Class09-11

September 14, 2020

0.1 Conditional probability

0.1.1 Definition

A **conditional probability** of an event A given B describes the chances that the event A occurs, having already observed an event B.

The conditional probability above can be represented in mathematical notation as

$$p(A|B) \tag{1}$$

For example, the probability of being admitted to the hospital given a patient tested positive for the novel coronavirus (COVID-19). This could be written

$$p(\text{Admitted to the hospital}|\text{Positive test}) \tag{2}$$

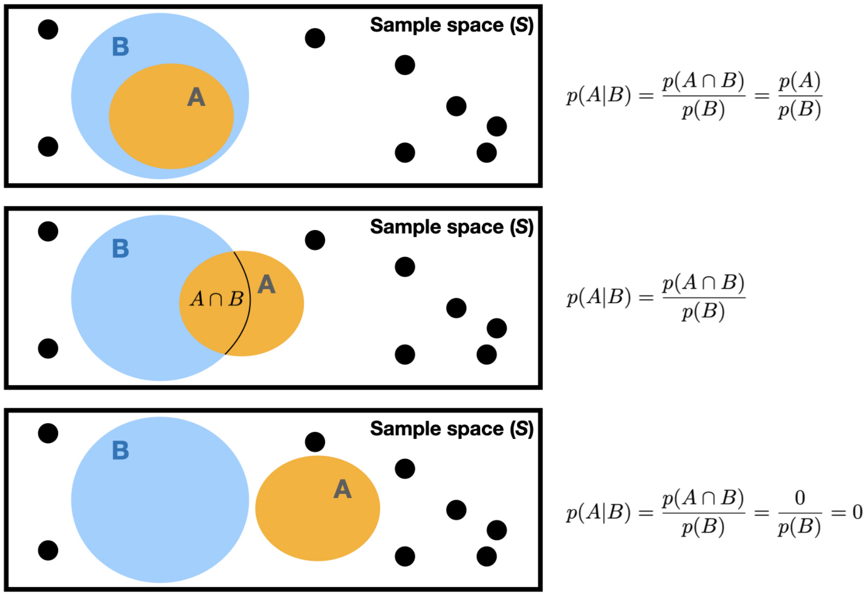
0.1.2 Computation

We can compute a conditional probability by

$$p(A|B) = \frac{p(A \cap B)}{p(B)} \tag{3}$$

The conditional probability is the probability that the events A and B occur simultaneously divided by the probability of event B.

There are three ways two events like A and B can interact to help us understand why we would compute conditional probabilities like this.



In the top panel, the event A only occurs if B occurs. The conditional probability computes the proportion of times A occurs relative to B The bottom panel shows the events A and B never occurring together. Since they never occur at the same time, if the event B occurs the event A will never occur: the conditional probability of A given B is zero. Te middle panel shows a common scenario. There is a subset of outcomes where A occurs when B happens. The conditional probability asks "how many outcomes include the event A and B relative to the number of times B occurs?"

0.1.3 Application

Below are two examples of conditional probabilities, the first more obvious than the second. Suppose we wanted to compute the probability of having SARS-COV-2 given a positive test. We estimate that the probability of having SARS-COV-2 **and** a test returning positive is 0.10. Next, suppose we estimate the probability of a test returning positive whether or not you have SARS-COV-2 is 0.50.

