# Module 10: Sleuthing with Probability and Statistics

My mother made me a scientist without ever intending to. Every Jewish mother in Brooklyn would ask her child after school: 'So? Did you learn anything today?' But not my mother. She always asked me a different question. 'Izzy,' she would say, 'did you ask a good question today?'

— Isidor Tabi (Nobel Prize in Physics, 1944; quotation given by John Barell in *Developing More Curious Minds*, 2003)

**Abstract**: Statistical sleuthing is concerned with the use of various probabilistic and statistical tools and methods to help explain or "tell the story" about some given situation. In this type of statistical detective work, a variety of probability distributions can prove useful as models for a given underlying process. These distributions include the Bernoulli, binomial, normal, Poisson (especially for spatial randomness and the assessment of "Poisson clumping"). Other elucidating probabilistic topics introduced include Benford's Law, the "birthday probability model," survival analysis and Kaplan-Meier curves, the Monty Hall problem, and what is called the "secretary problem" (or more pretentiously, the "theory of optimal stopping"). An amusing instance of the latter secretary problem is given as a *Car Talk* Puzzler called the "Three Slips of Paper"; a full listing of the script from the NPR show is included that aired on February 12, 2011.

## Contents

## 1 Sleuthing Interests and Basic Tools

Modern statistics is often divided into two parts: exploratory and confirmatory. Confirmatory methods were developed over the first half of the 20th century, principally by Karl Pearson and Ronald Fisher. This was, and remains, a remarkable intellectual accomplishment. The goal of confirmatory methods is largely judicial: they are used to weigh evidence and make decisions. The aim of exploratory methods is different. They are useful in what could be seen as detective work; data are gathered and clues are sought to enable us to learn what might have happened. Exploratory analysis generates the hypotheses that are tested by the confirmatory methods. Surprisingly, the codification, and indeed the naming of exploratory data analysis, came after the principal work on the development of confirmatory methods was complete. John Tukey's (1977) influential book changed everything. He taught us that we should understand what might be true before we learn how well we have measured it.

Some of the more enjoyable intellectual activities statisticians engage in might be called *statistical sleuthing*—the use of various statistical techniques and methods to help explain or "tell the story" about some given situation. We first give a flavor of several areas where such sleuthing has been of explanatory assistance:

(a) The irregularities encountered in Florida during the 2000 Presidential election and why; see, for example, Alan Agresti and Brett

Presnell, "Misvotes, Undervotes, and Overvotes: The 2000 Presidential Election in Florida" (*Statistical Science*, *17*, 2002, 436–440).

(b) The attribution of authorship for various primary sources; for example, we have the seminal work by Mosteller and Wallace (1964) on the disputed authorship of some of the Federalist Papers.

(c) Searching for causal factors and situations that might influence disease onset; for example, "Statistical Sleuthing During Epidemics: Maternal Influenza and Schizophrenia" (Nicholas J. Horton & Emily C. Shapiro, *Chance*, *18*, 2005, 11–18);

(d) Evidence of cheating and corruption, such as the Justin Wolfers (2006) article on point shaving in NCAA basketball as it pertains to the use of Las Vegas point spreads in betting (but, also see the more recent article by Bernhardt and Heston [2010] disputing Wolfers' conclusions);

(e) The observations of Quetelet's from the middle 1800s that based on the very close normal distribution approximations for human characteristics, there were systematic understatements of height (to below 5 feet, 2 inches) for French conscripts wishing to avoid the minimum height requirement needed to be drafted (Stigler, 1986, pp. 215–216);

(f) Defending someone against an accusation of cheating on a high-stakes exam when the "cheating" was identified by a "cold-hit" process of culling for coincidences, and with subsequent evidence provided by a selective search (that is, a confirmation bias). A defense that a false positive has probably occurred requires a little knowledge of Bayes' theorem and the positive predictive value.

(g) Demonstrating the reasonableness of results that seem "too good to be true" without needing an explanation of fraud or misconduct. An exemplar of this kind of argumentation is in the article, "A Little Ignorance: How Statistics Rescued a Damsel in Distress" (Peter Baldwin and Howard Wainer, *Chance*, 2009, *22*, 51–55).

