Probability with Discrete Variables

Ben Goodrich

January 24, 2019

Obligatory Disclosure

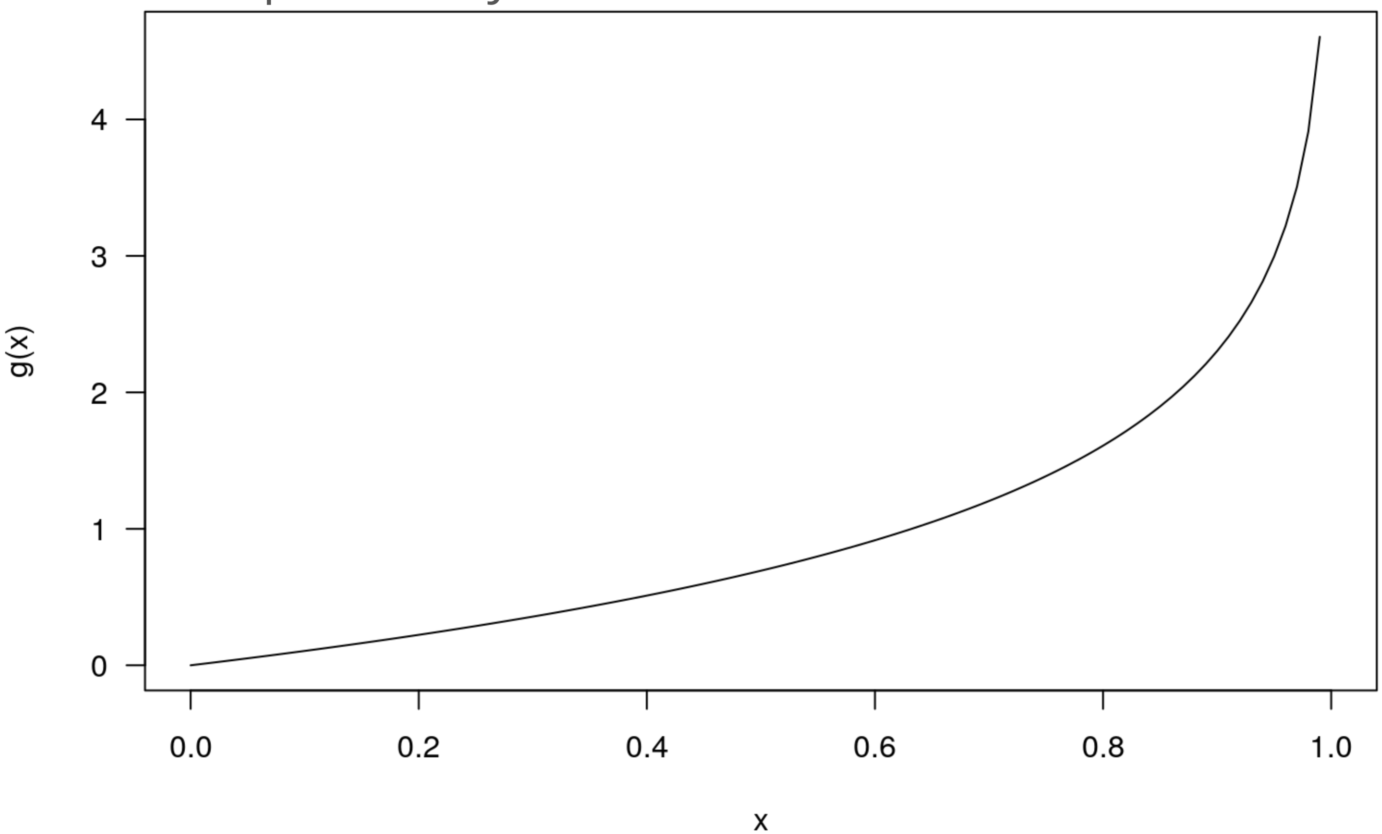
- Ben is an employee of Columbia University, which has received several research grants to develop Stan
- Ben is also a manager of GG Statistics LLC, which uses Stan for business purposes
- According to Columbia University policy, any such employee who has any equity stake in, a title (such as officer or director) with, or is expected to earn at least $\$5,000.00$ per year from a private company is required to disclose these facts in presentations