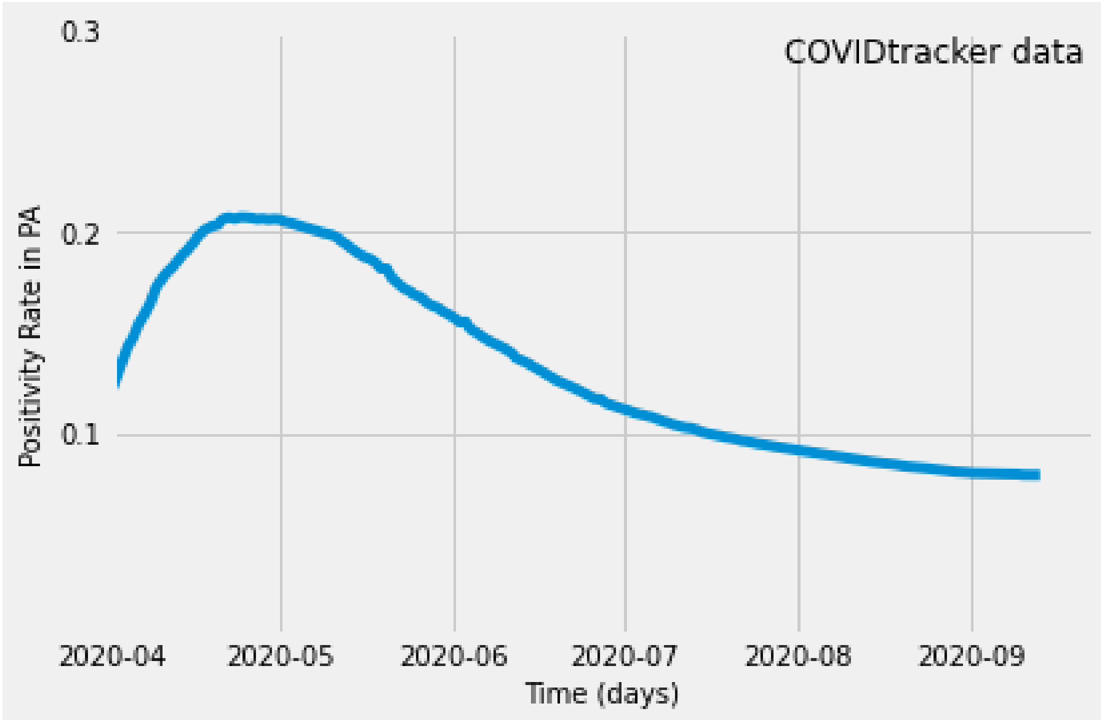
The conditional probability

$$
\begin{aligned}
p(\text{ SARS-COV-2 } \cap \text{ Test Pos. }) &= 0.10 && (4)\\
p(\text{ Test Pos. }) &= 0.50 && (5)\\
p(\text{ SARS-COV-2 } | \text{ Test Pos. }) &= 0.10/0.50 = 20\% && (6)\\
& && (7)
\end{aligned}
$$

A second example is below and a more subtle use of conditional probabilities. Data on COVID-19 positive rates, the probability of testing positive for SARS-COV-2, was taken from the COVID Tracking Project. The COVID tracking project is hostsed by the Atlantic. They scour as many news and iinformation sources on COVID-19 as possible to provide best possible estimates of SARS-COV-2/COVID-19 in the US.

Below is a plot of the number of positive tests divided by the total number of tests administered over time (in days) for the state of Pennsylvania. What is this proportion measuring?

[10]:
```python
covidData = pd.read_csv("https://covidtracking.com/data/download/all-states-history.csv") # downlaod data from the Covidtracker
covidData["positiveRate"] = covidData.positive/covidData.totalTestResults # compute positivity rate
covidData["date"] = [pd.to_datetime(x,format="%Y%m%d") for x in covidData.date] # convert integer date to date obj.

paData = covidData[covidData.state=="PA"] # subset to PA

plt.style.use("fivethirtyeight")
fig,ax = plt.subplots() # setup a plotting space

ax.plot(paData.date, paData.positiveRate ) # plot the date by positivity rate

# Format the x and ylimits
ax.set_xlim(pd.to_datetime("2020-04-01"),ax.get_xlim()[-1])
ax.set_ylim(0,0.30)

ax.set_ylabel("Positivity Rate in PA", fontsize=10)
ax.set_xlabel("Time (days)", fontsize=10)

ax.tick_params(labelsize=10)

ax.set_yticks([0.1,0.2,0.3])

ax.text(0.99,0.99,"COVIDtracker data",fontsize=12,transform=ax.transAxes,ha='right',va='top')

# a tightlayout asks python to move around objects on the graph for the "best" possible layout
fig.set_tight_layout(True)
plt.show()
```



0.2 Marginal probs from conditional probs

We can compute marginal probabilities (for example $p(A)$) by first finding a second set of events $(B_1, B_2, \cdots, B_N)$ that is a **partition** of A.

