

A Probability Exercise

Hamilton Hoxie Ackerman

<http://hhackerman.com/2009/07/a-probability-exercise/>

July 1, 2009

Q: Given n observations in a training set, suppose you uniformly sample n times. In the long run, approximately what proportion of the original observations will be left out of your sample?

A: If you think about it for a minute, you'll realize that the value that we're looking for is exactly the same as the probability of an arbitrary element in our training set being left out of a sample. Furthermore, since we're sampling uniformly with replacement, each element is equally like to be chosen on any draw. Let's formalize this a little bit:

Let \mathbf{X} = the number of times an arbitrary element of the training set is selected. Then $\mathbf{X} \sim \text{Bin}(n, \frac{1}{n})$. Then the pmf of \mathbf{X} is $f(x) = P(\mathbf{X} = x) = \binom{n}{x} \left(\frac{1}{n}\right)^x \left(1 - \frac{1}{n}\right)^{n-x}$. Thus,

$$P(\mathbf{X} = 0) = f(0) \tag{1}$$

$$= \binom{n}{0} \left(\frac{1}{n}\right)^0 \left(1 - \frac{1}{n}\right)^n \tag{2}$$

$$= 1 * 1 * \left(1 - \frac{1}{n}\right)^n \tag{3}$$

This last equation gives us the formula for the probability that an arbitrary element of the training set is excluded from the with-replacement sampling. To see where "about one-third" comes from, let's evaluate this expression for some values of n :

```
> # This is R, but you could do this in practically any language!  
> for (k in seq(from=5, to=95, by=5)) {cat("k =", k, "\t", (1-(1/k))^k, "\n")}  
k = 5    0.32768  
k = 10   0.3486784  
k = 15   0.3552644  
k = 20   0.3584859  
k = 25   0.3603967  
k = 30   0.3616615
```

```

k = 35  0.3625605
k = 40  0.3632324
k = 45  0.3637536
k = 50  0.3641697
k = 55  0.3645095
k = 60  0.3647923
k = 65  0.3650313
k = 70  0.365236
k = 75  0.3654132
k = 80  0.3655681
k = 85  0.3657048
k = 90  0.3658262
k = 95  0.3659347

```

At least empirically, we do seem to approach “about one-third,” or slightly more accurately, “about 37%”. Can we ground this result in some theory, particularly for large n ? Indeed, we can! Recall the following famous result from calculus / mathematical statistics / real analysis / all the other places I saw this:

$$\lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n = e^1,$$

with the superscript on e coming from the “1” in the numerator of the fraction. We can rewrite Equation 3 to look more like this as followings:

$$\begin{aligned} \lim_{n \rightarrow \infty} 1 * 1 * \left(1 - \frac{1}{n}\right)^n &= \lim_{n \rightarrow \infty} \left(1 + \frac{-1}{n}\right)^n \\ &= e^{-1} \\ &= 0.3678794\dots \end{aligned}$$

So indeed, we’ll leave about 37% of the observations in our training set when we uniformly sample with replacement to form a new sample of size n .

Food for thought: Why does this also produce the answer we’re looking for?

```

# dpois(0,1) = P(X=0) given X ~ Pois(1)
> dpois(0,1)
[1] 0.3678794

```

Thanks for reading!