week11-01–binomial-and-poisson

April 22, 2024

1 George McNinch Math 87 - Spring 2024

2 Week 11

3 Binomial & Poisson distributions

4 Intro

Recall that while modeling *Jane's Fish Tank Emporium*, we stipulated that "on average, there is one customer per week" meant that there was a `1/7` chance per day of a customer arriving. With this formulation, there could *never* be 2 customers in a day. On the other hand, that may not be a reasonable assumption.

In this notebook, we are going to talk about more sophisticated probabilistic descriptions.

We'll start by discussing the JFTE example again.

5 JFTE, revisited

Recall our asusmption is that the probability of a customer visiting the store each day is `p = 1/7`.

Suppose the store has a 4-hour morning shift and a 4-hour afternoon shift, and suppose that a customer is equally likely to come in the morning shift as in the afternoon shift.

Then the probability that a customer arrives in the morning shift is $\frac{1}{2} \cdot \frac{1}{7} = \frac{1}{14}$ and similarly the probability that a customer arrives in the afternoon shift is $\frac{1}{14}$.

But under this description, it is now possible to have 2 customers arrive in a day. In fact, the probability of seeing 0, 1 or 2 customers in a day is given by the following table:

# customers	probability
0	$\frac{13 \cdot 13}{14 \cdot 14}$
1	$\frac{2 \cdot 13}{14 \cdot 14}$
2	$\frac{1}{14 \cdot 14}$

Let's compute the *expected value* for the random variable A representing "number of customers arriving in a day":

$$E(A) = 0 \cdot \frac{13^2}{14^2} + 1 \cdot \frac{2 \cdot 13}{14 \cdot 14} + 2\frac{1}{14^2} = \frac{2}{14} = \frac{1}{7}$$

Thus the expected value for the day agrees with the earlier assumption.

6 More subdivision

Consider instead 4 shifts each of length 2-hours ("early morning", "late morning", "early afternoon", ...) and again suppose that the likelihood of customer arrival is the same for all four shifts. Thus the probability with which a customer arrives during any of the four 2-hour shifts is:

$$p_4 = \frac{1}{2} \cdot \frac{1}{14} = \frac{1}{2^2}\frac{1}{7} = \frac{1}{28}.$$

With this description, it is now possible to have 4 customers arrive during a day; this will happen with probabilty

$$(p_4)^4 = \frac{1}{28^4}$$

With what probability do we see 3 customers? Well, this situation will correspond to the arrival of a customer in all but one of the shifts. Now, the probability of having a customer in each shift except (say) the morning shift is

$$p_4^3 \cdot (1 - p_4) = \frac{1}{28^3} \cdot \left(\frac{27}{28}\right)$$

Thus the probability of having exactly 3 customers arrive in a day is

$$4 \cdot (p_4)^3 \cdot (1 - p_4)$$

7 The Binomial Theorem

Let X, Y be variables, and let $n \geq 1$ be a whole number. Then

$$(X+Y)^n = X^n + \binom{n}{1}X^{n-1}Y + \binom{n}{2}X^{n-2}Y^2 + \cdots + \binom{n}{n-2}X^2Y^{n-2} + \binom{n}{n-1}XY^{n-1} + Y^n$$

$$= \sum_{m=0}^{n} \binom{n}{m} X^{n-m}Y^m$$

where the *binomial coefficients* are given by the formula

$$\binom{n}{m} = \frac{n!}{m!(n-m)!}$$

For example, since $\binom{4}{2} = 6$ and $\binom{4}{3} = 4$, we have

$$(X+Y)^4 = X^4 + 4X^3Y + 6X^2Y^2 + 4XY^3 + Y^4$$

8 Return to example:

For n shifts each day, write $p_n = \frac{1}{n} \cdot \frac{1}{7} = \frac{1}{7n}$.

Then the probability that $0 \leq m \leq n$ customers arrive during the day is given by the product

$$(\clubsuit) \quad \binom{n}{m}(p_n)^m(1-p_n)^{n-m}$$

Indeed, let's represent the outcome symbolically as a list of length n, where each member of the list is either the symbol `1` or the symbol `0`.