Sets

- A set is a collection of elements
- Elements can be intervals and / or isolated elements
- One often-used set is the set of real numbers, $\mathbb{R}$
- Loosely, real numbers have decimal points
- Integers are a subset of $\mathbb{R}$, denoted $\mathbb{Z}$, where the decimal places are $.000\ldots$
- Often negative numbers are excluded from a set; e.g. $\mathbb{R}_+$
- Sets can be categorical
- In this session we are going to focus on some subset of $\mathbb{Z}$

Random Variables

- A function is a rule that UNIQUELY maps each element of an input set to some element of an output set
- A random variable is a FUNCTION from the sample space, Ω, to some subset of $\mathbb{R}$ with a probability-based rule



Sample Space

The sample space, denoted Ω, is the set of all possible outcomes of an observable random variable

- Suppose you roll a six-sided die. What is Ω?
- Do not conflate a REALIZATION of a random variable with the FUNCTION that generated it
- By convention, a capital letter, X, indicates a random variable and its lower-case counterpart, x, indicates a realization of X

First Roll in Bowling

- Each frame in bowling starts with $n = 10$ pins
- You get 2 rolls per frame to knock down pins
- What is Ω for your first roll?
- $|$ is read as "given"
- Hohn (2009) discusses a few distributions for the probability of knocking down $X \geq 0$ out of $n \geq X$ pins, including $\Pr\left(x|n\right) = \frac{\mathcal{F}_x}{-1+\mathcal{F}_{n+2}}$ where $\mathcal{F}_x$ is the x-th Fibonacci number, i.e. $\mathcal{F}_0 = 1$, $\mathcal{F}_1 = 1$, and otherwise $\mathcal{F}_x = \mathcal{F}_{x-1} + \mathcal{F}_{x-2}$
- First 13 Fibonacci numbers are 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, and 233
- Sum of the first 11 Fibonacci numbers is 232

source("https://tinyurl.com/y9ubz73j")

```r
# computes the x-th Fibonacci number without recursion and with vectorization
F <- function(x) {
  stopifnot(is.numeric(x), all(x == as.integer(x)))
  sqrt_5 <- sqrt(5) # defined once, used twice
  golden_ratio <- (1 + sqrt_5) / 2
  return(round(golden_ratio ^ (x + 1) / sqrt_5))
}
# probability of knocking down x out of n pins
Pr <- function(x, n = 10) return(ifelse(x > n, 0, F(x) / (-1 + F(n + 2))))

Omega <- 0:10 # 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
round(c(Pr(Omega), total = sum(Pr(Omega))), digits = 3)
```

```
##                                                                    total
## 0.004 0.004 0.009 0.013 0.022 0.034 0.056 0.091 0.147 0.237 0.384 1.000
```

```r
x <- sample(Omega, size = 1, prob = Pr(Omega)) # realization of random variable
```

Second Roll in Bowling

- How would you compute the probability of knocking down the remaining pins on your second roll?
- Let X_1 and X_2 respectively be the number of pins knocked down on the first and second rolls of a frame of bowling. What function yields the probability of knocking down x_2 pins on your second roll?
- $\Pr\left(x_2 \middle| X_1 = x_1, n = 10\right) = \frac{\mathcal{F}_{x_2}}{-1+\mathcal{F}_{10-x_1+2}} \times \mathbb{I}\left\{x_2 \leq 10 - x_1\right\}$
- $\mathbb{I}\left\{\cdot\right\}$ is an “indicator function” that equals 1 if it is true and 0 if it is false
- $\Pr\left(x_2 \middle| X_1 = x_1, n = 10\right)$ is a CONDITIONAL probability
- Conditioning is a fundamental idea that means, “do an operation only in the subset where the condition(s) hold(s)”

From Aristotelian Logic to Bivariate Probability

- In R, TRUE maps to 1 and FALSE maps to 0 when doing arithmetic operations

```
(TRUE & TRUE) == (TRUE * TRUE)
```

```
## [1] TRUE
```

```
(TRUE & FALSE) == (TRUE * FALSE)
```

```
## [1] TRUE
```

