Class12-14

September 28, 2020

0.1 Expectation, variance, and distribution of Random variables

Random variables assign numerical values to the outcome of an experiment or otherwise random phenomena. The study of random variables is central to describing the majority of subjects that arise in population health. The number of patients who experienced an adverse event, the proportion of a population susceptible to an infectious disease, and the association between cigarette sales and lung cancer in a community are all examples of assigning a numerical value to the outcome of a random phenomena—they are all examples of describing our world with random variables.

0.1.1 Discrete and continuous random variables

We can classify random variables into two types: (i) discrete and (ii) continuous. A **discrete** random variable takes a finite, distinct set of numerical values. For example, we can define a discrete r.v. to be the number of individuals infected with a respiratory disease in a fixed population. A **continuous** random variable takes an infinite number of numerical values. For example, we can defined a continuous r.v. to be the creatinine clearance of a patient, a continuous measurement of the rate of waste eliminated by the body through the kidneys.

Discrete random variables are often easier to conceptualize that continuous random variables.

0.1.2 Expectation and variance for discrete random variables

The **expectation** of a random variable is the sum of each value of the random variable weighted by the probability that value will occur.

If X is a random variable then the expectation of X is

$$\mathrm{E}(X) = x_1 p(x_1) + x_2 p(x_2) + \cdots + x_n p(x_n) \tag{1}$$

$$\sum_{i=1}^{N} x_i p(x_i) \tag{2}$$

where the random variable X can take any of n values from x_1 to x_n.

The **variance** of a random variable is the sum of squared differences between each value of the random variable and it's expectation weighted by the probability that value will occur.

$$\text{Var}(X) = [x_1 - E(X)]^2 p(x_1) + [x_2 - E(X)]^2 p(x_2) + \cdots + [x_n - E(X)]^2 p(x_n) \tag{3}$$

$$= \sum_{i=1}^{N} [x_i - E(X)]^2 p(x_i) \tag{4}$$

0.1.3 The Expectation of a function of a r.v.

We can generalize the expectation of a random $[E(X)]$ allowing us to compute the expectation of any function of a random variable. The expected value of a function of a r.v. X is

$$E[f(X)] = \sum_{i=1}^{N} f(x_i)p(x_i) \tag{5}$$

The expected value of $f(X)$ is just the value of each $f(x_i)$ weighted by how often the value x_i occurs. To see that this expectation is more general than our original definition, we can compute $E[f(X)]$ where f is the identity function ($f(x) = x$).

$$E[f(X)] = \sum_{i=1}^{N} f(x_i)p(x_i) \tag{6}$$

$$= \sum_{i=1}^{N} x_i p(x_i) \tag{7}$$

$$= E(X) \tag{8}$$

0.1.4 Probability distribution of random variables

Just like a set of disjoint outcomes and their associated probabilities defined a probability distribution, the values of a random variable and their associated probabilities defined a probability distribution of that r.v. Some probability distributions of random variables are so common that they have been standardized. Below we will discuss five random variables, their expectation and variance, and an example of how to apply them in real-world examples using Python.

0.1.5 Bernoulli

Definition A random variable X follows a **Bernoulli** distribution if X takes either the value 0 or 1, and

$$p(x) = \begin{cases} 0 & (1-\theta) \\ 1 & \theta \end{cases} \tag{9}$$

We write that $X \sim \text{Bern}(\theta)$, in words, that X follows a Bernoulli distribution with parameter theta.

Expectation and variance The expected value of X, if it is distributed Bernoulli is

$$\begin{aligned} \mathrm{E}(X) &= 1 \times p(X=1) + 0 \times p(X=0) && (10)\\ &= 1 \times \theta + 0 \times (1-\theta) && (11)\\ &= \theta && (12) \end{aligned}$$

The variance of X is

$$\begin{aligned} \mathrm{Var}(X) &= (1-\theta)^2 \times p(X=1) + (0-\theta)^2 \times p(X=0) && (13)\\ &= (1-\theta)^2 \times \theta + \theta^2 \times (1-\theta) && (14)\\ &= \theta(1-\theta)\left[(1-\theta)+\theta\right] && (15)\\ &= \theta(1-\theta) && (16) \end{aligned}$$

Application The US Food and Drug Administration (FDA) allows the voluntary addition of adverse event records for devices. This data base is called MAUDE: Manufacturer and User Facility Device Experience. MAUDE is a searchable database, and I searched and extracted events reported on the RESOLUTE ONYX Drug-eluting stent, manufactured by Medtronic.

