval cloud were the lowest, r(est) = 0.47. Generally, when the dispersion was small over most of the range of x values and the correlation was lowered by increasing the variance at a limited number of levels of x, the correlation estimates were high. Thus the intuitive estimates were relatively insensitive to the effect of outliers. Subjects seemed to depend in their estimates on the majority of x levels, where data points were concentrated around the regression line, and to ignore the higher dispersion at the upper end of the x axis.

The effect of the shape of the data point cloud interacted with the level of correlation, $F(4,108) = 2.79$, $p < 0.03$, $MS_e = 0.009$. As seen in Figure 3, the interaction was attributable primarily to the uniform distribution. The trumpet-shaped data point cloud yielded consistently higher estimates than did the oval point cloud, whereas the uniform distribution was perceived to be more like the oval cloud at low correlations and more like the trumpet-shaped cloud at high correlations. Interestingly, experts and novices were equally influenced by the shape of the data point cloud.

The presence of the regression line led subjects to increase the magnitudes of their estimates (and brought them closer to r^2) relative to estimates made from observing the same scatterplots without a regression line, r(est) = 0.53 vs. 0.46, $F(1,27) = 19.12$, $p < 0.0005$, $MS_e = 0.022$. A significant Dispersion × Line interaction indicated another effect of the regression line (see Figure 4): a decrease in the differences in estimates that were attributable to the types of dispersion, $F(2,54) = 3.96$, $p < 0.025$, $MS_e = 0.008$. Here, too, experts and novices were equally affected by a variable.

Results of the first experiment indicate that the typical underestimations of correlations and the insensitivity to outliers (as expressed by the differences between the various types of dispersion) seen in previous studies were replicated. In addition, we found that the presence of a regression line leads to increased estimates and that the differences in estimates between the types of dispersion were reduced when a regression line was displayed. It seems that the line serves as some kind of "perceptual center of gravity," creating the impression that the point cloud is relatively more condensed—and therefore more strongly correlated—while the dispersion of the point cloud becomes less important.

When we compared novices' and experts' estimates, we found that the two groups differed in the range of values used for estimates but that their estimates were equally affected by the dispersion of the point clouds and the presence of the regression line. Thus statisti-

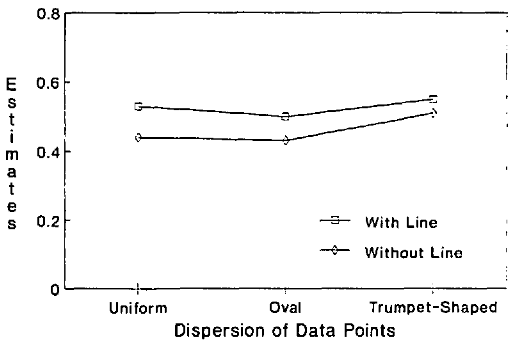

Figure 4. *Correlation estimates for the three types of dispersion of the point clouds, with and without the regression line.*

cal training and experience make a subject more likely to differentiate between the levels of correlation and to use higher estimates, which are closer to the accepted statistical coefficients. Still, statistical expertise does not change a subject's susceptibility to the influence of the more visual factors of the display: dispersion of the point cloud and presentation of the regression line.

EXPERIMENT 2

In order to replicate and extend the findings of the first experiment, we selected two additional groups of subjects—a group of fourth-year industrial engineering students and a group of high school students. Members of the former group had studied a number of courses in statistics and had applied their knowledge in additional courses. The high school students had no experience with statistics. Thus we had two groups with different levels of statistical knowledge, though both groups had considerably less knowledge than their counterparts in the first experiment.

