OKIEM EKSPERTA

Analiza bayesowska w Pythonie

Praktyczny przewodnik po modelowaniu probabilistycznym

Wydanie III

Osvaldo Martin

Helion

packt

Tytuł oryginału: Bayesian Analysis with Python: A practical guide to probabilistic modeling, 3rd Edition

Tłumaczenie: Piotr Pilch

ISBN: 978-83-289-3665-2

Drogi Czytelniku!
Jeżeli chcesz ocenić tę książkę, zajrzyj pod adres
helion.pl/user/opinie/anbap3
Możesz tam wpisać swoje uwagi, spostrzeżenia, recenzję.

Helion S.A.
ul. Kościuszki 1c, 44-100 Gliwice
tel. 32 230 98 63
e-mail: *helion@helion.pl*
WWW: *helion.pl* (księgarnia internetowa, katalog książek)

Printed in Poland.

Spis treści

Słowo wstępne **11**

O autorze **13**

O korektorze merytorycznym **13**

Przedmowa **14**

ROZDZIAŁ 1
Myślenie probabilistyczne **20**

Statystyki, modele i podejście zastosowane w książce 20
Praca z danymi 21
Modelowanie bayesowskie 22
Wprowadzenie do prawdopodobieństwa dla praktyków metod bayesowskich 23
Przestrzeń prób i zdarzenia 23
Zmienne losowe 26
Dyskretne zmienne losowe i ich rozkłady 27
Ciągłe zmienne losowe i ich rozkłady 31
Dystrybuanta 33
Prawdopodobieństwo warunkowe 34
Wartości oczekiwane 35
Twierdzenie Bayesa 36
Interpretacja prawdopodobieństwa 39
Prawdopodobieństwo, niepewność i logika 40
Wnioskowanie dotyczące jednego parametru 41
Problem rzutu monetą 41
Wybór funkcji wiarygodności 42
Wybór rozkładu a priori 43
Wyznaczanie rozkładu a posteriori 45
Wpływ rozkładu a priori 47
Sposób wyboru rozkładów a priori 48

Informowanie o wynikach analizy bayesowskiej 51
Notacja modeli i wizualizacja 51
Podsumowanie rozkładu a posteriori 52
Podsumowanie 52
Ćwiczenia 54

ROZDZIAŁ 2
Programowanie probabilistyczne 56
Programowanie probabilistyczne 56
Rzucanie monetą w wariancie biblioteki PyMC 57
Podsumowanie rozkładu a posteriori 60
Decyzje oparte na rozkładzie a posteriori 62
Stosunek gęstości Savage'a-Dickeya 62
Przedział praktycznej równoważności 64
Funkcje straty 65
Rozkłady gaussowskie na każdym poziomie 67
Wnioskowanie gaussowskie 68
Kontrole predykcyjne rozkładu a posteriori 71
Odporne wnioskowanie 72
Stopnie normalności 73
Wersja odporna modelu normalnego 74
Kontener InferenceData 77
Porównywanie grup 79
Zbiór danych tips 80
Współczynnik d Cohena 82
Prawdopodobieństwo przewagi 84
Analiza różnic średnich w rozkładzie a posteriori 84
Podsumowanie 85
Ćwiczenia 86

ROZDZIAŁ 3
Modele hierarchiczne 88
Udostępnianie informacji i rozkładów a priori 88
Przesunięcia hierarchiczne 89
Jakość wody 92
Kurczenie 95
Hierarchie na każdym poziomie 97
Podsumowanie 101
Ćwiczenia 101

ROZDZIAŁ 4
Modelowanie za pomocą linii 103
Prosta regresja liniowa 103
Rowery w ujęciu liniowym 105
Interpretacja średniej rozkładu a posteriori 107
Interpretacja predykcji z rozkładu a posteriori 109
Uogólnianie modelu liniowego 110
Liczenie rowerów 111
Regresja odporna 112
Regresja logistyczna 114
Model logistyczny 115
Klasyfikacja z użyciem regresji logistycznej 117
Interpretacja współczynników regresji logistycznej 119
Wariancja zmiennej 120
Hierarchiczna regresja liniowa 123
Modele hierarchiczne scentrowane i niescentrowane 125
Wieloraka regresja liniowa 127
Podsumowanie 129
Ćwiczenia 130

ROZDZIAŁ 5
Porównywanie modeli 131
Kontrole predykcyjne a posteriori 131
Równowaga między prostotą a dokładnością 136
Wiele parametrów może prowadzić do nadmiernego dopasowania 136
Zbyt mała liczba parametrów prowadzi do niedopasowania 138
Miary dokładności predykcyjnej 138
Kryteria informacyjne 139
Walidacja krzyżowa 141
Obliczanie dokładności predykcyjnej za pomocą biblioteki ArviZ 144
Uśrednianie modeli 146
Współczynniki Bayesa 147
Kilka spostrzeżeń 148
Obliczanie współczynników Bayesa 149
Współczynniki Bayesa i wnioskowanie 154
Regularyzacja rozkładów a priori 155
Podsumowanie 156
Ćwiczenia 157

ROZDZIAŁ 6
Modelowanie za pomocą interfejsu Bambi 159
Jedna składnia, by wszystkim zarządzać 159
Model rowerów w wersji opartej na interfejsie Bambi 163
Regresja wielomianowa 165
Funkcje sklejane 167
Modele rozkładowe 169
Predyktory kategoryczne 171
Pingwiny kategoryczne 171
Związek z modelami hierarchicznymi 173
Interakcje 174
Interpretacja modeli za pomocą interfejsu Bambi 176
Selekcja zmiennych 178
Wnioskowanie predykcyjne metodą projekcji 179
Predykcja projekcyjna z użyciem pakietu Kulprit 180
Podsumowanie 183
Ćwiczenia 184

ROZDZIAŁ 7
Modele mieszanin 185
Modele mieszanin 185
Skończone modele mieszanin 187
Rozkład kategoryczny 188
Rozkład Dirichleta 188
Mieszanina chemiczna 190
Nieidentyfikowalność modeli mieszanin 191
Metoda wyboru liczby rozkładów K 192
Modele z nadmiarem zer i modele progowe 195
Regresja Poissona z nadmiarem zer 196
Modele progowe 197
Modele mieszanin i grupowanie 200
Nieskończony model mieszaniny 200
Proces Dirichleta 201
Mieszaniny ciągłe 205
Niektóre popularne rozkłady to mieszaniny 205
Podsumowanie 206
Ćwiczenia 207

ROZDZIAŁ 8
Procesy gaussowskie 209
Modele liniowe i dane nieliniowe 209
Modelowanie funkcji 210
Wielowymiarowe rozkłady Gaussa i funkcje 212
Funkcje kowariancji i jądra 212
Procesy gaussowskie 214
Regresja procesów gaussowskich 215
Regresja procesów gaussowskich z użyciem biblioteki PyMC 216
Ustalanie rozkładów a priori dla skali długości 219
Klasyfikacja z użyciem procesów gaussowskich 220
Procesy gaussowskie w przypadku kosmicznej grypy 222
Procesy Coxa 223
Katastrofy w kopalniach węgla 224
Sekwoje 225
Regresja z autokorelacją przestrzenną 227
Procesy gaussowskie w przestrzeniach Hilberta 231
Proces HSGP w przypadku interfejsu Bambi 233
Podsumowanie 234
Ćwiczenia 235

ROZDZIAŁ 9
Bayesowskie addytywne drzewa regresyjne 236
Drzewa decyzyjne 236
Modele BART 238
Pingwiny bartiańskie 239
Wykresy zależności cząstkowej 240
Wykresy indywidualnej warunkowej wartości oczekiwanej 241
Selekcja zmiennych z użyciem modelu BART 242
Modele BART oparte na rozkładach 245
Odpowiedź stała i liniowa 247
Wybór liczby drzew 248
Podsumowanie 248
Ćwiczenia 249

ROZDZIAŁ 10
Silniki wnioskowania 250
Silniki wnioskowania 250
Metoda siatki 251
Metoda kwadratowa 254
Metody oparte na procesach Markowa 255
Monte Carlo 256
Łańcuch Markowa 257
Algorytm Metropolisa-Hastingsa 257
Hamiltonowska metoda Monte Carlo 261
Sekwencyjna metoda Monte Carlo 263
Diagnozowanie próbek 265
Zbieżność 266
Wykres śladu 266
Wykres rang 267
$\hat{R}$ (statystyka R-hat) 268
Efektywny rozmiar próby 270
Błąd standardowy Monte Carlo 271
Dywergencje 272
Zachowaj spokój i nie poddawaj się 274
Podsumowanie 275
Ćwiczenia 275

ROZDZIAŁ 11
Dalsze kierunki 277

Bibliografia 279

Rozdział 1

Myślenie probabilistyczne

Rachunek prawdopodobieństwa to nic innego jak zdrowy rozsądek sprowadzony do obliczeń.

— Pierre Simon Laplace

W niniejszym rozdziale poznasz podstawowe koncepcje statystyki bayesowskiej oraz niektóre instrumenty z bayesowskiego zestawu narzędzi. Będziesz używać kodu w języku Python, ale rozdział ten ma głównie charakter teoretyczny. Większość omawianych tutaj zagadnień będzie wielokrotnie powracać w całej książce. Niniejszy rozdział, który jest mocno osadzony w teorii, może wydawać się dla Ciebie jako programisty nieco niepokojący, ale myślę, że ułatwi Ci skuteczne stosowanie statystyki bayesowskiej do rozwiązywania problemów.

W tym rozdziale omawiam następujące zagadnienia:

- modelowanie statystyczne;
- prawdopodobieństwo i niepewność;
- twierdzenie Bayesa i wnioskowanie statystyczne;
- wnioskowanie z jednym parametrem i klasyczny problem rzutu monetą;
- dobór rozkładów a priori i wyjaśnienie, dlaczego ludzie często ich nie lubią, choć powinni;
- informowanie o wynikach analizy bayesowskiej.

Statystyki, modele i podejście zastosowane w książce

Statystyka zajmuje się gromadzeniem, organizowaniem, analizowaniem i interpretowaniem danych, dlatego wiedza z dziedziny statystyki jest niezbędna do analizy danych. W analizie danych stosuje się następujące dwie główne metody statystyczne:

- **Eksploracyjna analiza danych (ang. *Exploratory Data Analysis* — EDA)**. Dotyczy ona podsumowań liczbowych, takich jak średnia, dominanta, odchylenie standardowe i rozstępy międzykwartylowe. Analiza EDA polega również na wizualnej inspekcji danych przy użyciu narzędzi, które prawdopodobnie już znasz, takich jak histogramy i wykresy rozrzutu.

- **Statystyka inferencyjna**. Dotyczy ona formułowania stwierdzeń wykraczających poza aktualne dane. Może być wskazane zrozumienie jakiegoś konkretnego zjawiska, przewidzenie przyszłych (jeszcze nieobserwowanych) punktów danych lub wybranie spośród kilku konkurujących wyjaśnień tego samego zestawu obserwacji. Podsumowując: statystyka inferencyjna pozwala wyciągać sensowne wnioski z ograniczonego zestawu danych i podejmować świadome decyzje na podstawie wyników analizy.

Idealne połączenie

W książce skoncentrowano się na tym, jak korzystać z bayesowskiej statystyki inferencyjnej, ale będą również stosowane pomysły z zakresu eksploracyjnej analizy danych do podsumowywania, interpretowania, sprawdzania i przekazywania wyników wnioskowania bayesowskiego.

Większość wprowadzających kursów ze statystyki, przynajmniej dla osób nie będących statystykami, jest prowadzona jako zbiór przepisów, które wyglądają następująco: udaj się do statystycznej spiżarni, wybierz jedną puszkę i otwórz ją, dodaj dane do smaku i mieszaj, aż otrzymasz spójną wartość *p*, która najlepiej jest mniejsza niż 0,05. Głównym celem tych kursów jest nauczenie, jak wybrać właściwą puszkę. Nigdy nie lubiłem tego typu podejścia, głównie dlatego, że najczęstszym rezultatem jest grupa zdezorientowanych ludzi niezdolnych do uchwycenia jedności różnych poznanych metod (nawet na poziomie konceptualnym). Przyjmiemy inne podejście: nauczysz się kilku przepisów, ale będą to przepisy domowe, a nie gotowe dania z puszki. Dowiesz się, jak mieszać świeże składniki, które będą odpowiednie dla różnych okazji statystycznych i, co ważniejsze, pozwolą zastosować koncepcje daleko poza przykładami z tej książki.

Przyjęcie takiego podejścia jest możliwe z dwóch następujących powodów:

- **Ontologiczny**. Statystyka to forma modelowania ujednolicona w ramach matematycznych struktur teorii prawdopodobieństwa. Wykorzystanie podejścia probabilistycznego daje jednolity widok na to, co może wydawać się bardzo różnymi metodami. Metody statystyczne i metody uczenia maszynowego wyglądają znacznie bardziej podobnie z perspektywy probabilistycznej.
- **Techniczny**. Nowoczesne oprogramowanie, takie jak PyMC, pozwala praktykom, takim jak Ty i ja, stosunkowo łatwo definiować i rozwiązywać modele. Wiele z tych modeli było nierozwiązywalnych jeszcze kilka lat temu lub wymagało wysokiego poziomu zaawansowania matematycznego i technicznego.

Praca z danymi

Dane są podstawowym składnikiem statystyki i danologii. Pochodzą z kilku źródeł, takich jak eksperymenty, symulacje komputerowe, ankiety i obserwacje terenowe. Jeśli to my odpowiadamy za generowanie lub zbieranie danych, zawsze warto najpierw dokładnie przemyśleć pytania, na jakie chcemy odpowiedzieć, oraz metody, które będą używane, a dopiero potem przystąpić do gromadzenia danych. Istnieje cały dział statystyki zajmujący się zbieraniem danych znany jako planowanie doświadczeń. W erze zalewu danymi

czasami zapominamy, że gromadzenie danych nie zawsze jest tanie. Na przykład, choć prawdą jest, że Wielki Zderzacz Hadronów produkuje setki terabajtów dziennie, jego budowa zajęła lata pracy fizycznej i intelektualnej.

Zasadniczo o procesie generowania danych można myśleć jak o procesie stochastycznym, ponieważ istnieje niepewność ontologiczna, techniczna i/lub epistemiczna, co oznacza, że system jest z natury stochastyczny, występują problemy techniczne powodujące szum bądź ograniczające nas w pomiarach o dowolnej precyzji i/lub istnieją ograniczenia koncepcyjne zasłaniające przed nami szczegóły. Z tych wszystkich powodów zawsze trzeba interpretować dane w kontekście modeli, zarówno mentalnych, jak i formalnych. Dane „nie mówią" same za siebie, lecz za pośrednictwem modeli.

W książce założymy, że dane zostały już zebrane. Nasze dane będą też czyste i uporządkowane, co rzadko zdarza się w rzeczywistości. Przyjmiemy te założenia, aby skupić się na temacie książki. Chcę tylko podkreślić, szczególnie z myślą o osobach rozpoczynających przygodę z analizą danych, że jeśli nawet nie są omawiane w książce, istnieją ważne umiejętności, które należy opanować i rozwijać, aby móc z powodzeniem pracować z danymi.

Bardzo przydatną podczas analizowania danych umiejętnością jest tworzenie kodu w języku programowania, takim jak Python. Manipulowanie danymi jest zwykle konieczne, biorąc pod uwagę to, że żyjemy w chaotycznym świecie z jeszcze bardziej chaotycznymi danymi, a programowanie pomaga zrealizować różne rzeczy. Jeśli nawet masz szczęście i Twoje dane są bardzo czyste i uporządkowane, programowanie nadal będzie bardzo przydatne, ponieważ współczesna statystyka bayesowska jest stosowana głównie z wykorzystaniem takich języków programowania jak Python lub R. Jeżeli chcesz nauczyć się używać Pythona do czyszczenia danych oraz manipulowania nimi, dobre wprowadzenie możesz znaleźć w książce *Python w analizie danych* McKinneya (Helion, 2023).

Modelowanie bayesowskie

Modele to uproszczone opisy danego systemu lub procesu, którym z jakiegoś powodu się interesujemy. Opisy te są celowo zaprojektowane tak, aby uchwycić jedynie najistotniejsze aspekty systemu, a nie wyjaśniać każdy drobny szczegół. Jest to jeden z powodów, dla których bardziej złożony model nie zawsze jest lepszy. Istnieje wiele rodzajów modeli. W tej książce ograniczymy się do modeli bayesowskich. Proces modelowania bayesowskiego można podsumować w następujących trzech krokach:

1. Mając dane i pewne założenia dotyczące tego, jak one mogły zostać wygenerowane, projektujemy model przez łączenie elementów konstrukcyjnych znanych jako **rozkłady prawdopodobieństwa**. Najczęściej modele te są sporymi przybliżeniami, ale zazwyczaj to wszystko, co jest niezbędne.
2. Używamy twierdzenia Bayesa, aby dodać dane do naszych modeli i wyprowadzić logiczne konsekwencje połączenia danych z naszymi założeniami. Mówimy, że **warunkujemy** model na podstawie naszych danych.
3. Oceniamy model i jego prognozy według różnych kryteriów, w tym danych, naszej wiedzy na określony temat, a czasami drogą porównania z innymi modelami.