A variety of sleuthing approaches are available to help explain what might be occurring over a variety of different contexts. Some of those discussed in this monograph include Simpson's Paradox, Bayes' rule and base rates, regression toward the mean, the effects of culling on the identification of false positives and the subsequent inability to cross-validate, the operation of randomness and the difficulty in "faking" such a process, and confusions caused by misinterpreting conditional probabilities. We mention a few other tools below that may provide some additional assistance: the use of various discrete probability distributions, such as the binomial, Poisson, or those for runs, in constructing convincing explanations for some phenomena; the digit regularities suggested by what is named Benford's law (Benford, 1938); a reconception of some odd probability problems by considering pairs (what might be labeled as the "the birthday probability model"); and the use of the statistical techniques in survival analysis to model time-to-event processes.[1]

[1]There are several quantitative phenomena useful in sleuthing but which are less than transparent to understand. One particularly bedeviling result is called the Inspection Paradox. Suppose a light bulb now burning above your desk (with an average rated life of, say, 2000 hours), has been in operation for a year. It now has an expected life longer than 2000 hours because it has already been on for a while, and therefore cannot burn out at any earlier time than right now. The same is true for life spans in general. Because we have not, as they say, "crapped out" as yet, and we cannot die at any earlier time than right now, our lifespans have an expectancy longer than what they were when we were born. This is good news brought to you by Probability and Statistics!

The simplest probability distribution has only two event classes (for example, success/fail, live/die, head/tail, 1/0). A process that follows such a distribution is called Bernoulli; typically, our concern is with repeated and independent Bernoulli trials. Using an interpretation of the two event classes of heads ($H$) and tails ($T$), assume $P(H) = p$ and $P(T) = 1 - p$, with $p$ being invariant over repeated trials (that is, the process is stationary). The probability of any sequence of size $n$ that contains $k$ heads and $n-k$ tails is $p^k(1-p)^{n-k}$. Commonly, our interest is in the distribution of the number of heads (say, $X$) seen in the $n$ independent trials. This random variable follows the binomial distribution:

$$P(X = r) = \binom{n}{r} p^r (1-p)^{n-r} \ ,$$

where $0 \leq r \leq n$, and $\binom{n}{r}$ is the binomial coefficient:

$$\binom{n}{r} = \frac{n!}{(n-r)!r!} \ ,$$

using the standard factorial notation.

Both the binomial distribution and the underlying repeated Bernoulli process offer useful background models against which to compare observed data, and to evaluate whether a stationary Bernoulli process could have been responsible for its generation. For example, suppose a Bernoulli process produces a sequence of size $n$ with $r$ heads and $n - r$ tails. All arrangements of the $r$ $H$s and $n - r$ $T$s should be equally likely (cutting, say, various sequences of size $n$ all having $r$ $H$s and $n - r$ $T$s from a much longer process); if not, possibly the process is not stationary or the assumption of independence is inappropriate. A similar use of the binomial would first estimate $p$ from

the long sequence, and then use this value to find the expected number of heads in sequences of a smaller size $n$; a long sequence could be partitioned into segments of this size and the observed number of heads compared to what would be expected. Again, a lack of fit between the observed and expected might suggest lack of stationarity or trial dependence (a more formal assessment of fit could be based on the usual chi-square goodness-of-fit test).

A number of different discrete distributions prove useful in statistical sleuthing. We mention two others here, the Poisson and a distribution for the number of runs in a sequence. A discrete random variable, $X$, that can take on values 0, 1, 2, 3, ... , follows a Poisson distribution if

$$P(X = r) = \frac{e^{-\lambda}\lambda^r}{r!} \ ,$$

where $\lambda$ is an intensity parameter, and $r$ can take on any integer value from 0 onward. Although a Poisson distribution is usually considered a good way to model the number of occurrences for rare events, it also provides a model for spatial randomness as the example adapted from Feller (1968, Vol. 1, pp. 160–161) illustrates:

*Flying-bomb hits on London.* As an example of a spatial distribution of random points, consider the statistics of flying-bomb hits in the south of London during World War II. The entire area is divided into 576 small areas of 1/4 square kilometers each. Table 1 records the number of areas with exactly $k$ hits. The total number of hits is 537, so the average is .93 (giving an estimate for the intensity parameter, $\lambda$). The fit of the Poisson distribution is surprisingly good. As judged by the $\chi^2$-criterion, under ideal conditions, some 88