Again, subjects were required to estimate correlations from scatterplots, half of which had the regression line drawn in. A new variable in Experiment 2 was the slope of the regression line, a characteristic that a designer can easily manipulate by changing the scale of the axes without changing the correlation level. Although one previous study (Bobko and Karren, 1979) showed no significant effect of the slopes of the regression line, it dealt with the unique case of an exchange of the variables on the x and y axes, which created a mirror image around the 45 deg line. The usual way to change the slope of the line is to change the scale of one axis. An increase in the slope of the line brings about larger and more dispersed point clouds, which should in turn lead to a decrease in estimates (Cleveland et al., 1982; Lauer and Post, 1989). However, because people tend to draw perfect correlations with a slope of 45 deg, this angle might elicit the highest estimates, as Bobko and Karren (1979) suggested.

In the first experiment only three correlation levels were used, and therefore it was impossible to establish the function relating estimates to correlations. In the second experiment we tested six different levels of correlation in an attempt to compute reliable functions relating estimated correlations and objective measures. Because various slopes were used, it was necessary to use only scatterplots with correlations above a certain level. (Displaying data with $r^2 = 0.1$, for example, with a 60 deg slope of the regression line is extremely inconvenient and requires a very long y axis.) In order to generate correlation levels that could be displayed at various slopes and still compute a function, we used the *coefficient of alienation* as our point of departure. Following Cleveland et al. (1982), we considered that this measure, defined as $r' = 1 - \sqrt{(1 - r^2)}$, was a better approximation of the estimated correlation r(est) than was the correlation coefficient r^2.

Method

Subjects. Two groups of 49 subjects each participated in the experiment. Group 1 consisted of fourth-year industrial engineering and management students who had completed basic and advanced courses in theoretical and applied statistics. Group 2 consisted of twelfth-grade high school students who had no statistical training at all.

Apparatus and design. Each participant received a booklet that contained 36 different scatterplots similar to those in the first experiment (seven levels at the x axis and three points for each level). The variance of the data points at the different levels of x was held approximately constant. The 36 displays consisted of all possible combinations of the experimental variables:

(1) True correlation: Six different values were used, corresponding to $r' = 0.3, 0.4, 0.5, 0.6, 0.7$, and 0.8, where $r' = 1 - \sqrt{(1 - r^2)}$. The values of r^2 were 0.51, 0.64, 0.75, 0.84, 0.91, and 0.96, respectively.
(2) Slope of the regression line: Regression lines had a slope of 30, 45, or 60 deg.
(3) Presence or absence of the regression line: In half of the displays the line was drawn in and in half it was not.

Each scatterplot appeared on a separate page. On the bottom of each page a scale from 0 to 100 was drawn, on which the subjects were to mark their estimates. Examples of scatterplots similar to those used in Experiment 2 are shown in Figure 5.

Procedure. Both groups participated in the experiment as part of a regular classroom session and received instructions identical to those given in the first experiment. However, the high school students received a 15-min explanation about scatterplots and correlations but were given no examples of specific correlation coefficients' displays. Subjects required approximately 10 min to complete the task.

Results and Discussion

Estimates were analyzed in a four-way ANOVA in which the group was a between-subjects variable and correlation level, slope