A **partition** of an event A is a collection of sets such that their union equals A

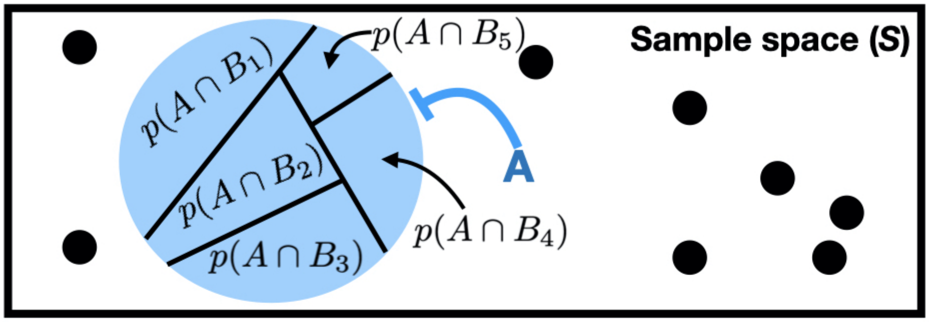
if

$$B_1 \cup B_2 \cup \cdots \cup B_N = A \tag{8}$$

then the collection of events B is a partition for A. We can compute $p(A)$ using a partition as

$$p(A) = p(A|B_1)p(B_1) + p(A|B_2)p(B_2) + \cdots p(A|B_N)p(B_N) \tag{9}$$

$$p(A) = \sum_{i=1}^{N} p(A|B_i)p(B_i) \tag{10}$$



$$p(A) = p(A \cap B_1) + p(A \cap B_2) + p(A \cap B_3) + p(A \cap B_4) + p(A \cap B_5)$$

$$\begin{aligned} p(A) = p(A \cap B_1)\frac{p(B_1)}{p(B_1)} + p(A \cap B_2)\frac{p(B_2)}{p(B_2)} + p(A \cap B_3)\frac{p(B_3)}{p(B_3)} \\ + p(A \cap B_4)\frac{p(B_4)}{p(B_4)} + p(A \cap B_5)\frac{p(B_5)}{p(B_5)} \end{aligned}$$

$$\begin{aligned} p(A) = p(A|B_1)p(B_1) + p(A|B_2)p(B_2) + p(A|B_3)p(B_3) \\ + p(A|B_4)p(B_4) + p(A|B_5)p(B_5) \end{aligned}$$

This equation can come in handy when there is more information about a set of conditional probabilities that partition an event A. A common case is when you know * the probability the event B occurs * the conditional probability of A when B occurs * the conditional probability of A when B does not occurs

One way we could compute the probability of SAR-COV-2 could be to estimate * the probability the a SARS-COV-2 test returns a positive result * the conditional probability of SARS-COV-2 when a test returns a positive result * the conditional probability of SARS-COV-2 when a test returns a negative result

$$\begin{aligned} p(\text{SARS-COV-2}) &= p(\text{SARS-COV-2}|+)p(+) + p(\text{SARS-COV-2}|-)p(-) && (11) \\ &= p(\text{SARS-COV-2}|+)p(+) + p(\text{SARS-COV-2}|-)(1 - p(+)) && (12) \\ & && (13) \end{aligned}$$

and it may be easier to find the probability of a positive and negative test in order to compute the probability of SARS-COV-2. We can use another event that we have data on to compute an event we're interested in.

