1. More Bounding

Imagine that X is the number of customers that enter a bank at a given hour. To simplify everything, in order to serve n customers you need at least n tellers. One less teller and you won't finish serving all of the customers by the end of the hour. You are the manager of the bank and you need to decide how many tellers there should be in your bank so that you finish serving all of the customers in time. You need to be sure that you finish in time with probability at least 95%.

a. Assume that from historical data you have found out that $\mathbf{E}[X] = 5$. How many tellers should you have?

Answer: You should apply Markov's. You get $\Pr[X \geq 100] \leq E[X]/100 = 0.05$. So you need 100 tellers.

b. Now assume that you have also found out that $\mathrm{Var}(X) = 5$. Now how many tellers do you need?

Answer: You should apply Chebyshev's. You get $\Pr[X > 5+10] \leq 5/(10^2) = 0.05$. So you need 15 tellers this time.

2. How Many Coupons?

Consider the coupon collecting problem covered in note 19. There are n distinct types of coupons that we wish to collect. Every time we buy a box, there is one coupon in it, with equal likelihood of being any one of the types of coupons. We want to figure out how many boxes we need to buy in order to get one of each coupon. For this problem, we want to bound the probability that we have to buy lots of coupons — say substantially more than $n \ln n$ coupons.

a. We represent X, the number of boxes we have to buy, as a sum of other random variables. Let X_i represent the number of boxes you buy to go from $i-1$ to i distinct coupons in your hand. The let $X = \sum_{i=1}^{n} X_i$. Argue that each X_i is an independent random variable with a geometric distribution.

Answer: Once we get $i-1$ coupons, there's a chance of $\frac{n-i+1}{n}$ chance of getting a coupon not already collected, if not we would keep on trying until we get a new coupon, thus X_i is a geometric random variable where

$$\text{Each } X_i \sim \text{Geom}\left(\frac{n-i+1}{n}\right)$$

b. Prove that $\mathbb{E}[X] \approx n \ln n$. Remember that the expectation of $\text{Geom}(p)$ is $\frac{1}{p}$.

Answer: We find the expectation of X using linearity of expectation and the expectation of a geometric random variable. Then we use the fact that $\sum_{i=1}^{n} \frac{1}{n} \approx \ln n$ (This sum is bounded both below and above by

$\int_{x=1}^{n} \frac{dx}{x} + c$ where c is a constant).

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{i=1}^{n} \mathbb{E}[X_i] \\
&= \sum_{i=1}^{n} \frac{n}{n-i+1} \\
&= n \sum_{i=1}^{n} \frac{1}{i} \\
&\approx n \ln n
\end{aligned}$$

c. We wish to use Chebyshev's inequality to bound the probability we have to buy substantially more than $\mathbb{E}[X] = n \ln n$ boxes. In order to do this, we need to compute the variance of a geometric random variable. We know from lecture that the variance of Geom(p) is $\frac{1-p}{p^2} \leq \frac{1}{p^2}$. Prove that $Var[X] \leq 2n^2$. (Note that $\sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{\pi^2}{6}$)

Answer: Since each X_i is linear, we know that the variance of their sum is linear.

$$\begin{aligned}
VarX &= \sum_{i=1}^{n} \frac{1-p_i}{p_i^2} \\
&\leq \sum_{i=1}^{n} \frac{1}{p_i^2} \\
&= \sum_{i=1}^{n} \frac{n^2}{(n-i+1)^2} \\
&\leq n^2 \frac{\pi^2}{6} \\
&< 2n^2
\end{aligned}$$

d. This means that the standard deviation for X scales like n and not like the expectation $n \ln n$. Use Chebyshev's inequality to show that $Pr[X \geq \alpha n \ln n]$ tends to zero for any $\alpha > 1$ as $n \to \infty$.

Answer:

$$\begin{aligned}
Pr[X \geq \alpha n \ln n] &\approx Pr[X \geq (\alpha - 1) n \ln n + \mathbb{E}[X]] \\
&= Pr[X - \mathbb{E}[X] \geq (\alpha - 1) n \ln n] \\
&\leq Pr[|X - \mathbb{E}[X]| \geq (\alpha - 1) n \ln n] \\
&\leq \frac{Var[X]}{((\alpha - 1) n \ln n)^2} \\
&\leq \frac{2}{(\alpha - 1)^2 (\ln n)^2}
\end{aligned}$$

$$\lim_{n \to \infty} \frac{2}{(\alpha - 1)^2 (\ln n)^2} = 0$$

3. Poission

a. A textbook has on average one misprint per page. You may assume that misprints are "rare events" that obey the Poisson distribution. What is the chance that you see exactly 4 misprints on page 1?

Answer: We assume that misprints are "rare events" that obey the Poisson distribution, with parameter $\lambda = 1$ (i.e. at a rate of one misprint per page). Hence

$$\Pr\left[\text{exactly 4 misprints on page 1}\right] = \frac{\lambda^4}{4!}e^{-\lambda} = \frac{1}{4!e} \approx 0.0153.$$

b. Suppose the box has 6 brown balls and 4,000 purple balls. A random sample of size n is selected with replacement and X= "number of brown balls selected". Write the distribution of X. Can this be closely approximated with a Poisson distribution? If so, write the approximate distribution. If not, explain why not. If so, but only under certain conditions, explain these conditions and write the approximate distribution.

Answer:

The distribution is Binomial($n, p = \frac{6}{4{,}006}$). Thus,

$$\Pr(X = k) \quad = \quad \binom{n}{k} p^k (1-p)^{n-k} = \binom{n}{k}\left(\frac{6}{4006}\right)^k \left(1 - \frac{6}{4006}\right)^{n-k} \quad \text{for } k = 0, 1, 2, \ldots, n$$

Because p is small, we can approximate this with Poisson(np) when n is large. For many applications, a good rule of thumb for this approximation is that it should be the case that $n \geq 20$ and $p \leq 0.05$. Here we have $p = \frac{6}{4006} = 0.0015$, so when $n \geq 20$ the distribution Poisson(np) will closely approximate Binomial(n, p). The Poisson distribution with parameter $\lambda = np$ is

$$\Pr(X = k) \quad = \quad \frac{\lambda^k}{k!}e^{-\lambda} = \frac{n^k p^k}{k!}e^{-np} \quad \text{for } k = 0, 1, 2, \ldots, n$$

The approximation is only good for small values of k relative to n. Because p is small, $\Pr(X = k) \approx 0$ for large values of k so we don't usually consider them. Some of the probabilities for small k in the Poisson approximation are

$$\begin{aligned}
\Pr(X = 0) &= \frac{n^0 p^0}{0!}e^{-np} = e^{-\frac{6n}{4006}} \\
\Pr(X = 1) &= \frac{n^1 p^1}{1!}e^{-np} = \frac{6}{4006} n e^{-\frac{6n}{4006}} \\
\Pr(X = 2) &= \frac{n^2 p^2}{2!}e^{-np} = \left(\frac{6}{4006}\right)^2 n^2 e^{-\frac{6n}{4006}} \\
\Pr(X = 3) &= \frac{n^3 p^3}{3!}e^{-np} = \left(\frac{6}{4006}\right)^3 n^3 e^{-\frac{6n}{4000}} \\
\Pr(X = 4) &= \frac{n^4 p^4}{4!}e^{-np} = \left(\frac{6}{4006}\right)^4 n^4 e^{-\frac{6n}{4000}} \\
&\vdots
\end{aligned}$$

The following table compares these probabilities for small values of k with $n = 20$.

	Binomial$(20, \frac{6}{4006})$	**Poisson**$(\frac{120}{4006})$
$\Pr(X = 0)$	0.970467	0.970489
$\Pr(X = 1)$	0.029114	0.0290711
$\Pr(X = 2)$	0.000414875	0.000435413
$\Pr(X = 3)$	$3.73387 * 10^{-6}$	$4.34761 * 10^{-6}$
$\Pr(X = 4)$	$2.38034 * 10^{-8}$	$3.25582 * 10^{-8}$