Ogólnie rzecz biorąc, wykonujemy te trzy kroki w sposób iteracyjny i nieliniowy. W każdym momencie można cofnąć się do poprzednich kroków: być może popełniliśmy głupi błąd podczas tworzenia kodu, znaleźliśmy sposób na zmianę i ulepszenie modelu lub zdaliśmy sobie sprawę, że trzeba dodać więcej danych bądź zebrać innego rodzaju dane.

Modele bayesowskie są również znane jako **modele probabilistyczne**, ponieważ są budowane z użyciem prawdopodobieństw. Dlaczego prawdopodobieństwa? Wynika to z tego, że są one bardzo użytecznym narzędziem do modelowania niepewności. Dysponujemy nawet dobrymi argumentami, aby stwierdzić, że są właściwą koncepcją matematyczną. Wybierzmy się zatem na spacer przez *ogród rozwidlających się ścieżek* (Borges, 1944).

Wprowadzenie do prawdopodobieństwa dla praktyków metod bayesowskich

W tym podrozdziale zostanie omówionych kilka ogólnych i ważnych koncepcji, które są kluczowe dla lepszego zrozumienia metod bayesowskich. Dodatkowe zagadnienia związane z prawdopodobieństwem będę wprowadzać lub rozwijać w kolejnych rozdziałach, w miarę jak będą nam potrzebne. Jeśli jednak chodzi o szczegółowe studium teorii prawdopodobieństwa, polecam zwłaszcza książkę *Introduction to Probability* Blitzsteina (2019). Ci, którzy są już zaznajomieni z podstawowymi elementami teorii prawdopodobieństwa, mogą pominąć ten podrozdział lub przeczytać go pobieżnie.

Przestrzeń prób i zdarzenia

Załóżmy, że prowadzimy badanie, aby sprawdzić, jak ludzie oceniają pogodę w swojej okolicy. Zapytaliśmy trzy osoby, czy lubią słoneczną pogodę, przy czym możliwe odpowiedzi to „tak” lub „nie”. Przestrzeń prób wszystkich możliwych wyników można oznaczyć literą *S*. Składa się ona z ośmiu możliwych kombinacji:

S = {(tak, tak, tak), (tak, tak, nie), (tak, nie, tak), (nie, tak, tak), (tak, nie, nie), (nie, tak, nie), (nie, nie, tak), (nie, nie, nie)}

W tym przypadku każdy element przestrzeni prób reprezentuje odpowiedzi trzech osób w kolejności, w jakiej zostały zapytane. Na przykład element „(tak, nie, tak)” oznacza, że pierwsza i trzecia osoba odpowiedziały „tak”, natomiast druga osoba odpowiedziała „nie”.

Zdarzenia można definiować jako podzbiory przestrzeni prób. Na przykład zdarzenie *A* występuje wtedy, gdy wszystkie trzy osoby odpowiedziały „tak”:

A = {(tak, tak, tak)}

Podobnie można zdefiniować zdarzenie *B* jako sytuację, w której przynajmniej jedna osoba odpowiedziała „nie”:

B = {(tak, tak, nie), (tak, nie, tak), (nie, tak, tak), (tak, nie, nie), (nie, tak, nie), (nie, nie, tak), (nie, nie, nie)}

Prawdopodobieństwa można wykorzystywać jako miarę tego, jak prawdopodobne są te zdarzenia. Zakładając, że wszystkie zdarzenia są jednakowo prawdopodobne, prawdopodobieństwo zdarzenia *A*, czyli sytuacji, w której wszystkie trzy osoby odpowiedziały „tak", wynosi:

$$P(A) = \frac{\text{liczba wyników w } A}{\text{łączna liczba wyników w } S}$$

W tym przypadku istnieje tylko jeden wynik w *A*, a w *S* jest osiem wyników. W związku z tym prawdopodobieństwo *A* wynosi:

$$P(A) = \frac{1}{8} = 0{,}125$$

Podobnie można obliczyć prawdopodobieństwo zdarzenia *B*, które polega na tym, że przynajmniej jedna osoba odpowiedziała „nie". Ponieważ w zdarzeniu tym jest siedem wyników, a w przestrzeni *S* znajduje się osiem wyników, prawdopodobieństwo zdarzenia *B* wynosi:

$$P(B) = \frac{7}{8} = 0{,}875$$

Założenie, że wszystkie zdarzenia są jednakowo prawdopodobne, to tylko szczególny przypadek, który ułatwia obliczanie prawdopodobieństw. Określa się to mianem naiwnej definicji prawdopodobieństwa, ponieważ jest ona ograniczona i opiera się na mocnych założeniach. Jest ona jednak nadal przydatna, jeśli używamy jej ostrożnie. Na przykład nieprawdą jest, że wszystkie pytania z odpowiedziami „tak" lub „nie" mają szansę „pół na pół". Oto inny przykład: jakie jest prawdopodobieństwo zobaczenia fioletowego konia? Właściwa odpowiedź może się bardzo różnić w zależności od tego, czy mowa jest o naturalnym kolorze prawdziwego konia, konia z kreskówki, konia przystrojonego na paradzie itd. W każdym razie niezależnie od tego, czy zdarzenia są jednakowo prawdopodobne, czy nie, prawdopodobieństwo całej przestrzeni prób zawsze równa się 1. Można to sprawdzić, obliczając:

$$P(S) = \frac{\text{liczba wyników w } S}{\text{łączna liczba wyników w } S}$$

1 to największa wartość, jaką może przyjąć prawdopodobieństwo. Stwierdzenie, że $P(S) = 1$, oznacza, że *S* jest nie tylko bardzo prawdopodobne, ale pewne. Jeśli wszystko, co może się zdarzyć, jest zdefiniowane przez przestrzeń *S*, to *S* będzie mieć miejsce.

Jeżeli zdarzenie jest niemożliwe, jego prawdopodobieństwo wynosi 0. Zdefiniujmy zdarzenie *C* jako polegające na tym, że trzy osoby mówią słowo „banan":

C = {(banan, banan, banan)}

Ponieważ zdarzenie *C* nie jest częścią przestrzeni *S*, z definicji nie może mieć miejsca. Można to sobie wyobrazić tak, że kwestionariusz z naszego badania zawiera tylko dwa pola: *tak* i *nie*. Z założenia nasze badanie ogranicza wszystkie inne możliwe opcje.

Można wykorzystać fakt, że język Python obsługuje zbiory, i zdefiniować jego funkcję do obliczania prawdopodobieństw zgodnie z ich naiwną definicją (listing 1.1):

Listing 1.1

```
def P(S, A):
    if set(A).issubset(set(S)):
        return len(A)/len(S)
    else:
        return 0
```

Pozostawiłem Ci przyjemność poeksperymentowania z tą funkcją.

Jednym z przydatnych sposobów konceptualizacji prawdopodobieństw jest traktowanie ich jako zachowanych wielkości rozłożonych w całej przestrzeni prób. Oznacza to, że jeśli prawdopodobieństwo jednego zdarzenia wzrasta, prawdopodobieństwo jakiegoś innego zdarzenia lub zdarzeń musi się zmniejszyć tak, aby całkowite prawdopodobieństwo pozostało równe 1. Można to zilustrować prostym przykładem.

Załóżmy, że pytamy jedną osobę, czy jutro będzie padać, a możliwe odpowiedzi to „tak" i „nie". Przestrzeń prób dla możliwych odpowiedzi jest określona przez S = {tak, nie}. Zdarzenie polegające na tym, że jutro będzie padać, reprezentowane jest przez A = {tak}. Jeśli $P(A)$ wynosi 0,5, prawdopodobieństwo dopełnienia zdarzenia A, oznaczone jako $P(A^c)$, również musi wynosić 0,5. Jeżeli z jakiegoś powodu $P(A)$ wzrasta do 0,8, to $P(A^c)$ musi się zmniejszyć do 0,2. Właściwość ta obowiązuje dla zdarzeń rozłącznych, czyli takich, które nie mogą wystąpić jednocześnie. Na przykład nie może jednocześnie jutro *padać* i *nie padać*. Możesz zaprotestować, że może padać rano, a nie padać po południu. Jest to prawda, ale to już inna przestrzeń prób!

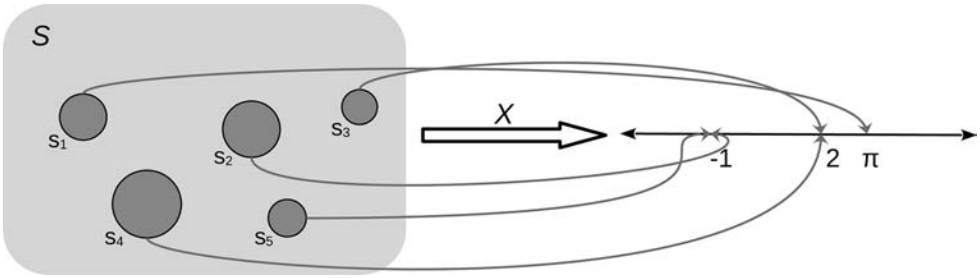
Do tej pory unikałem bezpośredniego definiowania prawdopodobieństw i pokazywałem jedynie niektóre ich właściwości i sposoby ich obliczania. Ogólna definicja prawdopodobieństwa, która sprawdza się w przypadku zdarzeń o nierównych szansach wystąpienia, jest następująca: gdy mamy przestrzeń prób S i zdarzenie A, które jest podzbiorem S, prawdopodobieństwo jest funkcją P, która przyjmuje A jako dane wejściowe i zwraca jako wynik liczbę rzeczywistą z zakresu od 0 do 1. Funkcja P ma pewne ograniczenia określone przez trzy poniższe aksjomaty. Pamiętaj, że aksjomat to stwierdzenie, które przyjmujemy za prawdziwe i którego używamy jako punktu wyjścia w ramach prowadzonego rozumowania:

1. Prawdopodobieństwo zdarzenia jest nieujemną liczbą rzeczywistą.
2. $P(S) = 1$.
3. Jeśli $A1, A2, \ldots$ to zdarzenia rozłączne, co oznacza, że nie mogą wystąpić jednocześnie, wtedy $P(A1, A2, \ldots) = P(A1) + P(A2) + \ldots$.

Gdyby to była książka o teorii prawdopodobieństwa, raczej przeznaczyłbym kilka stron na zademonstrowanie konsekwencji tych aksjomatów i przedstawienie ćwiczeń do manipulowania prawdopodobieństwami. Pomogłoby Ci to uzyskać biegłość w operowaniu prawdopodobieństwami. Jednakże naszym głównym celem nie są te zagadnienia. Moją motywacją do przedstawienia tych aksjomatów jest jedynie pokazanie, że prawdopodobieństwa to dobrze zdefiniowane pojęcia matematyczne z regułami, które nadzorują ich operacje. Są one szczególnym typem funkcji i nie ma w nich żadnej tajemnicy.

Zmienne losowe

Zmienna losowa to funkcja, która odwzorowuje przestrzeń prób na liczby rzeczywiste $\mathbb{R}$ (rysunek 1.1). Załóżmy, że zdarzeniami, które nas interesują, są liczby na kostce. Odwzorowanie jest wtedy bardzo proste: kojarzy się pierwszą ścianę ⚀ z liczbą 1, drugą ścianę ⚁ z liczbą 2 itd. Innym prostym przykładem jest odpowiedź na pytanie: czy jutro będzie padać? Można odwzorować odpowiedź „tak" na wartość 1, a „nie" na wartość 0. Powszechnie, choć nie zawsze, używa się dużej litery dla zmiennych losowych (np. X), a małej litery dla ich wyników (np. x). Jeśli na przykład X to pojedynczy rzut kostką, x reprezentuje konkretną liczbę całkowitą {1, 2, 3, 4, 5, 6}. Można zatem zapisać $P(X = 3)$, aby wskazać prawdopodobieństwo uzyskania wartości 3 przy rzucie kostką. Można też pozostawić x jako nieokreślone. Możliwe jest na przykład zapisanie $P(X = x)$) w celu wskazania prawdopodobieństwa uzyskania jakiejś wartości x lub $P(X \leq x)$, aby wskazać prawdopodobieństwo otrzymania wartości mniejszej lub równej x.



Rysunek 1.1. Zmienna losowa X zdefiniowana dla przestrzeni prób z 5 elementami $\{S_1, \ldots S_5\}$ oraz możliwymi wartościami –1, 2 i π

Możliwość odwzorowania symboli (np. ⚀) lub ciągów znaków (np. „tak") na liczby upraszcza analizę, ponieważ wiemy już, jak wykonywać operacje matematyczne na liczbach. Zmienne losowe są też przydatne, gdyż można je przetwarzać bez bezpośredniego myślenia w kategoriach przestrzeni próby. Cecha ta staje się coraz bardziej istotna, gdy przestrzeń próby staje się bardziej złożona. Na przykład podczas symulowania układów molekularnych trzeba określić pozycję i prędkość każdego atomu. W przypadku złożonych molekuł, takich jak białka, oznacza to konieczność śledzenia tysięcy, milionów, a nawet większej liczby parametrów. Zamiast tego można użyć zmiennych losowych do podsumowania pewnych właściwości układu, takich jak energia całkowita lub względne kąty między określonymi atomami układu.

Jeśli nadal jesteś zdezorientowany, w porządku. Pojęcie zmiennej losowej może brzmieć zbyt abstrakcyjnie na początku, ale w całej książce będzie mnóstwo przykładów, które pomogą Ci utrwalić te idee. Zanim przejdziemy dalej, spróbuję przedstawić analogię, która mam nadzieję okaże się przydatna. Zmienne losowe są użyteczne w podobny sposób jak funkcje języka Python. Często umieszcza się kod w funkcjach, żeby w pojedynczym wywołaniu móc przechowywać, ponownie stosować i *ukrywać* złożone modyfikacje danych. Co więcej, gdy dysponuje się już kilkoma funkcjami, można je czasem łączyć na wiele sposobów (na przykład dodawać wyniki dwóch funkcji lub używać wyniku jednej funkcji jako danych wejściowych innej). Wszystko to można robić bez funkcji, ale abstrakcyjne ujęcie

wewnętrznych mechanizmów nie tylko czyni kod czystszym, lecz także pomaga w zrozumieniu i rozwoju nowych pomysłów. Zmienne losowe odgrywają podobną rolę w statystyce.

Odwzorowanie między przestrzenią prób a $\mathbb{R}$ jest deterministyczne. Nie ma w nim żadnej losowości. Dlaczego zatem nazywa się to zmienną *losową*? Dlatego, że można „*prosić*" zmienną o wartości, a za każdym razem otrzyma się inną liczbę. Losowość wynika z prawdopodobieństwa związanego ze zdarzeniami. Na rysunku 1.1 przedstawiłem *P* jako rozmiar okręgów.

Dwa najczęstsze typy zmiennych losowych to zmienne dyskretne i ciągłe. Nie wgłębiając się w formalną definicję, można stwierdzić, że zmienne dyskretne przyjmują tylko wartości dyskretne i zwykle do ich reprezentacji używa się liczb całkowitych (np. 1, 5, 42). Zmienne ciągłe przyjmują wartości rzeczywiste, dlatego w ich przypadku stosuje się liczby zmiennoprzecinkowe (np. 3,1415, 1,01, 23,4214). To, jakiego typu zmiennej używamy, zależy od problemu. Jeśli zapytamy ludzi o ich ulubiony kolor, otrzymamy odpowiedzi takie jak „czerwony", „niebieski" i „zielony". Jest to przykład dyskretnej zmiennej losowej. Odpowiedzi to kategorie. Nie ma wartości pośrednich między odpowiedziami „czerwony" i „zielony". Z kolei gdy bada się właściwości absorpcji światła, wartości dyskretne, takie jak „czerwony" i „zielony", mogą być nieodpowiednie, a zamiast nich bardziej właściwe może być operowanie długością fali. Wówczas spodziewamy się takich wartości jak 650 nm i 510 nm oraz dowolnej liczby ulokowanej między nimi (włącznie z wartością 579,1).

Dyskretne zmienne losowe i ich rozkłady

Zamiast obliczać prawdopodobieństwo tego, że wszystkie trzy osoby odpowiedziały „tak", lub prawdopodobieństwo wyrzucenia trójki podczas rzutu kostką, możemy być bardziej zainteresowani ustaleniem *listy prawdopodobieństw* dla wszystkich możliwych odpowiedzi lub wszelkich możliwych liczb z kostki. Po obliczeniu takiej listy można ją sprawdzić wizualnie lub użyć jej do obliczenia innych wielkości, takich jak prawdopodobieństwo uzyskania co najmniej jednej odpowiedzi „nie", wyrzucenia liczby nieparzystej albo otrzymania liczby równej lub większej niż 5. Formalną nazwą tej *listy* jest **rozkład prawdopodobieństwa**.

