Elem. Math. 54 (1999) 80 – 85
0013-6018/99/020080-6 $ 1.50+0.20/0

Elemente der Mathematik

Statistical Independence and Model Choice: An Example

Osvaldo Marrero

Osvaldo Marrero studied mathematics at the University of Miami, and biometry and statistics at Yale University. His work experience includes positions in academia and in industry. He is currently a Professor at Villanova University. His main mathematical interests are in combinatorics and statistics. Outside mathematics he is particularly interested in languages: he is fluent in English, French, and Spanish, and is trying to increase his knowledge of Dutch and German.

1 Introduction

Suppose a friend says: "I've been playing the weekly lottery for 15 years, and I have never won; therefore, the next time I play, my chance of winning is going to be better". In general, statisticians will reply with something such as: "No, your chance remains the same; you are just as likely to win the next time you play as you were the first time you played". Perhaps surprisingly, one can obtain different values for the probability of winning at the next play, according to how one chooses to analyze the game.

The purpose of this paper is to present an analysis of such lottery situations by using two probability models. When comparing the results, one sees how different models lead the statistician to think about different aspects of the situation under study, sometimes with unexpected results. It becomes clear that the model determines the framework within which the situation is analyzed. Intended to be accessible to students in undergraduate mathematical-statistics courses, this paper contains more details than one normally finds in the research literature. Hopefully, this material will be useful in such courses.

In general it seems that beginning students get the impression that a probability model is fixed, and is not a matter of choice; it is as if the model came with the data, and

Bei der Beschreibung der Realität durch abstrakte Modelle, wie sie in der Wissenschaft seit Galilei verwendet wird, kommt der Modell*wahl* eine grosse Bedeutung zu. Denn es ist vor allem diese Wahl, welche die Qualität der Voraussagen bestimmt, die aus dem Modell gewonnen werden. Durch lange Gewöhnung übersieht man in gewissen Situationen gern, dass das Modell nicht durch die Realität vorgegeben ist, sondern dass in jedem konkreten Fall eine echte Wahl zu treffen ist. – Ein schönes derartiges Beispiel aus dem Bereich der Wahrscheinlichkeitsrechnung führt uns Osvaldo Marrero im vorliegenden Beitrag vor. *ust*

there is nothing one can do about it. This may be due to the fact that the usual way to teach undergraduate mathematical statistics is to present just one model for a given set of circumstances. Thus students get used to looking for *the* model that is appropriate for a given situation. However, one can sometimes learn more by using different models for the same situation. Students should have the opportunity to see early on how different models can bring out different aspects of a particular experiment.

Any standard textbook for an undergraduate mathematical-statistics course can serve as a reference for this paper; see, for example, [1].

2 The Experimental Situation

The concern is a weekly lottery with a binary outcome: win or lose. Each week the same m numbers $\{1, \ldots, m\}$ are available to players for choosing. At the end of the choosing period, one number is selected and declared the winning number. The winning number is drawn "at random", meaning that every effort is made to insure that each of the m numbers has the same chance of being selected. To keep the discussion as simple as possible, one assumes that, when playing, a person chooses just one number from $\{1, \ldots, m\}$. The random variable of interest is X_n, the number of losses after n plays.

3 A Frequentist Model

It would appear that the result from one drawing would not affect the outcome from another drawing. In this case it is reasonable to assume that the n plays are independent. Therefore, one has a binomial experiment with the probability of losing at a play equal to $p := 1 - 1/m$. The probability mass function for X_n is

$$\operatorname{pr}(X_n = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad \text{for} \quad k := 0, \ldots, n.$$

Suppose it is now known that a person has been playing the lottery for a long time, say n plays, and has never won. The concern is the outcome at the next play. Under this model, the probability of losing is the known constant $p := 1 - 1/m$, and the results in different drawings are independent. Therefore, the probability of losing at the next play given that one has a string of consecutive losses continues to equal p; that is,

$$\operatorname{pr}(X_{n+1} = n+1 \mid X_n = n) = \operatorname{pr}(X_{n+1} = n+1) = p.$$

Thus, again one uses a binomial probability model, the only difference being that now the number of trials is $n+1$.

4 A Bayesian Model

Suppose one doubts that p is really constant and equal to $1 - 1/m$. Perhaps the drawing mechanism does not work as intended, and then the winning number is not really drawn "at random". If one actually feels this way, then there is no information available about the probability of losing, which is therefore a random variable P in the interval $(0, 1)$.

With no information available, it makes sense to assume that P is a continuous random variable uniformly distributed on $(0,1)$, so that P has probability density function $f_P(p) = 1$ for $p \in (0,1)$. This is a *noninformative prior* distribution. In this case the probability mass function of X_n is conditional on the value of P, and it is given by

$$f_{X_n|P}(X_n = k \,|\, p) = \binom{n}{k} p^k (1-p)^{n-k},$$

for $0 < p < 1$ and $k := 0, \ldots, n$.

