6 Induction

Whereas estimation is inference of a quantity based on data directly bearing on that quantity, i.e. either a single random variable or function, or a set of independent random variables, induction is the process of combining observations of two or more random variables that may not be independent, in order to infer relationships between the variables, make predictions, infer causality, and revise beliefs. We will survey the theory and evidence about induction in four categories corresponding to these goals: generalization, prediction, explanation, and assimilation. But first, we begin with the mathematical concept known as correlation.

6.1 Correlation

In what follows, we will develop the statistical theory for dependencies between two random variables in a finitely additive probability space. The theory can be extended, with various levels of complication, to relationships between random variables in countably infinite or continuous spaces and between more than three random variables (see e.g. Mood, Graybill, and Boes, 1974).

DEFINITION 6.1.1. Let X and Y be two finite random variables. Then the *joint probability mass function* of X,Y is defined to be the probability mass function $p_{XY}(x,y) = P(X=x;Y=y)$ for all points x,y in the range of X,Y.

COROLLARY 6.1.2. If $p_{XY}(x,y)$ is the joint p.m.f. of two random variables $X, Y,$ then $\sum_{x,y} p_{XY}(x,y) = 1$.

Proof. Since $p_{XY}(x,y)$ is a p.m.f., by definition 5.1.3 its value at every point x,y in its domain is determined by summing the probabilities for each state in a sample space where $X=x$ and $Y=y$. By the definition of a random variable (5.1.1), X and Y must each assign values in their range for every state in the sample space. By definition 6.1.1, $p_{XY}(x,y) = P(X=x;Y=y)$ for all points x,y in the range of X,Y, so $\sum_{x,y} p_{XY}(x,y)$ is the sum of all the probabilities in a sample space, which by 4.1.7(a) is 1.

DEFINITION 6.1.3. Let X and Y be two finite random variables having a joint p.m.f. $p_{XY}(x,y)$. The *marginal (*or "unconditional") *probability mass functions* for X and Y are $p_X(x) = P(X=x)$ and $p_Y(y) = P(Y=y)$ for all points x,y in the range of X,Y.

DEFINITION 6.1.4. Let X and Y be two finite random variables having a joint p.m.f. $p_{XY}(x,y)$ and marginal p.m.f.'s $p_X(x)$ and $p_Y(y)$. The *conditional probability mass function* of Y given $X=x$ is defined by $p_{Y|X}(y|x) = p_{XY}(x,y) / p_X(x)$ if $p_X(x) > 0$ and is undefined if $p_X(x) = 0$, for all points x,y in the range of X,Y.

DEFINITION 6.1.5. Two finite random variables X and Y having a joint p.m.f. $p_{XY}(x,y)$ and marginal p.m.f.'s $p_X(x)$ and $p_Y(y)$ are *stochastically independent* (or just "independent") iff $p_{XY}(x,y) = p_X(x)p_Y(y)$ for all points x,y in the range of X,Y.

COROLLARY 6.1.6. If X and Y are stochastically independent, then $p_{Y|X}(y|x) = p_Y(y)$ for all points x,y in the range of X,Y provided $p_X(x) > 0$.

Proof. By 6.1.4, for all points x,y in the range of $X,Y,$ if $p_X(x) > 0$ then $p_{Y|X}(y|x) = p_{XY}(x,y) / p_X(x)$, hence $p_{XY}(x,y) = p_{Y|X}(y|x)p_X(x)$, but by 6.1.5, $p_{XY}(x,y) = p_X(x)p_Y(y)$, so $p_{Y|X}(y|x) = p_Y(y)$ by substitution.

DEFINITION 6.1.7. Let X and Y be any two finite random variables defined on the same probability

space. The *covariance* σ_{XY} of X and Y is defined as follows:

$\sigma_{XY} \stackrel{def}{=} Cov(X,Y) \stackrel{def}{=} M[(X-\mu_X)(Y-\mu_Y)]$.

COROLLARY 6.1.8. $Cov(X,Y) = M(XY) - \mu_X\mu_Y$.

Proof. $Cov(X,Y) = M[(X-\mu_X)(Y-\mu_Y)]$, by 6.1.1. $M[(X-\mu_X)(Y-\mu_Y)] = M(XY - \mu_XY - \mu_YX + \mu_X\mu_Y) = M(XY) - \mu_XM(Y) - \mu_YM(X) + \mu_X\mu_Y = M(XY) + \mu_X\mu_Y$.

