Math 425
Introduction to Probability
Lecture 18

Kenneth Harris
kaharri@umich.edu

Department of Mathematics
University of Michigan

March 1, 2009

Example

Example

In one of the earliest studies of the Poisson distribution, von Bortkiewicz (1898) considered deaths from mule kicks in the Prussian army corps. He collected data from 14 corps over a 20-year period.

# Deaths	# corps with x deaths in a given year
0	144
1	91
2	32
3	11
4	2

There are $n = 280$ corps years and 196 deaths. So, there is an average of $\lambda = \frac{196}{280}$.

How does this distribution compare to the Poisson approximation?

Example

☞ We compute the probabilities using the Poisson approximation with average $\lambda = \frac{196}{280} = 0.7$.

The random variable X counts deaths in a corp in a year.

$$\begin{aligned}
\mathbf{P}\{X = 0\} &= e^{-0.7} \approx 0.497 \\
\mathbf{P}\{X = 1\} &= e^{-0.7}(0.7) \approx 0.348 \\
\mathbf{P}\{X = 2\} &= e^{-0.7}\left(\frac{0.7^2}{2}\right) \approx 0.122 \\
\mathbf{P}\{X = 3\} &= e^{-0.7}\left(\frac{0.7^3}{3!}\right) \approx 0.028 \\
\mathbf{P}\{X = 4\} &= e^{-0.7}\left(\frac{0.7^4}{4!}\right) \approx 0.005
\end{aligned}$$

Example

☞ Comparing actual deaths from mule strikes (per corps) with Poisson prediction.

# Deaths	# corps with x deaths in a given year	Poisson prediction
0	144	139
1	91	97
2	32	34
3	11	8
4	2	1

Example

Example

William Feller discusses the statistics of bomb strikes in an area in the south of London during the Second world war. The area in question was divided into $24 \times 24 = 576$ small areas. There were 537 hits. The number of times an area of hit was as follows:

# of strikes	# areas with x strikes
0	229
1	211
2	93
3	35
4	7
≥ 5	1

Assuming the hits were purely random, use the Poisson approximation to find the probability that a given square would have exactly k hits.

Example

☞ We compute the probabilities using the Poisson approximation with average $\lambda = \frac{537}{596}$.

The random variable X counts bomb strikes in an area.

$$\begin{aligned}
\mathbf{P}\{X=0\} &= e^{-\lambda} \approx 0.406 \\
\mathbf{P}\{X=1\} &= e^{-\lambda}(\lambda) \approx 0.366 \\
\mathbf{P}\{X=2\} &= e^{-\lambda}\left(\frac{\lambda^2}{2}\right) \approx 0.165 \\
\mathbf{P}\{X=3\} &= e^{-\lambda}\left(\frac{\lambda^3}{3!}\right) \approx 0.05 \\
\mathbf{P}\{X=4\} &= e^{-\lambda}\left(\frac{\lambda^4}{4!}\right) \approx 0.011 \\
\mathbf{P}\{X\geq 5\} &= 1 - \mathbf{P}\{X<5\} \approx 0.0024
\end{aligned}$$

Example

☞ Comparing actual strikes with Poisson prediction.

# of strikes	# areas with x strikes	Poisson prediction
0	229	234
1	211	211
2	93	95
3	35	29
4	7	6
≥ 5	1	1

Demand

Example

The demand for a product (such as mangoes) in a grocery store typically follows a Poisson distribution. The average demand, λ, might be large, but the event of someone purchasing a mango is rare compared to the number of people who shop at the store.

☞ A grocery store purchases mangoes every week, and mango only lasts through the week before it must be discarded.

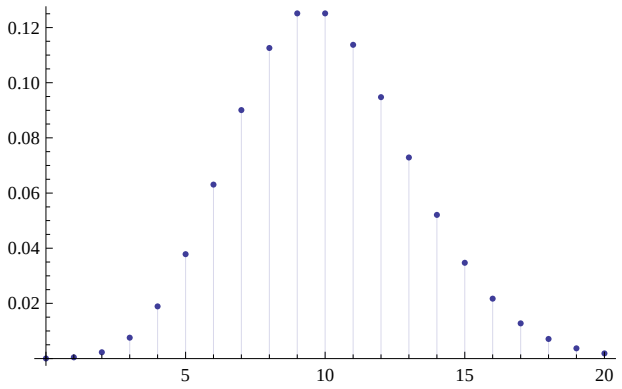
- The store averages a demand of λ mangoes,
- The store has a cost of c dollars per mango,
- The store clears a profit of d dollars for each mango sold.