Możliwe jest uzyskanie empirycznego rozkładu prawdopodobieństwa kostki w przypadku rzucenia nią kilka razy i zestawienia tego, ile razy otrzymano poszczególne liczny. Aby przekształcić każdą wartość w prawdopodobieństwo, a całą listę w prawidłowy rozkład prawdopodobieństwa, trzeba dokonać *normalizacji* liczb. Można to osiągnąć, dzieląc wartość otrzymaną dla każdej liczby przez liczbę rzutów kostką.

Rozkłady empiryczne są bardzo przydatne. Będziemy z nich intensywnie korzystać. Zamiast jednak rzucać kostkami ręcznie, użyjemy zaawansowanych metod obliczeniowych, które wykonają za nas ciężką pracę. Nie tylko pozwoli to nam zaoszczędzić czas i uniknąć znużenia, ale sprawi, że bez wysiłku uzyskamy próby z naprawdę złożonych rozkładów. Wyprzedzamy jednak fakty. Priorytetem jest skoncentrowanie się na rozkładach teoretycznych, które są kluczowe w statystyce, ponieważ umożliwiają między innymi konstruowanie modeli probabilistycznych.

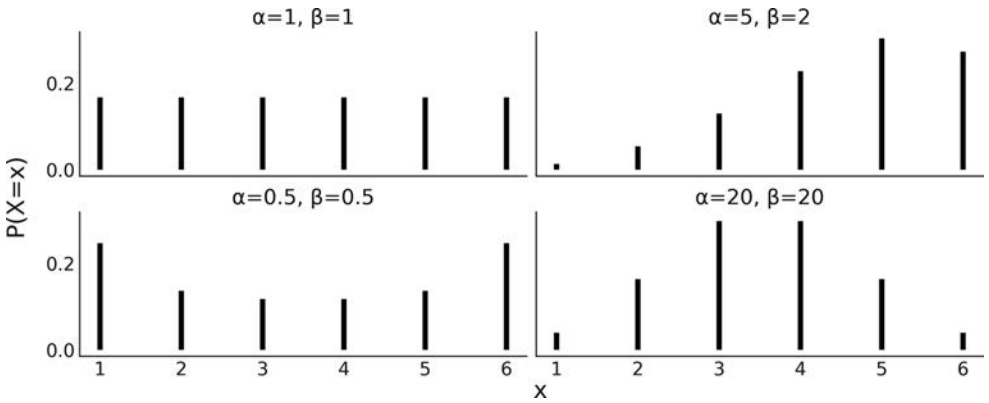
Jak już wspomniałem, nie ma nic losowego ani tajemniczego w zmiennych losowych. Są one po prostu typem funkcji matematycznej. To samo dotyczy teoretycznych rozkładów prawdopodobieństwa. Lubię porównywać rozkłady z okręgami. Ponieważ wszyscy jesteśmy zaznajomieni z okręgami jeszcze przed pójściem do szkoły, nie boimy się ich i nie wydają nam się tajemnicze. Można zdefiniować okrąg jako przestrzeń geometryczną punktów na płaszczyźnie, które znajdują się w jednakowej odległości od innego punktu zwanego środkiem. Idąc dalej, można podać wyrażenie matematyczne dla tej definicji. Jeśli założy się, że położenie środka jest nieistotne, okrąg o promieniu r można opisać jako zbiór wszystkich punktów (x, y) takich, że:

$$x^2 + y^2 = r^2$$

Na podstawie tego wyrażenia można stwierdzić, że dla danego **parametru *r*** okrąg jest całkowicie zdefiniowany. Jest to wszystko, czego potrzeba, aby go narysować i obliczyć takie właściwości jak obwód, który wynosi $2\pi r$.

Zauważmy tutaj, że wszystkie okręgi wyglądają bardzo podobnie do siebie, a ponadto że dowolne dwa okręgi o tej samej wartości r to zasadniczo te same obiekty. Można zatem pomyśleć o rodzinie okręgów, w ramach której każdy element różni się od pozostałych właśnie wartością promienia r.

Jak dotąd wszystko w porządku, ale dlaczego mowa jest o okręgach? Wynika to z tego, że wszystko to można bezpośrednio zastosować do rozkładów prawdopodobieństwa. Zarówno okręgi, jak i rozkłady prawdopodobieństwa mają definiujące je wyrażenia matematyczne, a te zawierają parametry, które można zmieniać, aby zdefiniować wszystkie elementy rodziny rozkładów prawdopodobieństwa. Na rysunku 1.2 pokazałem cztery elementy jednego rozkładu prawdopodobieństwa znanego jako rozkład beta-dwumianowy. Na rysunku 1.2 wysokość słupków reprezentuje prawdopodobieństwo każdej wartości x. Wartości x poniżej 1 lub powyżej 6 mają prawdopodobieństwo 0, ponieważ znajdują się poza nośnikiem rozkładu.



Rysunek 1.2. Cztery elementy rozkładu beta-dwumianowego z parametrami α i β

Oto wyrażenie matematyczne dla rozkładu beta-dwumianowego:

$$fmp(x) = \binom{n}{x} \frac{B(x+\alpha, n-x+\beta)}{B(\alpha,\beta)}$$

fmp to skrót od **funkcji masy prawdopodobieństwa**. W przypadku dyskretnych zmiennych losowych jest to funkcja zwracająca prawdopodobieństwa. Jeśli w zapisie matematycznym występuje zmienna losowa X, wtedy fmp(x) = $P(X = x)$.

Zrozumienie lub zapamiętanie funkcji masy prawdopodobieństwa rozkładu beta-dwumianowego nie ma dla nas żadnego znaczenia. Prezentuję ją tutaj tylko po to, aby pokazać, że to po prostu kolejna funkcja. Zapewniasz jedną liczbę, a dostajesz inną. Nic w tym dziwnego, a przynajmniej w teorii. Muszę przyznać, że do pełnego zrozumienia szczegółów rozkładu beta-dwumianowego trzeba wiedzieć, czym jest $\binom{n}{x}$, znane jako współczynnik dwumianowy, oraz czym jest B, czyli funkcja beta. Nie różni się to jednak zasadniczo od zaprezentowania równania $x^2 + y^2 = r^2$.

Wyrażenia matematyczne mogą być wyjątkowo przydatne, ponieważ są zwięzłe i można je wykorzystać do wyprowadzania właściwości. Czasami jednak może to być zbyt pracochłonne nawet wtedy, gdy dobrze radzimy sobie z matematyką. Wizualizacja może być dobrą alternatywą (lub uzupełnieniem) pomagającą zrozumieć rozkłady prawdopodobieństwa. Nie mogę w pełni pokazać tego na stronie książki, ale jeśli uruchomisz kod z listingu 1.2, otrzymasz interaktywny wykres, który będzie się aktualizował każdorazowo po przesunięciu suwaków parametrów `alpha`, `beta` i `n`:

Listing 1.2

```
pz.BetaBinomial(alpha=10, beta=10, n=6).plot_interactive()
```

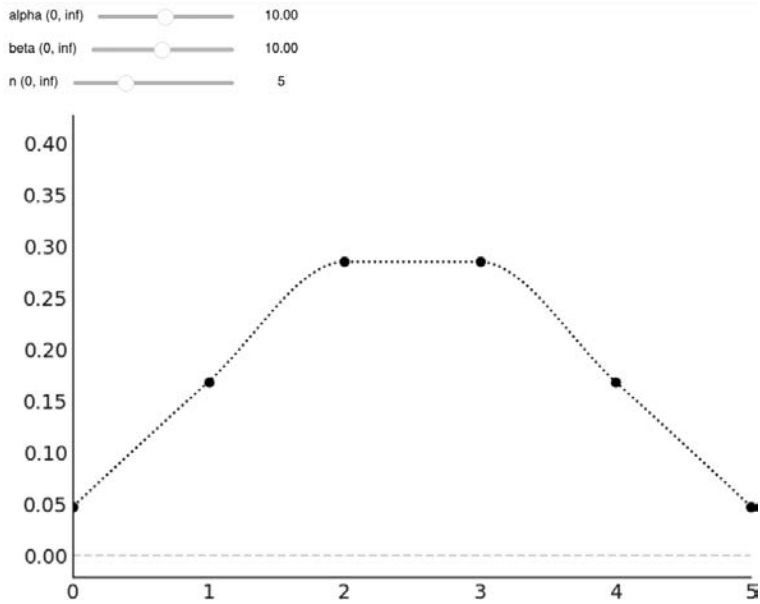
Rysunek 1.3 przedstawia statyczną wersję tego interaktywnego wykresu. Czarne kropki reprezentują prawdopodobieństwa dla każdej wartości zmiennej losowej, natomiast czarna linia przerywana służy jedynie jako pomoc wizualna.

Na osi *x* widoczny jest nośnik rozkładu beta-dwumianowego, czyli wartości, które mogą być następujące: x∈ {0, 1, 2, 3, 4, 5}. Na osi *y* znajdują się prawdopodobieństwa związane z każdą z tych wartości. Pełna lista została przedstawiona w tabeli 1.1.

Zauważ, że dla rozkładu reprezentowanego przez kod `BetaBinomial(alpha=10, beta=10, n=6)` prawdopodobieństwo, iż wartości są spoza zbioru {0, 1, 2, 3, 4, 5}, włączając w to takie wartości jak −1, 0,5, π i 42, wynosi 0.

Wcześniej wspomniałem, że można „*poprosić*" zmienną losową o wartości i za każdym razem otrzyma się inną liczbę. Można przeprowadzić symulację tego za pomocą biblioteki PreliZ (Icazatti i in., 2023) języka Python do określania rozkładów a priori. Weźmy na przykład następujący fragment kodu (listing 1.3).

Zapewni to liczbę całkowitą z przedziału do 0 do 5. A jaką? Tego nie wiadomo! Wykonajmy jednak następujący kod (listing 1.4).



Rysunek 1.3. Wynik działania kodu pz.BetaBinomial(alpha = 10, beta=10, n = 6).plot_interactive()

Tabela 1.1. Prawdopodobieństwa w przypadku kodu pz.BetaBinomial(alpha = 10, beta = 10, n = 6)

Wartość x	Prawdopodobieństwo
0	0,047
1	0,168
2	0,285
3	0,285
4	0,168
5	0,047

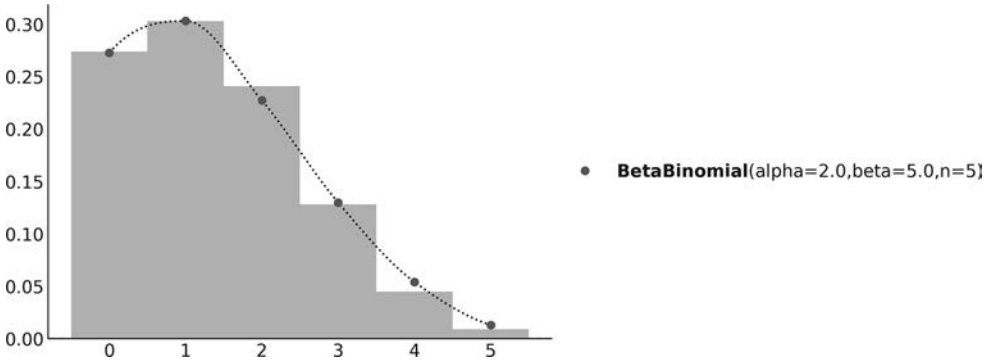
Listing 1.3

```
pz.BetaBinomial(alpha=10, beta=10, n=6).rvs()
```

Listing 1.4

```
1 plt.hist(pz.BetaBinomial(alpha=2, beta=5, n=5).rvs(1000))
2 pz.BetaBinomial(alpha=2, beta=5, n=5).plot_pdf();
```

Otrzymasz coś podobnego jak na rysunku 1.4. Nawet wtedy gdy nie można prognozować następnej wartości zmiennej losowej, można przewidzieć prawdopodobieństwo uzyskania dowolnej, konkretnej wartości, a tym samym, jeśli otrzyma się wiele wartości, można dokonać predykcji ich ogólnego rozkładu.



Rysunek 1.4. Szare kropki reprezentują funkcję masy prawdopodobieństwa próby rozkładu beta-dwumianowego. Kolorem jasnoszarym wyróżniono histogram 1000 losowań z tego rozkładu

W tej książce czasami będą znane parametry danego rozkładu, a ponadto wskazane będzie uzyskanie z niego losowych prób. Innym razem wystąpi odwrotna sytuacja: dostępny będzie zbiór prób i wskazane będzie oszacowanie parametrów rozkładu. Przełączanie się między tymi dwoma scenariuszami stanie się czymś naturalnym w miarę dalszej lektury książki.

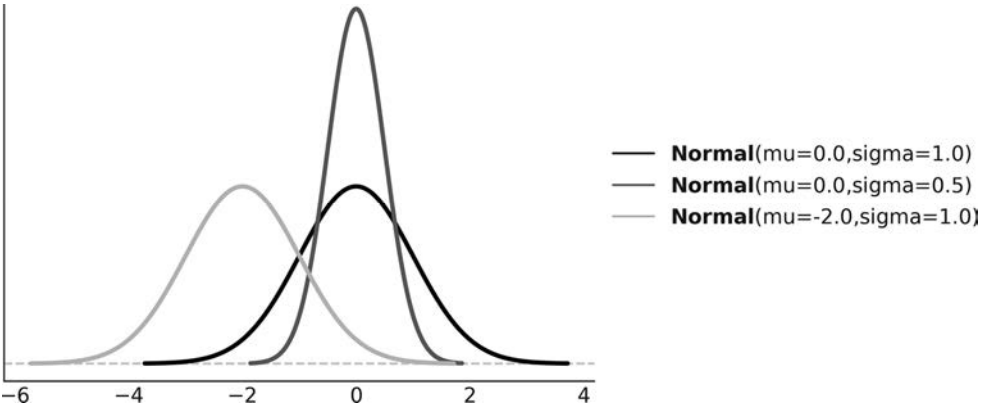
Ciągłe zmienne losowe i ich rozkłady

Prawdopodobnie najszerzej znanym ciągłym rozkładem prawdopodobieństwa jest **rozkład normalny** znany również jako **rozkład Gaussa**. Jego **funkcja gęstości prawdopodobieństwa** ma następującą postać:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

To wyrażenie również przedstawiam jedynie po to, by uchylić zasłonę tajemnicy. Nie ma potrzeby zwracać zbyt dużej uwagi na jego szczegóły, poza tym, że rozkład ten ma dwa parametry: μ (kontroluje położenie szczytu krzywej) oraz σ (decyduje o rozrzucie krzywej). Na rysunku 1.5 przedstawiłem trzy przykłady z rodziny rozkładów Gaussa. Aby dowiedzieć się więcej o tym rozkładzie, polecam obejrzeć film dostępny pod adresem *https://www.youtube.com/watch?v=cy8r7WSuT1I*.

Jeśli uważnie śledziłeś dotychczasową treść, mogłeś zauważyć, że użyłem terminu **funkcja gęstości prawdopodobieństwa** (**fgp**) zamiast **funkcja masy prawdopodobieństwa** (**fmp**). To nie była pomyłka, gdyż są to rzeczywiście dwa różne obiekty. Cofnijmy się o krok i zastanówmy się nad tym. Wynikiem dyskretnego rozkładu prawdopodobieństwa jest prawdopodobieństwo. Wysokość słupków na rysunku 1.2 lub wysokość punktów na rysunku 1.3 to prawdopodobieństwa. Każdy słupek lub punkt nigdy nie będzie wyższy niż 1, a jeśli zsumuje się wszystkie słupki lub punkty, zawsze uzyska się wartość 1. Zróbmy to samo, ale z krzywą z rysunku 1.5. Pierwszą rzeczą godną uwagi jest to, że nie ma słupków ani punktów. Dostępna jest ciągła i gładka krzywa. Można zatem pomyśleć,



Rysunek 1.5. Trzy przykłady z rodziny rozkładów Gaussa

że krzywa składa się z bardzo cienkich słupków, tak cienkich, że przypisuje się jeden słupek każdej rzeczywistej wartości nośnika rozkładów, mierzy się wysokość każdego słupka i wyznacza nieskończoną sumę. Jest to sensowne podejście, prawda?

Tak, ale nie jest od razu oczywiste, co z tego się uzyska. Czy ta suma zapewni dokładnie 1? Czy raczej dużą liczbę? Czy suma jest skończona? Czy wynik zależy od parametrów rozkładu?