DEFINITION 6.1.9. The *correlation coefficient* ρ_{XY} (also known as the *population correlation coefficient)* between two random variables X and Y is defined as follows:

$\rho_{XY} \stackrel{def}{=} Cov(X,Y) / \sigma_X\sigma_Y$.

THEOREM 6.1.10. If two random *variables* X and Y are stochastically independent, then:

$Cov(X,Y) = 0$ (= ρ_{XY} if $\sigma_X\sigma_Y > 0$).

Hence, whenever two variables are independent, they are uncorrelated. The reverse statement is not true, however.

EXAMPLE 6.1.11. *Uncorrelated random variables that are not independent.* Suppose that X is uniformly distributed over the interval {-1,1} (viz, X takes values -1, 0, and 1 with equal probability). Define $Y = X^2$. Thus, $p_Y(y) = 2/3$ for y=1, 1/3 for y=0, and 0 otherwise. But $p_{Y|X}(y|x=0) = 1$ if y=0 and 0 otherwise, $p_{Y|X}(y|x=1) = 1$ if y=1 and 0 otherwise, and $p_{Y|X}(y|x=-1) = 1$ if y=1 and 0 otherwise. So X and Y are not independent, because, for example, $p_{Y|X}(y=1|x=1) = 1 \neq 2/3 = p_Y(y=1)$. But $Cov(X,Y) = M(XY) - \mu_X\mu_Y = (1/3)[(-1)(1) + (0)(0) + (1)(1)] - [(1/3)(-1+0+1)][(2/3)(1)+(1/3)(0)] = (1/3)(0) - (0)(2/3) = 0 - 0 = 0$. Furthermore, $\sigma_X\sigma_Y = [\sqrt{(1/3)(1+0+1)}][\sqrt{\{(1/3)(0)+(2/3)(1)\}} > 0$, so X and Y are uncorrelated.

EXERCISE 6.1.12.. Prove 6.1.10.

LEMMA 6.1.13. *Cauchy-Schwartz inequality.* If X and Y are finite random variables, then

(a) $[M(XY)]^2 \leq M(X^2)M(Y^2)$, and

(b) there exists a constant c such that $P[Y=cX]=1$ iff $[M(XY)]^2 = M(X^2)M(Y^2)$.

Proof. (Mood, Graybill, and Boes, 1974). Define $0 \leq h(t) = M[(tX - Y)^2] = M(X^2)t^2 - 2M(XY)t + M(Y^2)$. But $h(t) \geq 0$ because it is the mean of a squared function. If $h(t) > 0$, then solving the quadratic equation in t implies that the roots of $h(t)$ are imaginary, so that $4[M(XY)]^2 - 4M(X^2)M(Y^2) < 0$, or $[M(XY)]^2 < M(X^2)M(Y^2)$. If $h(t) = 0$ for some t (call it c), then $M[(cX - Y)^2] = 0$, implying that $P(cX = Y) = 1$.

THEOREM 6.1.14. The range of the correlation coefficient is [-1,1] (the interval from -1 through 1), and $|\rho_{XY}| = 1$ iff there is a constant c such that $P[Y=cX]=1$.

Proof. (Mood, Graybill, and Boes, 1974). Rewrite 6.1.13 as $|M(UV)| \leq \sqrt{[M(U^2)M(V^2)}$, and let $U = X - \mu_X$ and $V = Y - \mu_Y$.

EXPERIMENT 6.1.15. *Intuitive correlation.* Jennings, Amabile, and Ross (1982) gave subjects lists of number pairs. The task was to assign a number between -100 and 100 (inclusive) to each list based on the strength of the relationship between the pairs. Quoting Baron (2000): "The true correlation coefficients of the numbers in each list differed from list to list: the range was from 0 to 1. Subjects gave ratings near 100 for correlations of 1, and ratings near 0 for correlations of 0. For correlations of .

5, subjects gave ratings of about 20. These results tell us that the naïve idea of 'degree of relationship', although it resembles correlation, is not quite the same as the relationship measured by the correlation coefficient ρ. The subjects' deviation from mathematical correlation was not an error, for ρ is only one of many possible measures of association, and there is no reason subjects should use this particular measure."

6.2 Generalization

Generalization involves the inference of a relationship between two or more variables. The relationship that is inferred might be that the variables are independent of each other, but there is no assumption beforehand that this will be case.

When we generalize, we infer a relationship between variables based on an incomplete sample of a population. In order to apply correlation to a sample, we must compute both the sample mean M' (defined in 5-Estimation) and also a quantity we have not yet defined: the sample variance.

DEFINITION 6.2.1. For a random sample $X_1,X_2,...,X_n$ from a sampling distribution $q(x)$, the *sample variance* is $S^2 = \sum_{i\in\{1,...,n\}} (X_i - M')^2 / (n - 1)$. The *sample standard deviation* is just $S = \sqrt{S^2}$.

LEMMA 6.2.2. $\sum (X_i - \mu)^2 = \sum (X_i - M')^2 + n(M' - \mu)^2$.

Proof. $\sum(X_i - \mu)^2 = \sum (X_i - M' + M' - \mu)^2 = \sum [(X_i - M') + (M' - \mu)]^2 = \sum [(X_i - M')^2 + 2(X_i - M')(M' - \mu) + (M' - \mu)^2] = \sum (X_i - M')^2 + 2(M' - \mu)\sum(X_i - M') + n(M' - \mu)^2 = \sum (X_i - M')^2 + n(M' - \mu)^2$ (Mood, Graybill, and Boes, 1974).

THEOREM 6.2.3. The sample variance S^2 of a random sample from a sampling distribution $q(x)$ is an unbiased estimator of the variance σ^2 of a p.m.f. $p(x)$ if $q(x) = p(x)$.

Proof. Applying lemma 6.2.2, $M(S^2) = M[\sum_{i\in\{1,...,n\}} (X_i - M')^2 / (n - 1)] = [1 / (n - 1)] M[\sum (X_i - M')^2 + n(M' - \mu)^2] = [1 / (n - 1)] \{[\sum M(X_i - M')^2 + nM[(M' - \mu)^2]\} = [1 / (n - 1)] [\sum \sigma^2 + nVar(M')] = [1 / (n - 1)] [n\sigma^2 + n\sigma^2 /n] = \sigma^2$ (Mood, Graybill, and Boes, 1974).

DEFINITION 6.2.4. For a random sample $X_1,X_2,...,X_n$ with sample mean M_X and sample standard deviation S_X, the *standard scores* (or *z-scores*) are defined by:

$$Z_{Xi} = (X_i - M_X)/S_X.$$

DEFINITION 6.2.5. For a random sample $\langle X_1,Y_1\rangle,\langle X_2,Y_2\rangle,...,\langle X_n,Y_n\rangle$ from a sampling distribution $q(x,y)$, with sample means M_X,M_Y and sample standard deviations S_X,S_Y, the *sample correlation coefficient* (or *Pearson product-moment correlation coefficient*) is defined by:

$$r_{XY} = \sum_i Z_{xi}Z_{Yi} / n.$$

The sample correlation coefficient r is an estimator for the true correlation coefficient ρ of a distribution $p(x,y)$ under the assumption that $q(x,y) = p(x,y)$, and it shares the same range as ρ. [Need to formulate unbiased estimation theorem and prove it here.]

EXPERIMENT 6.2.6. *Positive evidence bias*. Building on a study of 10-15 year old children by Inhelder and Piaget (1958), Smedslund (1963) found evidence that people attend selectively to positive (1,1) pairs when judging the correlation between boolean variables. Nurses were given 100 cards each supposedly describing the case of a patient. Nurses were told to determine "whether there was a

relationship (connection)" between a symptom and a disease. From Baron (2000): "Each card indicated whether the symptom was present or absent and whether the disease was ultimately found to be present or absent in each case." The following table shows the pattern of data seen by the nurses:

	Disease present	Disease absent
Symptom present	37	17
Symptom absent	33	13

For the above table, there is no positive correlation between presence of the disease and presence of the symptom, yet 85% of the nurses said there was a relationship. Data from various other such tables showed that the best predictor of whether subjects found a relationship was the proportion of cases in the present-present cell. Subjects tended to neglect the rest of the table and instead only look at the positive evidence for the hypothesis that a relationship existed.