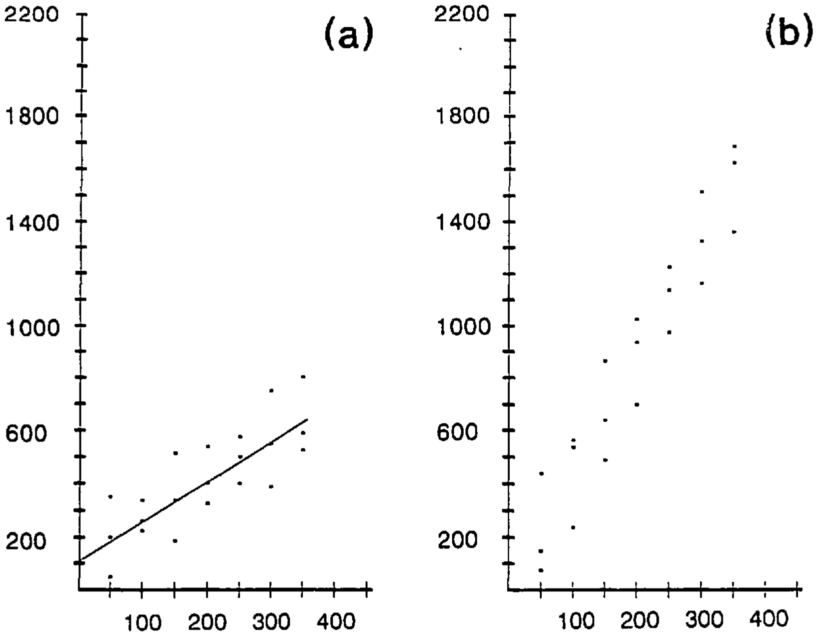

Figure 5. *Scatterplots similar to those used in Experiment 2: (a) data with $r^2 = 0.64$ ($r' = 0.4$), a 30-deg slope and a regression line; (b) data with $r^2 = 0.91$ ($r' = 0.7$), a 60-deg slope, and no regression line. Estimates for both displays are approximately equal, with $r(est) \approx 0.6$.*

of the regression line, and presence or absence of the regression line were within-subjects variables. The mean estimates and standard deviations for the various conditions in the experiment appear in Table 2.

The two groups differed significantly in their mean estimates, $F(1,96) = 17.78$, $p < 0.0005$, $MS_e = 0.412$, with university students' values somewhat higher—r(est) = 0.64—than those of the high school students, r(est) = 0.55. This is consistent with the results of the first experiment, in which estimates by subjects with more statistical training were higher and closer to accepted statistical measures (though they, too, fell below them). The finding in Experiment 1 that presentation of the regression line leads to an increase in magnitude of estimates was also replicated, r(est) = 0.62 versus 0.57, $F(1,96) = 22.16$, $p < 0.0001$, $MS_e = 0.068$.

The main effect of the objective correlation level was highly significant, $F(5,480) = 208.42$, $p < 0.0001$, $MS_e = 0.059$. Estimates fell below the level of r^2 but also differed from r', as can be seen in Figure 6. This lends some support to the Cleveland et al. (1982) claim

TABLE 2

Experts' and Novices' Mean Estimates (and Standard Deviations) in Experiment 2 as a Function of Correlation Level and of the Slope and Display of the Regression Line

Correlation Level (r')	*0.3*	*0.4*	*0.5*	*0.6*	*0.7*	*0.8*
(r^2)	*0.51*	*0.64*	*0.75*	*0.84*	*0.91*	*0.96*
			Experts			
Without regression line						
Slope 30°	0.49	0.59	0.68	0.72	0.76	0.78
	(0.182)	(0.175)	(0.167)	(0.145)	(0.127)	(0.135)
45°	0.44	0.56	0.63	0.67	0.72	0.79
	(0.222)	(0.227)	(0.190)	(0.186)	(0.184)	(0.187)
60°	0.33	0.45	0.55	0.64	0.69	0.76
	(0.218)	(0.203)	(0.200)	(0.155)	(0.169)	(0.151)
With regression line						
Slope 30°	0.55	0.64	0.71	0.74	0.79	0.84
	(0.196)	(0.194)	(0.167)	(0.149)	(0.151)	(0.168)
45°	0.44	0.58	0.62	0.69	0.73	0.80
	(0.200)	(0.169)	(0.192)	(0.139)	(0.151)	(0.136)
60°	0.41	0.52	0.60	0.69	0.69	0.78
	(0.203)	(0.192)	(0.173)	(0.159)	(0.167)	(0.115)
			Novices			
Without regression line						
Slope 30°	0.41	0.58	0.66	0.73	0.77	0.83
	(0.282)	(0.256)	(0.202)	(0.203)	(0.199)	(0.158)
45°	0.28	0.39	0.49	0.54	0.58	0.74
	(0.228)	(0.232)	(0.221)	(0.220)	(0.246)	(0.157)
60°	0.17	0.32	0.35	0.46	0.52	0.59
	(0.174)	(0.192)	(0.247)	(0.198)	(0.220)	(0.239)
With regression line						
Slope 30°	0.48	0.67	0.69	0.76	0.80	0.87
	(0.236)	(0.202)	(0.204)	(0.184)	(0.187)	(0.187)
45°	0.36	0.49	0.56	0.57	0.67	0.75
	(0.221)	(0.212)	(0.230)	(0.194)	(0.190)	(0.185)
60°	0.23	0.34	0.48	0.49	0.54	0.61
	(0.206)	(0.190)	(0.255)	(0.232)	(0.227)	(0.238)