How many mangoes should the store purchase to maximize their profit on mangoes for the week?

.

The shape of the Poisson distribution

☠ Warning. You might guess the store should stock λ mangoes, but this is incorrect. The reason is the shape of the Poisson distribution:



☞ Peak at λ, but slow drop-off near λ.

The shape of the Poisson distribution

☞ Let X be a r.v. with a Poisson distribution and mean λ. So,

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

☞ The distribution is at maximum probability at λ, but its rate of change for large λ is relatively small around λ:

$$\begin{aligned} p_X(k+1) - p_X(k) &= e^{-\lambda}\frac{\lambda^{k+1}}{(k+1)!} - e^{-\lambda}\frac{\lambda^k}{k!} \\ &= e^{-\lambda}\frac{\lambda^k}{k!}\Big(\frac{\lambda}{k+1} - 1\Big) \end{aligned}$$

$$\begin{aligned} p_X(k+1) &\geq p_X(k) \quad \text{when} \quad \lambda \geq k+1 \\ p_X(k+1) &\approx p_X(k) \quad \text{when} \quad k \approx \lambda \end{aligned}$$

Supply and Demand

☞ How many mangoes should our store stock?

The store stocks m mangoes; its profit, $\mathcal{P}(m)$, varies with demand X.
The expected profit depends on whether $X < m$ or $X \geq m$:

$$\begin{aligned} X < m: &\quad \overbrace{Xd}^{\text{sold}} - \overbrace{(m-X)c}^{\text{unsold}} \\ X \geq m: &\quad \underbrace{md}_{\text{sold}} \end{aligned}$$

Supply and Demand

☞ Computing expected profit when stocking m mangoes:

$$\begin{aligned} E[\mathcal{P}(m)] &= \sum_{k=0}^{m-1}\big(\overbrace{kd}^{\text{sold}} - \overbrace{(m-k)c}^{\text{unsold}}\big)p_X(k) + \overbrace{md}^{\text{sold}}\cdot\mathbf{P}\{X \geq m\} \\ &= \sum_{k=0}^{m-1}\big((d+c)k - mc\big)p_X(k) + md\cdot(1 - \mathbf{P}\{X < m\}) \\ &= (d+c)\sum_{k=0}^{m-1} kp_X(k) - c\sum_{k=0}^{m-1} mp_X(k) + md - d\sum_{k=0}^{m-1} mp_X(k) \\ &= md + (d+c)\sum_{k=0}^{m-1} kp_X(k) - (d+c)\sum_{k=0}^{m-1} mp_X(k) \\ &= md + (d+c)\sum_{k=0}^{m-1}(k-m)p_X(k) \end{aligned}$$

Supply and Demand

☞ Compute the difference between $E[\mathcal{P}(m+1)]$ and $E[\mathcal{P}(m)]$:

$$\begin{aligned} E[\mathcal{P}(m)] &= md + (d+c)\sum_{k=0}^{m-1}(k-m)p_X(k) \\ &= md + (d+c)\sum_{k=0}^{m}(k-m)p_X(k) \\ E[\mathcal{P}(m+1)] &= (m+1)d + (d+c)\sum_{k=0}^{m}(k-m-1)p_X(k) \\ &= (m+1)d + (d+c)\sum_{k=0}^{m}(k-m)p_X(k) - (d+c)\sum_{k=0}^{m}p_X(k) \end{aligned}$$

☞ The difference is

$$E[\mathcal{P}(m+1)] - E[\mathcal{P}(m)] = d - (d+c)\sum_{k=0}^{m}p_X(k)$$

Supply and Demand

$$\begin{aligned} E[\mathcal{P}(m+1)] - E[\mathcal{P}(m)] &= d - (d+c)\sum_{k=0}^{m}p_X(k) \\ &= d - (d+c)\cdot \mathbf{P}\{X \le m\} \end{aligned}$$

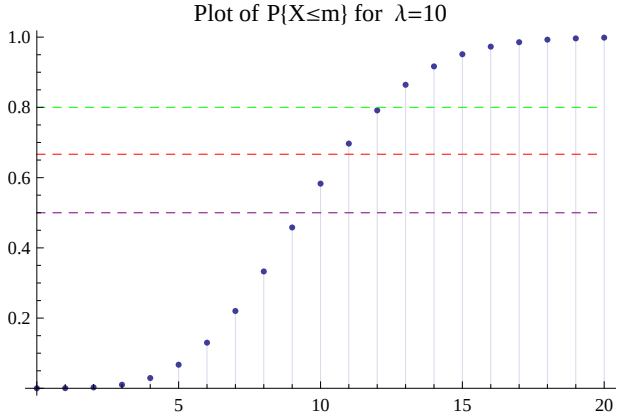
☞ Stocking $m+1$ mangoes is better than stocking m mangoes when

$$E[\mathcal{P}(m+1)] - E[\mathcal{P}(m)] > 0 \quad \text{equivalently} \quad \mathbf{P}\{X \le m\} < \frac{d}{d+c}.$$

☞ Since $\mathbf{P}\{X \le m\}$ increases, as m increases, we should look for when this value crosses the threshold $\frac{d}{d+c}$.

Supply and Demand

☞ Suppose demand is given by the Poisson r.v. X with average $\lambda = 10$.
At $d = 1$ and $c = 1$: buy 10 mangoes,
At $d = 1$ and $c = .5$: buy 11 mangoes,
At $d = 1$ and $c = .25$: buy 13 mangoes.



Divided populations

☞ A problem that arises in just about every division of science and industry is that of counting or assessing a divided population.

- Elections. Before November 4, what proportion of the population is going to vote for Obama and what proportion for McCain.
- Turkeys. UM is considering changing their fight song. They want to know how many fans like the new song and how many hate it.
- Mangoes. Some fruit are being attacked by the leaf-gall midge. A grower would like to know the extent of infestation in his crop.
- Widgits. Acme would like to know the proportion of their widgits that are defective.
- Fish. Some fish are normal and others are androgynous due to polluted water. What proportion are deformed.

Classical model of population

☞ The classical model of sampling is an urn containing red and green balls.

- The number of balls in the urn represents the size of the population.
- The two colors, red and green, represent two distinct groups which divide the population.
- Picking a ball from the urn at random corresponds to choosing a member of the population, where each member has the same chance of being chosen.

Sampling

☞ We often want to determine the proportion of a population that are reds and greens.

It is often practically impossible to count each member of the population. Instead, we randomly sample a part of the population, and extrapolate to the whole.

☞ There are two strategies for sampling:

1. Sampling with replacement. The color of a ball is noted and the ball is replaced in the urn, so that it may be picked again.

   A voter in a poll may be asked their political opinions any number of times.
2. Sampling without replacement. The color of a ball is noted and the ball is put aside, so that it cannot be picked again.

   A fish is disected to determine of it is androgynous – it cannot be easily returned to the pool.

Sampling with Replacement

Example. A population with two political groups, reds and greens, is being polled to determine the percentage of each group.

The population contains r reds and g greens. The poller chooses n people at random and ask their affilliation. The poll is otherwise anonymous.

☞ This experiment is a Bernoulli trials process;
the r.v. X counting reds is a binomial random variable.

$$\mathbf{P}\{X = k\} = \binom{n}{k} \left(\frac{r}{r+g}\right)^k \left(\frac{g}{r+g}\right)^{n-k}$$

Sampling without Replacement

Example. A population of fish from a polluted source are either normal or androgynous. Fish will be randomly removed from the population and dissected to determine if they are normal or androgynous.

☞ Suppose the population consists of r normal fish and g androgynous fish. Let $N = r + g$. The study will remove n fish. Let X be the random variable which counts the number of normal fish. Then,

$$\mathbf{P}\{X = k\} = \frac{\binom{r}{k}\binom{g}{n-k}}{\binom{N}{n}} \qquad \text{for } 0 \le k \le n.$$

Hypergeometric Random Variable

Definition

We say a random variable H is a hypergeometric random variable (with parameters N, m and n) if

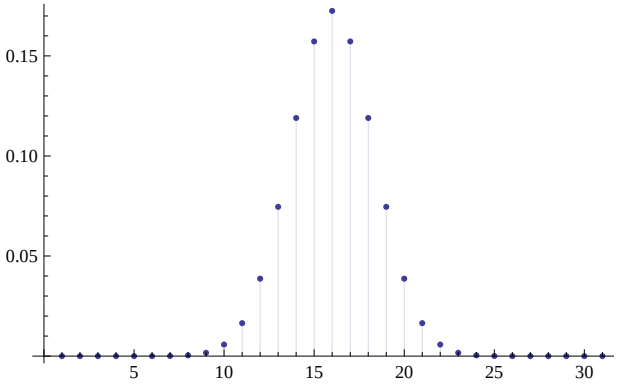
$$\mathbf{P}\{H = k\} = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$$

Consider a population of size N with m of some type and $N - m$ not of the type. If a sample of n members of the population are drawn without replacement, then the random variable H which counts the number the number k of this type from the sample of n is hypergeometric.

Graph of Hypergeometric distribution

☞ A distribution of a hypergeometric r.v. with parameters $N = 100$, $m = 50$ and $n = 30$.

Note how similar the distribution is to a binomial distribution.



Approximating a Hypergeometric r.v.

☞ Let H be a hypergeometric r.v. with parameters N, m and n. Suppose N and m are VERY LARGE compared to n.

Example. In polling n in thousands, but N and m will be in millions.

☞ Fix n, k and let $p = \frac{m}{N}$ and $(1-p) = \frac{N-m}{N}$.

$$\begin{aligned}
\mathbf{P}\{H = k\} &= \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} \\
&= \frac{m!}{(m-k)!k!} \cdot \frac{(N-m)!}{(N-m-n+k)!(n-k)!} \cdot \frac{(N-n)!n!}{N!} \\
&= \binom{n}{k} \underbrace{\left[\frac{m}{N} \cdots \frac{m-k+1}{N-k+1}\right]}_{\approx p^k} \underbrace{\left[\frac{N-m}{N-k} \cdots \frac{N-m-(n-k-1)}{N-k-(n-k-1)}\right]}_{\approx (1-p)^{n-k}} \\
&\approx \binom{n}{k} p^k (1-p)^{n-k}
\end{aligned}$$

Mean and Variance

☞ If H is hypergeometric r.v. with parameters N, m and n, then H is approximated by a binomial r.v. B with parameters n and $p = \frac{m}{N}$.

We expect the mean and variance of H to be close to the mean and variance of B when N is large compared to n.

$$E[B] = np \qquad Var(B) = np(1-p).$$

Theorem

Let H be a hypergeometric r.v. with parameters N, m and n.

The mean and variance of H are given by

$$E[H] = np \qquad \text{where } p = \frac{m}{N}$$

$$Var(H) = np(1-p) \cdot \left(1 - \frac{n-1}{N-1}\right)$$

Example

Example

There are an unknown number of moose on Isle Royale (a National Park in Lake Superior). To estimate the number of moose, 50 moose are captured and tagged. Six months later 200 moose are captured and it is found that 8 of these were tagged.

Estimate the number of moose on Isle Royale.

Solution to Example

☞ We have tagged $m = 50$ out of an unknown population N and captured $k = 8$ out of $n = 200$ previously tagged moose.

Intuitively, we would guess

$$\frac{m}{N} = \frac{k}{n} \quad \text{equivalently} \quad N = \frac{mn}{k}$$

In this case,

$$N = \frac{mn}{k} = \frac{10,000}{8} = 1250$$

So, there are 1250 moose.

☞ Why believe this guess? Is there an argument for why we should believe $N = 1250$ is the “best guess"?

Yes, the maximum likelihood estimate: what is the choice of N that makes the data we have collected most likely.

Maximum Likelihood Estimates

☞ A common problem for ecologists is to estimate the size N of a population. One method is the capture-recapture method:

(i) Capture m animals and tag (or mark) them;

(ii) Release the animals and wait some time for them to remix.

(iii) Capture n animals and count how many were previously tagged.

☞ Suppose the change in total and tagged population is negligible.

Let X_N (where N is an unknown) be the r.v. which counts the number of tagged animals in a sample of n animals caught.

X_N is hypergeometric:

$$\mathbf{P}\{X_N = k\} = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$$

☞ We estimate the population N by choosing it to maximize the probability $\mathbf{P}\{X_N = k\}$. This is the maximum likelihood estimate.

Maximum Likelihood Estimates

☞ We aim to choose N to maximize $\mathbf{P}\{X_N = k\}$:

$$\mathbf{P}\{X_N = k\} = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$$

☞ $\mathbf{P}\{X_N = k\}$ is greater than $\mathbf{P}\{X_{N-1} = k\}$ when

$$1 < \frac{\mathbf{P}\{X_N = k\}}{\mathbf{P}\{X_{N-1} = k\}} = \frac{(N-m)(N-n)}{N(N-m-n+k)}$$

Equivalently, we want the largest integer N satisfying

$$N \leq \frac{mn}{k}$$

☞ Our first intuitive guess is now justified.