Let's load the database into Python and look at how we can defined Bernoulli random variables to describe our data.

[11]:
```
# Load Data from a uniform resource locator (url)
import pandas as pd

maudeData = pd.read_excel("https://github.com/computationalUncertaintyLab/
 ↪2020F_PHDSI/blob/master/c5/maudeExcelReport13.xls?raw=true")

# Pandas is a Python module for working with data
#, and it is common to abbreviate the module pandas as pd.

# Above, i used the function read_csv() from pd to load our dataset.

# We can view the first few rows of our data by using the function "head"
maudeData.head()
```

[11]:

	Web Address	Report Number	\
0	https://www.accessdata.fda.gov/scripts/cdrh/cf...	9612164-2020-03220	
1	https://www.accessdata.fda.gov/scripts/cdrh/cf...	9612164-2020-03224	
2	https://www.accessdata.fda.gov/scripts/cdrh/cf...	9612164-2020-03225	
3	https://www.accessdata.fda.gov/scripts/cdrh/cf...	9612164-2020-03226	
4	https://www.accessdata.fda.gov/scripts/cdrh/cf...	9612164-2020-03227	

	Event Date	Event Type	Manufacturer	Date Received	Product Code	\
0	2020-08-30	Malfunction	MEDTRONIC IRELAND	2020-08-31	NIQ	
1	2020-04-27	Death	MEDTRONIC IRELAND	2020-08-31	NIQ	

```
2 2020-04-27        Death  MEDTRONIC IRELAND    2020-08-31          NIQ
3 2020-04-27        Death  MEDTRONIC IRELAND    2020-08-31          NIQ
4 2020-04-27        Death  MEDTRONIC IRELAND    2020-08-31          NIQ

         Brand Name                                      Device_Problem  \
0  RESOLUTE ONYX RX                                 Material Deformation
1  RESOLUTE ONYX RX  Adverse Event Without Identified Device or Use...
2  RESOLUTE ONYX RX  Adverse Event Without Identified Device or Use...
3  RESOLUTE ONYX RX  Adverse Event Without Identified Device or Use...
4  RESOLUTE ONYX RX  Adverse Event Without Identified Device or Use...

                                          Event text
0  Event Description: DURING A PROCEDURE AN ATTEM...
1  Event Description: DURING INDEX PROCEDURE, FOU...
2  Event Description: DURING INDEX PROCEDURE, FOU...
3  Event Description: DURING INDEX PROCEDURE, FOU...
4  Event Description: DURING INDEX PROCEDURE, FOU...
```

The MAUDE data related to the RESOLUTE ONYX device has 10 variables: * The unique web address that describes a particular event * The unique report number assigned to that event * The date of the event (YYYY-MM-DD) * The type of event * The Manufacturer of the device * The date the event was reported to the manufacturer * The product code * The brand name of the device * A longer description of the type of event * A detailed textual description of the event.

We can associate a Bernoulli random variable (M) that assigns the value 1 to observations identified as Malfunctions and the value 0 to all other event types. Since M is Bernoulli, we know it takes the value 1 with the probability θ. To estimate θ, lets use Python to 1 - map the variable "Event type" to a new variable M taking values 1 when the event is a malfunction and 0 otherwise. 2- Compute the proportion of observations that take the value 1 our data set.

