

## 2 - Small Worlds and Large Worlds

Antonio Jurlina

07/01/2020

### Easy

**2E1.** Which of the expressions below correspond to the statement: *the probability of rain on Monday*?

- (1)  $P(rain)$
- (2)  $P(rain|Monday)$
- (3)  $P(Monday|rain)$
- (4)  $P(rain, Monday)/P(Monday)$

The probability of rain on Monday is  $P(rain|Monday)$ .

The joint probability (probability that it is rain and it is a Monday) is  $P(rain, Monday) = P(rain|Monday)P(Monday)$ .

From there, we conclude that

$$\frac{P(rain, Monday)}{P(Monday)} = P(rain|Monday)$$

Therefore, correct answers are (2) and (4).

**2E2.** Which of the following statements corresponds to the expression:  $Pr(Monday|rain)$ ?

- (1) The probability of rain on Monday.
- (2) The probability of rain, given that it is Monday.
- (3) The probability that it is Monday, given that it is raining.
- (4) The probability that it is Monday and that it is raining.

$Pr(Monday|rain)$  can be understood as *the probability that it is Monday, given that it is raining*. Therefore, the answer is (3).

**2E3.** Which of the expressions below correspond to the statement: *the probability that it is Monday, given that it is raining*?

- (1)  $P(Monday|rain)$
- (2)  $P(rain|Monday)$
- (3)  $P(rain|Monday)P(Monday)$
- (4)  $P(rain|Monday)P(Monday)/P(rain)$
- (5)  $P(Monday|rain)P(rain)/P(Monday)$

According to Bayes' Theorem

$$P(\text{Monday}|\text{rain}) = \frac{P(\text{rain}|\text{Monday})P(\text{Monday})}{P(\text{rain})}$$

therefore, the correct answers are (1) and (4).

**2E4.** The Bayesian statistician Bruno de Finetti (1906–1985) began his book on probability theory with the declaration: “PROBABILITY DOES NOT EXIST.” The capitals appeared in the original, so I imagine de Finetti wanted us to shout this statement. What he meant is that probability is a device for describing uncertainty from the perspective of an observer with limited knowledge; it has no objective reality. Discuss the globe tossing example from the chapter, in light of this statement. What does it mean to say “the probability of water is 0.7”?

“The probability of water is 0.7” means the tosser is 70% certain his finger will land on water when catching the globe.

## Medium

**2M1.** Recall the globe tossing model from the chapter. Compute and plot the grid approximate posterior distribution for each of the following sets of observations. In each case, assume a uniform prior for  $p$ .

- (1) W,W,W
- (2) W,W,W,L
- (3) L,W,W,L,W,W,W

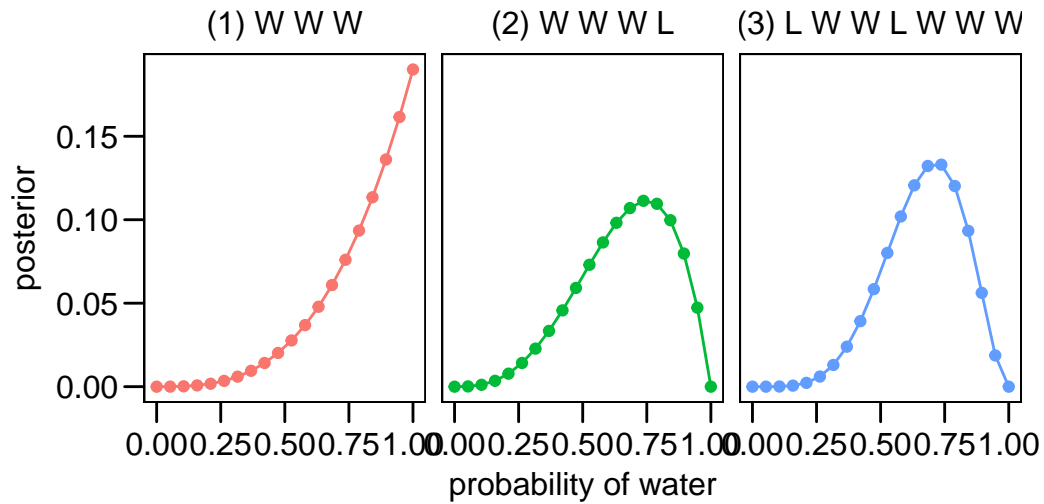
```
n_of_points <- 20

data <- tibble(p_grid = seq(0, 1, length.out = n_of_points),
               prior = 0.5,
               likelihood = dbinom(3, 3,
                                   prob = p_grid),
               part = "(1) W W W") %>%
bind_rows(tibble(p_grid = seq(0, 1, length.out = n_of_points),
                  prior = 0.5,
                  likelihood = dbinom(3, 4,
                                      prob = p_grid),
                  part = "(2) W W W L")) %>%
bind_rows(tibble(p_grid = seq(0, 1, length.out = n_of_points),
                  prior = 0.5,
                  likelihood = dbinom(5, 7,
                                      prob = p_grid),
                  part = "(3) L W W L W W W"))

data <- data %>%
  rename(`probability of water` = p_grid) %>%
  mutate(unstd_posterior = likelihood * prior) %>%
  group_by(part) %>%
  mutate(posterior = unstd_posterior / sum(unstd_posterior)) %>%
  ungroup()

ggplot(data, aes(x = `probability of water`, y = posterior, color = part)) +
  geom_point(show.legend = F) +
  geom_line(show.legend = F) +
```

```
facet_wrap(. ~ part) +
theme_par()
```