EXERCISE 6.2.7. Calculate the sample coefficient of correlation between the symptom and disease in the table in experiment 6.2.6.

EXPERIMENT 6.2.8. *Illusory correlation.* Chapman and Chapman (1971) gave college students drawings from the Draw-a-Person test, a tool used in clinical diagnosis of mental disorders. Each drawing was labeled with a psychological characteristic supposedly corresponding to the person who made the drawing. Example characteristics were "suspicious of other people" and "has had problems of sexual impotence". The labels were chosen so that there was no correlation between psychological characteristics and pictorial features in the drawings widely believed to be associated with these characteristics, e.g. "big eyes" for "suspicious of other people" and sexual features for "has had problems of sexual impotence". The subjects were asked to discover relationships between drawing features and psychological characteristics from these drawings, and reported the correlations that widely held prior beliefs predicted would exist. The authors called this "illusory correlation". The correlations that were found by subjects in the Draw-a-Person test experiment were the same as ones believed by clinicians who worked with actual patients, but no such correlations existed in the patient population. A similar study was done with Rorschach inkblot test data. Students found illusory positive correlations in Rorschach data between patients' actual responses and their clinical diagnoses when a relationship between the response and diagnosis is naively predicted. They also *failed* to find relationships that did exist when these were *not* expected.

EXPERIMENT 6.2.9. *Memory enhancement of illusory correlation.* Schweder and D'Andrade (1979) reanalyzed data from a study of extroversion and introversion in summer camp children done by Newcomb (1929). Counselors rated each child each day on a set of behavioral traits (e.g. "speaks of confidence of his own abilities", "takes the initiative in organizing games", and "spends more than an hour a day alone"). At the end of the whole summer, the counselors filled out ratings of the same children on the same items. The data permitted correlation coefficients to be computed for each pair of behaviors over the summer both for the daily reports and the end-of-summer reports. Scheder and D'Andrade found much higher correlations in the end-of-summer reports than in the daily reports between traits that University of Chicago undergraduates rated as conceptually similar. This indicates that the passage of time enhances the effect of prior belief about what traits should go together.

EXPERIMENT 6.2.10. *Distinctive events.* Hamilton and Gifford (1976) showed students at Southern Connecticut State College slides on members of "Group A" or "Group B" who were said to have done something desirable or undesirable. Myers (1983) writes: "For example, 'John, a member of Group A, visited a sick friend in the hospital.' Twice as many statements described members of Group A as Group

B, but both groups were associated with nine desirable behaviors for every four undesirable behaviors. Since both Group B and the undesirable acts were less frequent, their co-occurrence – for example, 'Allen, a member of Group B, dented the fender of a parked car and didn't leave his name' – was an infrequent combination that caught people's attention. The students therefore overestimated the frequency with which the 'minority' group (B) acted undesirably; consequently, they judged Group B more harshly. ... Evidently the joint occurrence of two distinctive events grabs attention."

EXPERIMENT 6.2.11. *Stereotyping.* Hamilton and Rose (1980) had University of California, Santa Barbara undergraduates read sentences describing members of various occupational groups, e.g. "Doug, an accountant, is timid and thoughtful." Each occupation was described equally often by each adjective, but students induced illusory correlations in the data, thinking they had seen more accountants who were timid, doctors who were wealthy, and salespeople who were talkative, in line with their prior beliefs or stereotypes. This shows that data are both interpreted in the light of and may fail to extinguish prior beliefs.

EXPERIMENT 6.2.12. *Prior beliefs and positive evidence bias.* Kuhn, Amsel, and O'Loughlin (1988) gave both children and adults data from experiments in which boarding school children were given different combinations of food by researchers to determine whether foods affected the probability of the children getting a cold. From Baron (2000): "Before seeing the data, each subject stated her own hypothesis about which of the foods would cause or prevent colds. Then subjects were shown the data (table by table) and were asked to describe what the data showed. When subjects were asked whether the data showed that the food made a difference, they interpreted the evidence in a way that was colored by their hypothesis. For example, one subject thought that the type of water (tap or bottled) would make a difference but that the type of breakfast roll would not. The evidence presented was identical for both variables, since tap water was always given with one type of roll and bottled water with another, yet the subject interpreted the evidence as showing that water made a difference and rolls did not. When another subject believed that mustard caused colds, she looked selectively for cases in which mustard had been eaten and colds had occurred, and she ignored cases in which mustard had been eaten and colds had not occurred."