1: Map Event type to a 0/1 variable

[12]:
```
#### The long way

# Define a new list M in python
M = []

# Loop through each observation, extract event type, and determine if it is a␣
 ↪malfunction
for (index, row) in maudeData.iterrows():
    eventType = row['Event Type'] # assign the variable eventType the String in␣
 ↪that row.

    if eventType=="Malfunction": # if the event type is the string "Malfunction"
        M.append(1)
    else:
        M.append(0)
```

```
# Print out the variable M to make sure we have 0s and 1s
print(M)

# Compute the proportion of 1s
propOf1s = np.mean(M)
print("Proportion of 1s {:.2f}".format(propOf1s))
```

```
[1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1,
1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1,
1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1,
1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0,
1, 1, 1, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1,
1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1,
1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1,
1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0,
0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0,
0, 0, 1, 0, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 1, 1,
1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0,
1, 0, 0, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0,
1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 0, 0,
1, 0, 0, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0]
Proportion of 1s 0.64
```

A best guess is that the random variable $M \sim \text{Bern}(0.64)$. Among observations of adverse events for this device, there is a 0.64 probability they are classified as a Malfunction.

0.1.6 Geometric

Definition A random variable has a geometric distribution if it is defined on all non-negative integers and the probability distribution of these values is

$$p(X = t) = (1 - p)^{t-1} p \tag{17}$$

This type of random variable describes a sequence of trials (experiments, outcomes, etc) where the first $t - 1$ trials "failed" and the final trial was a "success". For example, the number of shots until you make a free-throw on shot t or the number of negative COVID-19 tests until the t^{th} test is positive could both have a geometric distribution.

The values of a random variable are the number of attempts until a success.

Expectation and variance If a r.v. X has a geometric distribution with parameter p,

$$X \sim \text{Geom}(p) \tag{18}$$

The expected value is

$$E(X) = \frac{1}{p} \tag{19}$$

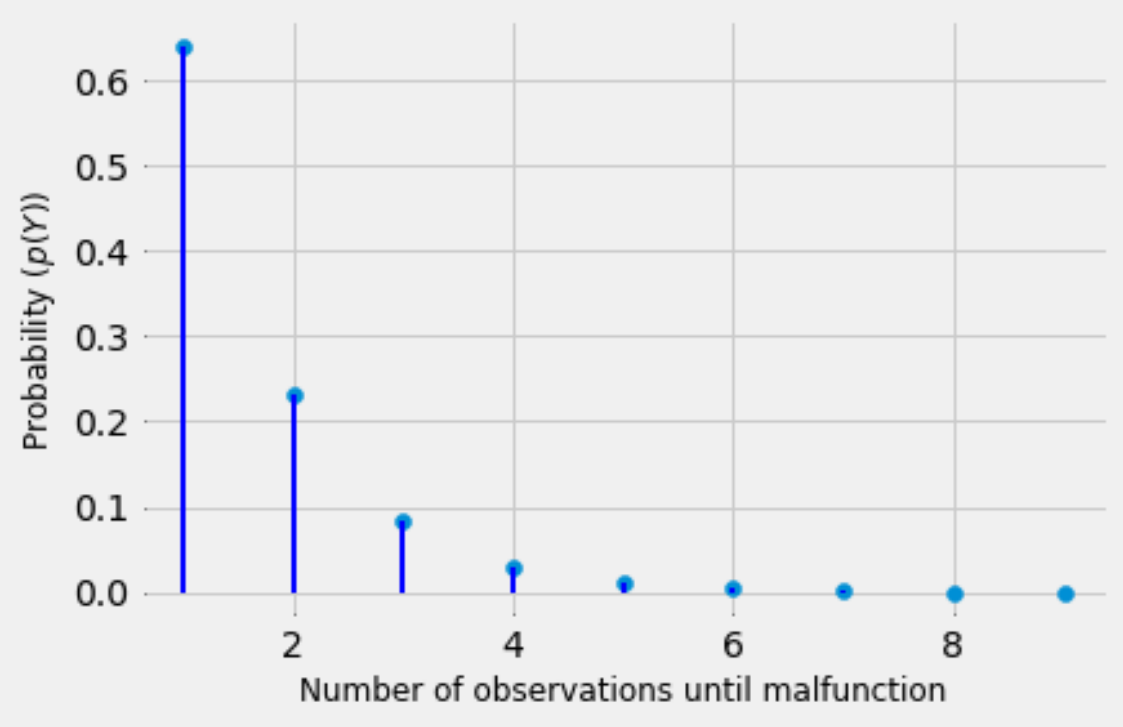
and the variance is

$$\text{Var}(X) = \frac{1}{p}\frac{(1-p)}{p} \tag{20}$$

Application (more MAUDE) We could consider a geometric variable (Y) that describes the probability of the number of adverse events up until we observe an AE classified as a malfunction. From the data, we estimated $p = 0.64$.