0.3 Independence and the multiplication rule

We can rearrange the conditional probability of A given B

$$p(A|B) = \frac{p(A \cap B)}{p(B)} \tag{14}$$

$$p(A \cap B) = p(A|B)p(B) \tag{15}$$

to compute the probability of A and B. This is called the **general multiplication rule**.

Two events are called **independent** when the occurrence of one event does not impact the probability of a second event occurring.

$$p(A|B) = p(A) \tag{16}$$

Given that B occurred does not change the probability of A. If two event are independent then computing the general multiplication rule is easier

$$p(A|B) = \frac{p(A \cap B)}{p(B)} \tag{17}$$

$$p(A \cap B) = p(A|B)p(B) \tag{18}$$

$$p(A \cap B) = p(A)p(B) \tag{19}$$

0.4 Baye's Theorem

Baye's Theorem (BT) relates two conditional probabilities to one another:

$$p(A|B) = p(B|A) \times \frac{p(A)}{p(B)} \tag{20}$$

0.4.1 A classic example of BT

A classic example of BT relates the reliability of a test to disease **prevalence**— the number or proportion of cases of a disease present in a population at a given time. Suppose a test is developed so that if you have the disease of interest it returns a positive result with 0.80 probability and if you don't have the rare disease it returns a positive result with probability 0.10.
Let's also assume the probability of having the disease is 0.001, this is a rare disease.

Given a positive test, do we have this rare disease? Can we compute the probability of having this rare disease (RD) given a positive test (+)?

BT says

$$p(\text{RD}|+) = p(+|\text{RD}) \times \frac{p(\text{RD})}{p(+)} \tag{21}$$

We know the probability the test returns a positive result if you have a disease ($p(+|\text{RD})$), and we also know the probability of having the disease ($p(\text{RD})$).

$$p(\text{RD}|+) = p(+|\text{RD}) \times \frac{p(\text{RD})}{p(+)} \qquad\qquad = 0.80 \times \frac{0.001}{p(+)} \tag{22}$$

But how do we compute the probability the test returns a positive result? Well we do know $p(+|\text{RD})$ and also $p(\text{RD})$.

From the above marginal probs section, we could compute $p(+)$ like this

$$p(+) = p(+|\text{RD})p(\text{RD}) + p(+|\text{Not RD})p(\text{Not RD}) \tag{23}$$

The first three terms are given to us

$$p(+) = 0.80 \times 0.001 + 0.10 \times p(\text{Not RD}) \tag{24}$$

and we can compute the fourth term $p(\text{Not RD}) = 1 - p(\text{ RD}) = 0.999$. So then the probability of a positive test is

$$p(+) = 0.80 \times 0.001 + 0.10 \times 0.999 = 0.1007 \tag{25}$$

We can finally find out the probability of having this rare disease given a positive test

$$p(\text{RD}|+) = p(+|\text{RD}) \times \frac{p(\text{RD})}{p(+)} \qquad\qquad = 0.80 \times \frac{0.001}{0.1007} = 0.008 = 0.8\% \tag{26}$$

Well whats going on? Our test has an 80% of returning a positive result when we have this rare disease. And it was positive. Why then is there only a 0.8% of actually having the disease? Because the disease itself is rare, a positive test is no guarantee.

0.4.2 BT as a way to learn from data

0.5 Random variables

0.5.1 Definition

A **random variable** assigns numerical values to the outcomes of a random process.

0.5.2 Example

Suppose we roll three die. We can define a random variable X to be the sum of all three die and can then take values from 3 up to 18.
Multiple outcomes (a roll of three die) correspond to a single value of our random variable X. There are several ways the three die can add up to the value 5: \$ { (1,3,1), (2,1,2), (2,2,1), $\cdots$ }\$.

We can define a random variable S to be the number of SARS-COV-2 infections present among PA residents who were tested. In this case, the random variable (r.v.) can take the values from 0 up to the number of tests, and again there are many different outcomes that correspond to the same r.v. values.