Table 1: Flying-bomb hits on London.

| Number of hits | 0 | 1 | 2 | 3 | 4 | 5 or more |
|---|---|---|---|---|---|---|
| Number of areas | 229 | 211 | 93 | 35 | 7 | 1 |
| Expected number | 226.74 | 211.39 | 98.54 | 30.62 | 7.14 | 1.57 |

per cent of comparable observations should show a worse agreement. It is interesting to note that most people believed in a tendency of the points of impact to cluster. If this were true, there would be a higher frequency of areas with either many hits or no hits and a deficiency in the intermediate classes. Table 1 indicates a randomness and homogeneity of the area, and therefore, we have an instructive illustration of the established fact that to the untrained eye, randomness appears as regularity or tendency to cluster (the appearance of this regularity in such a random process is sometimes referred to as "Poisson clumping").

To develop a distribution for the number of runs in a sequence, suppose we begin with two different kinds of objects (say, white (W) and black (B) balls) arranged randomly in a line. We count the number of runs, $R$, defined by consecutive sequences of all Ws or all Bs (including sequences of size 1). If there are $n_1$ W balls and $n_2$ B balls, the distribution for $R$ under randomness can be constructed. We note the expectation and variance of $R$, and the normal approximation:

$$E(R) = \frac{2n_1n_2}{n_1+n_2} + 1 \ ;$$

$$V(R) = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)} \ ;$$

and

$$\frac{R - E(R)}{\sqrt{V(R)}}$$

is approximately (standard) normal with mean zero and variance one. Based on this latter distributional approximation, an assessment can be made as to the randomness of the process that produced the sequence, and whether there are too many or too few runs for the continued credibility that the process is random. Run statistics have proved especially important in monitoring quality control in manufacturing, but these same ideas could be useful in a variety of statistical sleuthing tasks.

Besides the use of formal probability distributions, there are other related ideas that might be of value in the detection of fraud or other anomalies. One such notion, called Benford's law, has captured some popular attention; for example, see the article by Malcolm W. Browne, "Following Benford's Law, or Looking Out for No. 1" (*New York Times*, August 4, 1998). Benford's law gives a "probability distribution" for the first digits (1 to 9) found for many (naturally) occurring sets of numbers. If the digits in some collection (such as tax returns, campaign finances, (Iranian) election results, or company audits) do not follow this distribution, there is a *prima facie* indication of fraud.[2]

[2]The International Society for Clinical Biostatistics through its Subcommittee on Fraud published a position paper entitled "The Role of Biostatistics in the Prevention, Detection, and Treatment of Fraud in Clinical Trials" (Buyse et al., *Statistics in Medicine*, 1999, *18*, 3435–3451). Its purpose was to point out some of the ethical responsibilities the statistical community has in helping monitor clinical studies with public or personal health implications. The abstract is given below, but we still refer the reader directly to the article for more detail on a range of available statistical sleuthing tools (including Benford's law) that can assist in uncovering data fabrication and falsification:

Benford's law gives a discrete probability distribution over the digits 1 to 9 according to:

$$P(X = r) = \log_{10}(1 + \frac{1}{r}) \ ,$$

for $1 \leq r \leq 9$. Numerically, we have the following:

---

Recent cases of fraud in clinical trials have attracted considerable media attention, but relatively little reaction from the biostatistical community. In this paper we argue that biostatisticians should be involved in preventing fraud (as well as unintentional errors), detecting it, and quantifying its impact on the outcome of clinical trials. We use the term "fraud" specifically to refer to data fabrication (making up data values) and falsification (changing data values). Reported cases of such fraud involve cheating on inclusion criteria so that ineligible patients can enter the trial, and fabricating data so that no requested data are missing. Such types of fraud are partially preventable through a simplification of the eligibility criteria and through a reduction in the amount of data requested. These two measures are feasible and desirable in a surprisingly large number of clinical trials, and neither of them in any way jeopardizes the validity of the trial results. With regards to detection of fraud, a brute force approach has traditionally been used, whereby the participating centres undergo extensive monitoring involving up to 100 per cent verification of their case records. The cost-effectiveness of this approach seems highly debatable, since one could implement quality control through random sampling schemes, as is done in fields other than clinical medicine. Moreover, there are statistical techniques available (but insufficiently used) to detect "strange" patterns in the data including, but no limited to, techniques for studying outliers, inliers, overdispersion, underdispersion and correlations or lack thereof. These techniques all rest upon the premise that it is quite difficult to invent plausible data, particularly highly dimensional multivariate data. The multicentric nature of clinical trials also offers an opportunity to check the plausibility of the data submitted by one centre by comparing them with the data from all other centres. Finally, with fraud detected, it is essential to quantify its likely impact upon the outcome of the clinical trial. Many instances of fraud in clinical trials, although morally reprehensible, have a negligible impact on the trial's scientific conclusions. (pp. 3435–3436)

