1 Distributions and Densities

We are now interested in looking at important probability distributions, along with their applications.

1.1 Types of Discrete Probability Distributions

The following are some of the most important discrete probability distributions, some of which we've seen before.

- Uniform
- Bernoulli/Indicator
- Binomial
- Geometric
- Negative Binomial
- Poisson
- Hypergeometric

1.1.1 Uniform Distribution

The idea is that we put equal mass on every element in the set. More formally:

Definition 1.1: Uniform Distribution

X is **uniform** on a finite set A if $p(x) = \frac{1}{|A|}$ for all $x \in A$.

Remark: This distribution has the special property that, for all $B \subset A$, $P(X \in B) = \frac{|B|}{|A|}$.

1.1.2 Bernoulli Distribution

Essentially, we can only ever get two values: a 1 or a 0, with probability p and $1 - p$. More formally:

Definition 1.2: Bernoulli Distribution

X is **Bernoulli**(p) if $P(X = 1) = p$ and $P(X = 0) = 1 - p$.

Remark: In terms of Bernoulli, we usually think of it as a trial (where there's either a success or failure).

(Example: Indicator Random Variable.) Let E be an event. The random variable $\mathbf{1}_E$ is equal to 1 if E occurs and equal to 0 if E^C occurs. This is known as a **indicator** of E. Note that $\mathbf{1}_E$ is a Bernoulli random variable with parameter $p = \mathbb{P}(E)$.

1.1.3 Binomial Distribution

The binomial distribution involves Bernoulli trials.

Definition 1.3: Binomial Distribution

Suppose that n independent Bernoulli(p) trials are performed. Then, the number of successes observed during these n trials has the **Binomial**(n, p) distribution with

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

for $0 \leq k \leq n$.

Remark: In particular, for k successful trials, the idea is that we pick k trials to be successful (out of the n trials). Then, we have probability p^k for those trials to be successful, and probability $(1-p)^{n-k}$ for those trials to be failures. By independence, we can just multiply them out.

1.1.4 Geometric Distribution

The geometric distribution also involves Bernoulli trials.

Definition 1.4: Binomial Distribution

Suppose that we keep performing independent Bernoulli(p) trials until the first success is observed. Then, the total number of trials performed has the **Geometric**(p) distribution, with

$$p(k) = p(1-p)^{k-1}$$

for $k \geq 1$.

Remarks:

- Here, the kth trial is the successful trial, and all of the $k-1$ trials are thus failures.
- A related distribution is known as the **Shifted Geometric**(p) distribution. Instead of asking how many trials before the first success, we're asking how many *failures* before the first success. This has PMF
$$p(k) = p(1-p)^k$$
for $k \geq 0$.
- Here, X is geometric if and only if $X - 1$ is shifted geometric.

The geometric random variable is, in a sense, the discrete analogue of the (continuous) exponential random variable. Recall that the exponential random variable is the only continuous random variable with the *memoryless property*. The geometric random variable *also* has this property, and is the only *discrete random variable* that does.

Proof. Suppose that X is Geometric(p). Then[a],

$$\begin{aligned}
\mathbb{P}(X > k) &= \sum_{\ell=k+1}^{\infty} p(1-p)^{\ell-1} \\
&= p(1-p)^k \sum_{\ell=0} (1-p)^{\ell} \\
&= p(1-p)^k \frac{1}{1-(1-p)} \\
&= p(1-p)^k \frac{1}{p} \\
&= (1-p)^k.
\end{aligned}$$

Hence,

$$\begin{aligned}\mathbb{P}(X > k+\ell | X > k) &= \frac{\mathbb{P}(X > k+\ell)}{\mathbb{P}(X > k)} \\ &= \frac{(1-p)^{k+\ell}}{(1-p)^k} \\ &= (1-p)^\ell \\ &= \mathbb{P}(X > \ell).\end{aligned}$$

Thus, X has the memoryless property. □

[a] X is the number of trials until the first success. Intuitively, we note that $X > k$ occurs if and only all of the first k trials were failures.

1.1.5 Negative Binomial Distribution

The negative binomial distribution is a generalization of the geometric distribution. Instead of waiting for the first success, we wait for the kth success.

Definition 1.5

Suppose that we keep performing independent Bernoulli(p) trials until the kth success is observed. Then, the total number of trials performed has the **Negative Binomial**(k, p) distribution, with

$$p(n) = \binom{n-1}{k-1} p^k (1-p)^{n-k}$$

for $n \geq k$.

Remark: Why is this the PMF?