- Can generalize to probabilities on the $[0, 1]$ interval to compute the probability that two (or more) propositions are true simultaneously
- $\bigcap$ reads as “and”. **General Multiplication Rule**: $\Pr\left(A \bigcap B\right) = \Pr\left(B\right) \times \Pr\left(A \mid B\right) = \Pr\left(A\right) \times \Pr\left(B \mid A\right)$

Independence

- Loosely, A and B are independent propositions if A being true or false tells us nothing about the probability that B is true (and vice versa)
- Formally, A and B are independent iff $\Pr\left(A \middle| B\right) = \Pr\left(A\right)$ (and $\Pr\left(B \middle| A\right) = \Pr\left(B\right)$). Thus, $\Pr\left(A \bigcap B\right) = \Pr\left(A\right) \times \Pr\left(B\right)$.
- Why is it reasonable to think
  - Two rolls in the same frame are not independent?
  - Two rolls in different frames are independent?
  - Rolls by two different people are independent regardless of whether they are in the same frame?
- What is the probability of obtaining a turkey (3 consecutive strikes)?
- What is the probability of knocking down 9 pins on the first roll and 1 pin on the second roll?

Joint Probability of Two Rolls in Bowling

- How to obtain the joint probability, $\Pr\left(x_1 \bigcap x_2 \middle| n = 10\right)$, in general?

$$\Pr\left(x_1 \bigcap x_2 \middle| n = 10\right) = \Pr\left(x_1 \middle| n = 10\right) \times \Pr\left(x_2 \middle| X_1 = x_1, n = 10\right)$$
$$= \frac{\mathcal{F}_{x_1}}{-1 + \mathcal{F}_{10+2}} \times \frac{\mathcal{F}_{x_2}}{-1 + \mathcal{F}_{10-x_1+2}} \times \mathbb{I}\left\{x_2 \leq 10 - x_1\right\}$$

```
joint_Pr <- matrix(0, nrow = length(Omega), ncol = length(Omega))
rownames(joint_Pr) <- colnames(joint_Pr) <- as.character(Omega)
for (x1 in Omega) {
  Pr_x1 <- Pr(x1)
  for (x2 in 0:(10 - x1))
    joint_Pr[x1 + 1, x2 + 1] <- Pr_x1 * Pr(x2, 10 - x1)
}
sum(joint_Pr) # that sums to 1
```

```
## [1] 1
```

`joint_Pr`: Row is roll 1, Column is roll 2

	0	1	2	3	4	5	6	7	8	9	10
0	0.000019	0.000019	0.000037	0.000056	0.000093	0.000149	0.000242	0.00039	0.000632	0.001022	0.001654
1	0.00003	0.00003	0.00006	0.00009	0.000151	0.000241	0.000392	0.000633	0.001025	0.001658	0
2	0.000098	0.000098	0.000196	0.000294	0.00049	0.000784	0.001274	0.002057	0.003331	0	0
3	0.000239	0.000239	0.000479	0.000718	0.001197	0.001916	0.003113	0.005029	0	0	0
4	0.000653	0.000653	0.001306	0.001959	0.003265	0.005225	0.00849	0	0	0	0
5	0.001724	0.001724	0.003448	0.005172	0.008621	0.013793	0	0	0	0	0
6	0.00467	0.00467	0.009339	0.014009	0.023348	0	0	0	0	0	0
7	0.012931	0.012931	0.025862	0.038793	0	0	0	0	0	0	0
8	0.036638	0.036638	0.073276	0	0	0	0	0	0	0	0
9	0.118534	0.118534	0	0	0	0	0	0	0	0	0
10	0.383621	0	0	0	0	0	0	0	0	0	0

Composition

- The stochastic analogue to the **General Multiplication Rule** is composition
- Randomly draw a realization of x_1 and use that realization of x_1 when randomly drawing x_2 from its conditional distribution

```
S <- 10^6; yes <- 0
for (s in 1:S) {
  x1 <- sample(Omega, size = 1, prob = Pr(Omega))
  x2 <- sample(0:(10 - x1), size = 1, prob = Pr(0:(10 - x1), n = 10 - x1))
  if (x1 == 9 & x2 == 1) yes <- yes + 1
}
c(simulated = yes / S, truth = joint_Pr["9", "1"])
```

```
## simulated     truth
## 0.1188240 0.1185345
```