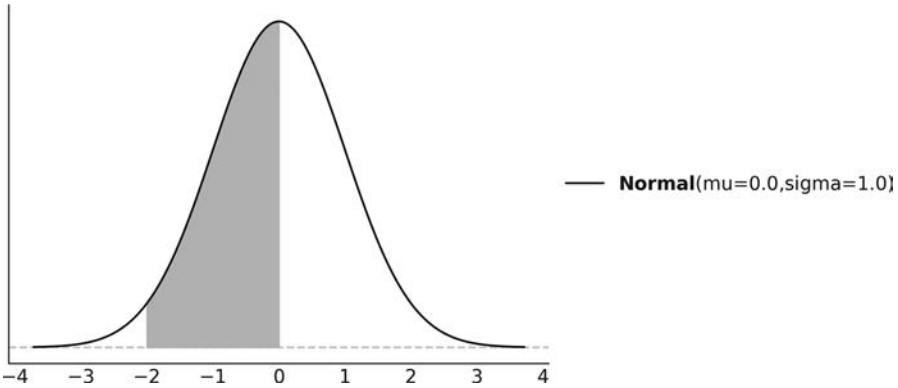
Właściwa odpowiedź na te pytania wymaga teorii miary, a to jest bardzo nieformalne wprowadzenie do prawdopodobieństwa, dlatego nie będziemy zagłębiać się w tę „króliczą norę". Odpowiedź w zasadzie brzmi jednak tak, że w przypadku ciągłej zmiennej losowej prawdopodobieństwo 0 można jedynie przypisać każdej pojedynczej wartości, którą może ona przyjąć. Zamiast tego można przypisać im gęstości, a następnie obliczyć prawdopodobieństwa dla zakresu wartości. A zatem w przypadku rozkładu Gaussa prawdopodobieństwo otrzymania dokładnie liczby –2, czyli takiej, po której następuje nieskończona liczba zer po przecinku, wynosi 0. Jednakże prawdopodobieństwo uzyskania liczby z przedziału od –2 do 0 to jakaś liczba większa od zera i mniejsza od 1. Aby znaleźć dokładną odpowiedź, trzeba wykonać obliczenia za pomocą następującego wzoru:

$$P(a < X < b) = \int_a^b f(x)dx$$

W celu przeprowadzenia tego obliczenia konieczne jest zastąpienie symboli konkretnymi wielkościami. Jeśli $f(x)$ zastąpi się rozkładem Normal(0, 1), gdy $a = -2$ i $b = 0$, uzyska się prawdopodobieństwo $P(-2 < X < 0) \approx 0{,}477$. Odpowiada to zacieniowanemu obszarowi na rysunku 1.6.

Możesz pamiętać, że całkę można przybliżyć, sumując pola prostokątów, a przybliżenie staje się coraz bardziej dokładne w miarę zmniejszania długości podstawy prostokątów (przeczytaj stronę serwisu Wikipedia dotyczącą całki Riemanna). Opierając się na tej koncepcji i korzystając z biblioteki PreliZ, można oszacować prawdopodobieństwo $P(-2 < X < 0)$ w następujący sposób (listing 1.5).

Po zwiększeniu wartości zmiennej `num` uzyska się lepsze przybliżenie.



Rysunek 1.6. Czarna linia reprezentuje funkcję gęstości prawdopodobieństwa rozkładu Gaussa z parametrami mu = 0 i sigma = 1, a szary obszar to prawdopodobieństwo tego, że wartość będzie większa niż –2 i mniejsza niż 0

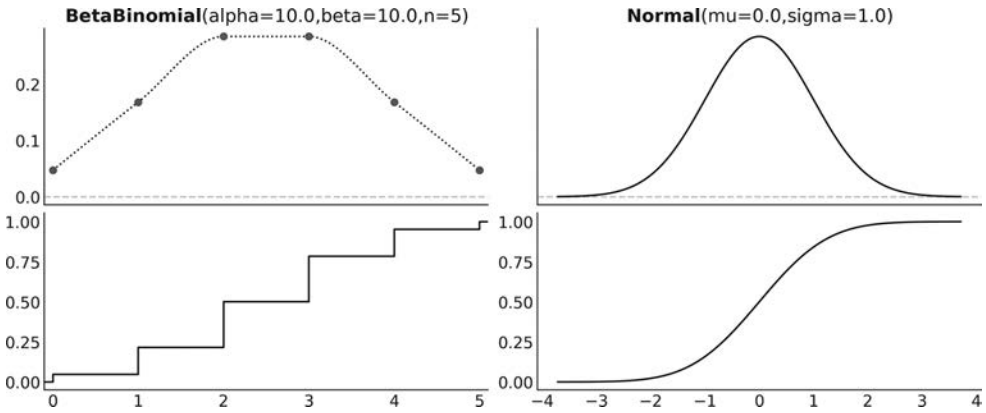
Listing 1.5

```
1 dist = pz.Normal(0, 1)
2 a = -2
3 b = 0
4 num = 10
5 x_s = np.linspace(a, b, num)
6 base = (b-a)/num
7 np.sum(dist.pdf(x_s) * base)
```

Dystrybuanta

Przedstawiłem już funkcję masy prawdopodobieństwa i funkcję gęstości prawdopodobieństwa, ale nie są to jedyne sposoby charakteryzowania rozkładów. Alternatywą jest **dystrybuanta** (ang. *cumulative distribution function*). Dystrybuanta zmiennej losowej X to funkcja F_X określona wzorem $F_X(x) = P(X \leq x)$. Mówiąc prościej, dystrybuanta odpowiada na następujące pytanie: jakie jest prawdopodobieństwo otrzymania liczby mniejszej lub równej x? W pierwszej kolumnie na rysunku 1.7 widoczna jest funkcja masy prawdopodobieństwa i dystrybuanta rozkładu beta-dwumianowego, a w drugiej kolumnie funkcja gęstości prawdopodobieństwa i dystrybuanta rozkładu Gaussa. Zwróć uwagę, jak dystrybuanta wyróżnia się *skokami* w przypadku zmiennej dyskretnej, ale jest gładka dla zmiennej ciągłej. Wysokość każdego skoku reprezentuje prawdopodobieństwo. Wystarczy porównać je z wysokością kropek. Wykres dystrybuanty zmiennej ciągłej można zastosować jako wizualny dowód na to, że prawdopodobieństwa są równe zero dla każdej wartości zmiennej ciągłej. Wystarczy zauważyć, że nie ma *skoków* dla zmiennych ciągłych, co jest równoznaczne z tym, iż wysokość skoków wynosi dokładnie zero.

Samo przyjrzenie się dystrybuancie pozwala łatwiej określić prawdopodobieństwo uzyskania liczby mniejszej niż na przykład 1. Wystarczy przejść do wartości 1 na osi x, przesunąć się w górę aż do przecięcia z czarną linią, a następnie sprawdzić wartość na osi y. Na przykład na rysunku 1.7 w przypadku rozkładu normalnego widać, że wartość mieści się w przedziale od 0,75 do 1. Załóżmy, że wynosi ona w przybliżeniu 0,85. W przypadku



Rysunek 1.7. Funkcja masy prawdopodobieństwa rozkładu beta-dwumianowego z odpowiadającą jej dystrybuantą oraz funkcja gęstości prawdopodobieństwa rozkładu normalnego z powiązaną z nią dystrybuantą

funkcji gęstości prawdopodobieństwa byłoby to znacznie trudniejsze, ponieważ w celu uzyskania odpowiedzi konieczne byłoby porównanie całego pola poniżej wartości 1 z całkowitym polem. Ludzie gorzej radzą sobie z ocenianiem pola powierzchni niż wysokości lub długości.

Prawdopodobieństwo warunkowe

Dla dwóch zdarzeń A i B, gdzie $P(B) > 0$, prawdopodobieństwo zdarzenia A pod warunkiem zajścia zdarzenia B, które zapisujemy jako $P(A \mid B)$, definiujemy następująco:

$$P(A|B) = \frac{P(A, B)}{P(B)}$$

$P(A, B)$ to prawdopodobieństwo tego, że wystąpi zarówno zdarzenie A, jak i zdarzenie B. $P(A \mid B)$ określa się mianem prawdopodobieństwa warunkowego. Jest to prawdopodobieństwo wystąpienia zdarzenia A *pod warunkiem*, że wiemy (lub zakładamy, wyobrażamy sobie, stawiamy hipotezę itp.), iż wystąpiło zdarzenie B. Na przykład prawdopodobieństwo, że chodnik jest mokry, różni się od prawdopodobieństwa, że chodnik jest mokry, jeśli wiadomo, że pada deszcz.

Prawdopodobieństwo warunkowe może być większe, mniejsze lub równe prawdopodobieństwu bezwarunkowemu. Jeśli fakt znajomości zdarzenia B nie zapewnia nam informacji o zdarzeniu A, to $P(A \mid B) = P(A)$. Będzie to prawdą tylko wtedy, gdy zdarzenia A i B są od siebie niezależne. Natomiast gdy z faktu znajomości zdarzenia B wynikają przydatne informacje o zdarzeniu A, prawdopodobieństwo warunkowe może być większe lub mniejsze od prawdopodobieństwa bezwarunkowego, w zależności od tego, czy znajomość zdarzenia B czyni zdarzenie A mniej, czy bardziej prawdopodobnym. Przeanalizujmy prosty przykład z użyciem uczciwej kostki sześciennej. Jakie jest prawdopodobieństwo uzyskania liczby 3 przy rzucie kostką? $P(\text{kostka} = 3) = {}^1/_6$, ponieważ każda z sześciu liczb ma taką samą szansę w przypadku uczciwej kostki sześciennej. Jakie jest prawdopodobieństwo otrzymania liczby 3, jeżeli wiadomo, że otrzymano liczbę nieparzystą? $P(\text{kostka} = 3 \mid$

kostka = {1, 3, 5}) = $^1/_3$, ponieważ w sytuacji gdy wiadomo, że uzyskano liczbę nieparzystą, jedynymi możliwymi liczbami są {1, 3, 5} i każda z nich ma taką samą szansę. I wreszcie: jakie jest prawdopodobieństwo uzyskania liczby 3, jeśli otrzymano liczbę parzystą? Jest to prawdopodobieństwo P(kostka = 3 | kostka = {2, 4, 6}) = 0, ponieważ gdy wiadomo, że liczba jest parzysta, jedyne możliwe liczby to {2, 4, 6}, a zatem uzyskanie trójki nie jest możliwe.

Jak widać w przypadku tych prostych przykładów, poprzez warunkowanie na podstawie obserwowanych danych zmienia się przestrzeń prób. Pytając o prawdopodobieństwo P(kostka = 3), trzeba wyznaczyć przestrzeń prób S = {1, 2, 3, 4, 5, 6}. Gdy jednak *warunkiem jest* otrzymanie liczby parzystej, nowa przestrzeń prób przyjmuje postać T = {2, 4, 6}.

Prawdopodobieństwa warunkowe są sednem statystyki, niezależnie od tego, czy Twój problem dotyczy rzucania kostką, czy budowania samochodów autonomicznych.

Środkowy panel na rysunku 1.8 reprezentuje prawdopodobieństwo $p(A, B)$ za pomocą skali szarości, w przypadku której ciemniejsze kolory oznaczają wyższe gęstości prawdopodobieństwa. Widać, że rozkład łączny jest wydłużony, co wskazuje, że im wyższa wartość zdarzenia A, tym wyższa jest wartość zdarzenia B, i odwrotnie. Znajomość wartości zdarzenia A dostarcza jakichś informacji o wartościach zdarzenia B, i odwrotnie. Na górnym i prawym *marginesie* na rysunku 1.8 znajdują się odpowiednio **rozkłady brzegowe** $p(A)$ i $p(B)$. Aby obliczyć rozkład brzegowy zdarzenia A, bierzemy prawdopodobieństwo $p(A, B)$ i uśredniamy po wszystkich wartościach zdarzenia B. Intuicyjnie można to porównać z użyciem dwuwymiarowego obiektu i rozkładu łącznego i rzutowaniem go na jeden wymiar. Rozkład brzegowy zdarzenia B oblicza się podobnie. Linie przerywane reprezentują **prawdopodobieństwo warunkowe** $p(A \mid B)$ dla trzech różnych wartości zdarzenia B. Otrzymuje się je przez przecięcie prawdopodobieństwa $p(A, B)$ rozkładu łącznego dla danej wartości zdarzenia B. Można to traktować jako rozkład zdarzenia A przy założeniu, że zaobserwowano konkretną wartość zdarzenia B.

Wartości oczekiwane

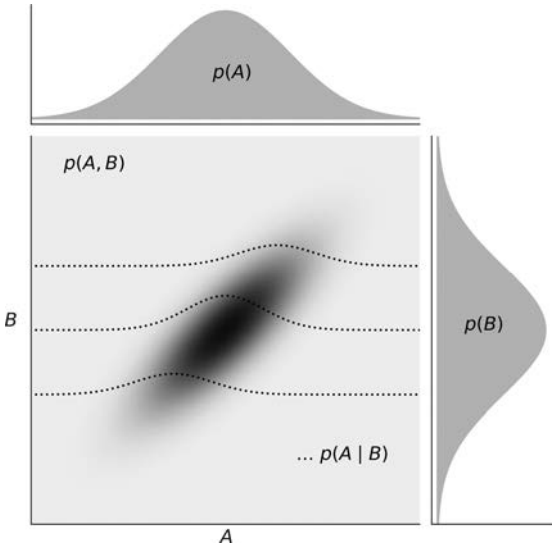
Jeśli X to dyskretna zmienna losowa, jej wartość oczekiwaną można obliczyć za pomocą następującego wzoru:

$$\mathbb{E}(X) = \sum_x xP(X = x)$$

Jest to po prostu średnia lub wartość przeciętna.

Prawdopodobnie jesteś przyzwyczajony do obliczania średnich z prób lub zbiorów liczb, ręcznie, na kalkulatorze lub przy użyciu kodu napisanego w Pythonie. Zauważ jednak, że tutaj nie jest mowa o średniej z kilku liczb, lecz o średniej rozkładu. Po zdefiniowaniu parametrów rozkładu w zasadzie można obliczyć jego wartości oczekiwane. Są to właściwości rozkładu tak samo, jak obwód jest właściwością koła, która zostaje określona po ustaleniu wartości promienia.

Inną wartością oczekiwaną jest wariancja, której można użyć do opisu rozrzutu rozkładu. Wariancja pojawia się *naturalnie* w wielu obliczeniach statystycznych, ale w praktyce często bardziej użyteczne jest stosowanie odchylenia standardowego, które jest



Rysunek 1.8. Przedstawienie relacji między prawdopodobieństwem p(A, B) rozkładu łącznego, prawdopodobieństwami p(A) i p(B) rozkładu brzegowego oraz prawdopodobieństwem warunkowym p(A | B)

pierwiastkiem kwadratowym z wariancji. Powodem jest to, że odchylenie standardowe ma te same jednostki co zmienna losowa.

Średnia i wariancja są często nazywane **momentami** rozkładu. Inne momenty to skośność, która informuje o asymetrii rozkładu, oraz kurtoza, informująca o zachowaniu się jego ogonów lub *wartości skrajnych* (Westfall, 2014). Na rysunku 1.9 pokazałem przykłady różnych rozkładów wraz z ich średnią μ, odchyleniem standardowym σ, skośnością γ i kurtozą κ. Zauważ, że dla niektórych rozkładów pewne momenty mogą być niezdefiniowane lub nieskończone. Biblioteka PreliZ pozwala obliczyć kurtozę nadmiarową, czyli kurtozę –3.

Gdy już przybliżyłem niektóre podstawowe pojęcia i słownictwo związane z teorią prawdopodobieństwa, możemy przejść do etapu, na który wszyscy czekali.

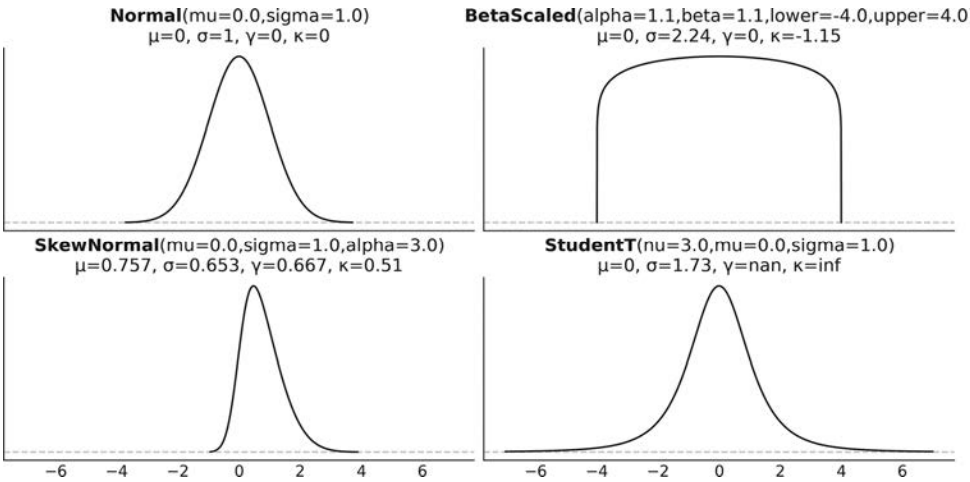
Twierdzenie Bayesa

Bez zbędnego wstępu spójrzmy na twierdzenie Bayesa w całej jego okazałości:

$$p(\theta|Y) = \frac{p(Y|\theta)p(\theta)}{p(Y)}$$

Nie robi ono specjalnego wrażenia, prawda? Wygląda jak wzór ze szkoły podstawowej, a jednak, parafrazując Richarda Feynmana, to wszystko, co musisz wiedzieć o statystyce bayesowskiej. Ustalenie, skąd się bierze twierdzenie Bayesa, pomoże nam zrozumieć jego znaczenie. Zgodnie z regułą iloczynu mamy:

$$p(\theta, Y) = p(\theta|Y)p(Y)$$



Rysunek 1.9. Cztery rozkłady z ich pierwszymi czterema momentami

Można to również zapisać w następującej postaci:

$$p(\theta, Y) = p(Y|\theta)p(\theta)$$

Ponieważ wyrażenia po lewej stronie są równe w obu równaniach, można je połączyć i zapisać następująco:

$$p(\theta|Y)p(Y) = p(Y|\theta)p(\theta)$$

Po zmianie kolejności uzyskuje się twierdzenie Bayesa:

$$p(\theta|Y) = \frac{p(Y|\theta)p(\theta)}{p(Y)}$$

Dlaczego twierdzenie Bayesa jest tak ważne? Dowiedzmy się.