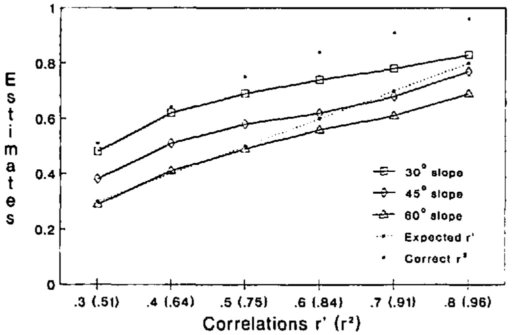

Figure 6. *Correlation estimates as a function of r' and the slope of the regression line. The coefficient of alienation r' has been suggested as a possible approximation of r(est). Also shown are the objective values of r^2.*

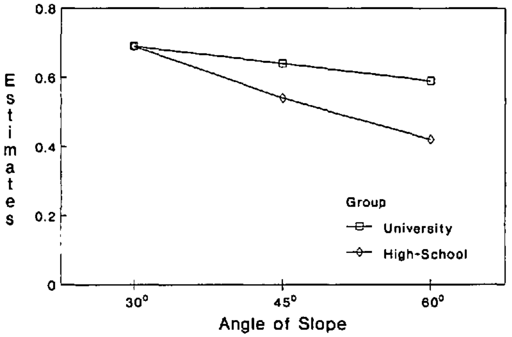

Figure 7. *Effect of the slope of the regression line on correlation estimates of university and high school students.*

that r' is not the correct function for relating subjective estimates to statistical measures. For intermediate levels of correlation estimates were above the corresponding values of r', and for high levels of correlations estimates fell beneath these values.

The slope of the regression line also strongly affected the estimates, $F(2,192) = 188.85$, $p < 0.0001$, $MS_e = 0.052$, with lower estimates for steeper angles of slope, $r(\text{est}) = 0.69, 0.59$, and 0.51 for angles of 30, 45, and 60 deg, respectively. This finding is contrary to Bobko and Karren's (1979) expectation that estimates will be highest when the centroid is at the 45 deg line. In addition, the slope of the regression line interacted with the objective correlation level, $F(10,960) = 1.86$, $p < 0.05$, $MS_e = 0.023$ (see Figure 6). Differences in slope were more pronounced at intermediate levels of correlation and decreased for the highest level of correlation. This is in line with Bobko and Karren's (1979) claim that variability should be minimal for endpoints—that is, for $r^2 = \pm 1$ the slope should have no effect.

The slope of the regression line affected the two groups of subjects differently, $F(2,192) = 38.66$, $p < 0.0001$, $MS_e = 0.052$ (see Figure 7), having more pronounced effects on the high school students than on the university students. This interaction in turn was moderated by the three-way interaction Group × Slope × Line, $F(2,192) = 3.25$, $p < 0.05$, $MS_e = 0.016$.

To understand these results better, we conducted separate analyses of the two groups. For the university students the interaction Slope × Line was significant, $F(2,96) = 6.04$, $p < 0.005$, $MS_e = 0.009$. At slopes of 30 and 60 deg, estimates made when a regression line was present were approximately 0.1 points higher than those made when no line was drawn in, whereas the presence of the line made no difference when the slope was 45 deg. For the high school students the effects of slope and line were additive ($F < 1$ for the interaction).

