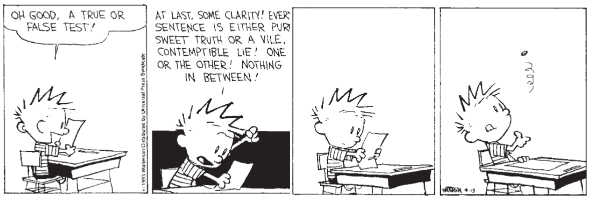
Lesson 30: What are bernoulli trials?



Bernoulli Trials

Random experiments with two outcomes

Success
probability of success is constant p

Failure
probability of failure is $q = (1-p)$

* trials are independent
* a bernoulli trial can be framed as a YES/NO question

Examples: coin flips, foul shots in basketball, true-false quiz, multiple choice test, rolling a die (six is a success, all else failure)

Several models can be used to estimate probabilities for bernoulli trials.

1. **Geometric Model**
   - When we want to know: How long till the first success in a series of bernoulli trials?
   - (Waiting time situations - till first success)
   - Requires one parameter *p* (probability of success) **Geom(p)**

x = number of trials until first success occurs

$P(X = x) = q^{x-1}p$

$$E(X) = \mu = \frac{1}{p} \qquad \sigma = \sqrt{\frac{q}{p^2}}$$

Recall chapter 11 simulation

- Cereal manufacturer puts pictures of athletes on cards in boxes of cereal.
- 20% boxes - Tiger woods
- 30% boxes - Becham
- 50% boxes - Serena Williams

You're a huge Tiger Woods fan. You don't care about completing the whole sports card collection, but you've just got to have the Tiger Woods picture. How many boxes do you expect you'll have to open before you find him?

Is this a bernoulli trial?

Open box -> success (find tiger), failure(not tiger)
Probability of success same for each trial (p=.2)
Trials are independent *****

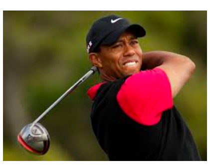
Independence

After you open a box, there is one less in circulation

so.....

We don't have infinite population
you are sampling without replacement
the probability of success changes
Oh no......

BUT.....
We can pretend the trials are independent:
10% condition: if independence assumption for bernoulli trials is violated, it is okay to proceed as long as the sample is smaller than 10% of the entire population that you are sampling from.



Let Y = # boxes we need to open to find woods card

P(Y=1) =

P(Y=2) =

.

.

.

P(Y=5) =

How many boxes do you expect to open until you find tiger?

E(Y) =

Spam and the Geometric model

Postini is a global company specializing in communications security. The company monitors over 1 billion Internet messages per day and recently reported that 91% of e-mails are spam! Let's assume that your e-mail is typical—91% spam. We'll also assume you aren't using a spam filter, so every message gets dumped in your inbox. And, since spam comes from many different sources, we'll consider your messages to be independent.

Questions: Overnight your inbox collects e-mail. When you first check your e-mail in the morning, about how many spam e-mails should you expect to have to wade through and discard before you find a real message? What's the probability that the 4th message in your inbox is the first one that isn't spam?

2. **Binomial Model**

- When we're interested in the probability of a number of successes in *n* trials
- Requires two parameters

Biom(*n*,*p*)

number of trials

probability of success

x = # of successes in n trials

$P(X = x) = {}_nC_x \, p^x q^{n-x}$

$\mu = np$

$\sigma = \sqrt{npq}$

The communications monitoring company *Postini* has reported that 91% of e-mail messages are spam. Suppose your inbox contains 25 messages.

Questions: What are the mean and standard deviation of the number of real messages you should expect to find in your inbox? What's the probability that you'll find only 1 or 2 real messages?

I assume that messages arrive independently and at random, with the probability of success (a real message) $p = 1 - 0.91 = 0.09$. Let X = the number of real messages among 25. I can use the model Binom(25, 0.09).

$$
\begin{aligned}
E(X) &= np = 25(0.09) = 2.25 \\
SD(X) &= \sqrt{npq} = \sqrt{25(0.09)(0.91)} = 1.43 \\
P(X = 1 \text{ or } 2) &= P(X = 1) + P(X = 2) \\
&= \binom{25}{1}(0.09)^1(0.91)^{24} + \binom{25}{2}(0.09)^2(0.91)^{23} \\
&= 0.2340 + 0.2777 \\
&= 0.5117
\end{aligned}
$$

Among 25 e-mail messages, I expect to find an average of 2.25 that aren't spam, with a standard deviation of 1.43 messages. There's just over a 50% chance that 1 or 2 of my 25 e-mails will be real messages.

3. **Normal Model**
   - When dealing with a large number of trials in a Binomial situation, making direct calculations of the probabilities becomes tedious (or outright impossible).
   - can be used to approximate binomial probability if we expect at least 10 successes and 10 failures.

$np \geq 10$ and $nq \geq 10$

The communications monitoring company *Postini* has reported that 91% of e-mail messages are spam. Recently, you installed a spam filter. You observe that over the past week it okayed only 151 of 1422 e-mails you received, classifying the rest as junk. Should you worry that the filtering is too aggressive?

What's the probability that no more than 151 of 1422 e-mails is a real message?

I assume that messages arrive randomly and independently, with a probability of success (a real message) $p = 0.09$. The model Binom(1422, 0.09) applies, but will be hard to work with. Checking conditions for the Normal approximation, I see that:

- ✓ These messages represent less than 10% of all e-mail traffic.
- ✓ I expect $np = (1422)(0.09) = 127.98$ real messages and $nq = (1422)(0.91) = 1294.02$ spam messages, both far greater than 10.

It's okay to approximate this binomial probability by using a Normal model.

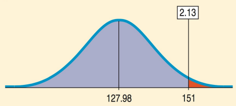
$$\mu = np = 1422(0.09) = 127.98$$

$$\sigma = \sqrt{npq} = \sqrt{1422(0.09)(0.91)} \approx 10.79$$

$$P(x \le 151) = P\left(z \le \frac{151 - 127.98}{10.79}\right)$$

$$= P(z \le 2.13)$$

$$= 0.9834$$



Among my 1422 e-mails, there's over a 98% chance that no more than 151 of them were real messages, so the filter may be working properly.