Przede wszystkim określa się w nim, że $p(\theta|\ Y)$ niekoniecznie jest tym samym co $p(Y|\ \theta)$. To bardzo ważny fakt, który łatwo przeoczyć w codziennych sytuacjach nawet osobom przeszkolonym z zakresu statystyki i prawdopodobieństwa. Użyjmy prostego przykładu, żeby wyjaśnić, dlaczego te wielkości niekoniecznie są takie same. Prawdopodobieństwo tego, że dana osoba jest papieżem, jeśli wiadomo, że jest Argentyńczykiem, nie jest tym samym co prawdopodobieństwo bycia Argentyńczykiem, gdy wiadomo, że dana osoba jest papieżem. Ponieważ na świecie żyje około 47 000 000 Argentyńczyków, a tylko jeden z nich był papieżem, mamy p(Papież | Argentyńczyk) ≈ $^1/_{47000000}$ i jednocześnie też p(Argentyńczyk | Papież) = 1.

Jeśli zastąpi się parametr θ słowem „hipoteza”, a literę Y słowem „dane”, twierdzenie Bayesa informuje, jak obliczyć prawdopodobieństwo hipotezy θ dla danych Y. Właśnie w ten sposób w wielu miejscach będzie wyjaśnione twierdzenie Bayesa. Jak jednak przekształcić hipotezę w coś, co można umieścić w twierdzeniu Bayesa? Realizuje się to za pomocą rozkładów prawdopodobieństwa. A zatem, ogólnie rzecz biorąc, nasza hipoteza jest hipotezą w bardzo, ale to bardzo wąskim sensie. Będziemy bardziej precyzyjni, jeśli będziemy mówić o znajdowaniu odpowiedniej wartości dla parametrów w naszych

modelach, czyli parametrów rozkładów prawdopodobieństwa. Nawiasem mówiąc, nie próbuj wiązać parametru θ z takimi stwierdzeniami jak „jednorożce istnieją", chyba że jesteś gotów zbudować realistyczny model probabilistyczny istnienia jednorożców!

Twierdzenie Bayesa stanowi fundament statystyki bayesowskiej. Jak się okaże w rozdziale 2., używanie takich narzędzi jak PyMC zwalnia nas z konieczności jawnego zapisywania twierdzenia Bayesa za każdym razem, gdy buduje się model bayesowski. Niemniej jednak ważne jest, żeby znać nazwy jego części, ponieważ będziemy się do nich stale odwoływać, a ponadto istotne jest rozumienie, co każda część oznacza, gdyż pomoże to w konceptualizacji modeli. Pozwolę sobie zatem zmodyfikować tutaj twierdzenie Bayesa z użyciem etykiet:

$$\overbrace{p(\theta|Y)}^{\text{rozkład a posteriori}} = \frac{\overbrace{p(Y|\theta)}^{\text{funkcja wiarygodności}} \overbrace{p(\theta)}^{\text{rozkład a priori}}}{\underbrace{p(Y)}_{\text{wiarygodność brzegowa}}}$$

Rozkład a priori powinien odzwierciedlać to, co wiadomo o wartości parametru θ przed ujrzeniem danych Y. Jeśli jak Jon Snow nic nie wiemy, moglibyśmy użyć płaskich rozkładów a priori, które nie zawierają zbyt wielu informacji. Ogólnie rzecz biorąc, jak się dowiesz z tej książki, jest lepsze rozwiązanie niż płaskie rozkłady a priori. Używanie tych rozkładów jest powodem, dla którego niektórzy ludzie nadal mówią o statystyce bayesowskiej jako subiektywnej nawet wtedy, gdy rozkłady a priori to po prostu kolejne założenie, jakie przyjmuje się podczas modelowania. W związku z tym są one równie subiektywne (lub obiektywne) jak każde inne założenie (na przykład dotyczące funkcji wiarygodności).

Funkcja wiarygodności określa sposób, w jaki wprowadza się dane do prowadzonej analizy. Jest to wyrażenie prawdopodobieństwa danych przy danych parametrach. W niektórych opracowaniach spotkasz się z „określeniem model próbkowania", „model statystyczny" lub po prostu „model". Pozostaniemy jednak przy terminie „funkcja wiarygodności" oraz będziemy modelować kombinację rozkładów a priori i tej funkcji.

Rozkład a posteriori to wynik analizy bayesowskiej odzwierciedlający wszystko, co wiadomo o problemie (w przypadku posiadanych danych i modelu). Rozkład a posteriori to rozkład prawdopodobieństwa dla parametrów w używanym modelu, a nie pojedyncza wartość. Rozkład ten stanowi równowagę między rozkładem a priori a funkcją wiarygodności. Oto dobrze znany żart: znawca bayesizmu to ktoś, kto trochę oczekując konia i dostrzegając osła, mocno wierzy w to, że zobaczył muła. Doskonałym sposobem na zepsucie dobrego nastroju po usłyszeniu tego żartu jest wyjaśnienie, że jeśli funkcja wiarygodności i rozkłady a priori są mgliste, otrzymasz rozkład a posteriori odzwierciedlający tego rodzaju przekonania dotyczące ujrzenia muła, a nie solidne. W każdym razie podoba mi się ten żart, a także to, jak oddaje on ideę rozkładu a posteriori jako pewnego kompromisu między rozkładem a priori a funkcją wiarygodności. Na poziomie pojęciowym rozkład a posteriori można traktować jak zaktualizowany rozkład a priori w świetle (nowych) danych. W teorii rozkład a posteriori z jednej analizy może zostać użyty jako rozkład a priori na potrzeby nowej analizy (w praktyce może się to okazać trudniejsze). Czyni to analizę bayesowską szczególnie odpowiednią do analizowania danych, które stają się dostępne w sposób sekwencyjny. Jednym z przykładów może być system wczesnego ostrzegania przed klęskami żywiołowymi, który przetwarza dane trybu online pochodzące

ze stacji meteorologicznych i satelitów. Aby uzyskać więcej szczegółów, poczytaj o metodach uczenia maszynowego trybu online.

Ostatni składnik to **wiarygodność brzegowa**, która czasami jest określana mianem **dowodu**. Formalnie wiarygodność ta to prawdopodobieństwo obserwowania danych uśrednione po wszystkich możliwych wartościach, jakie mogą przyjąć parametry (zgodnie z rozkładem a priori). Można to zapisać w następującej postaci: $\int_\Theta p(Y \mid \theta)\mathrm{p}(\theta)\mathrm{d}\theta$. Tak naprawdę nie będziemy się przejmować wiarygodnością brzegową aż do rozdziału 5. Na razie jednak można ją postrzegać jako czynnik normalizujący, który zapewnia, że rozkład a posteriori jest właściwą funkcją masy prawdopodobieństwa lub funkcją gęstości prawdopodobieństwa. Jeśli zignoruje się wiarygodność brzegową, twierdzenie Bayesa można zapisać jako proporcjonalność, co też jest powszechnym sposobem wyrażania go:

$$p(\theta|Y) \propto p(Y|\theta)p(\theta)$$

Zrozumienie dokładnej roli każdego składnika w twierdzeniu Bayesa wymaga trochę czasu i praktyki, a ponadto będzie wiązało się z koniecznością użycia kilku przykładów, ale temu służy reszta książki.

Interpretacja prawdopodobieństwa

Prawdopodobieństwa można interpretować na różne użyteczne sposoby. Na przykład można myśleć, że $P(A) = 0{,}125$ oznacza, iż gdyby powtarzano badanie wiele razy, spodziewano by się, że wszystkie trzy osoby odpowiedzą „tak" w około 12,5% przypadków. Prawdopodobieństwa interpretuje się jako wynik długotrwałych eksperymentów. Jest to bardzo powszechna i użyteczna interpretacja. Nie tylko pomaga ona nam myśleć o prawdopodobieństwach, ale może też dostarczyć empiryczną metodę szacowania prawdopodobieństw. Czy chcemy poznać prawdopodobieństwo eksplozji opony samochodowej napompowanej powietrzem powyżej zaleceń producenta? Wystarczy napompować około 120 opon i można uzyskać dobre przybliżenie. Określa się to zwykle mianem interpretacji częstościowej.

Inna interpretacja prawdopodobieństwa, która zazwyczaj jest nazywana interpretacją subiektywną lub bayesowską, głosi, że prawdopodobieństwa można interpretować jako miary niepewności jednostki wobec zdarzeń. W przypadku tej interpretacji prawdopodobieństwa dotyczą naszego stanu wiedzy o świecie i niekoniecznie opierają się na powtarzanych próbach. W ramach tej definicji prawdopodobieństwa zasadne i naturalne jest pytanie o prawdopodobieństwo życia na Marsie, prawdopodobieństwo tego, że masa elektronu wynosi $9{,}1 \cdot 10^{-31}$ kg, czy prawdopodobieństwo, że 9 lipca 1816 roku był słoneczny dzień w Buenos Aires. Są to zdarzenia jednorazowe. Nie można odtworzyć miliona wszechświatów, każdy z jednym Marsem, i sprawdzić, w ilu z nich rozwinie się życie. Oczywiście można to zrealizować jako eksperyment myślowy. Dzięki temu długookresowe częstości nadal mogą stanowić poprawną konstrukcję pojęciową.

Czasami interpretację bayesowską prawdopodobieństw opisuje się w kategoriach osobistych przekonań. Nie podoba mi się to. Myślę, że może to prowadzić do niepotrzebnego zamieszania, ponieważ przekonania są przeważnie kojarzone z pojęciem wiary lub nieuzasadnionych twierdzeń. Takie skojarzenie może łatwo prowadzić ludzi do myślenia, że

prawdopodobieństwa bayesowskie, a przez to statystyka bayesowska, są mniej obiektywne lub mniej naukowe niż alternatywy. Uważam też, że sprzyja to tworzeniu zamieszania wokół roli wcześniejszej wiedzy w statystyce i sprawia, iż ludzie myślą, że bycie obiektywnym lub racjonalnym oznacza nieużywanie wcześniejszych informacji.

Metody bayesowskie są tak samo subiektywne (lub obiektywne) jak każda inna dostępna i dobrze ugruntowana metoda naukowa. Pozwolę sobie wyjaśnić to na przykładzie: życie na Marsie istnieje albo nie. Wynik jest binarny, czyli pytanie ma odpowiedź „tak" lub „nie". Skoro jednak nie jesteśmy pewni tego faktu, rozsądnym działaniem jest próba ustalenia, jak prawdopodobne jest życie na Marsie. Aby odpowiedzieć na to pytanie, każda uczciwa i naukowo nastawiona osoba wykorzysta wszystkie istotne dane geofizyczne o Marsie, całą istotną wiedzę biochemiczną o warunkach niezbędnych do życia itd. Odpowiedź będzie z konieczności dotyczyć naszego epistemicznego stanu wiedzy, a inni mogą się nie zgadzać, a nawet uzyskać różne prawdopodobieństwa. Przynajmniej jednak wszyscy będą mogli w zasadzie przedstawić argumenty popierające zastosowane przez siebie dane, metody, decyzje modelowe itd. Naukowa i racjonalna debata o życiu na Marsie nie dopuszcza takich *argumentów* jak „anioł powiedział mi o malutkich, zielonych stworzeniach". Statystyka bayesowska jest jednak tylko procedurą formułowania naukowych stwierdzeń przy użyciu prawdopodobieństw będących elementami konstrukcyjnymi.

Prawdopodobieństwo, niepewność i logika

Prawdopodobieństwa pomagają określić niepewność w sposób ilościowy. Jeśli nie są dostępne informacje o problemie, rozsądne jest stwierdzenie, że każde możliwe zdarzenie jest jednakowo prawdopodobne. Jest to równoznaczne z przypisaniem tego samego prawdopodobieństwa każdemu możliwemu zdarzeniu. Przy braku informacji nasza niepewność jest maksymalna — i nie używam tego określenia w znaczeniu potocznym. Jest to coś, co można obliczyć za pomocą prawdopodobieństw. Jeżeli wiadomo natomiast, że niektóre zdarzenia są bardziej prawdopodobne, można to formalnie przedstawić przez przypisanie wyższego prawdopodobieństwa tym zdarzeniom, a niższego pozostałym. Zauważ, że kiedy w języku statystyki jest mowa o zdarzeniach, nie ograniczamy się do rzeczy, które mogą się zdarzyć, takich jak uderzenie asteroidy w Ziemię lub 60. urodziny mojej cioci. Zdarzenie to po prostu dowolna z możliwych wartości (lub podzbiór wartości), jakie może przyjąć zmienna (na przykład zdarzenie polegające na tym, że masz więcej niż 30 lat, cena tortu Sachera czy liczba rowerów, jakie zostaną sprzedane w przyszłym roku na całym świecie).

Pojęcie prawdopodobieństwa jest również związane z zagadnieniem logiki. W logice klasycznej można mieć tylko stwierdzenia, które przyjmują wartości „prawda" lub „fałsz". Zgodnie z bayesowską definicją prawdopodobieństwa pewność jest tylko szczególnym przypadkiem: prawdziwe stwierdzenie ma prawdopodobieństwo równe 1, a prawdopodobieństwo fałszywego stwierdzenia wynosi 0. Przypisalibyśmy prawdopodobieństwo równe 1 stwierdzeniu, że na Marsie istnieje życie, dopiero po uzyskaniu rozstrzygających danych wskazujących, że coś rośnie, rozmnaża się i przejawia inne aktywności, które kojarzy się z organizmami żywymi.

Zauważ jednak, że przypisanie prawdopodobieństwa równego 0 jest trudniejsze, ponieważ zawsze można pomyśleć, że istnieje jakieś niezbadane miejsce na Marsie, że popełniono błędy w niektórych eksperymentach albo że istnieje kilka innych powodów, które mogłyby prowadzić nas do fałszywego przekonania, iż życie nie istnieje na Marsie, nawet wtedy, gdy tak nie jest. Wiąże się to z regułą Cromwella, zgodnie z którą prawdopodobieństwa 0 lub 1 powinno się zarezerwować dla logicznie prawdziwych lub fałszywych stwierdzeń. Co jest dość ciekawe, można wykazać, że jeśli chcemy rozszerzyć logikę o niepewność, trzeba używać prawdopodobieństw i teorii prawdopodobieństwa.

Jak się wkrótce okaże, twierdzenie Bayesa to po prostu logiczna konsekwencja reguł prawdopodobieństwa. Można zatem myśleć o statystyce bayesowskiej jako o rozszerzeniu logiki, które jest przydatne zawsze wtedy, gdy ma się do czynienia z niepewnością. Jednym ze sposobów uzasadnienia stosowania metody bayesowskiej jest uznanie, że niepewność jest powszechna. Zasadniczo trzeba radzić sobie z niepełnymi lub „zaszumionymi" danymi. Ponadto jesteśmy z natury ograniczeni przez ukształtowany drogą ewolucji mózg naczelnych itd.

Etos bayesowski

Prawdopodobieństwa służą do mierzenia niepewności, jaką mamy w odniesieniu do parametrów, a twierdzenie Bayesa jest mechanizmem prawidłowego aktualizowania tych prawdopodobieństw w świetle nowych danych, co miejmy nadzieję zmniejsza naszą niepewność.

Wnioskowanie dotyczące jednego parametru

Gdy już wiadomo, czym jest statystyka bayesowska, zobaczmy na prostym przykładzie, jak ją stosować. Zaczniemy od wnioskowania dotyczącego jednego, nieznanego parametru.

Problem rzutu monetą

Problem rzutu monetą (czyli model beta-dwumianowy, jeśli chcesz zabłysnąć na przyjęciach) to klasyczny problem w statystyce, który wygląda następująco: rzucamy monetą kilka razy i zapisujemy, ile razy wypadł orzeł, a ile razy reszka. Na podstawie tych danych staramy się odpowiedzieć na pytania typu: czy moneta nie budzi wątpliwości? Ewentualnie bardziej ogólnie: jak bardzo moneta jest obciążona? Choć ten problem może wydawać się mało ciekawy, nie powinno się go lekceważyć.

