Topics: Expectation, Variance

1 Life Insurance[1]

As an extended example of probability, we analyze a simple life insurance system. A real system would be too cumbersome to look at, so we make many simplifications here.

Here are the basic rules for our system:

1. You pay b dollars to the insurance company when you are born. You never have to pay again.
2. If you die before age c, the company pays your beneficiaries d dollars.
3. The insurance company is non-profit, so just wants to break even.

Given these rules, what should the insurance company set as the values of b and d, in terms of c? Let X be the age at which a person dies. The fraction of its customers the insurance pays is then the fraction of those that die before age c, or $\Pr[X < c]$. Then b and d are related by $b = d \cdot \Pr[X < c]$.

Let's do a detailed example, where $c = 60$ and $d = \$1,000,000$. We need to compute $\Pr[X < 60]$.

1.1 Distribution of Death

Before we can calculate $\Pr[X < 60]$, we need to know what the distribution of X looks like. First, let's assume that nobody lives past 100. Now we can't just take the distribution to be uniform in the range $\{1, \cdots, 100\}$, since a person is more likely to dies as they get older. So let's assume a linear distribution, $\Pr[X = k] = k/N$ for $k \in \{1, \cdots, 100\}$. We calculate the constant N in order to ensure the probabilities sum to 1:

$$\begin{aligned}\sum_{i=1}^{100} \Pr[X = k] &= \sum_{i=1}^{100} i/N \\ &= 1/N \cdot \sum i = 1^{100} i \\ &= 1/N \cdot 5050 \\ &= 1,\end{aligned}$$

so $N = 5050$.

1.2 Life Expectancy

The first thing we should calculate is the expected age at which a person dies. We have

$$\begin{aligned}\mathbb{E}[X] &= \sum_{i=1}^{100} i \times \Pr[X = i] \\ &= \sum_{i=1}^{100} i \times i/N\end{aligned}$$

[1] This section is so blatantly ripped off of Felix Wu's notes that I have to give him credit here.

$$= 1/N \cdot \sum_{i=1}^{100} i^2$$

$$= 1/N \cdot \frac{100 \cdot (100+1) \cdot (2 \cdot 100 + 1)}{6}$$

$$= 67,$$

where we used the identity

$$\sum_{i=1}^{n} i^2 = \frac{n \cdot (n+1) \cdot (2n+1)}{6}$$

in the fourth line.

Knowing just the expectation is not enough to calculate $\Pr[X < 60]$. Consider the two distributions A where $\Pr[X = 67] = 1$ and B where $\Pr[X = 55] = \Pr[X = 79] = 0.5$. In A, $\Pr[X < 60] = 0$, whereas in B, $\Pr[X < 60] = 0.5$.

The variance is what makes the difference in the above distributions. It is variance that makes insurance useful. If there were no variance, everyone would know when they would die and thus no one would need life insurance.

1.3 Variance and Chebyshev's Inequality

We proceed by calculating the variance of the age at which a person dies. We have

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2,$$

so we need to first calculate $\mathbb{E}[X^2]$:

$$\mathbb{E}[X^2] = \sum_{i=1}^{100} i^2 \times \Pr[X^2 = i^2]$$

$$= \sum_{i=1}^{100} i^2 \times \Pr[X = i]$$

$$= \sum_{i=1}^{100} i^2 \times i/N$$

$$= 1/N \cdot \sum_{i=1}^{100} i^3$$

$$= 1/N \cdot N^2$$

$$= 5050,$$

where in the fifth line, we used the identity

$$\sum_{i=1}^{n} i^3 = (\sum_{i=1}^{n} i)^2.$$

Then

$$\mathrm{Var}[X] = 5050 - 67^2$$

$$= 561.$$

Now recall Chebyshev's inequality

$$\Pr[|X - \mathbb{E}[X]| > r] \leq \frac{\text{Var}[X]}{r^2}.$$

We want to calculate $\Pr[X < 60] = \Pr[\mathbb{E}[X] - X > 7]$. So in order to use Chebyshev's, we need to plug in $r = 7$. Note that this also allows $X > 74$, but we can't do any better with Chebyshev's. So we have

$$\begin{aligned}
\Pr[X < 60] &= \Pr[67 - X > 7] \\
&\leq \Pr[|X - 67| > 7] \\
&\leq \frac{\text{Var}[X]}{7^2} \\
&= \frac{561}{49} \\
&= 11.45.
\end{aligned}$$

But notice a problem here. Probabilities are always at most 1. Chebyshev's tells us that $\Pr[X < 60] \leq 11.45$, which is less than we already knew!

Notice that in order for Chebyshev's to give us a bound less than 1, we must have $r > \sigma(X)$ so that $\frac{\text{Var}[X]}{r^2} = \frac{\sigma(x)^2}{r^2} < 1$. Thus the inequality gives us no information when we are looking within a standard deviation from the mean.