Thus if $n = 5$, the list $[1, 0, 0, 1, 0]$ represents the outcome "a customer arrived in the first and fourth shifts, and no customers arrived in the remaining shifts".

Now, the probability with which a given list $[a_1, a_2, \ldots, a_n]$ occurs is $(p_n)^m(1-p_n)^{n-m}$ where m is the number of indices i for which $a_i = 1$. And to find the probability that m customers arrive during a given day, one needs to add the probabilities for all such lists' this sum is evidently just the product of the *number* of such lists and the quantity $(p_n)^m(1-p_n)^{n-m}$.

Thus, ($\clubsuit$) amounts to the assertion that $\binom{n}{m}$ is equal to the number of lists $[a_1, a_2, \ldots, a_n]$ with $a_i \in \{0, 1\}$ for which $m = \#\{i \mid a_i = 1\}$.

Remark: Thus the binomial coefficient $\binom{n}{m}$ "counts" the number of ways of choosing m things from a list of n things.

Remark: Think about "FOILing" the expression $(X+Y)^n = (X+Y)(X+Y)^{n-1}$. The coefficient of X^mY^{n-m} in this expression is a sum of terms which arise from lists $[a_1, a_2, \ldots, a_n]$ as above, where $m = \#\{i \mid a_i = 1\}$.

E.g. if $n = 7$, the list $[0, 0, 1, 1, 1, 0, 1]$ determines the terrm $Y \cdot Y \cdot X \cdot X \cdot X \cdot Y \cdot X = X^4Y^3$.

Observe that ($\clubsuit$) together with the binomial Theorem tell us that the probabilities we have found sum to 1:

$$\sum_{m=0}^{n} \binom{n}{m}(p_n)^{n-m}(1-p_n)^m = (1-p_n+p_n)^n = 1^n = 1$$

(use $X = p_n$ and $Y = (1 - p_n)$).

Consider the *Y-partial derivative* of the expression for $(X + Y)^n$ given by the binomial theorem. One finds that

$$n(X+Y)^{n-1} = \sum_{m=1}^{n} m\binom{n}{m} X^{n-m}Y^{m-1} = \sum_{m=0}^{n} m\binom{n}{m} X^{n-m}Y^{m-1}$$

so that

$$(\diamondsuit) \quad nY(X+Y)^{n-1} = \sum_{m=0}^{n} m\binom{n}{m} X^{n-m}Y^{m}$$

For n shifts in a day, use ($\clubsuit$) to see that the expected value for the number of customers arriving in a day is given by

$$E(A) = \sum_{i=0}^{n} i \cdot \binom{n}{i} (p_n)^i (1-p_n)^{n-i}$$

Using ($\diamondsuit$) with $Y = p_n$ and $X = (1 - p_n)$, find that

$$E(A) = n \cdot p_n = p_1$$

i.e. in the example $E(A) = \dfrac{1}{7}$ is the "daily probability" we've seen before.

9 Binomial distribution

Consider a random variable B representing the outcome of a "binomial experiment" – thus there are two outcomes: "succeed" and "fail", with "succeed" occuring with probability $0 < p < 1$ and "fail" occuring with probability $1 - p$.

Now, we consider n trials of the binomial experiment, and we write X_n for the discrete random variable that represents the number of successes from the n trials.

Just like our "customer arrival" setting, X_n is determined by the binomial distribution. Namely, the probability $P(X_n = m)$ representing the probability of m successes in n trials is given by the formula

$$P(X_n = m) = \binom{n}{m} p^m (1-p)^{n-m}.$$

A good example of the binomial distribution arises when B is the toss of a fair coin – so that $p = \dfrac{1}{2}$. In this case, the value of X_n reflects (say) the number of heads in a trial of n coin tosses.

Such a binomial distribution is not *quite* the same as our customer arrival example, though. In our case, the probability of customer arrival depends on the number n of "trials" (or rather, "shifts").

Consider a random variable B describing a binomial outcome as before, where the "success" outcome has probability $0 < p < 1$. Now consider instead the random variable Y_n which counts the number

of successes in n trials of a binomial experiment with success probability $\frac{p}{n}$. One still often refers to Y_n is a "binomial distribution" – it satisfies

$$P(Y_n = m) = \binom{n}{m} \left(\frac{p}{n}\right)^m \left(1 - \frac{p}{n}\right)^{n-m}.$$

10 A limit of binomial distributions

Let's keep working in the setting of the binomial distribution Y_n just described.