Experiment 2 showed again that statistical training and the presentation of a regression line led to increased estimates. In addition, the slope of the regression line affected estimates, with lower estimates for steeper slopes (and more widely spread point clouds). Although a systematic function seems to exist that relates correlation levels and estimates when statistical expertise and the presentation of the regression line are held constant, this function is moderated by the slope of the regression line. Thus any attempt to formulate such a function has to take into account the perceptual characteristics of the display, such as the slope of the point cloud.

GENERAL DISCUSSION

The two experiments were designed to clarify and integrate issues related to the intuitive estimation of correlations from scatterplots. Our research replicated two results obtained in the majority of previous studies—first, that people are able to estimate correlations from scatterplots and, second, that the intuitive estimates consistently fall short of the generally used coefficients of correlation r and r^2. This does not mean that intuitive estimates should be considered wrong. Formal statistical coefficients are one way of expressing linear relations between pairs of variables, but they are not the only way. In fact, ·an infinite number of equally valid measures could be devised to express this correlation (e.g., any linear or exponential transformation of the product-moment coefficient). There is no compelling reason for subjects to base their intuitive estimates on any one of these coefficients. The importance of understanding intuitive estimates lies in their relation to the commonly used coefficients of association. As Bobko and Karren (1979) have pointed out, subjects might have misleading notions about the meaning of correlation values. The consistent underestimation of correlation values indicates that when subjects receive a statistical estimate for the level of correlation between two variables (e.g., r^2), they imagine a point cloud that is denser (more highly correlated) than the one leading to the estimate. If, for example, the validity of a psychometric exam is reported as $r = 0.6$, subjects probably imagine a point cloud that corresponds to a much higher correlation level. (In Experiment 2 intuitive estimates of $r(\text{est}) = 0.6$ were given when $r^2 \approx 0.9$). Thus the use of a correlation coefficient might cause overconfidence in the validity of the test and, accordingly, inappropriate decisions.

One way to deal with subjects' biased view of correlation coefficients is to display scatterplots of bivariate relations. As Wainer and Thissen (1979) have pointed out, the use of these graphic displays can lead to a more accurate understanding of the relation between two variables. Especially important are the effects of outliers. The statistical product-moment coefficient is highly sensitive to these outliers, and a single data point with deviant values of the two variables can substantially lower the correlation coefficient, whereas the intuitive estimates remain relatively unaffected by the existence of outliers (e.g., Bobko and Karren, 1979). Our findings support this view. It seems that subjects base their estimates on the density of the point cloud around the majority of x levels. When the points are placed close to the regression line on the majority of values of x (as in the trumpet-shaped dispersions in Experiment 1), correlation estimates are high, even though the wide dispersion at a limited number of levels of x lowers the formal correlation coefficient considerably.

Although the graphic presentation of bivariate relations has definite advantages, one has to remember that scatterplots are graphic displays. As such, impressions based on them may be biased by display characteristics that are irrelevant to the data itself and on which the person creating the scatterplot decides arbitrarily. In the present study two such characteristics were studied. The first was the presentation of the regression line. Our results show that the presence of a regression line has a significant impact on subjective estimates of correlation. It seems that the line serves as some kind of perceptual center that increases the apparent correlation between variables. Considering that subjects' estimates fall consistently below accepted statistical measures, it might be advisable to use regression lines wherever applicable and thereby minimize the difference between subjective estimates and statistic coefficients.

