5 Estimation

As we have seen, one can study people's confidence and beliefs without reference to how much they tell us about reality outside of someone's head, by studying how well beliefs and judgments cohere with each other and with a priori principles of inference. In this section, we relate confidence and probability to observable or calculable quantities which are the objects of belief.

We build a theory of agent estimation from the elementary theory of statistics, as developed in Mood, Graybill, and Boes (1974).

5.1 Random variables

DEFINITION 5.1.1. For a given probability space $<S,E,P>$, $X: S \rightarrow \mathbb{R}$ (a function that maps the sample space into the real numbers) is a *random variable* on $<S,E,P>$ iff $\forall r \in \mathbb{R}\ A_r = \{s: X(s) \leq r\} \in E$.

In what follows we will restrict our attention to random variables on finite sample spaces, which we will call *finite random variables.*

EXAMPLE 5.1.2. We can define a random variable X on the sample space for coin flips as $X(Heads)=1$ and $X(Tails)=0$.

DEFINITION 5.1.3. Let X be a random variable on a finitely additive probability space $<S,E,P>$, let x represent numbers in the range of X, and let $I: S \times X \rightarrow \{0,1\}$ be the indicator function $I(s,x) = 1$ if $X(s)=x$, and 0 otherwise. Then we say that X has a *probability mass function* $p(x)$ defined by $p(x) = \sum_{s \in S} I(s,x)P(\{s\})$.

When we have a probability mass function ("p.m.f.") for a random variable X, we can often dispense with references to the probability space and refer just to the values x in the range of X, writing $x \in X$. We can also treat mathematical statements about random variables as events in a probability space, and use the function P to refer to the probability measure over such statements as defined by a p.m.f. A probability mass function is not the same as a probability measure because it is defined only for particular values of a random variable and not over an algebra of events. We can easily extend the notion of a p.m.f. to a function of multiple random variables, called a *joint probability mass function.*

DEFINITION 5.1.4. Let x be a variable taking on values in a finite set of n numbers. Then p is a *uniform probability mass function* on x iff $p(x) = 1/n$ for all x.

EXAMPLE 5.1.5. Let $S = \{1,2,3,4,5,6\}$ correspond to the faces of a six-sided die. Let X be the absolute value of the difference between the number represented on a face of the die and 3, e.g. $X(s=1)=|1-3|=2$. Then the probability mass function $p(x)$ is 1/6 for x=0, 1/3 for x=1, 1/3 for x=2, 1/6 for x=3 and 0 for all other *x*.

EXERCISE 5.1.6. What is the probability mass function for last-name lengths in example 5.1.8?

DEFINITION 5.1.7. Let *X* be a finite random variable with a probability mass function *p(x).* The *mean* (or *average)* of X, $M(X) = \sum_{x \in X} xp(x)$, the probability-weighted sum of the values of *x* in *X.* (We will generally write the mean of *X* as μ_X or simply μ.)

EXAMPLE 5.1.8. Let *S* be the set of students in a class, *{Joel,Yonah,Jessika,Tim}*. Let *X* be the number of letters in each student's last name: *X(Joel)*=10, *X(Yonah)*=8, *X(Jessika)*=4, and *X(Tim)*=6. Let each student be weighted equally, so that $P(\{s_i\})=1/4$ for all *i*. Then the mean *M(X)* is (10+8+4+6)/ 4 = 7.

DEFINITION 5.1.9. Let *X* be a finite random variable with a probability mass function *p(x)*, and let $f: \mathbb{R} \rightarrow \mathbb{R}$. The *average value of the function f* applied to *X*, $M[f(X)] = \sum_{x \in X} f(x)p(x)$.

THEOREM 5.1.10. The mean has the following properties:
(a) If *c* is a constant, then $M(c) = c$.
(b) If *c* is a constant and *f* is a function defined on *X*, then $M[cf(X)] = cM[f(x)]$.
(c) If *c* and *d* are constants and *f* and *g* are functions defined on *X*, then $M[cf(x) + dg(x)] = cM[f(x)] + dM[g(x)]$.
Proof.
(a) $M[f(X)] = M(c) = \sum_{x \in X} cp(x) = c\sum_{x \in X} p(x) = (c)(1) = c$.
(b) $M[cf(X)] = \sum_{x \in X} cf(x)p(x) = c\sum_{x \in X} f(x)p(x) = cM[f(x)]$.
(c) $M[cf(x) + dg(x)] = \sum_{x \in X} [cf(x)+dg(x)]p(x) = c\sum_{x \in X} f(x)p(x) + d\sum_{x \in X} g(x)p(x) = cM[f(x)] + dM[g(X)]$.