6.3 Prediction

When we use the value of one variable to infer the value of another variable that is not yet observed, we call this "prediction". When two random variables are independent, the best prediction for each one given a value for the other is just the mean of the random variable that is being predicted.

EXPERIMENT 6.3.1. *The "hot hand" - the flipside of the gambler's fallacy.* Gilovich, Vallone, and Tversky (1985, handed out) recruited 100 basketball fans from the student bodies of Stanford and Cornell Universities to fill out a questionnaire. The sample included 50 captains of intramural basketball teams, and all subjects played basketball at least "occasionally" (65% played "regularly"). All watched at least 5 games per year and 73% watched more than 15 games a year. The authors write:

> The fans were asked to consider a hypothetical player who shoots 50% from the field. Their average estimate of his field goal percentage was 61% 'after having just made a shot,' and 42% 'after having just missed a shot.' Moreover, the former estimate was greater than or equal to the latter for every

> respondent. When asked to consider a hypothetical player who shoots 70% from the free-throw line, the average estimate of his free-throw percentage was 74% "for second free throws after having made the first," and 66% "for second free throws after having missed the first." Thus, our survey revealed that basketball fans believe in "streak shooting."

Gilovich et al. collected data from National Basketball Association (NBA) games for both field goals and free throws. In both cases, they found no evidence for streak shooting, also known as the "hot hand" hypothesis. Serial correlations (the correlations between boolean outcomes on a shot given the previous shot) were negative on average, the opposite of what people predict, and there were no more streaks in the shooting data than would be expected by chance.

EXPERIMENT 6.3.2. *Base rate neglect in prediction.* Kahneman and Tversky (1973) showed that people overuse nondiagnostic information in making predictions, essentially an application of base rate neglect. Subjects were divided into a base-rate group, a similarity group, and a prediction group. The base rate group were asked:

> Consider all first-year graduate students in the U.S. today. Please write down your best guesses about the percentage of these students who are now enrolled in each of the following nine fields of specialization.

The nine fields are listed in the table, reproduced from the article. The similarity group was asked to rank the nine fields in terms of how similar the subject of following personality sketch was to each field:

> Tom W. is of high intelligence, although lacking in true creativity. He has a need for order and clarity, and for neat and tidy systems in which every detail finds its appropriate place. His writing is rather dull and mechanical, occasionally enlivened by somewhat corny puns and by flashes of imagination of the sci-fi type. He has a strong drive for competence. He seems to have little feel and little sympathy for other people and does not enjoy interacting with others. Self-centered, he nontheless has a deep moral sense.

The third, or prediction, group was given both the above personality sketch and and the following additional information:

> The preceding personality sketch of Tom W. was written during Tom's senior year in high school by a psychologist, on the basis of projective tests. Tom W. is currently a graduate student. Please rank the following nine fields of graduate specialization in order of the likelihood that Tom W. is now a graduate student in each of these fields.

ESTIMATED BASE RATES OF THE NINE AREAS OF GRADUATE SPECIALIZATION AND SUMMARY OF SIMILARITY AND PREDICTION DATA FOR TOM W.

Graduate specialization area	Mean judged base rate (in %)	Mean similarity rank	Mean likelihood rank
Business Administration	15	3.9	4.3
Computer Science	7	2.1	2.5
Engineering	9	2.9	2.6
Humanities and Education	20	7.2	7.6
Law	9	5.9	5.2
Library Science	3	4.2	4.7
Medicine	8	5.9	5.8
Physical and Life Sciences	12	4.5	4.3
Social Science and Social Work	17	8.2	8.0

The table displays data from each of the three groups. The ranks were inverted for the second and third columns so that higher numbers would correspond to higher similarity and likelihood scores. The correlation between judged likelihood and similarity was .97, while the correlation between judged likelihood and estimated base rates was -.67, indicating that likelihoods were judged almost entirely based on similarity. Overcoming base rates is normatively justified only if the description is a highly reliable predictor of graduate study field. But subjects did not believe it was. Respondents in the prediction group were asked, following the prediction task itself, to estimate the percentage of correct first choices among the nine fields which could be achieved with different types of information. The median estimate was 23% for predictions based on personality sketches of the type given above, from what are called “projective tests”, compared to 53% “for predications based on high school seniors' reports' of their interests and plans”. Kahneman and Tversky summarize:

> In their exclusive reliance on the personality sketch, the subjects in the prediction group apparently ignored the following considerations. First, given the notorious invalidity of projective personality tests, it is very likely that Tom W. was never in fact as compulsive and as aloof as his description suggests. Second, even if the description was valid when Tom W. was in high school, it may no longer be valid now that he is in graduate school. Finally, even if the description is still valid, there are probably more people who fit that description among students of humanities and education than among students of computer science, simply because there are so many more students in the former than in the latter field.

The authors attribute the results to the representativeness heuristic, i.e. subjects ignore base rates and judge only on the basis of how similar, or representative, the description is relative to a field.

DEFINITION 6.3.3. For a random sample $<X_1,Y_1>,<X_2,Y_2>,...,<X_n,Y_n>$ from a sampling distribution $q(x,y)$, with sample means M_X,M_Y and sample standard deviations S_X,S_Y, the *simple linear model* for predicting the standard score Z_{Yi} of Y_i from the calculated standard score Z_{Xi} of an observed value of X_i is defined by:

$Z_{Yi}' = a + bZ_{Xi}$ for all $i \in \{1,...,n\}$.

The calculated value Z_{Yi}' is called the *predicted value* of the standard score Z_{yi}.

DEFINITION 6.3.4. For random variables $Y_1,Y_2,...,Y_n$, the *least squares estimator* Y_i' is the function that minimizes $\sum_i (Y_i' - Y_i)^2/n$.

LEMMA 6.3.5. For a random sample $<X_1,Y_1>,<X_2,Y_2>,...,<X_n,Y_n>$ from a sampling distribution $q(x,y)$, with sample means M_X,M_Y and sample standard deviations S_X,S_Y, the least squares estimator for Z_{Yi}' in the simple linear model $Z_{Yi}' = a + bZ_{Xi}$ is of the form $Z_{Yi}' = bZ_{Xi}$, i.e. $a = 0$.

Proof. See Hays (1980, handed out).

THEOREM 6.3.6. The least squares estimator Z_{Yi}' in the simple linear model $Z_{Yi}' = bZ_{Xi}$ is $Z_{Yi}' = r_{XY}Z_{Xi}$, where r_{XY} is the sample correlation coefficient of the random sample $<X_1,Y_1>,<X_2,Y_2>,...,<X_n,Y_n>$.

Proof. See Hays (1980, handed out).

DEFINITION 6.3.7. For random variables $Y_1,Y_2,...,Y_n$, the *best linear unbiased estimator* Y_i' is the function that minimizes $Var(Y_i')$ among all unbiased linear estimators for Y_i.

THEOREM 6.3.8. *Gauss-Markov theorem.* The simple linear model $Z_{Yi}' = r_{XY}Z_{Xi}$, where r_{XY} is the sample correlation coefficient of the random sample $<X_1,Y_1>,<X_2,Y_2>,...,<X_n,Y_n>$ is the best linear unbiased estimator for Z_{Yi}.

Proof. See Ruud (1995) at http://emlab.berkeley.edu/GMTheorem/index.html.

COROLLARY 6.3.9. *Regression toward the mean.* The best linear unbiased estimate (predicted value) for Y_i based on a random sample $<X_1,Y_1>,<X_2,Y_2>,...,<X_n,Y_n>$ with $|r_{XY}| < 1$ is closer to M_Y in units of sample standard deviation S_Y than X_i is to M_X in units of sample standard deviation S_X (Galton, 1886).

EXERCISE 6.3.10. Prove 6.3.9.

REMARK 6.3.11. Regression toward the mean is often stated as a mere consequence of imperfect correlation. But it also depends on the assumption of a linear estimate, i.e. a linear model as defined above. For nonlinear models (e.g. cubic models), an unbiased estimate of one variable based on another may not imply regression to the mean. Regression to the mean is therefore an empirical phenomenon which is implied by the linear model. It may not occur in reality if a nonlinear model better fits observations than a linear one.