Problem rzucania monetą to doskonały przykład do nauki podstaw statystyki bayesowskiej, ponieważ jest to prosty model, który można łatwo rozwiązać i obliczyć. Poza tym wiele rzeczywistych problemów składa się z binarnych i wzajemnie wykluczających się wyników, takich jak 0 lub 1, pozytywny lub negatywny, parzyste lub nieparzyste, spam lub niespam, hotdog lub niehotdog, kot lub pies, bezpieczne lub niebezpieczne oraz zdrowe

lub niezdrowe. A zatem nawet wtedy, gdy jest mowa o monetach, model ten ma zastosowanie do każdego z tych problemów. Aby oszacować obciążenie monety i ogólnie odpowiedzieć na jakiekolwiek pytania w kontekście bayesowskim, niezbędne będą dane i model probabilistyczny. W tym przykładzie założymy, że już rzucono monetą kilka razy i istnieje zapis liczby zaobserwowanych orzełków, dlatego etap zbierania danych jest już za nami. Stworzenie modelu będzie wymagało nieco więcej wysiłku. Ponieważ jest to nasz pierwszy model, wprost zapiszemy twierdzenie Bayesa i wykonamy wszystkie niezbędne obliczenia matematyczne (nie obawiaj się, obiecuję, że będzie to bezbolesne), a następnie będziemy postępować bardzo powoli. Począwszy od rozdziału 2., będziemy używać biblioteki PyMC i komputera do automatycznego wykonywania obliczeń.

Pierwszą rzeczą do zrobienia jest uogólnienie pojęcia obciążenia. Stwierdzimy, że moneta z obciążeniem 1 zawsze wypadnie orzełkiem, moneta z obciążeniem 0 każdorazowo wypadnie reszką, a moneta z obciążeniem 0,5 będzie wypadać orzełkiem przez połowę przypadków, a reszką przez drugą połowę. Do reprezentowania obciążenia zostanie użyty parametr θ, a do reprezentacji całkowitej liczby orzełków w kilku rzutach posłuży zmienna Y. Zgodnie z twierdzeniem Bayesa trzeba określić rozkład a priori $p(\theta)$ i funkcję wiarygodności $p(Y \mid \theta)$, których będziemy używać. Zacznijmy od funkcji wiarygodności.

Wybór funkcji wiarygodności

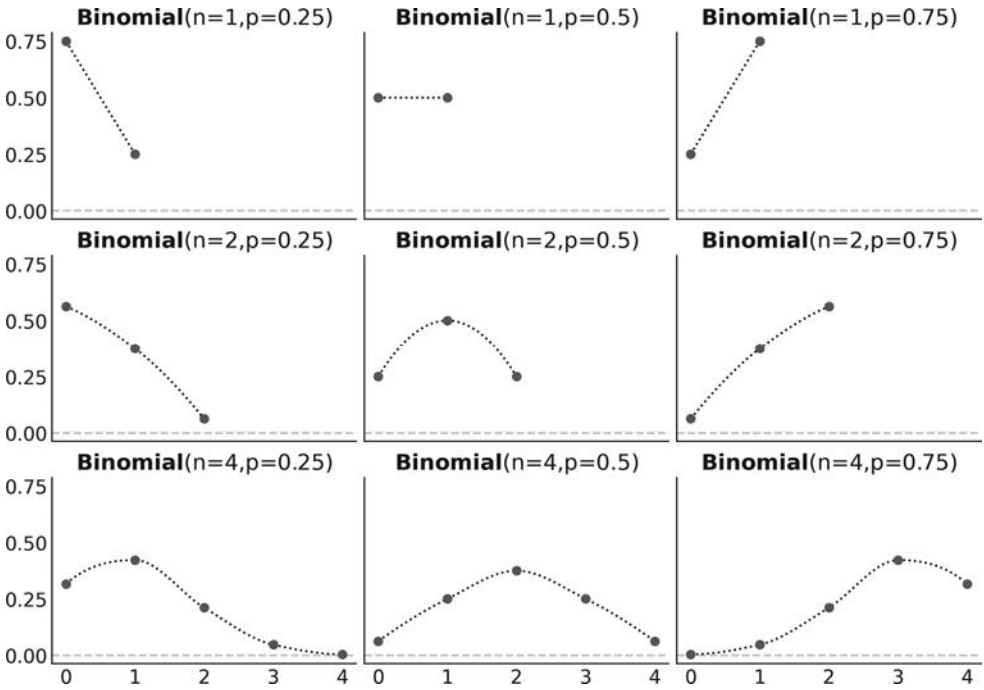
Załóżmy, że możliwe są tylko dwa wyniki, czyli orzełek lub reszka, a ponadto przyjmijmy, iż jeden rzut monetą nie wpływa na inne rzuty. Zakładamy zatem, że rzuty monetą są od siebie niezależne. Dodatkowo przyjmiemy, że wszystkie rzuty monetą pochodzą z tego samego rozkładu. Dzięki temu zmienna losowa rzutu monetą jest przykładem zmiennej **niezależnej i identycznie rozłożonej**. Mam nadzieję, że zgodzisz się, iż są to bardzo rozsądne założenia z punktu widzenia naszego problemu. Przy tych założeniach dobrym kandydatem na funkcję wiarygodności jest rozkład dwumianowy:

$$p(\mathrm{Y} \mid \theta) = \underbrace{\frac{N!}{y!\,(\mathrm{N}-\mathrm{y})!}}_{\text{stała normalizująca}} \theta^{y}(1-\theta)^{N-y}$$

Jest to rozkład dyskretny, który zwraca prawdopodobieństwo uzyskania y orzełków (lub ogólnie sukcesów) w przypadku N rzutów monetą (albo ogólnie prób lub eksperymentów) przy ustalonej wartości θ.

Rysunek 1.10 przedstawia dziewięć rozkładów z rodziny dwumianowej. Każdy wykres ma własną legendę wskazującą wartości parametrów. Zauważ, że w przypadku tych wykresów nie pominąłem wartości na osi y. Postąpiłem tak, abyś mógł sam sprawdzić, że jeśli zsumuje się wysokość wszystkich słupków, otrzyma się wartość 1. Oznacza to, że w odniesieniu do rozkładów dyskretnych wysokość słupków reprezentuje rzeczywiste prawdopodobieństwa.

Rozkład dwumianowy to rozsądny wybór w przypadku funkcji wiarygodności. Widać, że parametr θ wskazuje prawdopodobieństwo wyrzucenia reszki podczas rzutu monetą. Łatwiej to zauważyć, gdy $N = 1$, ale jest to prawdziwe dla dowolnej wartości N. Wystarczy porównać dla $y = 1$ (reszka) wartość parametru θ z wysokością słupka.



Rysunek 1.10. Dziewięć rozkładów z rodziny dwumianowej

Wybór rozkładu a priori

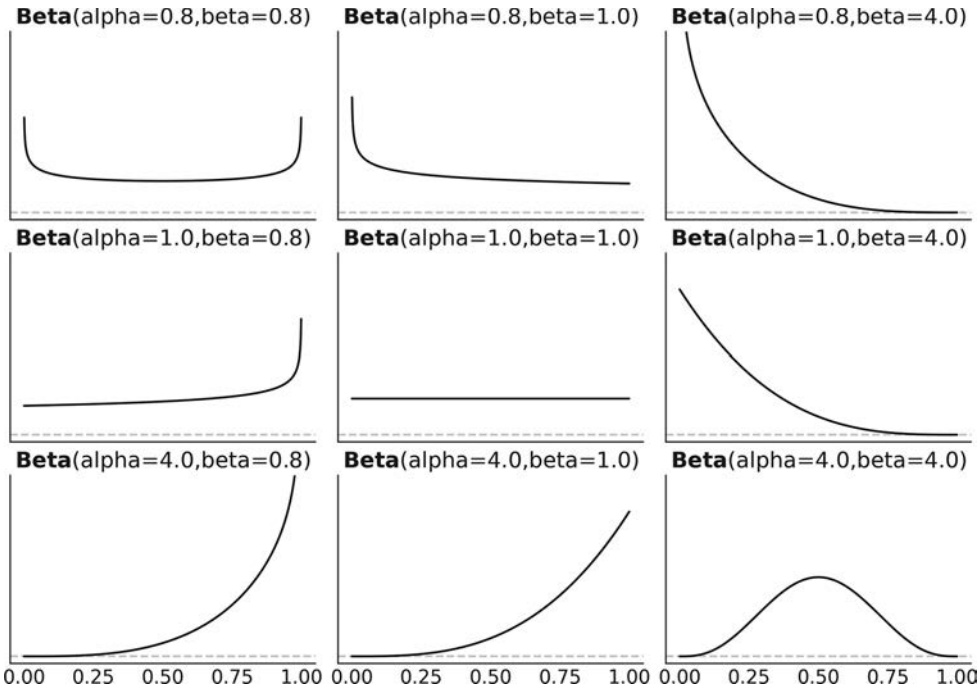
Jako rozkład a priori zostanie użyty rozkład beta, który jest bardzo popularny w statystyce bayesowskiej i ma następującą postać:

$$p(\theta) = \underbrace{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)}}_{\text{stała normalizująca}} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Jeśli przyjrzysz się uważnie, dostrzeżesz, że rozkład beta wygląda podobnie do rozkładu dwumianowego, z wyjątkiem pierwszego składnika. Γ to duża litera grecka gamma reprezentująca funkcję gamma, ale tak naprawdę nie jest to istotne. Ważne jest dla nas to, że pierwszy składnik to stała normalizacyjna, która zapewnia, iż całka z rozkładu równa się 1. Na podstawie powyższego wzoru można stwierdzić, że rozkład beta ma dwa parametry: α i β. Na rysunku 1.11 przedstawiłem dziewięć przykładów z rodziny rozkładów beta.

Podoba mi się rozkład beta i wszystkie kształty, które można z niego uzyskać, ale dlaczego używamy go w naszym modelu? Istnieje wiele powodów, dla których warto stosować rozkład beta w tym i innych problemach.

Jednym z nich jest to, że rozkład beta jest ograniczony do przedziału od 0 do 1, tak samo jak parametr θ. Ogólnie rzecz biorąc, rozkładu beta używa się, gdy zamierza się modelować proporcje zmiennej dwumianowej. Kolejnym powodem jest jego wszechstronność.



Rysunek 1.11. Dziewięć przykładów z rodziny rozkładów beta

Jak widać na rysunku 1.11, rozkład przyjmuje różne kształty (wszystkie ograniczone są do przedziału [0, 1]), w tym rozkład jednostajny, rozkłady podobne do normalnego oraz rozkłady w kształcie litery U.

Trzecim powodem jest to, że rozkład beta jest sprzężonym rozkładem a priori rozkładu dwumianowego (który został użyty jako funkcja wiarygodności). Sprzężony rozkład a priori dla funkcji wiarygodności to rozkład a priori, który w połączeniu z daną funkcją wiarygodnością zwraca rozkład a posteriori o tej samej postaci funkcyjnej co rozkład a priori. Mówiąc prościej, za każdym razem gdy używa się rozkładu beta jako rozkładu a priori i rozkładu dwumianowego jako funkcji wiarygodności, otrzymuje się rozkład beta jako rozkład a posteriori. Istnieją inne pary sprzężonych rozkładów a priori. Na przykład rozkład normalny jest sprzężonym rozkładem a priori samego siebie. Przez wiele lat analiza bayesowska była ograniczona do stosowania sprzężonych rozkładów a priori. Sprzężenie zapewnia matematyczną łatwość obliczeń rozkładu a posteriori, co jest ważne, biorąc pod uwagę, że powszechnym problemem w statystyce bayesowskiej jest otrzymanie rozkładu a posteriori, którego nie można rozwiązać analitycznie. Stanowiło to poważną przeszkodę przed opracowaniem odpowiednich metod obliczeniowych do rozwiązywania problemów probabilistycznych. Począwszy od rozdziału 2., zaczniesz się uczyć używać nowoczesnych metod obliczeniowych do rozwiązywania problemów bayesowskich, niezależnie od tego, czy wybierze się sprzężony rozkład a priori, czy nie.

Wyznaczanie rozkładu a posteriori

Przypomnijmy sobie, że zgodnie z twierdzeniem Bayesa rozkład a posteriori jest proporcjonalny do funkcji wiarygodności pomnożonej przez rozkład a priori. W ramach rozpatrywanego problemu trzeba zatem pomnożyć rozkłady dwumianowy i beta:

$$p(\theta \mid Y) = \overbrace{\frac{N!}{y!\,(N-y)!}\theta^{y}(1-\theta)^{N-y}}^{\text{funkcja wiarygodności}} \overbrace{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}}^{\text{rozkład a priori}}$$

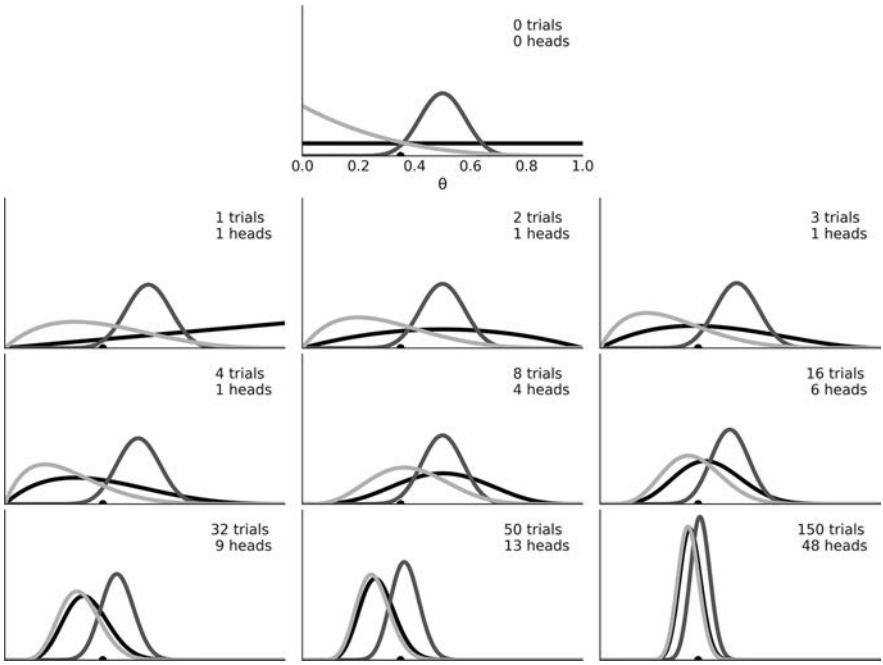
Wyrażenie to można uprościć, usuwając wszystkie składniki, które nie są zależne od parametru θ. Dzięki temu wyniki pozostaną poprawne. W związku z tym można użyć następującego zapisu:

$$p(\theta \mid Y) \overbrace{\propto \theta^{y}(1-\theta)^{N-y}}^{\text{funkcja wiarygodności}} \overbrace{\theta^{\alpha-1}(1-\theta)^{\beta-1}}^{\text{rozkład a priori}}$$

Jeśli zmienimy kolejność składników i zauważymy, że ma to postać rozkładu beta, otrzymamy:

$$p(\theta \mid Y) = Beta(\alpha_{priori} + y, \beta_{priori+N-y})$$

Na podstawie tego wyrażenia analitycznego można obliczyć rozkład a posteriori. Rysunek 1.12 przedstawia wyniki dla trzech rozkładów a priori i różnej liczby prób. Blok kodu z listingu 1.6 prezentuje podstawowy sposób generowania zawartości rysunku 1.12 (z pominięciem kodu niezbędnego do tworzenia wykresów).



Rysunek 1.12. Pierwszy wykres przedstawia trzy rozkłady a priori. Pozostałe wykresy prezentują kolejne aktualizacje w miarę otrzymywania nowych danych

Listing 1.6

```
n_trials = [0, 1, 2, 3, 4, 8, 16, 32, 50, 150]
n_heads = [0, 1, 1, 1, 1, 4, 6, 9, 13, 48]
beta_params = [(1, 1), (20, 20), (1, 4)]

x = np.linspace(0, 1, 2000)
for idx, N in enumerate(n_trials):
    y = n_heads[idx]
    for (α_prior, β_prior) in beta_params:
        posterior = pz.Beta(α_prior + y, β_prior + N - y).pdf(x)
```

Na pierwszym wykresie na rysunku 1.12 występuje zero prób, dlatego trzy krzywe reprezentują następujące rozkłady a priori:

- Jednorodny rozkład a priori (czarny): reprezentuje wszystkie możliwe wartości obciążenia jako równie prawdopodobne a priori.
- Rozkład a priori podobny do gaussowskiego (ciemnoszary): jest wyśrodkowany i skoncentrowany wokół wartości 0,5, dlatego rozkład ten jest zgodny z informacją wskazującą, że w przypadku monety są mniej więcej takie same szanse na wypadnięcie orzełka lub reszki. Można również stwierdzić, że ten rozkład a priori jest zgodny z wiedzą o tym, iż monety nie budzą wątpliwości.
- Skośny rozkład a priori (jasnoszary): przypisuje większość wagi wynikom faworyzującym skrajne wartości.