We can describe the probability distribution of Y by plotting on the horizontal axis the number of observation until a malfunction and on the vertical axis the corresponding probability.

[20]:
```
def prob(p,n):
    return p*(1-p)**(n-1)

xs = np.arange(1,10)
probs = []

for x in xs:
    p = prob(0.64,x)
    probs.append(p)

plt.style.use("fivethirtyeight")
fig,ax = plt.subplots()

ax.scatter(xs,probs,alpha=1.)

for (x,prb) in zip(xs,probs):
    ax.plot([x]*2, [0,prb],"b-",lw=2.)

ax.set_xlabel("Number of observations until malfunction",fontsize=12)
ax.set_ylabel(r"Probability $(p(Y))$",fontsize=12)

ax.tick_params(size=2.,direction="in")
```



0.1.7 Probability mass function and the cumulative mass function

When we look at the figure above, we see that we plotted a pair of points (x,p(x)). For every disjoint outcome of the random variable $Y = y_1, y_2, \cdots, y_n$ there is a corresponding, unique, probability $p_1, p_2, \cdots, p_n$.

This function from the disjoint values of a random variable to the probability of those outcomes is called a **probability mass function** or **p.m.f**.

The pmf for a random variable that follows a geometric distribution is

$$f(y) = (1-p)^{y-1}p \tag{21}$$

Here, y is the outcome—the number of trials until a success—and function produces the probability of our the random will have that value $f(y)$.

0.1.8 Uniform (discrete)

Definition A random variable has a uniform discrete distribution—$X \sim U(a, b)$—if it is defined for all values between two parameters a and b, and the probability of values at or between a and b is

$$p(x) = \frac{1}{N} \tag{22}$$

where N is the number of values the random variable can be. The uniform discrete distribution assigns an equal probability to all possible values of a random variable.

Expectation and variance The expectation of a r.v. following a uniform discrete distribution is the average of the two endpoints

$$E(X) = \frac{a+b}{2} \tag{23}$$

The variance is

$$Var(X) = \frac{(b-a+1)^2 - 1}{12} \tag{24}$$

0.1.9 A bit about continuous probability distributions.

Continuous probability distributions, unlike discrete distributions, do not assign probabilities to every possible outcome. This is because the probability of any single value a continuous r.v. could take is zero. To see why, lets consider the uniform discrete distribution. With 10 possible values, the probability assigns to each outcome of a uniform discrete r.v. is $\frac{1}{10}$. With 100 values the probability assigned to each value would be $\frac{1}{100}$ and with $1,000$ the probability assigned would be $\frac{1}{1,000}$. We could then reason that as the number of values grows towards infinity, the probability assigned would get closer and closer to zero.

Then how do we talk about probability with continuous r.v.s? We associate probabilities with a **probability density function** (pdf), defining probability on a continuous *interval* as the areas under the pdf. When our r.v. was discrete, the sum of the probabilities of all possible events had to equal one. There is a similar rule for continuous r.v.s and densities. The sum of the area under the curve of a pdf associated with a r.v. X over the largest possible interval must equal 1.

Like the cumulative mass function for discrete r.v.s, continuous random variables have a **cumulative density function (CDF)**. The CDF is a function that assigns a *probability* from the smallest possible value of a random variable up to a user-specified input (say x). Because the input to a *CDF* is an interval—(smallest possible value, x)—the output is a probability.