**2M2.** Now assume a prior for  $p$  that is equal to zero when  $p < 0.5$  and is a positive constant when  $p \geq 0.5$ . Again compute and plot the grid approximate posterior distribution for each of the sets of observations in the problem just above.

*Note: I rewrote code for this problem from scratch so that it stands on its own, but I could have just reused the data from the previous one with a slight modification.*

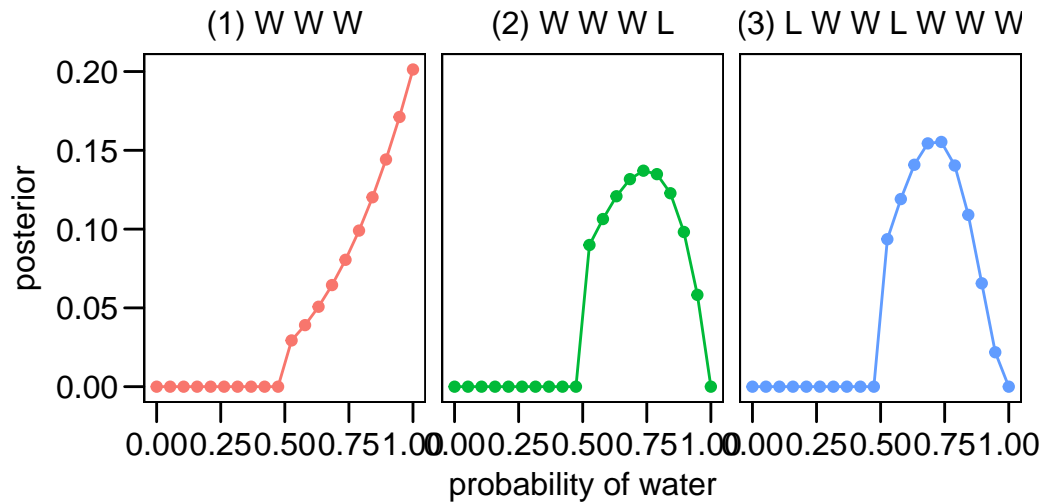
```
n_of_points <- 20

data <- tibble(p_grid = seq(0, 1, length.out = n_of_points),
  prior = if_else(p_grid < 0.5, 0, 1),
  likelihood = dbinom(3, 3,
    prob = p_grid),
  part = "(1) W W W") %>%
bind_rows(tibble(p_grid = seq(0, 1, length.out = n_of_points),
  prior = if_else(p_grid < 0.5, 0, 1),
  likelihood = dbinom(3, 4,
    prob = p_grid),
  part = "(2) W W W L")) %>%
bind_rows(tibble(p_grid = seq(0, 1, length.out = n_of_points),
  prior = if_else(p_grid < 0.5, 0, 1),
  likelihood = dbinom(5, 7,
    prob = p_grid),
  part = "(3) L W W L W W W"))

data <- data %>%
  rename(`probability of water` = p_grid) %>%
  mutate(unstd_posterior = likelihood * prior) %>%
  group_by(part) %>%
  mutate(posterior = unstd_posterior / sum(unstd_posterior)) %>%
  ungroup()

ggplot(data, aes(x = `probability of water`, y = posterior, color = part)) +
  geom_point(show.legend = F) +
  geom_line(show.legend = F) +
```

```
facet_wrap(. ~ part) +
theme_par()
```