- For the $p^k(1-p)^{n-k}$ part, there are k successes, and there are $n-k$ failures like usual.
- For the binomial coefficient, the idea is that if the nth trial is the kth success, then there's no choice as to when the kth success will be; it has to be $\binom{n}{n}$. But, what about the $n-1$ trials, of which $k-1$ of them are successes? Well, there were exactly $k-1$ successes during the first $n-1$ trials. These could have occurred in any of $\binom{n-1}{k-1}$ ways.

1.1.6 Poisson Distribution

This is the most important distributions in all of probability theory.

Definition 1.6

A **Poisson** RV with rate $\lambda > 0$ has PMF

$$p(k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

for $k \geq 0$.

Remark: One way to remember this PMF is to recall the Taylor expansion $e^\lambda = \sum_{k=0}^\infty \frac{\lambda^k}{k!}$, which can be used to show that this is indeed a PMF, $\sum_{k=0}^\infty p(k) = 1$.

The Poisson is related to the exponential and binomial distributions. In particular, the connection with the binomial is that if N is the Binomial$(n, \lambda/n)$, then as $n \mapsto \infty$, it "converges" to Poisson(λ). Note, if

$$\mathbb{P}(N = k) = \binom{n}{k} (\lambda/n)^k (1-\lambda/n)^{n-k}.$$

For any fixed integer $k \geq 0$, we have

$$\binom{n}{k} \approx n^k/k!$$

and

$$(1-\lambda/n)^{n-k} \approx e^{-\lambda},$$

so it follows that

$$\mathbb{P}(N-k) \approx e^{-k}\frac{\lambda^k}{k!},$$

which is the PMF of a Poisson(λ). Thus, since a Binomial(n,p) is approximately Poisson(λ), when $p \approx \lambda/n$, the Poisson is useful for studying the occurrence of **rare events**.

If p is small compared with n, then

$$\boxed{\mathbb{P}(\text{Binomial}(n,p)=k) \approx \mathbb{P}(\text{Poisson}(np)=k)}.$$

A good "rule of thumb" is $n \geq 100$ and $np \leq 10$ in order for the approximation to be reasonable.

(Example.) On average, a typist makes one typo in every 1000 words. Approximate the probability of having at most 2 typos on the first 3 pages (≈ 300 words) of a manuscript they have typed.

We could use a Binomial($300, 0.001$) random variable here, but it would lead to the somewhat length calculation

$$\sum_{k=0}^{2}\binom{300}{k}(0.001)^k(0.999)^{300-k} \approx 99.6429\%.$$

Here, the calculation is essentially:

- The probability that there is at most 0 typos, and
- The probability that there is at most 1 typo, and
- The probability that there is at most 2 typos.

Note that $n=300$ and $p=\frac{1}{1000}$, so $n=300>100$ and $np=300(0.001)=0.3<10$, so it follows that we will get a good approximation; in particular, we expect a Poisson(0.3) random variable X to give a good approximation. Indeed,

$$\mathbb{P}(X \leq 2) = \sum_{k=0}^{2} e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{2}\frac{\lambda^k}{k!} = e^{-0.3}\left(1+0.3+\frac{(0.3)^2}{2}\right) \approx 99.6401\%.$$

This is nearly as good as what we found by using the Binomial.

1.1.7 Hypergeometric Distribution

This distribution is useful for sampling *without* replacement.

(Example.) Suppose that a lake contains N fish, where k of them are big and $N-k$ of them are small. Suppose that a fisherman catches n fish. Let X be the number of them that are big. Then, with probability

$$\mathbb{P}(X=x) = \frac{\binom{k}{n}\binom{N-k}{n-x}}{\binom{N}{n}},$$

where x of them will be big.

Note that if we instead sampled *with* replacement, then the probability that k of the n fish are big would just be the Binomial probability

$$b(n, k/N, x) = \binom{n}{x}(k/N)^x(1 - k/N)^{n-x}.$$

If both of N and k are much larger than n, then there is not much difference. Intuitively, because then there is little chance that we would catch the same fish twice anyhow, so sampling with or without replacement would be essentially the same.

1.1.8 Benford Distribution

In many "real life" datasets, the first digits tend to **not** be uniform on $\{1, 2, \ldots, 9\}$. In particular, digit 1 tends to be the most likely to occur. The **Benford distribution**, in many cases, does a better job of modeling this distribution of the leading digits occurring in "typical" data.

(Example.) The Benford distribution for the *first* digit (base 10) has PMF

$$p(k) = \log_{10}(1 + 1/k)$$

for $1 \leq k \leq 9$.

In particular, the first digit of most "typical" datasets is a 1 with about a 30% chance.