LEMMA 5.1.11. *Markov's inequality.* Let *X* be a finite random variable with a probability mass function *p(x)*, and let $f: \mathbb{R} \rightarrow \mathbb{R}^{\geq 0}$ (a nonnegative function on real numbers). Then $\forall k>0$, $P[f(X) \geq k] \leq M[f(X)] / k$.
Proof. By 5.1.7, $M[f(X)] = \sum_{x \in X} f(x)p(x)$. We can decompose this into $\sum_{X: f(x) \geq k} f(x)p(x) + \sum_{X: f(x) < k} f(x)p(x)$. Since *f* is nonnegative, this sum must be greater than or equal to the first term $\sum_{X: f(x) \geq k} f(x)p(x)$. Since the terms of this sum are all such that $f(X) \geq k$, $\sum_{X: f(x) \geq k} f(x)p(x) \geq \sum_{X: f(x) \geq k} kp(x) = kP[f(X) \geq k]$. The result follows if we divide by *k* (which we can do since *k*>0).

DEFINITION 5.1.12. Let *X* be a finite random variable with a probability mass function *p(x),* and mean μ_X. The *variance* of *X* (denoted by σ_X^2 or *Var(X))* is defined by $Var(X) = M\{(x - \mu_X)^2] = \sum_{x \in X} (x - \mu_X)^2 p(x)$.

COROLLARY 5.1.13. If *X* is a random variable, then $Var(X) = M\{(x - \mu_X)^2] = M(X^2) - [M(X)]^2$.
Proof. $Var(X) = M\{(x - \mu_X)^2]$ by 5.1.11, so by expansion, $Var(X) = M[X^2 - 2XM(X) + (M(X))^2] = M(X^2) - 2\{M(X)]^2 + [M(X)]^2 = M(X^2) - \{M(X)]^2$, applying 5.10.

DEFINITION 5.1.14. If X is a random variable and $Var(X)$ is finite, then the *standard deviation* of X, $\sigma_X = \sqrt{Var(X)}$, the square root of the variance.

EXAMPLE 5.1.15. If X is last-name length, and is applied to the students in example 5.1.8, then $Var(X) = [(10-7)^2 + (8-7)^2 + (4-7)^2 + (6-7)^2] /4 = (9 + 1 + 9 + 1)/4 = 5$, and $\sigma_X = \sqrt{5} = 2.236$.

THEOREM 5.1.16. *Chebyshev's inequality*. If X is a finite random variable with finite mean μ_X and a finite standard deviation σ_X, then $\forall r>0,\ P(|X - \mu_X| \geq r\sigma_X) \leq 1/r^2$.
Proof. Plugging $f(X) = (X - \mu_X)^2$ and $k = r^2\sigma_X^2$ into Markov's inequality (5.1.11) yields $P[(X - \mu_X)^2 \geq r^2\sigma_X^2] \leq M[(X - \mu_X)^2] / r^2\sigma_X^2 = 1/ r^2$. The result follows from dividing both sides of the inequality in $P[(X - \mu_X)^2 \geq r^2\sigma_X^2]$ by $r\sigma_X$.

Chebyshev's inequality provides an upper bound of $1/ r^2$ on the chances that a random variable will differ from its mean by more than a given multiple r of its standard deviation. It holds regardless of the probability mass function and, although we have not proven so, it even extends to infinite and continuous random variables, provided only that the random variable has a finite mean and variance.

EXERCISE 5.1.17. According to Chebyshev's inequality, what is the upper bound on the probability that a student from the class in example 5.1.8 will have a last-name length differing from the class mean by 3 letters or more?

5.2 Samples

Concepts like random variable, and probability mass function refer to the reality underlying what can be observed. Actual observations, on the other hand, might give only partial information about these quantities. In such cases, we must reason from what can actually be observed. The theory of sampling gives us ways to infer how close our observations are to the true probability mass function.

DEFINITION 5.2.1. The finite random variables $X_1,X_2,...,X_n$ having a joint probability mass function $q_{X1,X2,...,Xn}(x_1,x_2,...,x_n)$ are a *sample* of size n (and their joint p.m.f. is a *joint sampling distribution*) iff a value v_i is assigned to each random variable X_i in the set. We call $X_1,X_2,...,X_n$ the *draws* or *trials* of the sample.