- As $S \uparrow \infty$, this process converges to $\Pr\left(X_1 = 9 \bigcap X_2 = 1\right)$

Aristotelian Logic to Probability of Alternatives

- What is the probability you fail to get a strike on this frame or the next one?
- Can generalize Aristotelian logic to probabilities on the $[0,1]$ interval to compute the probability that one of two (or more) propositions is true
- $\bigcup$ is read as “or”. **General Addition Rule**: $\Pr\left(A\bigcup B\right)=\Pr\left(A\right)+\Pr\left(B\right)-\Pr\left(A\bigcap B\right)$
- If $\Pr\left(A\bigcap B\right)=0$, A and B are mutually exclusive (disjoint)

What is the probability that $X_2 = 9$?

	0	1	2	3	4	5	6	7	8	9	10
0	0.000019	0.000019	0.000037	0.000056	0.000093	0.000149	0.000242	0.00039	0.000632	0.001022	0.001654
1	0.00003	0.00003	0.00006	0.00009	0.000151	0.000241	0.000392	0.000633	0.001025	0.001658	0
2	0.000098	0.000098	0.000196	0.000294	0.00049	0.000784	0.001274	0.002057	0.003331	0	0
3	0.000239	0.000239	0.000479	0.000718	0.001197	0.001916	0.003113	0.005029	0	0	0
4	0.000653	0.000653	0.001306	0.001959	0.003265	0.005225	0.00849	0	0	0	0
5	0.001724	0.001724	0.003448	0.005172	0.008621	0.013793	0	0	0	0	0
6	0.00467	0.00467	0.009339	0.014009	0.023348	0	0	0	0	0	0
7	0.012931	0.012931	0.025862	0.038793	0	0	0	0	0	0	0
8	0.036638	0.036638	0.073276	0	0	0	0	0	0	0	0
9	0.118534	0.118534	0	0	0	0	0	0	0	0	0
10	0.383621	0	0	0	0	0	0	0	0	0	0

Marginal Distribution of Second Roll in Bowling

- How to obtain $\Pr\left(x_2 \middle| n = 10\right)$ irrespective of x_1?
- Since events in the first roll are mutually exclusive, use the easy form of the General Addition Rule to "marginalize":

$$\begin{aligned}\Pr\left(x_2 \middle| n = 10\right) &= \sum_{i:x_i \in \Omega_{X_1}} \Pr\left(x_i \bigcap x_2 \middle| n = 10\right) \\ &= \sum_{i:x_i \in \Omega_{X_1}} \Pr\left(x_2 \middle| X_1 = x_i, n = 10\right) \times \Pr\left(x_i \middle| n = 10\right)\end{aligned}$$

```
round(rbind(Pr_X1 = Pr(Omega), margin1 = rowSums(joint_Pr), margin2 = colSums(joint_Pr)), 3)
```

```
            0     1     2     3     4     5     6     7     8     9    10
Pr_X1   0.004 0.004 0.009 0.013 0.022 0.034 0.056 0.091 0.147 0.237 0.384
margin1 0.004 0.004 0.009 0.013 0.022 0.034 0.056 0.091 0.147 0.237 0.384
margin2 0.559 0.176 0.114 0.061 0.037 0.022 0.014 0.008 0.005 0.003 0.002
```

Marginal, Conditional, and Joint Probabilities

- To compose a joint (in this case bivariate) probability, MULTIPLY a marginal probability by a conditional probability
- To decompose a joint (in this case bivariate) probability, ADD the relevant joint probabilities to obtain a marginal probability
- To obtain a conditional probability, DIVIDE the relevant joint probability by the relevant marginal probability since

$$\Pr\left(A \bigcap B\right) = \Pr\left(B\right) \times \Pr\left(A \mid B\right) = \Pr\left(A\right) \times \Pr\left(B \mid A\right)$$

$$\Pr\left(A \mid B\right) = \frac{\Pr\left(A\right) \times \Pr\left(B \mid A\right)}{\Pr\left(B\right)} \text{ if } \Pr\left(B\right) > 0$$

- This is Bayes Rule
- What is $\Pr\left(X_1 = 3 \mid X_2 = 4, n = 10\right)$?