Fix m, and consider the probability of m successes in n trials, where we allow $n \to \infty$.

We have

$$P(Y_n = m) = \binom{n}{m} \left(\frac{p}{n}\right)^m \left(1 - \frac{p}{n}\right)^{n-m}.$$

Thus

$$(\heartsuit) = \lim_{n\to\infty} P(Y_n = m) = \lim_{n\to\infty} \binom{n}{m} \left(\frac{p}{n}\right)^m \left(1 - \frac{p}{n}\right)^{n-m}$$

$$= \lim_{n\to\infty} \binom{n}{m} \left(\frac{p}{n}\right)^m \left(1 - \frac{p}{n}\right)^{n} \left(1 - \frac{p}{n}\right)^{-m} = A \cdot B \cdot C$$

where

$$A = \lim_{n\to\infty} \binom{n}{m} \cdot \left(\frac{p}{n}\right)^m \quad \text{and} \quad B = \lim_{n\to\infty} \left(1 - \frac{p}{n}\right)^n \quad \text{and} \quad C = \lim_{n\to\infty} \left(1 - \frac{p}{n}\right)^{-m}$$

$(\heartsuit) = A \cdot B \cdot C$ where $A = \lim_{n\to\infty} \binom{n}{m} \cdot \left(\frac{p}{n}\right)^m$, $B = \lim_{n\to\infty} \left(1 - \frac{p}{n}\right)^n$, $C = \lim_{n\to\infty} \left(1 - \frac{p}{n}\right)^{-m}$

First notice that since m is fixed, we have:

$$\left(1 - \frac{p}{n}\right)^{-m} \to 1 \quad \text{as } n \to \infty$$

so $C = 1$.

Now recall that $\left(1 + \frac{x}{n}\right)^n \to e^x$ as $n \to \infty$ (calculus!) so that $B = e^{-p}$.

Finally,

$$\binom{n}{m} \frac{p^m}{n^m} = \frac{p^m}{m!} \frac{n!}{(n-m)!n^m} = \frac{p^m}{m!} \frac{n(n-1)\cdots(n-(m-1))}{n^m}$$

$$= \frac{p^m}{m!} 1 \cdot \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{m-1}{n}\right)$$

$$\to \frac{p^m}{m!}$$

as $n \to \infty$.

This shows that $A = \frac{p^m}{m!}$ so that the

$$(\heartsuit) = \frac{p^m e^{-p}}{m!}.$$

11 Poisson distribution

The limiting distribution described in the previous cell is called the Poisson distribution. It is a discrete random variable with a countably infinite set of outcomes. More precisely, the Poisson distribution describes a random variable X_{poisson} whose outcomes are $m = 0, 1, 2, ...$ and for which

$$P(X_{\text{poisson}} = m) = \frac{p^m e^{-p}}{m!}$$

for $m = 0, 1, 2, \cdots$.

Now, if this is really a probability distribution, it should be the case that

$$\sum_{m \geq 0} P(X_{\text{poisson}} = m) = 1$$

.

Well,

$$\sum_{m \geq 0} P(X_{\text{poisson}} = m) = \sum_{m \geq 0} \frac{p^m e^{-p}}{m!} = e^{-p} \sum_{m \geq 0} \frac{p^m}{m!} = e^{-p} e^{p} = 1.$$

Let's compute the *expected value* $E(X_{\text{poisson}})$. Well,

$$E(X_{\text{poisson}}) = \sum_{m \geq 0} m \cdot P(X_{\text{poisson}} = m)$$

$$= \sum_{m \geq 0} m \frac{p^m e^{-p}}{m!} = p e^{p} e^{-p} = p.$$

12 Return to JFTE

We may use the Poisson distribution to model customer arrival X – thus the probability that m customers arrive over the course of 1 day is given by

$$P(X = m) = \frac{p^m \cdot e^{-p}}{m!}$$

where $p = \frac{1}{7}$.