The stipulation that each random variable in a sample have a value assigned to it is a way of saying that there is an *observation* or *data point* for each trial. This distinguishes a sample from random variables that are not observed or are unobservable.

DEFINITION 5.2.2. A sample $X_1,X_2,...,X_n$ is a *random sample* iff there is some probability mass function $q(x)$ which is the shared sampling distribution of each of the n random variables in the set, i.e. $q_{X1,X2,...,Xn}(x_1,x_2,...,x_n) = q(x_1)q(x_2)...q(x_n)$. In this case, we say that the draws $X_1,X_2,...,X_n$ are *independent and identically distributed,* or *i.i.d.*

The above definition is carefully worded not to imply that we know what the common p.m.f. of the random sample is -- only that we know that there is one.

DEFINITION 5.2.3. A sample $X_1, X_2, \ldots, X_n$ is an *unbiased sample* of a p.m.f. $p_{X1,X2,\ldots,Xn}(x_1,x_2,\ldots,x_n)$ iff the joint sampling distribution $q_{X1,X2,\ldots,Xn}(x_1,x_2,\ldots,x_n) = p_{X1,X2,\ldots,Xn}(x_1,x_2,\ldots,x_n)$.

DEFINITION 5.2.4. A function $s: X_1 \times X_2 \times \ldots \times X_n \rightarrow \mathbb{R}$ is a *statistic* on the sample $X_1, X_2, \ldots, X_n$ iff s contains no unobserved parameters.

DEFINITION 5.2.5. A statistic $s: X_1 \times X_2 \times \ldots \times X_n \rightarrow \mathbb{R}$ on a random sample $X_1, X_2, \ldots, X_n$ from a sampling distribution $q(x)$ is an *unbiased estimator* of a function $g(X)$ on a finite random variable X with a p.m.f. $p(x)$ iff $M[s(X_1,X_2,\ldots,X_n)] = \sum_{<x1,x2,\ldots xn> \in X1 \times X2 \times \ldots \times Xn} s(x_1,x_2,\ldots,x_n) q(x_1,) q(x_2,) \ldots q(x_n) = g(X)$.

The idea of an unbiased estimator is that both the statistic and the sampling distribution of the random sample should be such that the long-run average value of the statistic will tend to equal the function being estimated.

DEFINITION 5.2.6. For a random sample $X_1, X_2, \ldots, X_n$ from a sampling distribution $q(x)$, the *sample mean* $M' = (\sum_{i \in \{1,\ldots,n\}} X_i)/n$.

THEOREM 5.2.7. The sample mean M' of a random sample $X_1, X_2, \ldots, X_n$ from a sampling distribution $q(x)$ is an unbiased estimator for the mean μ_X of a random variable X with p.m.f. $p(x)$ if $q(x) = p(x)$.
Proof. $M(M') = M[(\sum_{i \in \{1,\ldots,n\}} X_i)/n] = (1/n) M(\sum_{i \in \{1,\ldots,n\}} X_i) = (1/n)[\sum_{i \in \{1,\ldots,n\}} M(X_i)] = (1/n)(\sum_{i \in \{1,\ldots,n\}} \mu_X) = (1/n)(n\mu_X) = \mu_X$. Thus, the mean of the sample mean equals the mean of the random variable X, which meets the definition (5.2.5) for an unbiased estimator of the mean of X.

THEOREM 5.2.8. The variance $Var(M')$ of the sample mean of a random sample $X_1, X_2, \ldots, X_n$ from a sampling distribution $q(x)$ is related to the variance $Var(X)$ of a random variable X with p.m.f. $p(x)$ by $Var(M') = Var(X)/n$ if $q(x) = p(x)$.
Proof. $Var(M') = Var[(\sum_{i \in \{1,\ldots,n\}} X_i)/n] = M[(\sum_{i \in \{1,\ldots,n\}} X_i)^2 / n^2] - [M(\sum_{i \in \{1,\ldots,n\}} X_i)/n]^2 = (1/n^2) \sum_{i \in \{1,\ldots,n\}} [M(X_i^2) - (M(X_i))^2] = (n/n^2)\ [M(X^2) - (M(X))^2] = [M(X^2) - (M(X))^2] / n = Var(X)/n$.

The above theorems and definitions all help us to establish the following result, which tells us for any desired probability 1-α and margin of error ϵ, how large the sample size n must be to guarantee these values.