Conditioning on $X_2 = 4$

	0	1	2	3	4	5	6	7	8	9	10
0	0.000019	0.000019	0.000037	0.000056	0.000093	0.000149	0.000242	0.00039	0.000632	0.001022	0.001654
1	0.00003	0.00003	0.00006	0.00009	0.000151	0.000241	0.000392	0.000633	0.001025	0.001658	0
2	0.000098	0.000098	0.000196	0.000294	0.00049	0.000784	0.001274	0.002057	0.003331	0	0
3	0.000239	0.000239	0.000479	0.000718	0.001197	0.001916	0.003113	0.005029	0	0	0
4	0.000653	0.000653	0.001306	0.001959	0.003265	0.005225	0.00849	0	0	0	0
5	0.001724	0.001724	0.003448	0.005172	0.008621	0.013793	0	0	0	0	0
6	0.00467	0.00467	0.009339	0.014009	0.023348	0	0	0	0	0	0
7	0.012931	0.012931	0.025862	0.038793	0	0	0	0	0	0	0
8	0.036638	0.036638	0.073276	0	0	0	0	0	0	0	0
9	0.118534	0.118534	0	0	0	0	0	0	0	0	0
10	0.383621	0	0	0	0	0	0	0	0	0	0

Example of Bayes Rule

```
joint_Pr["3", "4"] / sum(joint_Pr[,"4"])
```

```
## [1] 0.03221668
```

- Bayesians generalize this by taking A to be “beliefs about whatever you do not know” and B to be whatever you do know in

$$\Pr\left(A|\,B\right) = \frac{\Pr\left(A\right) \times \Pr\left(B|\,A\right)}{\Pr\left(B\right)} \text{ if } \Pr\left(B\right) > 0$$

- Frequentists accept Bayes Rule but object to using the language of probability to describe beliefs about unknown propositions and insist that probability is a property of a process that can be defined as a limit

$$\Pr\left(A\right) = \lim_{S \uparrow \infty} \frac{\text{times that } A \text{ occurs in } S \text{ independent tries}}{S}$$

Probability in Football

- What is the probability that the Patriots beat the Rams next Sunday?
- To a frequentist, it is infeasible to answer this question objectively and it should not be answered subjectively
- One way of understanding it from a Bayesian perspective is via betting: Do you want to risk \$6 to gain \$4 if the Patrios win? If so, you believe the probability the Patriots win is greater than 0.6.

$$\mathrm{Odds}\left(A\right) = \frac{\Pr\left(A\right)}{1 - \Pr\left(A\right)}$$

- Once you commit to a probability, the decision to bet is straightforward
- Everyone understands what you mean if you say the probability the Patriots beat the Rams is greater than 0.6. Why must science be different?

Objectivity and Subjectivity

- Under weak and not particularly controversial assumptions, Bayesian inference is THE objective way to update your beliefs about (functions of) θ in light of new data $y_1, y_2, \ldots, y_N$
- Nevertheless, the Bayesian approach is labeled subjective because it does not say what your beliefs about θ should be before you receive $y_1, y_2, \ldots, y_N$
- Thus, if you currently believe something absurd about θ now, your beliefs about θ will merely be less absurd after updating them with $y_1, y_2, \ldots, y_N$
- The big problem is not that people believe wrong things now, but that they do not update their beliefs about θ according to Bayesian principles when they observe $y_1, y_2, \ldots, y_N$
- In fact, in some situations, observing data that contradicts people's previous beliefs makes them believe in their wrong beliefs more strongly
- Bayesian principles are also used in formal models, but as an assumption about how people should behave rather than a behavioral description

(Dis)Advantages of Bayesian Inference

- Bayesian inference remains useful in situations other paradigms specialize in:
  - Experiments: What are your beliefs about the ATE after seeing the data?
  - Repeated designs: Bayesian estimates have correct frequentist properties
  - Predictive modeling: If you only care about predictions, use the posterior predictive distribution
- Bayesian inference is very useful when you are using the results to make a decision or take an action; other paradigms are not
- Bayesian inference is orders of magnitude more difficult for your computer because it is attempting to answer a more ambitious question
- The Bayesian approach is better suited for convincing yourself of something than convincing other people