| $r$ | Probability | $r$ | Probability |
|---|---|---|---|
| 1 | .301 | 6 | .067 |
| 2 | .176 | 7 | .058 |
| 3 | .125 | 8 | .051 |
| 4 | .097 | 9 | .046 |
| 5 | .079 | | |

Although there may be many examples of using Benford's law for detecting various monetary irregularities, one of the most recent applications is to election fraud, such as in the 2009 Iranian Presidential decision. A recent popular account of this type of sleuthing is Carl Bialik's article, "Rise and Flaw of Internet Election-Fraud Hunters" (*Wall Street Journal*, July 1, 2009). It is always prudent to remember, however, that heuristics, such as Benford's law and other digit regularities, might point to a potentially anomalous situation that should be studied further, but violations of these presumed regularities should never be considered definitive "proof."

Another helpful explanatory probability result is commonly referred to as the "birthday problem": what is the probability that in a room of $n$ people, at least one pair of individuals will have the same birthday. As an approximation, we have $1 - e^{-n^2/(2\times 365)}$; for example, when $k = 23$, the probability is .507; when $k = 30$, it is .706. These surprisingly large probability values result from the need to consider matchings over all pairs of individuals in the room; that is, there are $\binom{n}{2}$ chances to consider for a matching, and these inflate the probability beyond what we might intuitively expect. We give an example from Leonard Mlodinow's book, *The Drunkard's Walk* (2009):

Another lottery mystery that raised many eyebrows occurred in Germany on June 21, 1995. The freak event happened in a lottery named Lotto 6/49, which means that the winning six numbers are drawn from the numbers 1 to 49. On the day in question the winning numbers were 15-25-27-30-42-48. The very same sequence had been drawn previously, on December 20, 1986. It was the first time in 3,016 drawings that a winning sequence had been repeated. What were the chances of that? Not as bad as you'd think. When you do the math, the chance of a repeat at some point over the years comes out to around 28 per cent. (p. 65)

## 2 Survival Analysis

The area of statistics that models the time to the occurrence of an event, such as death or failure, is called *survival analysis*. Some of the questions survival analysis is concerned with include: what is the proportion of a population that will survive beyond a particular time; among the survivors, at what (hazard) rate will they die (or fail); how do the circumstances and characteristics of the population change the odds of survival; can multiple causes of death (or failure) be taken into account. The primary object of interest is the survival function, specifying the probability that time of death (the term to be used generically from now on), is later than some specified time. Formally, we define the survival function as: $S(t) = P(T > t)$, where $t$ is some time, and $T$ is a random variable denoting the time of death. The function must be nonincreasing, so: $S(u) \leq S(v)$, when $v \leq u$. This reflects the idea that survival to some later time requires survival at all earlier times as well.

The most common way to estimate $S(t)$ is through the now ubiquitous Kaplan–Meier estimator, which allows a certain (important)

type of right-censoring of the data. This censoring is where the corresponding objects have either been lost to observation or their lifetimes are still ongoing when the data were analyzed. Explicitly, let the observed times of death for the $N$ members under study be $t_1 \leq t_2 \leq \cdots \leq t_N$. Corresponding to each $t_i$ is the number of members, $n_i$, "at risk" just prior to $t_i$; $d_i$ is the number of deaths at time $t_i$. The Kaplan–Meier nonparametric maximum likelihood estimator, $\widehat{S(t)}$, is a product:

$$\widehat{S(t)} = \prod_{t_i \leq t} (1 - \frac{d_i}{n_i}) \ .$$

When there is no right-censoring, $n_i$ is just the number of survivors prior to time $t_i$; otherwise, $n_i$ is the number of survivors minus the number of censored cases (by that time $t_i$). Only those surviving cases are still being observed (that is, not yet censored), and thus at risk of death. The function $\widehat{S(t)}$ is a nonincreasing step function, with steps at $t_i, 1 \leq i \leq N$; it is also usual to indicate the censored observations with tick marks on the graph of $\widehat{S(t)}$.