THEOREM 5.2.9. *Weak law of large numbers.* Let M' be the sample mean of a random sample of size n from a sampling distribution $q(x)$. Let X be a random variable with p.m.f. $p(x)$ and mean μ_X, and assume $q(x)=p(x)$. Then $\forall \epsilon > 0\ \forall \alpha$ such that $0 < \alpha < 1$, if $n \geq \sigma_X^2/(\epsilon^2 \alpha)$, then $P[-\epsilon < M' - \mu_X < \epsilon] = P(|M' - \mu_X| < \epsilon) \geq 1 - \alpha$.
Proof. By Chebyshev's inequality (5.1.16), $\forall r > 0$, $P(|X - \mu_X| \geq r\sigma_X) \leq 1/r^2$. Applying binary complementarity, $1 - P(|X - \mu_X| \geq r\sigma_X) = P(|X - \mu_X| < r\sigma_X)$. From algebra, $x \leq y \Rightarrow 1-x \geq 1-y$, so $P(|X - \mu_X| < r\sigma_X) \geq 1 - 1/r^2$. For $X = M'$, $\mu_{M'} = \mu_X$ by theorem 5.2.7, and $Var(M') = \sigma_X^2/n$ by 5.2.8.

Squaring both sides of the resulting inequality inside the probability function yields $P(|M' - \mu_X|^2 < r^2\sigma_X^2/n) \geq 1 - 1/r^2$. We can get an expression for the margin of error ϵ by setting $r^2 = n\epsilon^2/\sigma_X^2$, so that $P[|M' - \mu_X|^2 < n\epsilon^2\sigma_X^2/(n\sigma_X^2)] \geq 1 - \sigma_X^2/(n\epsilon^2)$. This reduces to $P[(M'-\mu_X)^2 < \epsilon^2] = P(|M'-\mu_X| < \epsilon) \geq 1 - \sigma_X^2/(n\epsilon^2) \geq 1 - \alpha$, provided $\alpha \geq \sigma_X^2/(n\epsilon^2)$, which can be rewritten as $n \geq \sigma_X^2/(\epsilon^2\alpha)$.

EXERCISE 5.2.10. Suppose you want to estimate the average grade point average of Stanford students. Specifically, you want to be at least 95% sure that the real average is within 0.1 grade points of your estimate. Give a sample size and provide an argument that a random sample of that size or greater will give you an estimate within your desired margin of error.

5.3 Calibration

DEFINITION 5.3.1. Let the random variables $E_1,E_2,...,E_n$ be a sample of *n estimates* of *target* random variables $X_1,X_2,...,X_n$. Then the *error* e_i associated with each estimate is defined by $e_i = E_i - X_i$. The *bias* associated with the set of estimates is defined by $b = M_E - M_X$, where M_E is the sample mean of the estimates E_i and M_X is the mean of the targets X_i. When e_i=0 for all *i,* we say the estimates are *perfectly calibrated.* When $b = 0$, we say that the estimates are an *unbiased set.*

The above definition gives us a notion of a set of estimates, each of which has an associated error, and the group of which has a bias. Calibration can be defined either in terms of the error of individual estimates or in terms of average error. In both cases, the above definition treats calibration as a feature of an enumerated set of estimates.

EXAMPLE 5.3.2. Gillian wakes up each morning and estimates the temperature in degrees celsius. The data vary each day, so the sample space is defined by the set of trials, which simply refer to each day's guessing event. Gillian's guesses over a 5-day period are 15, 17, 17, 19, and 15. The actual temperatures at the times of these trials are, respectively, 17, 16, 17, 20, and 13. Immediately, we can say that Gillian's estimates are not perfectly calibrated, because they sometimes differ from the actual temperature. But the mean of her errors is (-2+1+0-1+2)/5 = 0, so these estimates are an unbiased set. We cannot say for sure whether Gillian herself is unbiased, however, because this set of trials is not exhaustive. Future trials may reveal a bias either toward too high or too low estimates.

COROLLARY 5.3.3. The bias of a set of estimates is equal to the average error.

EXERCISE 5.3.4. Prove 5.3.3.

5.4 Calculation

DEFINITION 5.4.1. Assume that an agent is given an n-ary arithmetic function f: $D \rightarrow R$, where $D \subseteq \mathbb{R}^n$ is the domain, and $R \subseteq \mathbb{R}$ is the range. If the agent is asked to calculate $f(\boldsymbol{d})$ for $\boldsymbol{d} \in D$, then the agent's *calculation error* is defined by $e_C = E_f(\mathbf{d}) - f(\mathbf{d})$.

EXAMPLE 5.4.2. Sam is asked to compute the square root of 4, and his calculation is $E_{\surd}(4) = 2$. His calculation error is therefore $2 - \sqrt{4} = 0$.

Although one could argue that a perfectly rational agent should be a perfectly calibrated calculator, this assumes unlimited computational resources and no error. Even the most powerful computers do not meet this standard. We might still wonder, however, if people are unbiased calculators, i.e. if the errors they make tend to cancel each other out over time so that they are correct on average. In this case, we can apply definition 5.3.1 to a set of calculations and calculate the agent's *calculation bias* for that set of calculations.

We already saw from experiments comparing confidence to probability theory that degrees of belief do not obey the rules of the probability calculus. We might wonder whether people's estimates of values that can be calculated from statistical theory are biased as well, and if so in what ways.

We first define two types of probability mass functions that will help us calculate percentage outcomes in frequency estimation problems.

DEFINITION 5.4.3. *Bernoulli distribution.* A random variable X has a *bernoulli distribution* iff the probability mass function of X is given by

$$p(x) = a^x(1-a)^{1-x} \text{ for } x = 0 \text{ or } 1, \text{ and } 0 \text{ otherwise,}$$

where the parameter $a \in [0,1]$ represents the probability that $x = 1$.

A bernoulli distribution, named for the early probability theorist Jacob Bernoulli (who proved the first version of the law of large numbers), has a very simple probability mass function in which all of the mass sits on two possibilities. Another way of writing this is to say $P(x=1) = a$, and $P(x=0) = 1-a$. Since x represents values of the random variable (or function) X for different outcomes in a sample space, we can think of X as partitioning the sample space into two sets of outcomes, one of which is often labeled “success” (x=1) and the other of which is often labeled “failure” (x=0) on a so-called “bernoulli trial”.

THEOREM 5.4.4. If X has a bernoulli distribution, then $\mu_X = a$ and $Var(X) = a(1-a)$.
Proof. $M(X)$ $\mu_X = 0(1-a) + 1(a) = a$. $Var(X) = M(X^2) - [M(X)]^2 = 0^2(1-a) + 1^2(a) - a^2 = a(1-a)$.

Theorem 5.4.4 gives us expressions for the mean and variance of a bernoulli-distributed random variable. In these calculations, the range of X is important (0 or 1), since the values appear in the formulae for the mean and variance. The mean is simply the probability of success (a), while the variance is a measure of how well a predicts what will happen. The variance is at its maximum (.25) when a=.5, i.e. when the outcome of the bernoulli trial is maximally uncertain. When certainty is maximized (a=0 or 1), the variance is 0.

We might ask what is the distribution of the number of successes within a given number n of bernoulli trials. This is equivalent to a random sample of size n of bernoulli trials with probability of success on each trial *a.* The answer is called the "binomial distribution".

DEFINITION 5.4.5. *Binomial distribution.* A random variable *X* has a *binomial distribution* iff the probability mass function of *X* is given by

$p(x) = C(n,x)a^x(1\text{-}a)^{n-x}$ for $x = 0,1,...,n,$ and 0 otherwise,

where the parameter $a \in [0,1]$ and $n \in \mathbb{N}$ (the natural numbers), and $C(n,x) = n!/[(n-x)!x!]$.

LEMMA 5.4.6. If $X_1,X_2,...,X_n$ is a set of *n* independent and identically distributed random variables, then $Var(\sum_{i\in\{1,...,n\}} X_i) = \sum_{i\in\{1,...,n\}} Var(X_i)$.

EXERCISE 5.4.7. Prove lemma 5.4.6.

THEOREM 5.4.8. If $X_1,X_2,...,X_n$ is a set of *n* independent and identically distributed bernoulli trials with probability of success on each trial *a*, then the random variable *X* representing $\sum_{i\in\{1,...,n\}} X_i$, the sum of the boolean values observed each of the *n* trials, has a binomial distribution with $M(X) = na$ and $Var(X) = na(1\text{-}a)$.

Proof. In *n* trials, for any number $k \in \{0,1...,n\}$, then by the definition of combinations from the theory of combinatorics, there are $C(n,k)$ exact sequences of $X_1,X_2,...,X_n$ resulting in *k* successes. Since the trials are independent and all drawn from a bernoulli distribution with parameter *a,* each one of these sequences has a probability of $a^x(1\text{-}a)^{n-x}$ of occurring. Thus, the probability that one of them will occur is just the probability of the union of each of these exact sequences, which, since the sequences are disjoint, is just the sum of the probabilities of each sequence. Since the probability for each such sequence is the same, the sum is just the number of such sequences $C(n,k)$ multiplied by the probability $a^x(1\text{-}a)^{n-x}$. Thus the distribution is binomial by 5.4.5. The mean $M(X) = \sum_{i\in\{1,...,n\}} M(X_i) = \sum_{i\in\{1,...,n\}} a = na$. The variance $Var(X) = Var(\sum_{i\in\{1,...,n\}} X_i) = na(1\text{-}a)$.

EXPERIMENT 5.4.9. *Hospital problem.* Tversky and Kahneman (1974) gave undergraduates the following question:

> A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50 percent of all babies are boys. However, the exact percentage varies from day to day. Sometimes it will be higher than 50 percent, sometimes lower.
>
> For a period of 1 year, each hospital recorded the days on which more than 60 percent of the babies born were boys. Which hospital do you think recorded more such days?
>
> The larger hospital
> The smaller hospital
> About the same (that is, within 5 percent of each other)

Out of 95 subjects, 21 said the larger hospital, 21 said the smaller one, and 53 said about the same. The theorems we have proven can be applied to estimating the number of days on which each hospital records more than 60 percent boys among its newborns. The calculation is left as an exercise below, but it shows that the best estimate is much larger in the small hospital than in the large hospital, because, as the authors say, "a large sample is less likely to stray from 50 percent." Tversky and Kahneman give this as an example of insensitivity to sample size in such calculations.

Of course, the actual calculation in 5.4.9 is somewhat complicated and requires quite a bit of statistical theory to do it correctly. That people are not perfectly calibrated on such calculations is thus not surprising. What is noteworthy is that the errors are not random, which would result in most subjects in this experiment answering "the smaller hospital". Instead, *estimates* are biased in the direction of neglecting differences in sample sizes. The discussion of base rate neglect in our notes on confidence is another example in which people's calculations are biased away from true estimates. In both cases, people appear to apply *heuristics* ("rules of thumb) to the calculation, simplifying it based on what appears to be the most directly pertinent evidence. In the hospital problem, subjects may be applying the heuristic of representativeness.

EXERCISE 5.4.10. Apply the statistical theory we have outlined to make an argument that the small hospital in 5.4.9 is likely to have more days in a given year with 60% or greater male births than is the large hospital.

EXERCISE 5.4.11. How can the results of experiment 5.4.9 be explained by the heuristic of representativeness?

EXPERIMENT 5.4.12. *Multiplication problem.* Tversky and Kahneman (1974) asked two groups of high school students to estimate, within 5 seconds, the value of a numerical expression written on a blackboard. One group estimated the value of

$$8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$$

and the other group estimated the value of

$$1 \times 2 \times 3 \times 4 \times 5 \times 6 \times 7 \times 8.$$

Tversky and Kahneman describe the results as an application of *anchoring and adjustment*: "To rapidly answer such questions, people may perform a few steps of computation and estimate the product by extrapolation or adjustment. Because adjustments are typically insufficient, this procedure should lead to underestimation. Furthermore because the result of the first few steps of multiplication (performed from left to right) is higher in the descending sequence than in the ascending sequence, the former expression should be judged larger than the latter. Both predictions were confirmed. The median estimate for the ascending sequence was 512, while the median estimate for the descending sequence was 2,250. The correct answer is 40,320." Thus, subjects show a calculation bias in this experiment of 38,070 for the descending sequence and 39,808 for the ascending sequence, or 38,939 on average across the two conditions. The bias is composed of general underestimation of 38,939 plus a positive or negative bias introduced by the order of presentation of numerals.

In the above experiment, subjects were given the exact function they were being asked to calculate, and the researchers found that quick calculation was systematically biased. The noteworthy result is not that students make errors in fast calculations, since almost everyone would predict that. Rather, the important result is that calculations are biased – the errors are systematic and can be predicted. Another noteworthy feature of experiment 5.4.12 is that it involves a relatively simple calculation requiring no advanced mathematics, showing that errors and biases are not always dependent on a lack of understanding of the mathematics that is required for solving a problem, although of course in the multiplication problem experiment subjects had little time to do the calculation.

EXAMPLE 5.4.13. *Committees problem.* Tversky and Kahneman (1974) had subjects estimate the number of different committees of *k* members that can be formed among a group of 10 people, for different values of *k*. The correct answer in each case is $C(10,k)$, which is maximal for $k=5$, when $C(10,k) = 252$. The authors found that subjects' estimates were biased by a form of availability heuristic they call *imaginability.* "For example," they write, "the median estimate of the number of committees of 2 members was 70, while the estimate for committees of 8 members was 20 (the correct answer is 45 in both cases)." Note that $C(10,k) = C(10,10\text{-}k)$ for all *k* in the range from 0 to 10, because each committee of *k* defines a committee of $10\text{-}k$ "nonmembers", and vice versa. The proposed explanation of subject's estimation bias is that committees of a small number of members (e.g. 2) are more available than are committees of many members (e.g. 8). The authors write: "The simplest scheme for the construction of committees is a partition of the group into disjoint sets. One readily sees that it is easy to construct five disjoint committees of 2 members, while it is impossible to generate even two disjoint committees of 8 members. Consequently, if frequency is assessed by imaginability, or by availability for construction, the small committees will appear more numerous than larger committees, in contrast to the correct bell-shaped function."

5.5 Subjective Measurement

In some cases, estimation can rely entirely on a calculation – all of the information needed to make an estimate using statistical theory or arithmetic is available. In other cases, however, we may be estimating a quantity which is observable but not yet observed. The process of doing this observation we will call "measurement". Measurement can involve procedures applied to the external world, as when we read a thermometer in order to observe the temperature. It can also be applied, for example by social scientists, to observations of human behavior, as when we observe a subject's estimate of a quantity by noting the value they report. We are interested here in measurement as a process that is internal to an agent, i.e. measurement that is based on introspective observations. Confidence reports ("I believe the chances it will rain tomorrow are 30%") are an example of such measurements, as are estimates based on perception, such as the loudness of a noise or how many jelly beans there are in a glass jar. In each case, we can ask how well calibrated people's measurements are, and whether they show systematic biases in one direction or another.

EXPERIMENT 5.5.1. Tversky and Kahneman (1973) showed that availability biases the subjective measurement of frequency. Tversky and Kahneman (1974) write: "When the size of a class is judged by the availability of its instances, a class whose instances are easily retrieved will appear more numerous than a class of equal frequency whose instances are less retrievable. In an elementary demonstration of this effect, subjects heard a list of well-known personalities of both sexes and were subsequently asked to judge whether the list contained more names of men than of women. Different lists were presented to different groups of subjects. In some of the lists, the men were relatively more famous than the women, and in others the women were relatively more famous than the men. In each of the lists, the subjects erroneously judged that the class (sex) that had the more famous personalities was the more numerous."

EXPERIMENT 5.5.2. *Conjunction fallacy for frequency estimation.* Tversky and Kahneman (1973) asked subjects the following: "In four pages of a novel (about 2,000 words), how many words would you expect to find that have the form _ _ _ _ *ing* (seven-letter words that end with 'ing')?" In a second version, subjects were asked to estimate the frequency of words of the form _ _ _ _ _ *n* _. The median estimates were 13.4 for *ing* words and 4.7 for words with *n* in the sixth position. The results are explained as an instance of availability bias. The estimates closely mirror results of another study done in which subjects were asked to list the words they could think of having each form, in 60 seconds, and on average produced lists of 6.4 words for *ing* words and 2.9 words for seven-letter words with *n* in the sixth position. The results violate a form of the conjunction rule for frequencies however, since every word of seven letters that ends in *ing* also has *n* in the sixth position, but there are some words that are in the latter category only (e.g. "latrine").

EXERCISE 5.5.3. Give an original example of two quantities for which you would expect people's estimates to show a reversal of the inequality that actually exists between the two quantities.

EXPERIMENT 5.5.4. *Overconfidence*. Oskamp (1962) showed subjects psychological profiles from the Minnesota Multiphasic Personality Inventory (MMPI), half of which came from patients at a Veterans Administration (VA) hospital who had been admitted for psychiatric reasons, and half of which had been admitted to the VA for purely medical reasons. As Lichteinstein, Fischhoff, and Phillips (1982) write: "The subjects' task was to decide, for each profile, whether the patient's status was psychiatric or medical and to state the probability that their decision was correct." The task can be seen as an instance of subjective measurement, in which the people's stated probabilities can be taken as estimating a quantity, namely the bernoulli probability *a* that their answer is correct. We can assess whether these are biased or not by comparing subjects' average confidence across a set of trials with the percentage of profiles they classify correctly. Oskamp found that subjects' stated confidence was not an unbiased estimate of their accuracy: the average confidence of subjects prior to training for better accuracy was 78%, whereas their accuracy was only 70%. Training for accuracy reduced this bias, but did not eliminate it. Similar results have been found in numerous studies showing that subjects tend to be overconfident, although the effect varies somewhat cross-culturally.

5.6 Comparison

Estimation tasks often appear to us as comparison questions, e.g. "Is the value of X higher or lower than 20?", and "Billy found out he got a 78 on the test – I wonder what I got." A normative principle of comparison is adherence to a principle we will call "anchor independence", which is based on statistical concepts of conditionality.

DEFINITION 5.6.1. Let X and Y be finite random variables with a joint p.m.f. $p(x,y)$. The *conditional probability mass function* $p_{Y|X}(y|x)$ is defined by $p_{Y|X}(y|x) = p(x,y)/p_X(x)$, where $p_X(x)$ is the *marginal probability mass function* for X.

DEFINITION 5.6.2. Two random variables X and Y are *statistically independent* iff $p_{Y|X}(y|x) = p_Y(y)$.

DEFINITION 5.6.3. *Anchor independence.* An estimation function E for a target random variable X is *anchor independent* iff for any random variables A and B which are statistically independent of X,

$E(X|A) = E(X|B)$.

EXPERIMENT 5.6.4. *Insufficient adjustment.* Tversky and Kahneman (1974) write: "In a demonstration of the anchoring effect, subjects were asked to estimate various quantities, stated in percentages (for example, the percentage of African countries in the United Nations). For each quantity, a number between 0 and 100 was determined by spinning a wheel of fortune in the subjects' presence. The subjects were instructed to indicate first whether that number was higher or lower than the value of the quantity by moving upward or downward from the given number. Different groups were given different numbers for each quantity, and these arbitrary numbers had a marked effect on estimates. For example, the median estimates of the percentage of African countries in the United Nations were 25 and 45 for groups that received 10 and 65, respectively, as starting points. Payoffs for accuracy did not reduce the anchoring effect."

EXERCISE 5.6.5. Imagine an experiment in which subjects are asked to estimate the age of a famous movie star who is 65 years old. One group is first asked whether the star's real age is greater or lower than 80, and the other is asked whether it is greater or lower than 50. What pattern of results would you expect in this experiment and why? Is this a violation of anchor independence? Why or why not?

EXPERIMENT 5.6.5. *Compound event probabilities.* Tversky and Kahneman (1974) write: "Studies of choice among gambles and judgments of probability indicate that people tend to overestimate the probability of conjunctive events (Cohen, Chesnick, & Haran, 1972, 24) and to underestimate the probability of disjunctive events. These biases are readily explained as effects of anchoring. The stated probability of the elementary event (success at any stage) provides a natural starting point for the estimation of the probabilities of both conjunctive and disjunctive events. Since adjustment from a starting point is typically insufficient, the final estimates remain too close to the probabilities of the elementary events in both cases."

APPLICATION 5.6.6. Tversky and Kahneman (1974) write: "The general tendency to overestimate the probability of conjunctive events leads to unwarranted optimism in the evaluation of the likelihood that a plan will succeed or that a project will be completed on time. Conversely, disjunctive structures are typically encountered in the evaluation of risks. A complex system, such as a nuclear reactor or the human body, will malfunction if any of its essential components fails. Even when the likelihood of failure in each component is slight, the probability of an overall failure can be high if many of the components are involved. Because of anchoring, people will tend to underestimate the probabilies of failure in complex systems. Thus, the direction of the anchoring bias can sometimes be inferred from the structure of the event."

EXPERIMENT 5.6.7. *Cross-modality anchoring.* Oppenheimer, LeBoeuf, and Brewer (2007) recently extended the scope of anchoring across modalities. They write: "An initial study showed that participants drawing long "anchor" lines made higher numerical estimates of target lengths than did those drawing shorter lines. We then replicated this finding, showing that a similar pattern was obtained even when the target estimates were not in the dimension of length. A third study showed that an anchor's length relative to its context, and not its absolute length, is the key to predicting the anchor's impact on judgments. A final study demonstrated that magnitude priming (priming a sense of largeness or smallness) is a plausible mechanism underlying the reported effects. We conclude that the boundary conditions of anchoring effects may be much looser than previously thought, with anchors operating across modalities and dimensions to bias judgment."

EXERCISE 5.6.8. Having read about the pervasiveness of anchoring biases as explored by

Oppenheimer et al. in 5.6.7, what corrective procedure might an individual apply to counteract this effect in everyday estimation tasks. Give an example.