EXPERIMENT 6.3.12. *Nonregressive prediction – insensitivity to predictability.* A correlation of 1 or -1 implies perfect predictability of one variable based on another. In mathematical terms, this means that one standard deviation of difference between an observed value and the mean of one variable X predicts exactly one standard deviation of difference between a target value and the mean of the target variable Y. When the absolute value of the correlation is less than 1, however, predictability is imperfect, and an unbiased linear estimate of Y based on an observed value of X would reflect what is known as *regression to the mean.* If the observed value for a given individual on variable X is $\mu_X + \sigma_X$, and r_{XY} is between 0 and 1, then the conditional mean of Y given that $X = \mu_X + \sigma_X$ is $\mu_Y + r_{XY}\,\sigma_Y$, so that the observed value of Y is likely to be closer to the mean of Y than the observed value of X is to the

mean of *X*. Subjects in a study from Kahneman and Tversky (1973) violated this principle. Tversky and Kahneman (1974) write:

> In one of these studies, subjects were presented with several paragraphs, each describing the performance of a student teacher during a particular practice lessons. Some subjects were asked to *evaluate* the quality of the lesson described in the paragraph in percentile scores, relative to a specified population. Other subjects were asked to *predict*, also in percentile scores, the standing of each student teacher 5 years after the practice lesson. The judgments made under the two conditions were identical. That is, the prediction of a remote criterion (success of a teacher after 5 years) was identical to the evaluation of the information on which the prediction was based (the quality of the practice lesson). The students who made these predictions were undoubtedly aware of the limited predictability of teaching competence on the basis of a single trial lesson 5 years earlier; nevertheless, their predictions were as extreme as their evaluations.

APPLICATION 6.3.13. *Failure to recognize regression effects*. Kahneman (2002) describes the following anecdote:

> I had the most satisfying Eureka experience of my career while attempting to teach flight instructors that praise is more effective than punishment for promoting skill-learning. When I had finished my enthusiastic speech, one of the most seasoned instructors in the audience raised his hand and made his own short speech, which began by conceding that positive reinforcement might be good for the birds, but went on to deny that it was optimal for flight cadets. He said, "On many occasions I have praised flight cadets for clean execution of some aerobatic maneuver, and in general when they try it again, they do worse. On the other hand, I have often screamed at cadets for bad execution, and in general they do better the next time. So please don't tell us that reinforcement works and punishment does not, because the opposite is the case." This was a joyous moment, in which I understood an important truth about the world: because we tend to reward others when they do well and punish them when they do badly, and because there is regression to the mean, it is part of the human condition that we are statistically punished for rewarding others and rewarded for punishing them. I immediately arranged a demonstration in which each participant tossed two coins at a target behind his back, without any feedback. We measured the distances from the target and could see that those who had done best the first time had mostly deteriorated on their second try, and vice versa. But I knew that this demonstration would not undo the effects of lifelong exposure to a perverse contingency.

The example shows the power of nonregressive prediction in influencing everyday behavior. Tversky and Kahneman (1974) note:

> Consequently, the human condition is such that, by chance alone, one is most often rewarded for punishing others and most often punished for rewarding them. People are generally not aware of this contingency. In fact, the elusive role of regression in determining the apparent consequences of reward and punishment seems to have escaped the notice of students of this area.

6.4 Explanation

Explanation is the process of inferring which of two or more possibilities is most likely given observed data. It can be seen as the flip side of prediction. Whereas prediction involves estimating the most

likely effect from a given cause, explanation (also sometimes called “abduction”, diagnosis, attribution, or “postdiction”) involves estimating the most likely cause given an observed effect.

EXPERIMENT 6.4.1. *Conservatism in bayesian updating.* Edwards (1968) presented subjects with the following problem:

> This bookbag contains 1,000 poker chips. I started out with two such bags, one containing 700 red and 300 blue chips, the other containing 300 red and 700 blue. I flipped a fair coin to determine which one to use. Thus, if your opinions are like mine, your probability at the moment tthat this is the predominantly red bookbag is 0.5. Now, you sample, randomly, with replacement after each chip. In 12 samples, you get 8 reds and 4 blues. Now, on the basis of everything you know, what is the probability that this is the predominantly red bag?