0.1.10 Uniform (continuous)

Definition A r.v. has a continuous uniform distribution $X \sim U(a, b)$ if the probability density is

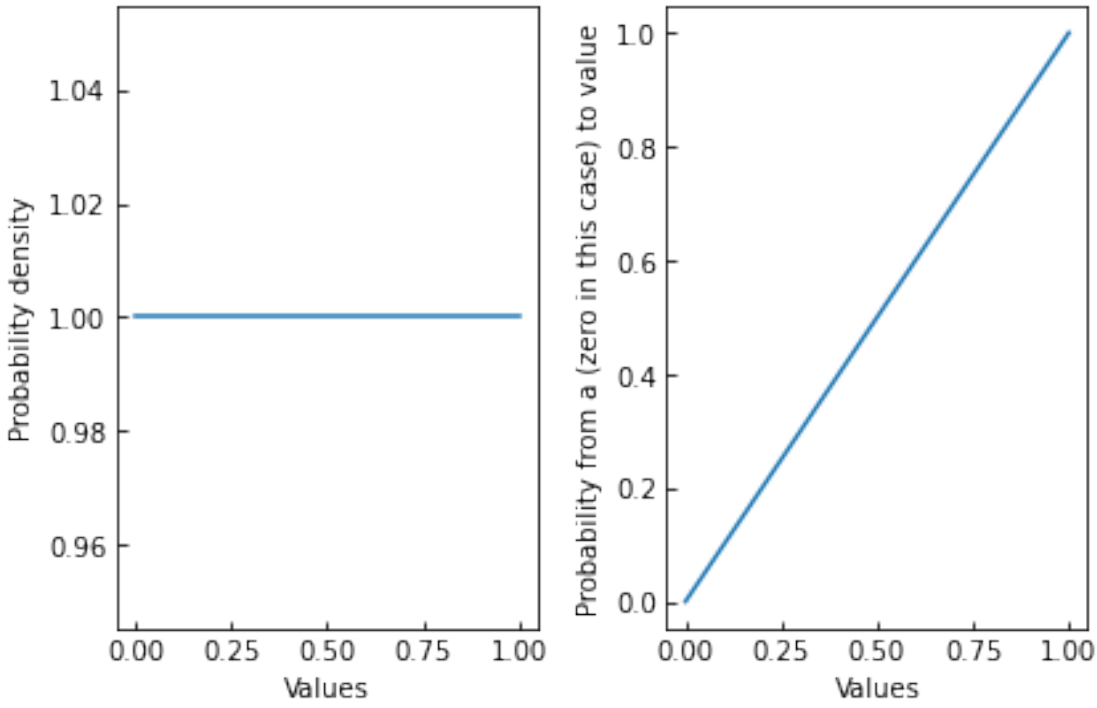
$$f(x) = \frac{1}{b-a} \tag{25}$$

The probability is defined for intervals between a and b. The CDF for a continuous uniform distribution is

$$F(x) = \frac{x}{b-a} \tag{26}$$

[9]:
```
def uniformContinuousPDF(x,a,b):
    '''
    Inputs: x and y are the endpoints of an interval
            a and b are parameters of the Uniform cont distribution
    Output:  The PDF value at x
    '''
    return 1./(b-a)

def uniformContinuousProb(x,y,a,b):
    '''
    Inputs: x and y are the endpoints of an interval
            a and b are parameters of the Uniform cont distribution
    Output: a float value that is the probability between x and y
    '''
    return (y-x)/(b-a)

a,b = 0,1
values = np.linspace(a,b,100)

pdfVals = []
for v in values:
    f = uniformContinuousPDF(v,a,b)
    pdfVals.append(f)

fig,axs = plt.subplots(1,2)

#first plot
ax = axs[0]
ax.plot( values, pdfVals  )

ax.set_xlabel("Values")
ax.set_ylabel("Probability density")
ax.tick_params(direction="in")

# compute probabilitiues from on end point (a) to a value (v)
prbVals = []
for v in values:
    f = uniformContinuousProb(a,v,a,b)
    prbVals.append(f)

#second plot
ax = axs[1]
ax.plot( values, prbVals  )

ax.set_xlabel("Values")
ax.set_ylabel("Probability from a (zero in this case) to value")
ax.tick_params(direction="in")
```

```
fig.set_tight_layout(True)

plt.show()
```



Expectation and variance The expected value (expectation) equals

$$E(X) = \frac{a+b}{2} \tag{27}$$

and the variance equals

$$Var(X) = \frac{(b-a)^2}{12} \tag{28}$$