The second characteristic of the display that was shown to influence the correlation estimates was the slope of the regression line, which could be manipulated by altering the scales of the axes. Contrary to the claim raised by Bobko and Karren (1979), we found a definite relation between the slope of the regression line and the correlation estimates. Generally, estimates were higher for shallower slopes. The most plausible explanation is that subjects were influenced by the density of the point cloud, as has been demonstrated in the past (Cleveland et al., 1982; Lauer and Post, 1989). When the point cloud is denser, as is the case when the slope is shallow, the correlation between the variables appears to be higher. This is another factor that the graphic designer has to consider when deciding which format and scales to use for a scatterplot display. These considerations with regard to the design of the display seem to be of major importance because they may lead to serious misconceptions about the relative strength of association between pairs of variables. The two scatterplots in Figure 5 may serve as an example. Both lead to estimates of approximately r(est) = 0.6, though the correlation coefficients for Panels A and B are $r^2 = 0.64$ and $r^2 = 0.91$, respectively. The fact that the lower correlation is displayed with a regression line and at a shallow slope leads to the impression that the two data sets are equally correlated.

A novel issue that has not been studied systematically is the influence of statistical experience and training on subjective estimates of correlation. Our two experiments showed a consistent effect of statistical training on subjective estimates of correlation. Having some knowledge of statistics seems to lead to somewhat higher estimates, a wider range of values, and therefore estimates that are closer to the objective measures. Still, even for people with extensive knowledge in statistics, estimates fell below accepted measures.

A major finding of our study was that people with and without training were equally influenced by perceptual and configural aspects of the display. Statistics courses do not eliminate one's tendency to be influenced by statistically irrelevant but perceptually compelling aspects of the graphic display, such as the existence and slope of a regression line, nor do they sensitize one to the effect of outliers in the data set on the correlation coefficient. Thus the conclusion may be justified that subjective estimates of correlations from scatterplots use a different set of cognitive processes than do those serving regular statistical reasoning.

Our findings indicate that a major component in the estimation of correlations from scatterplots is perceptually based and affects estimates independently from cognitive processes acquired through formal statistical training. Nisbett and Ross's (1980) distinction between theory-driven and data-driven estimates may require some modifications. It seems plausible to conclude that data-driven estimates are the result of two different components: one is related to characteristics of the data (e.g., the dispersion of the point cloud), whereas the second results from purely graphic properties of the display. Both components affect the visual appearance of the display. The final estimate is the result of some combination of statistical knowledge, which influences the range of values of estimates, and a perceptual process that influences the placement of a given display within this range and that is relatively independent from statistical training. The precise nature of the perceptual process remains to be determined, but one promising direction may be an attempt to analyze it in terms of some direct perceptual process analogous to ideas proposed by Gibson (1979), in which people are sensitive to the relations between certain characteristics of the display (e.g., differences in point density in various areas of the display).

Although we still do not know enough about the processes involved in estimating correlations from scatterplots to define the correct function relating estimates and statistical measures of correlation, we have some indications of the reasons previous attempts have failed and what should be the form of such a function. If, as our findings suggest, estimates are the result of some combination of higher cognitive functions influenced by statistical training and perceptual characteristics of the display resulting from the data set and the way the data are shown, any function has to consider these perceptual variables. Some previous attempts were made in this direction, mainly through relating estimates to geometric properties of the point cloud (Cleveland et al., 1982; Collyer, 1988). However, our results show that seemingly irrelevant aspects of the display, such as the presence of the regression line, might affect estimates. Thus future studies should attempt to reveal perceptual relations in the display on which subjects base their estimates.

The study of estimates of correlations from scatterplots may also serve, in a broader theoretical context, as an interesting case for the study of the relation and independence or interaction of various levels of cognitive processes influenced by expertise. The questions raised here concerning the existence of various levels of cognitive processes that are influenced to different degrees by statistical training might also add to our understanding of expertise in general and, thus, may have broader implications.

On a more practical level, we can make a number of recommendations for the presentation of bivariate relations. First, one must keep in mind the differences between the statistical measure and the intuitive estimate. If a task requires an understanding of the general relational pattern of the variables, a scatterplot might be more appropriate, especially when it is possible to inquire more specifically into the characteristics of deviating data points. However, when a task requires a more formal approach to correlations (e.g., statistical inference), there seems no point in displaying the graphical scatterplot, which may mislead subjects by causing them to ignore either existing (relatively weak) correlations or the possible effects of outliers. In addition, if one decides to present the scatterplot graphically, we recommend that the point cloud have a shallow slope (e.g., a 30-deg angle) and that the regression line be displayed in order to create impressions that approach the accepted statistical measures. It might also be advantageous to combine the scatterplot with a display of the statistical measure, thus allowing the user to benefit from both types of information.