Even in general, Chebyshev's still gives us a weak bound. It's usefulness is due to the fact that it is easy to compute and only requires knowledge of the expectation and variance of a random variable.

1.4 Exact Solution

In this case, since the distribution is so simple, we can compute $\Pr[X < 60]$ directly. We have

$$\begin{aligned}
\Pr[X < 60] &= \sum_{i=1}^{59} \Pr[X = i] \\
&= \sum_{i=1}^{59} i/N \\
&= 1/N \cdot 1770 \\
&= 0.35.
\end{aligned}$$

Thus the insurance company should set $b = 0.35 \cdot \$1,000,000 = \$350,000$, quite a large sum of money!

2 The Florida Debacle

Recall the 2000 presidential election. At the center of the scandal were the infamous "butterfly ballots" of Palm Beach County. Many people claimed that the format of these ballots resulted in many votes intended for Al Gore to go to Pat Buchanan. Here we will analyze the statistical significance of the number of votes Buchanan received in that county.

The percentages of votes cast for each of the candidates in the entire state of Florida were as follows:

Gore	Bush	Buchanan	Nader	Browne	Others
48.8%	48.9%	0.3%	1.6%	0.3%	0.1%

In Palm Beach County, the actual votes cast (before the recounts began) were as follows:

Gore	Bush	Buchanan	Nader	Browne	Others	Total
268945	152846	3407	5564	743	781	432286

To model this situation probabilistically, we need to make some assumptions. Let's model the vote cast by each voter in Palm Beach County as a random variable X_i, where X_i takes on each of the six possible values (five candidates or "Others") with probabilities corresponding to the Florida percentages. (Thus, e.g., $\Pr[X_i = \text{Gore}] = 0.488$.) There are a total of $n = 432286$ voters, and their votes are assumed to be mutually independent. Let the r.v. B denote the total votes cast for Buchanan in Palm Beach County (i.e., the number of voters i for which $X_i = \text{Buchanan}$).

We first compute the expectation and variance of B. Let B_i be a random variable representing whether the ith person voted for Buchanan, i.e., $B_i = 1$ if and only if $X_i = \text{Buchanan}$. Note that the B_i's are independently and identically distributed, with $\mathbb{E}[B_i] = 0.003$ and $\text{Var}[B_i] = 0.003 \times (1 - 0.003) = 0.002991$. Moreover, by linearity of expectation and independence, we find that $\mathbb{E}[B] = \sum_{i=1}^{n} \mathbb{E}[B_i] = 432286 \times 0.003 \approx 1297$ and $\text{Var}[B] = \sum_{i=1}^{n} \text{Var}[B_i] = 432286 \times 0.002991 \approx 1293$.

Now we use Chebyshev's inequality to compute an upper bound b on the probability that Buchanan receives at least 3407 votes, so that

$$\Pr[B \geq 3407] \leq b.$$

Chebyshev's inequality promises that

$$\Pr[|B - \mathbb{E}[B]| \geq a] \leq \text{Var}[B]/a^2.$$

In our case $\mathbb{E}[B] = 1297$, $\text{Var}[B] = 1293$, so if we take $a = 2110$, we find that $\Pr[|B - 1297| \geq 2110] \leq 1293/2110^2 \approx 0.0003$. Now note that the condition $|B - 1297| < 2110$ is equivalent to the condition $-813 < B < 3407$, and since B is non-negative, we find that $\Pr[B > 3407] \leq 0.0003$ (roughly), so we can take $b \approx 0.0003$. In other words, receiving 3407 votes for Buchanan in Palm Beach County seems very unlikely to happen by chance, under this simple model.

We can use the Central Limit Theorem to achieve a better bound on this probability. Note that the value 3407 is about $(3407 - 1297)/\sqrt{1293} \approx 58.7$ standard deviations above the mean. The probability that a normally distributed r.v. is at least 58 standard deviations above the mean is incredibly small. The probability of being 6 standard deviations above the mean is already 10^{-9}, and decreasing exponentially, so the probability for 58 standard deviations would be too small to compute on a calculator. However, one can show using what's known as a "Chernoff bound" that the answer will be at most $\exp\{-.3 \times (2110/1297)^2 \times 1297\} \approx e^{-1030} \approx 1/10^{447}$, which as you see is exceedingly tiny.

Now let's take another look at our assumptions to see how they affect this result. We assumed that everyone votes at random according to a probability distribution when they go to the polls, but in reality most people have already made up their mind by then. If we assume that only 20% of the population votes randomly and the rest exactly according to the Florida percentages, the probability that Buchanan received 3407 votes decreases further. The assumption that Palm Beach votes according to the Florida percentages is another unreasonable simplification. Notice that it is a left-leaning county. Considering that Buchanan is right-wing, it would be even more unlikely that he receive so many votes.

Does this affect your opinion of who should have won the presidency?