The following information will be helpful below. Let $\alpha > 0$ and $\beta > 0$. Suppose the continuous random variable Y follows the Beta(α, β) distribution. Then Y has shape parameters α and β, and its probability density function is

$$f_Y(y \,|\, \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}, \quad \text{for} \quad 0 \le y \le 1;$$

therefore, in particular,

$$\int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}\, dy = 1.$$

In the preceding two displays, $\Gamma(\cdot)$ is the *gamma function*, defined by

$$\Gamma(\alpha) := \int_0^\infty t^{\alpha-1} e^{-t}\, dt$$

for each positive real number α. Also, if k is a positive integer, then $\Gamma(k) = (k-1)!$.

To obtain the probability mass function of X_n for $k := 0, \ldots, n$, one uses the law of total probability to compute

$$\begin{aligned}
\mathrm{pr}(X_n = k) &= \int_0^1 f_{X_n|P}(X_n = k \,|\, p) f_P(p)\, dp \\
&= \int_0^1 \binom{n}{k} p^k (1-p)^{n-k}\ 1\, dp \\
&= \binom{n}{k} \int_0^1 p^{(k+1)-1}(1-p)^{(n+1-k)-1}\, dp \\
&= \binom{n}{k} \frac{\Gamma(k+1)\Gamma(n+1-k)}{\Gamma(k+1+n+1-k)} \\
&= \frac{1}{n+1}.
\end{aligned}$$

Therefore, X_n has the discrete uniform distribution on $\{0, \ldots, n\}$. This makes sense; the prior distribution is noninformative, and then all possible outcomes are equiprobable.

As before, suppose it is now known that a person has been playing the lottery for a long time, say n plays, and has never won. The concern is the outcome at the next play. Under this model, the probability of losing is the random variable P. From the previous paragraph, X_{n+1} has the discrete uniform distribution on $\{0, \ldots, n+1\}$. The probability of losing at the $(n+1)$st play given that one has lost at the preceding n plays is given by

$$\begin{aligned}\operatorname{pr}(X_{n+1} = n+1 \mid X_n = n) &= \frac{\operatorname{pr}\{(X_{n+1} = n+1) \text{ and } (X_n = n)\}}{\operatorname{pr}(X_n = n)} \\ &= \frac{\operatorname{pr}(X_{n+1} = n+1)}{\operatorname{pr}(X_n = n)} \\ &= \frac{1/(n+2)}{1/(n+1)} \\ &\neq \frac{1}{n+2} \\ &= \operatorname{pr}(X_{n+1} = n+1).\end{aligned}$$

Therefore, under this model, the outcomes from different drawings are *not* independent. Moreover,

$$\lim_{n\to\infty} \operatorname{pr}(X_{n+1} = n+1 \mid X_n = n) = 1;$$

thus, the longer the string of consecutive losses, the more likely a person is to lose at the next play.

In the preceding paragraph one sees how the prior information on losing affects the outcome at the next play. A long history of losing indicates that it is very difficult to win, and so one would anticipate a loss at the next play. But there is also a more rigorous statistical explanation, which one can obtain from the posterior distribution of P. First one computes the posterior cumulative distribution function

$$\begin{aligned}\operatorname{pr}(P \le p \mid X_n = k) &= \frac{\displaystyle\int_0^p \operatorname{pr}(X_n = k \mid P = u) f_P(u)\, du}{\operatorname{pr}(X_n = k)} \\ &= \frac{\displaystyle\int_0^p \binom{n}{k} u^k (1-u)^{n-k}\ 1\ du}{1/(n+1)} \\ &= (n+1)\binom{n}{k} \int_0^p u^k (1-u)^{n-k}\ du \\ &= \frac{\Gamma(n+2)}{\Gamma(k+1)\Gamma(n+1-k)} \int_0^p u^k (1-u)^{n-k}\ du.\end{aligned}$$

By differentiating the preceding expression with respect to p, one sees that the posterior probability density function of P is given by

$$f_{P|X_n=k}(p \mid x_n = k) = \frac{\Gamma(n+2)}{\Gamma(k+1)\Gamma(n+1-k)} p^k (1-p)^{n-k}, \quad \text{for} \quad 0 < p < 1;$$

this distribution is the Beta$(k+1, n+1-k)$.

In particular, when $k = n$, then $X_n = n$ refers to the case of n losses after n plays. In this case one has

$$f_{P|X_n=n}(p \mid x_n = n) = (n+1)p^n, \quad \text{for} \quad 0 < p < 1.$$

As one can see in Figure 1, this density function has most of its probability mass near 1, even for values of n as small as 10. The mean value is given by

$$E(P \mid X_n = n) = \int_0^1 p f_{P|X_n=n}(p \mid x_n = n)\, dp = \frac{n+1}{n+2},$$

so that

$$\lim_{n\to\infty} E(P \mid X_n = n) = 1.$$

Moreover, for $\epsilon > 0$,

$$\text{pr}(P > 1-\epsilon \mid X_n = n) = \int_{1-\epsilon}^1 f_{P|X_n=n}(p \mid x_n = n)\, dp = 1 - (1-\epsilon)^{n+1};$$

hence, for all $\epsilon > 0$,

$$\lim_{n\to\infty} \text{pr}(P > 1-\epsilon \mid X_n = n) = 1.$$

This confirms what one would infer from Figure 1: as the number of consecutive losses increases, it is more probable that the values of P are very close to 1; this makes it very likely that the outcome at the next play will be another loss.

5 Concluding Remarks

The two models differ with respect to independence of events; this is the most salient and perhaps unexpected difference between them. However, as happens very often, one obtains the same overall conclusion from either model. For the friend in the Introduction the message from either model is the same: it is very difficult to win. This message is delivered clearly by the graph in Figure 1; this shows the posterior distribution that one uses to make inferences in a Bayesian analysis. In the frequentist approach one can get the same clear message by computing some probabilities, and by computing the expected number of plays needed to obtain the first win. For example, if there are $m := 1000$ numbers available to play the lottery, and one is going to play 50 times, the probability that one will never win is 0.9512; if the number of plays doubles, the corresponding probability decreases a little, to 0.9048. For the same m and assuming independence of drawings, the expected number of plays needed to win for the first time is 1000.

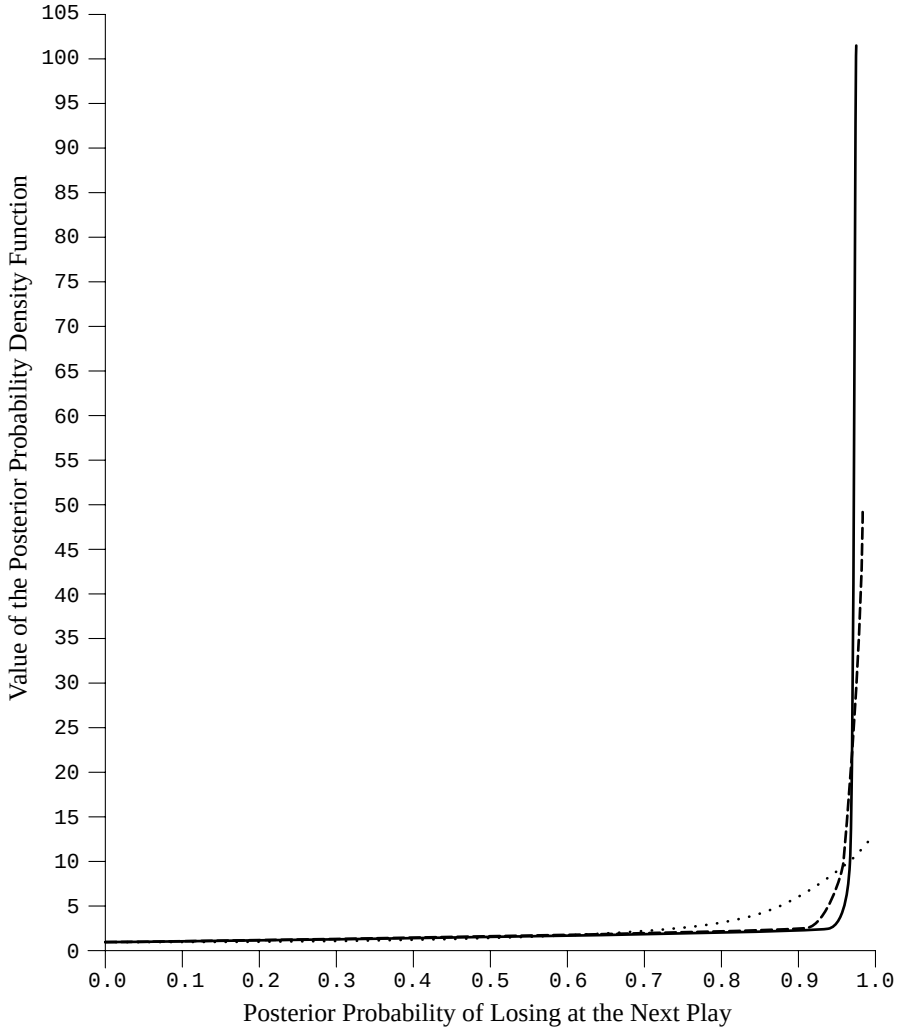

Fig. 1 Posterior distribution of the probability of losing at the next play given 10 (dotted line), or 50 (dashed line), or 100 (solid line) consecutive previous losses.

Reference

1 J. A. Rice, *Mathematical Statistics and Data Analysis*, 2d edition, Wadsworth, Belmont, California, 1995.

Osvaldo Marrero
Department of Mathematical Sciences
Villanova University
800 Lancaster Avenue
Villanova, Pennsylvania 19085-1699
USA