Typical responses ranged from 0.7 to 0.8. But the correct answer based on probability theory is 0.97. Griffin and Tversky (1992) explain this finding as a result of neglect of the effects of sample size: Subjects are underconfident in this experiment because they look mostly at the proportion of reds in the sample, which is 67%. Subjects could also be engaging in what Kahneman and Frederick (2000) call *attribute substitution*, or using the proportion of reds as a heuristic in estimating the chances that the bookbag is one favoring red chips.

EXPERIMENT 6.4.2. *Sample size underweighting.* Griffin and Tversky (1992) gave students two hypotheses for the bias of a coin: 3/5 favoring heads, or 3/5 favoring tails. Data were provided from samples of different sizes. The posterior probability that the coin is biased toward heads rather than tails depends only on the difference between the number of heads and the number of tails in a sample (*#H - #T)*. Thus, 6 heads and 3 tails leads to the same posterior probability that the coin is biased toward heads as does a sample of 10 heads and 7 tails (0.77). Subjects' confidence in the hypothesis largely ignored sample size, however. Thus, they were overconfident when sample sizes were small, and underconfident when sample sizes were large. Equating for a posterior probability of 0.8 favoring heads bias hypotheses, subjects assigned posterior probabilities of only 0.71 (< 0.8, representing underconfidence) on average for sample sizes of 17, but of 0.87 (> 0.8, representing overconfidence) for sample sizes of 5. Subjects thus underweight or neglect sample size in making such judgments.

EXPERIMENT 6.4.3. *Cause-effect asymmetries.* Tversky and Kahneman (1980) asked students at the University of Oregon to consider the following

> Let A be the event that before the end of next year, Peter will have installed a home burglar system in his home. Let B denote the event that Peter's home will be burglarized before the end of next year. Let ¬A and ¬B denote the negations of A and B, respectively.
>
> (a) Which of the two conditional probabilities, P(A|B) or P(A|¬B), is higher?
>
> (b) Which of the two conditional probabilities, P(B|A) or P(B|¬A), is higher?

Most subjects (132/161) said that P(A|B) > P(A|¬B) and P(B|A) < P(B|¬A). But this violates the laws of probability. The authors interpret the results as evidence that people impose a causal schema on the interpretation of evidence, which introduces asymmetries that are not consistent with probability. For example, subjects are more likely to say it is more probable that a girl will have blue eyes if her mother has blue eyes than the reverse, although the conditional probabilities are the same.

6.5 Assimilation

Assimilation is the process of updating one's prior beliefs on the basis of new evidence.

EXPERIMENT 6.5.1. *Biased assimilation.* In an experiment conducted by Lord, Lepper and Ross (1979), when shown two research studies supporting opposite sides of the capital punishment debate, both students in favor of and those opposing the death penalty accepted evidence consistent with their prior beliefs but were very critical of opposing evidence. In fact, the same evidence caused *both* groups to become more confident of their views. Normatively, mixed evidence should bring the two groups closer together, but instead it pushed them further apart. This phenomenon is known as biased assimilation, and it has also been demonstrated observationally with the audiences for presidential debates in the U.S. Supporters of a candidate are overwhelmingly more likely than opponents to see their favored candidate as having won a debate, and debates tend to strengthen each group's belief in their favored candidate. When people were asked to be "as *objective* and *unbiased* as possible", they were just as biased as before (Lord, Lepper, and Preston, 1984).

The following two experiments violate a general principle of *order independence*.

EXPERIMENT 6.5.2. *Primacy effect in persuasion.* Miller and Campbell (1959, cited in Myers, 2006) gave students at Northwestern University a condensed transcript from a civil trial. The plaintiff's testimony and arguments were placed in one block, while those for the defense were placed in another. Students read both blocks. A week later, most students sided with whatever information they had read first. Similar effects have been observed in other experiments. Asch (1946) found that the order in which ajectives are listed describing a person affects the positivity of overall ratings by subjects. Candidates listed first on a ballot tend to garner more votes (Moore, 2004), and success before failure is seen as a better sign of capability than failure before success (Myers, 2006).

EXPERIMENT 6.5.3. *Recency effect in persuasion.* Recency, the opposite of primacy, predominates when time or other memory limiting factors favor information presented later. Miller and Campbell (1959, cited in Myers, 2006) gave a group of students one block of testimony to read one week, and the second block a week later. Opinions right after reading the second block a week later showed a recency effect: students favored the second block's arguments.