Pozostałe wykresy przedstawiają rozkłady a posteriori dla kolejnych prób. Liczba prób (lub rzutów monetą) oraz liczba orzełków są widoczne w legendzie każdego wykresu. Na wykresach znajduje się również czarna kropka przy wartości 0,35, reprezentująca prawdziwą wartość parametru θ. Oczywiście przy zajmowaniu się rzeczywistymi problemami nie jest znana ta wartość. Jest ona tutaj użyta tylko w celach dydaktycznych. Zawartość rysunku 1.12 może nas wiele nauczyć o analizie bayesowskiej, dlatego sięgnij po kawę, herbatę lub swój ulubiony napój i poświęć chwilę na jej zrozumienie:

- Wynikiem analizy bayesowskiej jest rozkład a posteriori. Nie jest to pojedyncza wartość, lecz rozkład prawdopodobnych wartości w przypadku konkretnych danych i używanego modelu.
- Najbardziej prawdopodobną wartość zapewnia dominanta rozkładu a posteriori (szczyt rozkładu).
- Rozrzut rozkładu a posteriori jest proporcjonalny do niepewności dotyczącej wartości parametru. Im większy rozrzut rozkładu, tym mniej jesteśmy pewni.
- Intuicyjnie jesteśmy bardziej pewni wyniku, gdy zaobserwowano więcej danych potwierdzających go. A zatem nawet wtedy, kiedy pod względem liczbowym 1/2 = 4/8 = 0,5, wypadnięcie czterech orzełków w ośmiu próbach daje nam większą pewność tego, że systematyczne odchylenie wynosi 0,5, niż zaobserwowanie jednego orzełka w dwóch próbach. Taka intuicja znajduje odzwierciedlenie w rozkładzie a posteriori. Możesz to sprawdzić samodzielnie, zwracając uwagę na (czarny) rozkład a posteriori na trzecim i szóstym wykresie. Choć dominanta jest taka sama, rozrzut (niepewność) jest większy w przypadku trzeciego niż szóstego wykresu.

- Przy wystarczająco dużej ilości danych dwa lub więcej modeli bayesowskich z różnymi rozkładami a priori będzie dążyć do tego samego wyniku. W granicy nieskończonej ilości danych, bez względu na to, jakiego użyje się rozkładu a priori, wszystkie zapewnią ten sam rozkład a posteriori.
- Pamiętaj, że nieskończoność to granica, a nie liczba, dlatego z praktycznego punktu widzenia można uzyskać praktycznie równoważne rozkłady a posteriori dla skończonej i stosunkowo małej liczby punktów danych.
- To, jak szybko rozkłady a posteriori stają się zbieżne z tym samym rozkładem, zależy od danych i modelu. Można dostrzec, że rozkłady a posteriori wywodzące się z czarnego rozkładu a priori (jednorodnego) i jasnoszarego rozkładu a priori (faworyzującego reszkę) zbiegają się szybciej do niemal identycznego rozkładu, natomiast ciemnoszary rozkład a posteriori (wywodzący się ze skoncentrowanego rozkładu a priori) potrzebuje więcej czasu. Nawet po 150 próbach dość łatwo można rozpoznać ciemnoszary rozkład a posteriori jako odmienny od dwóch pozostałych.
- Na rysunku nie jest oczywiste to, że zostanie uzyskany ten sam rezultat niezależnie od tego, czy aktualizuje się rozkład a posteriori sekwencyjnie, czy robi się to wszystko naraz. Można obliczyć rozkład a posteriori 150 razy, dodając za każdym razem jeszcze jedną obserwację i używając otrzymanego rozkładu a posteriori jako nowego rozkładu a priori, albo można po prostu obliczyć jeden rozkład a posteriori dla wszystkich 150 rzutów jednocześnie. Wynik będzie dokładnie taki sam. Właściwość ta nie tylko ma całkowity sens, ale również prowadzi do naturalnego sposobu aktualizowania oszacowań, gdy otrzymuje się nowe dane (sytuacja powszechna w wielu problemach związanych z analizą danych).

Wpływ rozkładu a priori

Z poprzedniego przykładu jasno wynika, że rozkłady a priori mogą wpływać na wnioskowanie. Jest to w porządku, gdyż rozkłady a priori mają tak działać. Może lepiej byłoby w ogóle nie mieć rozkładów a priori? Ułatwiłoby to modelowanie, prawda? No niekoniecznie. Jeśli nie definiujesz rozkładu a priori, ktoś inny zrobi to za Ciebie. Czasami nie stanowi to żadnego problemu. *Domyślne rozkłady a priori* mogą być użyteczne i mają swoje miejsce, ale czasami lepiej jest mieć większą kontrolę. Pozwól, że to wyjaśnię.

Można myśleć, że każdy model (statystyczny), bayesowski czy nie, uwzględnia jakiś rodzaj rozkładu a priori nawet wtedy, gdy nie jest on ustawiony jawnie. Na przykład wiele procedur używanych zwykle w statystyce częstościowej można postrzegać jako szczególne przypadki modelu bayesowskiego w określonych warunkach, takich jak płaskie rozkłady a priori. Jednym z powszechnych sposobów szacowania parametrów jest metoda największej wiarygodności. Unika ona określania rozkładu a priori, a jej działanie polega na znalezieniu pojedynczej wartości maksymalizującej wiarygodność. Wartość ta jest zwykle oznaczana przez dodanie niewielkiego daszka nad nazwą szacowanego parametru (np. $\hat{\theta}$). W przeciwieństwie do szacowania a posteriori, które jest rozkładem, parametr $\hat{\theta}$ to oszacowanie punktowe w postaci liczby. W przypadku problemu rzutu monetą można to obliczyć analitycznie:

$$\hat{\theta} = \frac{y}{N}$$

Jeśli powrócisz do rysunku 1.12, będziesz mógł sam sprawdzić, że dominanta czarnego rozkładu a posteriori (odpowiadającego jednorodnemu/płaskiemu rozkładowi a priori) zgadza się z wartościami parametru $\hat{\theta}$ obliczonymi dla każdego wykresu. Nie jest to przypadek, lecz konsekwencja faktu, że ustawienie jednorodnego rozkładu a priori, a następnie użycie dominanty rozkładu a posteriori jest równoważne metodzie największej wiarygodności.

Nie można uniknąć rozkładów a priori, ale jeśli uwzględni się je w prowadzonej analizie, można uzyskać pewne potencjalne korzyści. Najbardziej bezpośrednią z nich jest to, że otrzymuje się rozkład a posteriori, który jest rozkładem prawdopodobnych wartości, a nie tylko najbardziej prawdopodobnych. Uzyskanie rozkładu może być bardziej informacyjne niż pojedyncze oszacowanie punktowe, ponieważ, jak już wspomniałem, szerokość rozkładu jest związana z niepewnością, jaką mamy w kwestii oszacowania. Inną korzyścią jest to, że obliczanie rozkładów a posteriori oznacza uśrednianie względem rozkładu a priori. Może to prowadzić do bardziej solidnych predykcji i modeli, które są trudniejsze do przeuczenia (Wilson i Izmailov, 2022).

Rozkłady a priori mogą zapewnić inne korzyści. Począwszy od następnego rozdziału, do uzyskiwania rozkładów a posteriori będziemy używać metod numerycznych. Metody te wydają się magiczne, dopóki nie przestają takie być. W ramach nieformalnego twierdzenia obliczeń statystycznych podaje się: „Gdy masz problemy obliczeniowe, często występuje problem z Twoim modelem" (Gelman, 2008). Czasami mądry wybór rozkładu a priori może uczynić wnioskowanie łatwiejszym lub szybszym. Ważne jest, żeby zaznaczyć, że nie opowiadamy się za ustawianiem rozkładów a priori specjalnie po to, aby uczynić wnioskowanie szybszym, ale często zdarza się, że myśląc o rozkładach a priori, można uzyskać szybsze modele.

Jedną z zalet rozkładów a priori, która jest czasami pomijana, jest to, że konieczność myślenia o rozkładach a priori może *zmusić* nas do głębszego zastanowienia się nad problemem, który próbujemy rozwiązać, i dostępnymi danymi. Czasami sam proces modelowania prowadzi do lepszego zrozumienia, niezależnie od tego, jak dobrze dopasowuje się dane lub tworzy predykcje. Poprzez jawne określenie rozkładów a priori otrzymuje się bardziej przejrzyste modele, co oznacza, że można je łatwiej recenzować, debugować (w szerokim znaczeniu tego słowa), objaśniać innym i, miejmy nadzieję, ulepszać.

Sposób wyboru rozkładów a priori

Osoby rozpoczynające korzystanie z analizy bayesowskiej (jak również krytycy tego rozwiązania) zazwyczaj odczuwają pewien niepokój związany z wyborem rozkładów a priori. Obawiają się zwykle, że rozkład a priori nie pozwoli danym „mówić" samym za siebie! To zrozumiałe, ale musimy pamiętać, że dane „nie mówią". W najlepszym razie dane „szepczą". Można nadać sens danym jedynie w kontekście naszych modeli, w tym modeli matematycznych i mentalnych. W historii nauki znajdziemy wiele przykładów, w których

te same dane prowadziły ludzi do różnych wniosków na temat tych samych zagadnień. Może się to zdarzyć nawet wtedy, gdy opieramy swoje opinie na modelach formalnych.

Niektórzy lubią ideę używania nieinformacyjnych rozkładów a priori (znanych również jako płaskie, niejasne lub rozproszone). Takie rozkłady a priori mają najmniejszy możliwy wpływ na analizę. Choć można używać ich w niektórych problemach, uzyskanie prawdziwie nieinformacyjnych rozkładów a priori może być trudne lub po prostu niemożliwe. Ponadto zazwyczaj można postąpić lepiej, ponieważ przeważnie dostępne są jakieś wcześniejsze informacje.

W całej książce będziemy postępować zgodnie z zaleceniami Gelmana, McElreatha, Kruschkego i wielu innych osób, a ponadto będziemy preferować słabo informacyjne rozkłady a priori. W przypadku wielu problemów często wiemy coś o wartościach, jakie parametr może przyjmować. Można wiedzieć, że parametr jest ograniczony do wartości dodatnich, albo można znać przybliżony zakres, jaki może on przyjmować, albo można się spodziewać, że wartość będzie bliska zeru lub poniżej/powyżej jakiejś wartości. W takich przypadkach można zastosować rozkłady a priori, aby wprowadzić kiepskie informacje do modeli bez obawy o zbyt duży wpływ. Ponieważ te rozkłady a priori działają tak, by utrzymać rozkład a posteriori w pewnych rozsądnych granicach, są one również znane jako regularyzujące rozkłady a priori.

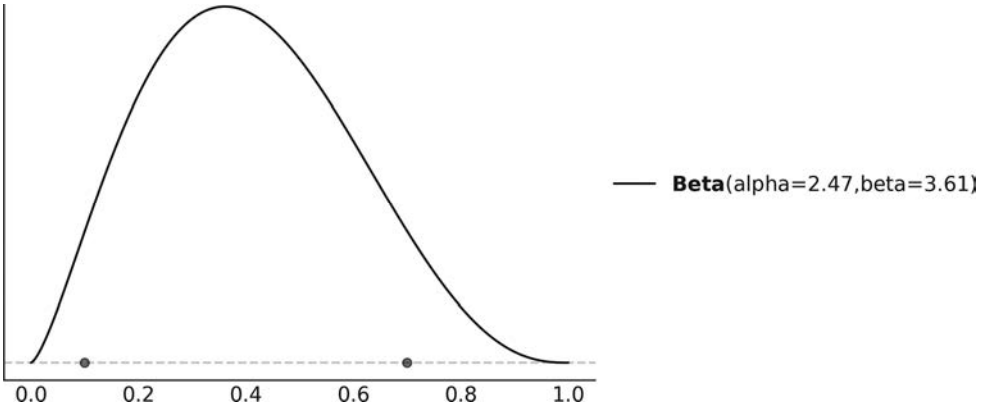
Informacyjne rozkłady a priori to bardzo silne rozkłady, które przekazują wiele informacji. Używanie ich również stanowi poprawną opcję. W zależności od problemu znalezienie dobrej jakości informacji w zasobach wiedzy domenowej i przekształcenie ich w rozkłady a priori może być łatwe albo nie. Zajmowałem się kiedyś zawodowo bioinformatyką strukturalną. W tej dziedzinie ludzie używali, w sposób bayesowski i niebayesowski, wszystkich informacji a priori, jakie mogli uzyskać, aby badać i przewidywać strukturę białek. Jest to rozsądne, ponieważ od dziesięcioleci zbieramy dane z tysięcy starannie opracowanych eksperymentów i dlatego mamy do dyspozycji ogromną ilość wiarygodnych informacji a priori. Zrezygnowanie z nich byłoby absurdalne! Nie ma nic „obiektywnego" ani „naukowego" w pozbywaniu się cennych informacji. Jeśli masz wiarygodne informacje a priori, powinieneś z nich skorzystać. Wyobraź sobie inżyniera z branży motoryzacyjnej, który za każdym razem, kiedy miałby za zadanie zaprojektować nowy samochód, musiałby zaczynać od zera i na nowo odkrywać silnik spalinowy, koło, a nawet całą koncepcję pojazdu.

PreliZ to bardzo nowa biblioteka języka Python służąca do wyznaczania rozkładów a priori (Mikkola i in., 2023; Icazatti i in., 2023). Jej celem jest zapewnienie pomocy w ustalaniu, reprezentowaniu i wizualizacji posiadanej wiedzy apriorycznej. Można na przykład zażądać od biblioteki PreliZ obliczenia parametrów rozkładu spełniającego zestaw ograniczeń. Przyjmijmy, że ma zostać znaleziony rozkład beta z 90% masy prawdopodobieństwa w przedziale od 0,1 do 0,7. W tym przypadku kod będzie mieć następującą postać (listing 1.7):

Listing 1.7

```
1 dist = pz.Beta()
2 pz.maxent(dist, 0.1, 0.7, 0.9)
```

Wynikiem jest rozkład beta z parametrami $\alpha = 2{,}5$ i $\beta = 3{,}6$ (zaokrąglone do pierwszego miejsca po przecinku). Funkcja `pz.maxent` oblicza rozkład o **maksymalnej entropii** przy podanych ograniczeniach, które określono. Dlaczego jest to rozkład o maksymalnej entropii? Dlatego, że jest to równoważne z obliczeniem najmniej informatywnego rozkładu przy tych ograniczeniach. Domyślnie biblioteka PreliZ utworzy wykres rozkładu widoczny na rysunku 1.13.



Rysunek 1.13. Rozkład beta o maksymalnej entropii z 90% masy prawdopodobieństwa w przedziale od 0,1 do 0,7

Jako że wyznaczanie rozkładu a priori ma wiele aspektów, biblioteka PreliZ oferuje też inne metody ustalania go. Jeśli chcesz dowiedzieć się więcej o tej bibliotece, zajrzyj do dokumentacji dostępnej na stronie *https://preliz.readthedocs.io*.

Budowanie modeli to proces iteracyjny. Czasami jedna iteracja trwa kilka minut, a czasami może potrwać lata. Odtwarzalność ma kluczowe znaczenie, a przejrzyste założenia w modelu przyczyniają się do niej. W przypadku danej analizy można swobodnie używać więcej niż jednego rozkładu a priori (lub funkcji wiarygodności), jeśli nie jesteśmy pewni co do żadnej konkretnej opcji wyboru. Badanie wpływu różnych rozkładów a priori może również dostarczyć cennych informacji. Część procesu modelowania polega na kwestionowaniu założeń, a rozkłady a priori (i funkcje wiarygodności) są właśnie takimi założeniami. Różne założenia prowadzą do odmiennych modeli i prawdopodobnie rozmaitych wyników. Korzystając z danych i wiedzy eksperckiej związanej z problemem, będziemy w stanie porównać modele, a jeśli to konieczne, wybrać zwycięzcę. Rozdział 5. będzie poświęcony tej kwestii. Ponieważ rozkłady a priori odgrywają centralną rolę w statystyce bayesowskiej, będziemy je omawiać w miarę napotykania nowych problemów. Jeśli zatem masz wątpliwości i czujesz się nieco zdezorientowany tym omówieniem, zwyczajnie zachowaj spokój i się nie martw. Ludzie są zdezorientowani od dekad, a dyskusja wciąż trwa.

Informowanie o wynikach analizy bayesowskiej

Tworzenie raportów i informowanie o wynikach ma kluczowe znaczenie z punktu widzenia praktyki statystycznej i danologii. W tym podrozdziale w skrócie omawiam niektóre osobliwości tego zadania realizowanego podczas pracy z modelami bayesowskimi. W kolejnych rozdziałach będziemy nadal analizować przykłady tej ważnej kwestii.