The original Kaplan and Meir article that appeared in 1958 (Kaplan, E. L., & Meier, P., "Nonparametric Estimation From Incomplete Observations," *Journal of the American Statistical Association*, *53*, 457–481), is one of the most heavily cited papers in all of the sciences. It was featured as a "Citation Classic" in the June 13, 1983 issue of *Current Contents: Life Sciences*. As part of this recognition, Edward Kaplan wrote a short retrospective that we excerpt below:

This paper began in 1952 when Paul Meier at Johns Hopkins University (now at the University of Chicago) encountered Greenwood's paper on the duration

of cancer. A year later at Bell Telephone Laboratories I became interested in the lifetimes of vacuum tubes in the repeaters in telephone cables buried in the ocean. When I showed my manuscript to John W. Tukey, he informed me of Meier's work, which already was circulating among some of our colleagues. Both manuscripts were submitted to the *Journal of the American Statistical Association*, which recommended a joint paper. Much correspondence over four years was required to reconcile our differing approaches, and we were concerned that meanwhile someone else might publish the idea.

The nonparametric estimate specifies a discrete distribution, with all the probability concentrated at a finite number of points, or else (for a large sample) an actuarial approximation thereto, giving the probability in each of a number of successive intervals. This paper considers how such estimates are affected when some of the lifetimes are unavailable (censored) because the corresponding items have been lost to observation, or their lifetimes are still in progress when the data are analyzed. Such items cannot simply be ignored because they may tend to be longer-lived than the average. (p. 14)

To indicate the importance of the Kaplan–Meier estimator in sleuthing within the medical/pharmaceutical areas and elsewhere, we give the two opening paragraphs of Malcolm Gladwell's *New Yorker* article (May 17, 2010), entitled "The Treatment: Why Is It So Difficult to Develop Drugs for Cancer?":

In the world of cancer research, there is something called a Kaplan–Meier curve, which tracks the health of patients in the trial of an experimental drug. In its simplest version, it consists of two lines. The first follows the patients in the "control arm," the second the patients in the "treatment arm." In most cases, those two lines are virtually identical. That is the sad fact of cancer research: nine times out of ten, there is no difference in survival between those who were given the new drug and those who were not. But every now and again—after millions of dollars have been spent, and tens of thousands of pages of data collected, and patients followed, and toxicological issues examined, and safety issues resolved, and manufacturing processes fine-tuned—the patients in the treatment arm will live longer than the patients in

the control arm, and the two lines on the Kaplan–Meier will start to diverge.

Seven years ago, for example, a team from Genentech presented the results of a colorectal-cancer drug trial at the annual meeting of the American Society of Clinical Oncology—a conference attended by virtually every major cancer researcher in the world. The lead Genentech researcher took the audience through one slide after another—click, click, click—laying out the design and scope of the study, until he came to the crucial moment: the Kaplan–Meier. At that point, what he said became irrelevant. The members of the audience saw daylight between the two lines, for a patient population in which that almost never happened, and they leaped to their feet and gave him an ovation. Every drug researcher in the world dreams of standing in front of thousands of people at ASCO and clicking on a Kaplan–Meier like that. "It is why we are in this business," Safi Bahcall says. Once he thought that this dream would come true for him. It was in the late summer of 2006, and is among the greatest moments of his life. (p. 69)

A great deal of additional statistical material involving survival functions can be helpful in our sleuthing endeavors. Survival functions may be compared over samples (for example, the log-rank test), and generalized to accommodate different forms of censoring; the Kaplan–Meier estimator has a closed-form variance estimator (for example, the Greenwood formula); various survival models can incorporate a mechanism for including covariates (for example, the proportional hazard models introduced by Sir David Cox; see Cox and Oakes (1984): *Analysis of Survival Data*). All of the usual commercial software (SAS, SPSS, SYSTAT) include modules for survival analysis. And, as might be expected, a plethora of cutting edge routines are in R, as well as in the Statistics Toolbox in MATLAB.

## 3 Sleuthing in the Media

One of the trite quantitative sayings that may at times drive individuals "up a wall" is when someone says condescendingly, "just do the math." This saying can become a little less obnoxious when reinterpreted to mean working through a situation formally rather than just giving a quick answer based on first impressions. We give two examples of this that may help: one is called the Monty Hall problem; the second is termed the Secretary problem.

In 1990, Craig Whitaker wrote a letter to Marilyn vos Savant's column in *Parade* magazine stating what has been named the Monty Hall problem:[3]

> Suppose you're on a game show, and you're given the choice of three doors. Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, 'Do you want to pick door No. 2?' Is it to your advantage to switch your choice? (p. 16)

The answer almost universally given to this problem is that switching does not matter, presumably with the reasoning that there is no way for the player to know which of the two unopened doors is the winner, and each of these must then have an equal probability of being the winner. By writing down three doors hiding one car and two goats, and working through the options in a short simulation, it becomes clear quickly that the opening of a goat door changes the information one has about the original situation, and that always changing doors

[3]As an interesting historical note, the "Monty Hall" problem has been a fixture of probability theory from at least the 1890s; it was named the problem of the "three caskets" by Henri Poincaré, and is more generally known as (Joseph) Bertrand's Box Paradox

doubles the probability of winning from 1/3 to 2/3.[4]

An enjoyable diversion on Saturday mornings is the NPR radio show, *Car Talk*, with Click and Clack, The Tappet Brothers (aka Ray and Tom Magliozzi). A regular feature of the show, besides giving advice on cars, is The Puzzler; a recent example on Febuary 12, 2011 gives us another chance to "do the math." It is called the Three Slips of Paper, and it is stated as follows on the Car Talk website:

Three different numbers are chosen at random, and one is written on each of three slips of paper. The slips are then placed face down on the table. You have to choose the slip with the largest number. How can you improve your odds?

The answer given on the show:

Ray: This is from Norm Leyden from Franktown, Colorado. The date on it is 1974—I'm a little behind.

[4]To show the reach of the Monty Hall problem, we give the abstract for an article by Herbranson and Schroeder (2010): "Are Birds Smarter Than Mathematicians? Pigeons (*Columba livia*) Perform Optimally on a Version of the Monty Hall Dilemma" (*Journal of Comparative Psychology*, *124*, 1–13):

The "Monty Hall Dilemma" (MHD) is a well known probability puzzle in which a player tries to guess which of three doors conceals a desirable prize. After an initial choice is made, one of the remaining doors is opened, revealing no prize. The player is then given the option of staying with their initial guess or switching to the other unopened door. Most people opt to stay with their initial guess, despite the fact that switching doubles the probability of winning. A series of experiments investigated whether pigeons (*Columba livia*), like most humans, would fail to maximize their expected winnings in a version of the MHD. Birds completed multiple trials of a standard MHD, with the three response keys in an operant chamber serving as the three doors and access to mixed grain as the prize. Across experiments, the probability of gaining reinforcement for switching and staying was manipulated, and birds adjusted their probability of switching and staying to approximate the optimal strategy. Replication of the procedure with human participants showed that humans failed to adopt optimal strategies, even with extensive training. (p. 1)

Three different numbers are chosen at random, and one is written on each of three slips of paper. The slips are then placed face down on the table. The objective is to choose the slip upon which is written the largest number.

Here are the rules: You can turn over any slip of paper and look at the amount written on it. If for any reason you think this is the largest, you're done; you keep it. Otherwise you discard it and turn over a second slip. Again, if you think this is the one with the biggest number, you keep that one and the game is over. If you don't, you discard that one too.

Tommy: And you're stuck with the third. I get it.

Ray: The chance of getting the highest number is one in three. Or is it? Is there a strategy by which you can improve the odds?

Ray: Well, it turns out there is a way to improve the odds—and leave it to our pal Vinnie to figure out how to do it. Vinnie's strategy changes the odds to one in two. Here's how he does it: First, he picks one of the three slips of paper at random and looks at the number. No matter what the number is, he throws the slip of paper away. But he remembers that number. If the second slip he chooses has a higher number than the first, he sticks with that one. If the number on the second slip is lower than the first number, he goes on to the third slip.

Here's an example. Let's say for the sake of simplicity that the three slips are numbered 1000, 500, and 10.