**2M3.** Suppose there are two globes, one for Earth and one for Mars. The Earth globe is 70% covered in water. The Mars globe is 100% land. Further suppose that one of these globes—you don’t know which—was tossed in the air and produced a “land” observation. Assume that each globe was equally likely to be tossed. Show that the posterior probability that the globe was the Earth, conditional on seeing “land” ( $Pr(Earth|land)$ ), is 0.23.

From the problem statement we know that  $P(land|Earth) = 0.3$ ,  $P(land|Mars) = 1$ , and  $P(E) = P(M) = 0.5$ .

$$\begin{aligned}
 P(land) &= P(land|Earth)P(Earth) + P(land|Mars)P(Mars) \\
 P(land) &= 0.3 * 0.5 + 1 * 0.5 \\
 P(land) &= 0.65
 \end{aligned}$$

From Bayes’ theorem we know that

$$\begin{aligned}
 P(Earth|land) &= \frac{P(land|Earth)P(Earth)}{P(land)} \\
 P(Earth|land) &= \frac{0.3 * 0.5}{0.65} \\
 P(Earth|land) &= 0.23
 \end{aligned}$$

**2M4.** Suppose you have a deck with only three cards. Each card has two sides, and each side is either black or white. One card has two black sides. The second card has one black and one white side. The third card has two white sides. Now suppose all three cards are placed in a bag and shuffled. Someone reaches into the bag and pulls out a card and places it flat on a table. A black side is shown facing up, but you don’t know the color of the side facing down. Show that the probability that the other side is also black is  $2/3$ . Use the counting method (Section 2 of the chapter) to approach this problem. This means counting up the ways that each card could produce the observed data (a black side facing up on the table).

```
cards <- tibble(card = c("BB", "BW", "WW"),
               ways = c(2, 1, 0),
               prob = round(ways / sum(ways), 3))

kable(cards)
```

card	ways	prob
BB	2	0.667
BW	1	0.333
WW	0	0.000

**2M5.** Now suppose there are four cards: B/B, B/W, W/W, and another B/B. Again suppose a card is drawn from the bag and a black side appears face up. Again calculate the probability that the other side is black.

```
cards <- tibble(card = c("BB", "BW", "WW", "BB"),
               ways = c(2, 1, 0, 2),
               prob = round(ways / sum(ways), 3)) %>%
  group_by(card) %>%
  summarize(ways = sum(ways),
            prob = sum(prob))

kable(cards)
```

card	ways	prob
BB	4	0.8
BW	1	0.2
WW	0	0.0

**2M6.** Imagine that black ink is heavy, and so cards with black sides are heavier than cards with white sides. As a result, it's less likely that a card with black sides is pulled from the bag. So again assume there are three cards: B/B, B/W, and W/W. After experimenting a number of times, you conclude that for every way to pull the B/B card from the bag, there are 2 ways to pull the B/W card and 3 ways to pull the W/W card. Again suppose that a card is pulled and a black side appears face up. Show that the probability the other side is black is now 0.5. Use the counting method, as before.

```
cards <- tibble(card = c("BB", "BW", "WW"),
               ways = c(2, 1, 0),
               weight = c(1, 2, 3),
               updated_ways = ways * weight,
               prob = round(updated_ways / sum(updated_ways), 3))

kable(cards)
```

card	ways	weight	updated_ways	prob
BB	2	1	2	0.5
BW	1	2	2	0.5
WW	0	3	0	0.0

**2M7.** Assume again the original card problem, with a single card showing a black side face up. Before looking at the other side, we draw another card from the bag and lay it face up on the table. The face that is shown on the new card is white. Show that the probability that the first card, the one showing a black side, has black on its other side is now 0.75. Use the counting method, if you can. *Hint: Treat this like the sequence of globe tosses, counting all the ways to see each observation, for each possible first card.*

```
cards <- tibble(card = c("BB", "BW", "WW"),
               ways_black = c(2, 1, 0),
```

```
ways_white = c(3, 2, 0),
updated_ways = ways_black * ways_white,
prob = round(updated_ways / sum(updated_ways), 3))
```

```
kable(cards)
```

card	ways_black	ways_white	updated_ways	prob
BB	2	3	6	0.75
BW	1	2	2	0.25
WW	0	0	0	0.00

## Hard

**2H1.** Suppose there are two species of panda bear. Both are equally common in the wild and live in the same places. They look exactly alike and eat the same food, and there is yet no genetic assay capable of telling them apart. They differ however in their family sizes. Species A gives birth to twins 10% of the time, otherwise birthing a single infant. Species B births twins 20% of the time, otherwise birthing singleton infants. Assume these numbers are known with certainty, from many years of field research. Now suppose you are managing a captive panda breeding program. You have a new female panda of unknown species, and she has just given birth to twins. What is the probability that her next birth will also be twins?

From the problem statement we know that  $P(\text{twins}|A) = 0.1$ ,  $P(\text{twins}|B) = 0.2$ , and  $P(A) = P(B) = 0.5$ .

$$P(\text{twins}) = P(\text{twins}|A)P(A) + P(\text{twins}|B)P(B)$$

$$P(\text{twins}) = 0.5(0.1 + 0.2)$$

$$P(\text{twins}) = 0.15$$

$$P(A|\text{twins}) = \frac{P(\text{twins}|A)P(A)}{P(\text{twins})}$$

$$P(A|\text{twins}) = \frac{0.1 * 0.5}{0.15}$$

$$P(A|\text{twins}) = \frac{1}{3}$$

$$P(B|\text{twins}) = \frac{P(\text{twins}|B)P(B)}{P(\text{twins})}$$

$$P(B|\text{twins}) = \frac{0.2 * 0.5}{0.15}$$

$$P(B|\text{twins}) = \frac{2}{3}$$

$$P(\text{twins}) = P(\text{twins}|A)P(A|\text{twins}) + P(\text{twins}|B)P(B|\text{twins})$$

$$P(\text{twins}) = 0.1 * \frac{1}{3} + 0.2 * \frac{2}{3}$$

$$P(\text{twins}) = \frac{1}{6}$$

**2H2.** Recall all the facts from the problem above. Now compute the probability that the panda we have is from species A assuming we have observed only the first birth and that it was twins.

$$P(\text{twins}) = P(\text{twins}|A)P(A) + P(\text{twins}|B)P(B)$$

$$P(twins) = 0.5(0.1 + 0.2)$$

$$P(twins) = 0.15$$

$$P(A|twins) = \frac{P(twins|A)P(A)}{P(twins)}$$

$$P(A|twins) = \frac{0.1 * 0.5}{0.15}$$

$$P(A|twins) = \frac{1}{3}$$

**2H3.** Continuing on from the previous problem suppose the same panda mother has a second birth and that it is not twins but a singleton infant. Compute the posterior probability that this panda is species A.

From the problem statement we know that  $P(twins|A) = 0.1$ ,  $P(twins|B) = 0.2$ , and  $P(A) = P(B) = 0.5$ . Also, from 2H1 we know that  $P(A|twins) = \frac{1}{3}$  and  $P(B|twins) = \frac{2}{3}$ . Finally, we can rewrite  $P(twins|A)$  and  $P(twins|B)$  as  $P(single|A) = 0.9$  and  $P(single|B) = 0.8$ .

$$P(single) = P(single|A)P(A) + P(single|B)P(B)$$

If the prior of having already had a twin is included,  $P(A)$  becomes  $P(A|twins)$  and  $P(B)$  becomes  $P(B|twins)$ .

$$P(single) = P(single|A)P(A|twins) + P(single|B)P(B|twins)$$

$$P(single) = 0.9 * \frac{1}{3} + 0.8 * \frac{2}{3}$$

$$P(single) = \frac{5}{6}$$

$$P(A|single) = \frac{P(single|A)P(A)}{P(single)}$$

$$P(A|single) = \frac{0.9 * \frac{1}{3}}{\frac{5}{6}}$$

$$P(A|single) = 0.36$$

**2H4.** A common boast of Bayesian statisticians is that Bayesian inference makes it easy to use all of the data, even if the data are of different types. So suppose now that a veterinarian comes along who has a new genetic test that she claims can identify the species of our mother panda. But the test, like all tests, is imperfect. This is the information you have about the test:

- The probability it correctly identifies a species A panda is 0.8.
- The probability it correctly identifies a species B panda is 0.65.

The vet administers the test to your panda and tells you that the test is positive for species A. First ignore your previous information from the births and compute the posterior probability that your panda is species A. Then redo your calculation, now using the birth data as well.

From the problem statement we know that  $P(positive|A) = 0.80$ ,  $P(positive|B) = 0.65$ , and  $P(A) = P(B) = 0.5$ .

$$P(positive) = P(positive|A)P(A) + P(positive|B)P(B)$$

$$P(positive) = 0.5(0.8 + 0.65)$$

$$P(positive) = 0.725$$

$$P(A|positive) = \frac{P(positive|A)P(A)}{P(positive)}$$

$$P(A|positive) = \frac{0.8 * 0.5}{0.725}$$

$$P(A|positive) = 0.551724$$

Now, including birth data: We know from 2H3 that the posterior probability a panda is species A, given twins followed by a single birth, is 0.36, which also makes the probability a panda is species B, under same circumstances,  $1 - 0.36 = 0.64$ .

$$P(positive) = P(positive|A)P(A) + P(positive|B)P(B)$$

$$P(positive) = 0.8 * 0.36 + 0.65 * 0.64$$

$$P(positive) = 0.704$$

$$P(A|positive) = \frac{P(positive|A)P(A)}{P(positive)}$$

$$P(A|positive) = \frac{0.8 * 0.36}{0.704}$$

$$P(A|positive) = 0.409$$



## sessionInfo()

```
## R version 3.6.2 (2019-12-12)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS Sierra 10.12.6
##
## Matrix products: default
## BLAS: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats      graphics  grDevices  utils      datasets  methods   base
##
## other attached packages:
## [1] knitr_1.28      here_0.1      ggthemes_4.2.0 forcats_0.5.0
## [5] stringr_1.4.0  dplyr_0.8.5   purrr_0.3.4   readr_1.3.1
## [9] tidyr_1.0.3     tibble_3.0.3  ggplot2_3.3.2 tidyverse_1.3.0
##
## loaded via a namespace (and not attached):
## [1] tidyselect_1.1.0 xfun_0.13      haven_2.2.0    lattice_0.20-41
## [5] colorspace_1.4-1 vctrs_0.3.2    generics_0.0.2 htmltools_0.4.0
## [9] yaml_2.2.1        rlang_0.4.7    pillar_1.4.6   glue_1.4.1
## [13] withr_2.2.0       DBI_1.1.0      dbplyr_1.4.3   modelr_0.1.7
## [17] readxl_1.3.1      lifecycle_0.2.0 munsell_0.5.0  gtable_0.3.0
## [21] cellranger_1.1.0 rvest_0.3.5    evaluate_0.14  labeling_0.3
## [25] fansi_0.4.1       highr_0.8      broom_0.5.6    Rcpp_1.0.5
## [29] scales_1.1.1      backports_1.1.8 jsonlite_1.7.0 farver_2.0.3
## [33] fs_1.4.1          hms_0.5.3      digest_0.6.25  stringi_1.4.6
## [37] rprojroot_1.3-2   grid_3.6.2     cli_2.0.2      tools_3.6.2
## [41] magrittr_1.5      crayon_1.3.4   pkgconfig_2.0.3 ellipsis_0.3.1
## [45] xml2_1.3.2        reprex_0.3.0   lubridate_1.7.8 assertthat_0.2.1
## [49] rmarkdown_2.1     httr_1.4.1     rstudioapi_0.11 R6_2.4.1
## [53] nlme_3.1-147      compiler_3.6.2
```