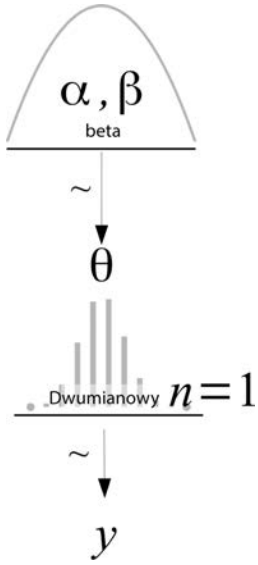
Notacja modeli i wizualizacja

Jeśli chcesz przekazać wyniki analizy, powinieneś również poinformować o zastosowanym modelu. Oto powszechna notacja pozwalającą zwięźle reprezentować modele probabilistyczne:

$$\theta \sim \text{Beta}(\alpha, \beta)$$

$$y \sim \text{Bin}(n = 1, p = \theta)$$

Jest to dokładnie ten model, którego użyto w przykładzie z rzutem monetą. Jak być może pamiętasz, symbol ~ oznacza, że zmienna po jego lewej stronie jest zmienną losową o rozkładzie określonym po prawej stronie. W wielu kontekstach symbol ten służy do wskazania, że zmienna przyjmuje *w przybliżeniu* pewną wartość. Gdy jednak jest mowa o modelach probabilistycznych, symbol ten będzie odczytywany jako „*jest rozłożone według*". Można zatem stwierdzić, że model θ ma rozkład beta z parametrami α i β, a model y ma rozkład dwumianowy z parametrami $n = 1$ i $p = \theta$. Ten sam model można przedstawić graficznie za pomocą diagramów Kruschkego (rysunek 1.14).



Rysunek 1.14. Diagram Kruschkego modelu beta-dwumianowego

Na pierwszym poziomie znajduje się rozkład a priori, który generuje wartości dla modelu θ, następnie pojawia się funkcja wiarygodności, a w ostatnim wierszu dane modelu y. Strzałki wskazują relacje między zmiennymi, a symbol ~ oznacza stochastyczną naturę zmiennych. Wszystkie diagramy Kruschkego zawarte w tej książce zostały utworzone przy użyciu szablonów udostępnionych przez Rasmusa Bååtha (*http://www.sumsar.net/blog/2013/10/diy-kruschke-style-diagrams/*).

Podsumowanie rozkładu a posteriori

Wynikiem analizy bayesowskiej jest rozkład a posteriori, a wszystkie informacje o parametrach (dla danego modelu i zbioru danych) są zawarte w rozkładzie a posteriori. W związku z tym, podsumowując rozkład a posteriori, zestawia się ze sobą logiczne konsekwencje modelu i danych. Powszechną praktyką jest raportowanie dla każdego parametru średniej (albo dominanty lub mediany), aby mieć wyobrażenie o położeniu rozkładu, a także pewnej miary rozproszenia, takiej jak odchylenie standardowe, w celu ustalenia stopnia niepewności posiadanych oszacowań. Odchylenie standardowe sprawdza się dobrze w przypadku rozkładów podobnych do normalnego, ale może być mylące w ramach innych typów rozkładów, takich jak rozkłady skośne.

Powszechnie stosowane rozwiązanie do podsumowania rozproszenia rozkładu a posteriori polega na wykorzystaniu **przedziału najwyższej gęstości** (ang. ***Highest-Density Interval* — HDI**). HDI to najkrótszy przedział zawierający daną część gęstości prawdopodobieństwa. Jeśli mówi się, że 95% przedział HDI dla pewnej analizy wynosi [2, 5], oznacza to, iż zgodnie z używanymi danymi i modelem badany parametr znajduje się między wartościami 2 a 5 z prawdopodobieństwem 0,95. Nie ma nic szczególnego w wyborze wartości 95%, 50% czy jakiejkolwiek innej. Jeśli jest to wskazane, można swobodnie wybrać 82% przedział HDI. W idealnej sytuacji uzasadnienia powinny być zależne od kontekstu, a nie automatyczne, ale można się zadowolić jakąś typową wartością, taką jak 95%. W ramach uprzejmego przypomnienia o arbitralności tego wyboru należy zaznaczyć, że domyślna wartość w pakiecie ArviZ to 94%.

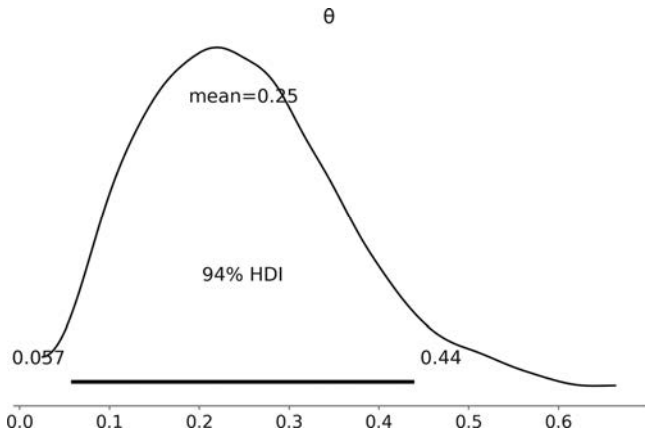
ArviZ to pakiet języka Python służący do eksploracyjnej analizy modeli bayesowskich, który zawiera wiele funkcji pomagających podsumować rozkład a posteriori. Jedna z tych funkcji, `az.plot_posterior`, umożliwia wygenerowanie wykresu ze średnią i przedziałem HDI modelu θ (listing 1.8). Rozkład nie musi być rozkładem a posteriori. Sprawdzi się każdy rozkład. Na rysunku 1.15 pokazałem wynik dla losowej próby z rozkładu beta.

Listing 1.8

```
1 np.random.seed(1)
2 az.plot_posterior({'θ':pz.Beta(4, 12).rvs(1000)})
```

Podsumowanie

Naszą bayesowską przygodę rozpoczęliśmy od bardzo krótkiego omówienia modelowania statystycznego, prawdopodobieństwa, prawdopodobieństwa warunkowego, zmiennych losowych, rozkładów prawdopodobieństwa i twierdzenia Bayesa. W dalszej kolejności



Rysunek 1.15. Jądrowa estymacja gęstości próby z rozkładu beta z jego średnią i 94% przedziałem HDI

To nie są przedziały ufności

Jeśli jesteś zaznajomiony z paradygmatem częstościowym, pamiętaj, że przedział HDI to nie to samo co przedziały ufności. W podejściu częstościowym parametry są z założenia stałe. Częstościowy przedział ufności zawiera prawdziwą wartość parametru albo nie. W podejściu bayesowskim parametry są zmiennymi losowymi, dlatego można mówić o prawdopodobieństwie tego, że parametr ma określone wartości lub znajduje się w pewnym przedziale. Nieintuicyjna natura przedziałów ufności sprawia, że łatwo je błędnie interpretować, a ludzie często mówią o częstościowych przedziałach ufności tak, jakby były to bayesowskie przedziały wiarygodności.

użyliśmy problemu rzutu monetą jako pretekstu do zaprezentowania podstawowych aspektów modelowania bayesowskiego i analizy danych. Ten klasyczny przykład posłużył do przekazania niektórych najważniejszych idei statystyki bayesowskiej, takich jak stosowanie rozkładów prawdopodobieństwa do budowania modeli i reprezentowania niepewności. Objaśniłem kwestię użycia rozkładów a priori i postawiłem je na równi z innymi elementami, które są częścią procesu modelowania, takimi jak funkcja wiarygodności, czy nawet bardziej metapytaniami (na przykład dlaczego w ogóle próbuje się rozwiązać konkretny problem?).

Rozdział zakończyłem omówieniem interpretacji wyników analizy bayesowskiej i informowania o nich. Założyłem, że istnieje prawdziwy rozkład, który zasadniczo jest nieznany (i w zasadzie również niepoznawalny), z którego otrzymuje się skończoną próbę, drogą eksperymentu, badania, obserwacji bądź symulacji. Aby dowiedzieć się czegoś o prawdziwym rozkładzie, mając do dyspozycji jedynie obserwowaną próbę, buduje się model probabilistyczny. Zawiera on dwa podstawowe składniki: rozkład a priori i funkcję wiarygodności. Używając modelu i próby, przeprowadza się wnioskowanie bayesowskie i uzyskuje rozkład a posteriori. Zawiera on w sobie całą informację o problemie, biorąc pod uwagę użyty model i dane.

Z perspektywy bayesowskiej rozkład a posteriori jest głównym obiektem zainteresowania, a wszystko inne jest z niego wyprowadzane, w tym predykcje w postaci rozkładu

predykcyjnego a posteriori. Ponieważ rozkład a posteriori (i dowolna inna wielkość z niego wyprowadzona) jest konsekwencją modelu i danych, przydatność wnioskowań bayesowskich jest ograniczona przez jakość modeli i danych. Na koniec w skrócie podsumowałem główne aspekty prowadzania bayesowskiej analizy danych. W pozostałej części książki będziemy powracać do tych idei, aby je przyswoić i wykorzystać jako fundament bardziej zaawansowanych zagadnień.

W następnym rozdziale zaprezentuję bibliotekę PyMC języka Python służącą do modelowania bayesowskiego i probabilistycznego uczenia maszynowego. Ponadto użyjemy więcej funkcji biblioteki ArviZ Pythona do eksploracyjnej analizy modeli bayesowskich, a także biblioteki PreliZ tego języka do określania rozkładów a priori.

Ćwiczenia

Nie wiemy, czy mózg działa w sposób bayesowski, w przybliżeniu w ten sposób, czy może korzysta z pewnych ewolucyjnie (mniej lub bardziej) zoptymalizowanych heurystyk. Niemniej jednak wiemy, że uczymy się poprzez ekspozycję na działanie danych, przykładów i ćwiczeń. Można powiedzieć, że ludzie wcale się nie uczą, biorąc pod uwagę naszą historię jako gatunku w takich kwestiach jak wojny czy systemy ekonomiczne, które przedkładają zysk nad dobrostan ludzi… W każdym razie polecam wykonanie następujących ćwiczeń proponowanych na końcu każdego rozdziału:

1. Załóżmy, że masz słoik z czterema żelkami: dwiema o smaku truskawkowym, jedną o smaku jagodowym i jedną o smaku cynamonowym. Losowo wyciągasz jedną żelkę ze słoika.
   a) Jaka jest przestrzeń prób dla tego eksperymentu?
   b) Definiujemy zdarzenie *A* jako *wylosowanie żelki truskawkowej*, a zdarzenie *B* jako *wylosowanie żelki, która nie jest cynamonowa*.
   Jakie są prawdopodobieństwa zdarzeń *A* i *B*?
   c) Czy zdarzenia *A* i *B* wykluczają się wzajemnie? Dlaczego tak lub dlaczego nie?
2. Wcześniej zdefiniowano funkcję `P` języka Python do obliczania prawdopodobieństwa zdarzenia przy użyciu naiwnej definicji prawdopodobieństwa. Uogólnij tę funkcję, aby obliczać prawdopodobieństwo zdarzeń, gdy nie wszystkie są jednakowo prawdopodobne. Użyj tej nowej funkcji do obliczenia prawdopodobieństwa zdarzeń *A* i *B* z poprzedniego ćwiczenia. Wskazówka: możesz przekazać trzeci argument z prawdopodobieństwem każdego zdarzenia.
3. Użyj biblioteki PreliZ do zbadania różnych parametrów rozkładów beta-dwumianowego i Gaussa. Zastosuj metody `plot_pdf`, `plot_cdf` i `plot_interactive`.
4. Omówiłem funkcje masy i gęstości prawdopodobieństwa oraz dystrybuantę. Istnieją jednak inne sposoby reprezentacji funkcji, takie jak funkcja punktu percentylowego. Używając metody `plot_ppf` biblioteki PreliZ, utwórz wykres funkcji punktu percentylowego dla rozkładów beta-dwumianowego i Gaussa.

Czy potrafisz wyjaśnić, jak ta funkcja jest powiązana z dystrybuantą oraz funkcjami masy i gęstości prawdopodobieństwa?

5. Które z poniższych wyrażeń odpowiada prawdopodobieństwu pogodnego dnia pod warunkiem, że jest 9 lipca 1816 roku?
   a) $p(\text{pogodnie})$.
   b) $p(\text{pogodnie}|\text{lipiec})$.
   c) $p(\text{pogodnie}|\text{9 lipca 1816})$.
   d) $p(\text{9 lipca 1816}|\text{pogodnie})$.
   e) $\frac{p(\text{pogodnie}|\text{9 lipca 1816})}{p(\text{9 lipca 1816})}$
6. Pokazałem, że prawdopodobieństwo wylosowania przypadkowej osoby, która okaże się papieżem, nie jest takie samo jak prawdopodobieństwo tego, że papież jest człowiekiem. W serialu animowanym *Futurama* papież (kosmiczny) jest gadem. Jak to zmienia poprzednie obliczenia?
7. Wzorując się na przykładzie z rysunku 1.9, użyj biblioteki PreliZ do obliczenia momentów rozkładu skośno-normalnego dla innej kombinacji parametrów. Wygeneruj próby losowe o różnych rozmiarach (np. 10, 100 i 1000) i sprawdź, czy na podstawie prób możesz odtworzyć wartości pierwszych dwóch momentów (średniej i wariancji). Co obserwujesz?
8. Powtórz poprzednie ćwiczenie dla rozkładu t-Studenta. Wypróbuj wartości parametru ν, takie jak 2, 3 i 500. Co obserwujesz?
9. W następującej definicji modelu probabilistycznego zidentyfikuj rozkład a priori i funkcję wiarygodności:

$$Y \sim \text{Normal}(\mu, \sigma)$$
$$\mu \sim \text{Normal}(0, 2)$$
$$\sigma \sim \text{HalfNormal}(0.75)$$

10. Ile parametrów w poprzednim modelu będzie mieć rozkład a posteriori? Porównaj go z modelem problemu rzutu monetą.
11. Zapisz twierdzenie Bayesa dla modelu z ćwiczenia 9.
12. Załóżmy, że mamy dwie monety. Gdy rzucamy pierwszą, w połowie przypadków wypada reszka, a w połowie orzełek. Druga moneta jest obciążona i zawsze wypada na niej orzełek. Jeśli losowo wybierze się jedną z monet i wypadnie orzełek, jakie jest prawdopodobieństwo, że to jest obciążona moneta?
13. Spróbuj ponownie utworzyć wykresy z rysunku 1.12, używając innych rozkładów a priori (`beta_params`) oraz innych danych (`trials` i `data`).
14. Przeczytaj o regule Cromwella na stronie serwisu Wikipedia o adresie *https://en.wikipedia.org/wiki/Cromwell%27s_rule*.
15. Przeczytaj o prawdopodobieństwach i holenderskiej księdze zakładów na stronie serwisu Wikipedia o adresie *https://en.wikipedia.org/wiki/Dutch_book*.

Jasne i zwięzłe wprowadzenie do metod bayesowskich i biblioteki PyMC.

— Christopher Fonnesbeck i Thomas Wiecki

W ostatnich dekadach statystyka bayesowska zyskała ogromne znaczenie w nauce i inżynierii. Współczesna analiza bayesowska to w dużej mierze statystyka obliczeniowa — elastyczna, przejrzysta i umożliwiająca intuicyjną interpretację wyników. Dzięki rozwojowi bibliotek języka Python koncepcje bayesowskie stały się praktycznym narzędziem do realizacji zaawansowanych scenariuszy analitycznych.

Książka stanowi kompleksowe wprowadzenie do stosowanego wnioskowania bayesowskiego i jego implementacji w Pythonie. Autor używa nowoczesnej biblioteki PyMC do programowania probabilistycznego, a ArviZ do analizy i diagnostyki modeli. Omawia także inne narzędzia ekosystemu bayesowskiego, takie jak Bambi, PreliZ i Kulprit. Zapoznasz się z zagadnieniami bayesowskich addytywnych drzew regresyjnych (BART), selekcji zmiennych, konstrukcji rozkładów a priori i porównywania modeli. Ponadto dowiesz się, jak budować, analizować i interpretować modele probabilistyczne w projektach z zakresu data science.

W książce między innymi:

- budowa modeli probabilistycznych z użyciem PyMC
- analiza i diagnostyka modeli w ArviZ
- modele hierarchiczne — zalety i ograniczenia
- porównywanie modeli i wybór najlepszych rozwiązań
- interpretacja wyników w kontekście rzeczywistych problemów
- myślenie probabilistyczne w ujęciu bayesowskim

Osvaldo Martin jest badaczem specjalizującym się w metodach obliczeniowych. Zajmował się bioinformatyką strukturalną i symulacją układów molekularnych, obecnie koncentruje się na statystyce bayesowskiej i programowaniu probabilistycznym. Prowadził kursy z zakresu bioinformatyki, data science i analizy bayesowskiej. Współtworzył projekty open source związane z bibliotekami: ArviZ, Bambi, Kulprit, PreliZ i PyMC.

Helion
helion.pl
HELION S.A.
ul. Kościuszki 1c
44-100 Gliwice
tel.: 32 230 98 63
helion@helion.pl
KOD KORZYŚCI
Sięgnij po więcej!
ISBN 978-83-289-3665-2
9 788328 936652
Cena: 89,00 zł
packt