Let's say Vinnie picks the slip with the 1000. We know he can't possibly win because, according to his rules, he's going to throw that slip out. No matter what he does he loses, whether he picks 500 next or 10. So, Vinnie loses—twice.

Now, let's look at what happens if Vinnie starts with the slip with the 500 on it. If he picks the 10 next, according to his rules, he throws that slip away and goes to the 1000.

Tommy: Whopee! He wins.

Ray: Right. And if Vinnie picks the 1000 next, he wins again!

Finally, if he picks up the slip with the 10 on it first, he'll do, what?

Tommy: Throw it out. Those are his rules.

Ray: Right. And if he should be unfortunate enough to pick up the one that says 500 next, he's going to keep it and he's going to lose. However, if his second choice is not the 500 one but the 1000 one, he's gonna keep that slip—and he'll win.

If you look at all six scenarios, Tommy will win one in three times, while Vinnie will win three times out of six.

Tommy: That's almost half!

Ray: In some countries.

One particularly rich area in probability theory that extends the type of *Car Talk* example just given is in the applied probability topic known as optimal stopping, or more colloquially, "the secretary problem." We paraphrase the simplest form of this problem from Thomas Ferguson's review paper in *Statistical Science* (1989), "Who Solved the Secretary Problem?": There is one secretarial position to be filled from among $n$ applicants who are interviewed sequentially and in a random order. All applicants can be ranked from best to worse, with the choice of accepting an applicant based only on the relative ranks of those interviewed thus far. Once an applicant has been rejected, that decision is irreversible. Assuming the goal is to maximize the probability of selecting the absolute best applicant,

it can be shown that the selection rules can be restricted to a class of strategies defined as follows: for some integer $r \geq 1$, reject the first $r-1$ applicants and select the next who is best in the relative ranking of the applicants interviewed thus far. The probability of selecting the best applicant is $1/n$ for $r = 1$; for $r > 1$, it is

$$(\frac{r-1}{n}) \sum_{j=r}^{n} \frac{1}{j-1} \ .$$

For example, when there are 5 $(= n)$ applicants, the probabilities of choosing the best for values of $r$ from 1 to 5 are given in the following table:

| $r$ | Probability |
|---|---|
| 1 | $1/5 = .20$ |
| 2 | $5/12 \approx .42$ |
| 3 | $13/30 \approx .43$ |
| 4 | $7/20 = .35$ |
| 5 | $1/5 = .20$ |

Thus, because an $r$ value of 3 leads to the largest probability of about .43, it is best to interview and reject the first two applicants and then pick the next relatively best one. For large $n$, it is (approximately) optimal to wait until about 37% ($\approx 1/e$) of the applicants have been interviewed and then select the next relatively best one. This also gives the probability of selecting the best applicant as .37 (again, $\approx 1/e$).

In the *Car Talk* Puzzler discussed above, $n = 3$ and Vinnie uses the rule of rejecting the first "interviewee" but then selects the next

that is relatively better. The probability of choosing the best therefore increases from 1/3 to 1/2.

Any beginning statistics class should always include a number of formal tools to help work through puzzling situations. Several of these are mentioned elsewhere in this monograph: Bayes' theorem and implications for screening using sensitivities, specificities, and prior probabilities; conditional probabilities more generally and how probabilistic reasoning might work for facilitative and inhibitive events; sample sizes and variability in, say, a sample mean, and how a confidence interval might be constructed that could be made as accurate as necessary by just increasing the sample size, and without any need to consider the size of the original population of interest; how statistical independence operates or doesn't; the pervasiveness of natural variability and the use of simple probability models (such as the binomial) to generate stochastic processes.

## References


[1] Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society*, *78*, 551–572.

[2] Bernhardt, D., & Heston, S. (2010). Point shaving in college basketball: A cautionary tale for forensic economics. *Economic Inquiry*, *48*, 14–25.

[3] Feller, W. (1968). *An introduction to probability theory and its applications* (3rd ed., Vol. 1). New York: Wiley.

[4] Mosteller, F., & Wallace, D. L. (1964). *Inference and disputed authorship: The Federalist*. Reading, MA: Addison-Wesley.

[5] Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: Belknap Press / Harvard University Press.

[6] Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

[7] Wolfers, J. (2006). Point shaving: Corruption in NCAA basketball. *American Economic Review*, *96*, 279–283.