13 Remarks:

- Poisson distributions are important in queuing theory and other areas, as they describe probabilities of independent events, such as the arrival of customers.
- The first practical application was due to Ladislaus Bortkiewicz. In 1898, he investigated the number of soldiers in the Prussian army who died each year from being kicked by a horse. Poisson distributions are ideal for modeling events that have a really low probability of occurring, but many opportunities to occur.

14 Implementation

How can we use the Poisson distribution in practice? e.g. with our JFTE simulation??

Let's compute the probabilities for $m = 0, 1, 2, ...$ for customer arrival, as before:

[1]:
```
import numpy as np
import math as math

def poisson(p,m):
    return (1.*p**m/ math.factorial(m))*np.exp(-p)

#print("\n".join([f"m = {m} -- q_{m+1} = P(X={m}) = {poisson(1./7,m):.8f}" for␣
 ↪m in range(6)]))

# make a dictionary of the probabilities for various values of m
pdict={ m: poisson(1./7,m) for  m in range(8) }

# display those probabilities
[ f"P(X={m}) = {pdict[m]:.10f}" for m in pdict.keys() ]
```

[1]:
```
['P(X=0) = 0.8668778998',
 'P(X=1) = 0.1238397000',
 'P(X=2) = 0.0088456929',
 'P(X=3) = 0.0004212235',
 'P(X=4) = 0.0000150437',
 'P(X=5) = 0.0000004298',
 'P(X=6) = 0.0000000102',
 'P(X=7) = 0.0000000002']
```

15 Addendum

We want to simulate arrival of customers according to the Poisson distribution. In fact, we only approximate this, because while the Poisson distribution allows for any number of customers, our simulation is going to impose an upper bound on the number.

So: we desire a `python` function which takes as arguments `p` the base probability of an event and `M` the maximum number of events to consider.

In our case we are modeling customer arrivals, so we'll call this function `arrival`. Our function will compute the first `M-1` probabilities `q0,q1,...,q{M-1}` for the Poisson distribution. We then set `qM` to be `1 - q0 - q1 - ...`

Remember that the `numpy` function `choose` makes a pseudo-random choice: `choose(ls,ps)` chooses an element from the list ls, where the probability of choosing `ls[i]` is given by `p[i]`.

[2]:
```
from numpy.random import default_rng
rng=default_rng()

def arrival(p=1./7,M = 10,rng=default_rng()):
    qq = list(map(lambda m:poisson(p,m),range(M)))
    qq = qq + [1-sum(qq,0)]

    return rng.choice(list(range(M+1)),p=qq)
```

The function `arrival` just introduced makes it possible to simulate customer arrival using the Poisson distribution.

[3]:
```
# sample: 10 trials of 25 arrivals (with default parameters)

short_trials = [[ arrival() for _ in range(25) ] for _ in range(10) ]
```

[4]:
```
# 6 months of data
customers = [arrival(p=1./7,M=10) for n in range(6*4*7)]
```

If you inspect the lists `short_trials` or `customers` from the preceding cells, you should see that the lists mainly – perhaps even exclusively – contain the entries `0` and `1`. (Remember, the entries were pseudo-randomly generated!)

To make larger numbers of customer arrivals likely to appear in our arrival data, we need to wait longer!!

Let's use a `pandas` DataFrame to keep track of the frequency of customer counts:

[13]:
```
import pandas as pd

def get_customers(p,N):
    return [arrival(p,10) for n in range(N)]

year = 52*7

# let's make data for 10, 100 , 1000, 10000 years
# this can take a little while...

data = { n: pd.DataFrame(get_customers(1./7,n*year)) for n in [10,100,1000] }
```

[14]:
```
# a pandas `DataFrame` has a method `value_counts` which returns a table␣
 ↪indicating the number of times a value appears
```

```
# values for 10 years
data[10].value_counts()
```

[14]:
```
0    3184
1     414
2      40
3       2
Name: count, dtype: int64
```

[15]:
```
## values for 100 years
data[100].value_counts()
```

[15]:
```
0    31561
1     4523
2      294
3       22
Name: count, dtype: int64
```

[16]:
```
## values for 1,000 years
data[1000].value_counts()
```

[16]:
```
0    315493
1     45034
2      3299
3       168
4         6
Name: count, dtype: int64
```

[ ]:
```

```