ACKNOWLEDGMENTS

Parts of this study were presented at the Fourth Israeli Conference on Judgment and Decision Making, Beer-Sheva, December 1990, and at the Human Factors Society 35th Annual Meeting, San Francisco, September 1991. Work on this project was supported in part by a grant from the Human Factors Branch of the Israeli Ministry of Defence, and we would like to thank contract technical monitor I. Shelach for his support. We would also like to thank Adi Cohen, Arie Lev, and Avner Shalom for their help in preparing and running the experiments and R. Thein and two anonymous reviewers for their comments on an earlier draft of the manuscript.

REFERENCES

Anscombe, F. J. (1973). Graphics in statistical analysis. *American Statistician, 27,* 17–21.

Bobko, P., and Karren, R. (1979). The perception of Pearson product-moment correlations from bivariate scatterplots. *Personnel Psychology, 32,* 313–325.

Chapman, L. J. (1967). Illusory correlation in observational reports. *Journal of Verbal Learning and Verbal Behavior, 6,* 151–155.

Chapman, L. J., and Chapman, J. P. (1967). Genesis of popular but erroneous diagnostic observations. *Journal of Abnormal Psychology, 72,* 193–204.

Chapman, L. J., and Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology, 74,* 271–280.

Cleveland, W. S., Diaconis, P., and McGill, R. (1982). Variables on scatterplots look more highly correlated when the scales are increased. *Science, 216,* 1138–1141.

Cleveland, W. S., Harris, C. S., and McGill, R. (1983). Experiments on quantitative judgments of graphs and maps. *Bell System Technical Journal, 62,* 1659–1674.

Collyer, C. E. (1988, November). *Perceiving scattergrams: Visual line fitting and direct estimation of correlation.* Paper presented at the Annual Meeting of the Psychonomic Society, Chicago, IL.

Collyer, C. E., Stanley, K. A., and Bowater, C. (1990). Psychology of the scientist: LXII. Perceiving scattergrams: Is visual line fitting related to estimation of the correlation coefficient? *Perceptual and Motor Skills, 71,* 371–378.

Gibson, J. J. (1979). *The ecological approach to visual perception.* Boston: Houghton Mifflin.

Jennings, D., Amabile, T. M., and Ross, L. (1982). Informal covariation assessment: Data-based vs. theory-based judgments. In D. Kahneman, P. Slovic, and A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 211–230). New York: Cambridge University Press.

Lane, D. M., Anderson, C. A., and Kellam, K. L. (1985). Judging the relatedness of variables: The psychophysics of covariation detection. *Journal of Experimental Psychology: Human Perception and Performance, 11,* 640–649.

Lauer, T. W., and Post, G. V. (1989). Density in scatterplots and the estimation of correlation. *Behavior and Information Technology, 8,* 235–244.

Lewandowsky, S., and Spence, I. (1989). Discriminating strata in scatterplots. *Journal of the American Statistical Association, 84,* 682–688.

Mosteller, F., Siegel, A. F., Trapido, E., and Youtz, C. (1981). Eye fitting straight lines. *American Statistician, 35,* 150–152.

Nisbett, R., and Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment.* Englewood Cliffs, NJ: Prentice-Hall.

Pollack, I. (1960). Identification of visual correlational scatterplots. *Journal of Experimental Psychology, 59,* 351–360.

Strahan, R. S., and Hansen, C. J. (1978). Underestimating correlation from scatterplots. *Applied Psychological Measurement, 2,* 543–550.

Wainer, H., and Thissen, D. (1979). On the robustness of a class of naive estimators. *Applied Psychological Measurement, 3,* 543–551.