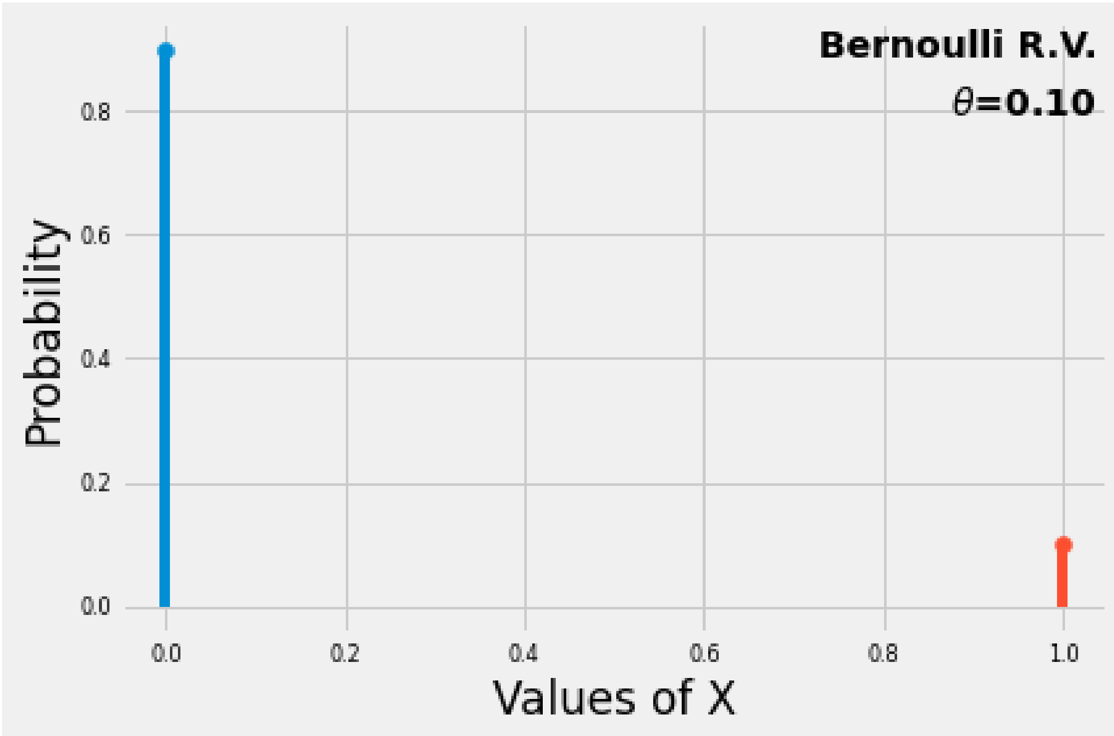
0.5.3 Standard R.V.s

The probability distribution of a random variable is (like any prob dist) the disjoint values the r.v. can take and the associated probabilities. There are a number of random variables that have standard probability distributions.

0.6 Bernoulli Random variable

[11]:

```
p=0.1

def Bernoulli(p):
    plt.style.use("fivethirtyeight")
    fig,ax = plt.subplots()

    ax.plot([0]*2,[0,1-p])
    ax.scatter(0,1-p)

    ax.plot([1]*2,[0,p])
    ax.scatter(1,p)

    ax.set_xlabel("Values of X")
    ax.set_ylabel("Probability")
    ax.tick_params(labelsize=8)

    ax.text(0.99,0.99,"Bernoulli R.V.",ha="right",va="top",transform=ax.
 ↪transAxes,weight="bold")
    ax.text(0.99,0.90,r"$\theta$={:.2f}".
 ↪format(p),ha="right",va="top",transform=ax.transAxes,weight="bold")

    plt.show()
Bernoulli(p)
```


Bernoulli R.V.
θ=0.10
Probability
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Values of X

[ ]:

[ ]:

[ ]:

[ ]: