NORTHWESTERN UNIVERSITY | SCHOOL OF CONTINUING STUDIES

Handout: Binomial Distribution
PREDICT 401: Introduction to Statistical Analysis

Binomial Probability Distributions

I. Probability Distributions

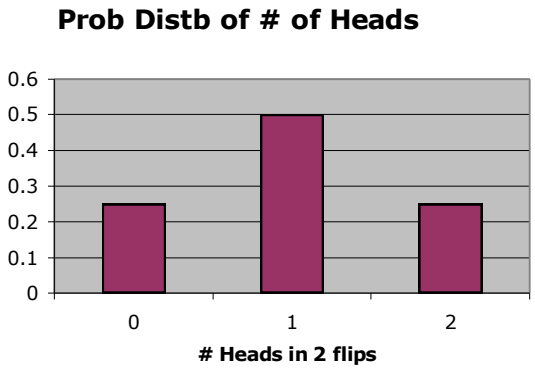
Why do we care about probability in predictive analytics?

- To understand the nature of uncertainty. In all systems, whether comprised of people, molecules, or anything else, there are MANY, MANY unknowns. If we can manage that uncertainty, we can better analyze those systems.
- For statistics, we need to be able to build **probability distributions**. They represent a key building block for the entire discipline of statistics.

A probability distribution can be shown as a graph where the horizontal axis includes the range of "outcomes" or "values" of interest, and the vertical axis is the probability of each "outcome" or "value" occurring.

Example:

You are interested in how many "heads" occurs when you flip a fair coin twice.
What are the chances that 0, 1, and 2 heads occur?
Possible outcomes: HH, HT, TH, and TT.
Each outcome happens 25% of the time.
P(0 heads) = .25
P(1 head) = .25+.25 = .50
P(2 heads) = .25
The probability distribution looks like:



II. The Binomial Distribution

Before we can start to use distributions, we need to understand a little more about how they are built. **The binomial distribution** is an important construct to the study of statistics.

This distribution also is a good context for introducing the concept of the **expected value**, which is the mean of a distribution. It can be a very useful tool in statistics and in decision-making.

Mathematically, the expected value is:

$$E(X) = \mu = \sum_{i=1}^{n}(x_i * p(x_i))$$

where x_i is the value of the i^{th} event, and $p(x_i)$ is the probability of the i^{th} event occurring.

> Don't let the term "expected value" mislead you. It is just an **average**. The expected value doesn't have to actually be one of the possible values. It is merely an average.

Example: Imagine you work for a national lobbying organization that is supporting one congressional candidate in each of four states.

You estimate that if you use your own strategy (Strategy 1), the probability of each candidate winning is .55. A co-worker suggests a strategy (Strategy 2) that re-allocates resources to ensure more wins. She argues that if your resources were concentrated in two states, you could raise the probability of wins in those two states to .75. The "abandoned" candidates would then have probability of winning of .3

Which strategy is better?
We will answer this using three tools:

- The probability distribution.
- The expected value of wins in each distribution.
- The standard deviation of number of wins in each distribution.

Strategy 1

1) Build a probability distribution. What are the possible outcomes? What is the probability of each outcome occurring?
Lose all four states (x=0)
Win exactly one state (x=1)
Win exactly two states (x=2)
Win exactly three states (x=3)
Win all four states (x=4)

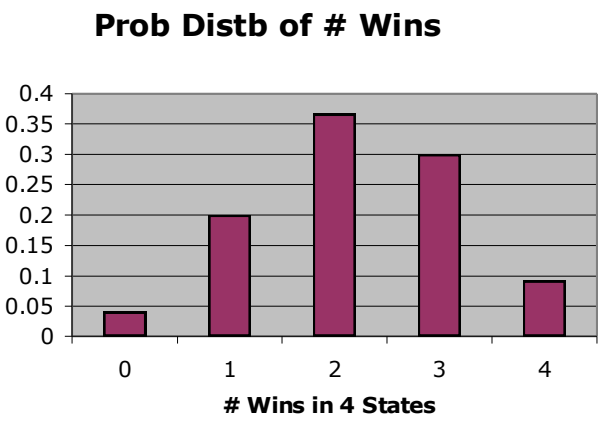
P(x=0) + P(x=1) + P(x=2) + P(x=3) + P(x=4) = 1.0

	State A	**State B**	**State C**	**State D**	
Outcome	**P(win) = .55 P(lose) = .45**	**P(win) = .55 P(lose) = .45**	**P(win) = .55 P(lose) = .45**	**P(win) = .55 P(lose) = .45**	**Probability**
0 wins	.45	.45	.45	.45	**= .041**
	The outcome = "0 wins" means your candidate loses in every state. The total probability is .45*.45*.45*.45 =				
1 win	.55 (win)	.45	.45	.45	= .05
	.45	.55 (win)	.45	.45	= .05
	.45	.45	.55 (win)	.45	= .05
	.45	.45	.45	.55 (win)	= .05
	There are four ways to get one win: You could win in state A, state B, state C, or state D. The top row is the probability of winning only state A. The second row is the probability of winning only state B, and so on. Each row is the product of the four probabilities in that row.				**= 0.20**

	The total probability is then just the sum of the individual rows.				
2 wins	.55	.55	.45	.45	= .061
	.45	.55	.55	.45	= .061
	.45	.45	.55	.55	= .061
	.55	.45	.55	.45	= .061
	.55	.45	.45	.55	= .061
	.45	.55	.45	.55	= .061
	There are six ways to get two wins: You could win in states A and B, B and C, C and D, A and C, A and D, and B and D. The probabilities are calculated the same as above.				**= .366**
3 wins	.55	.55	.55	.45	= .075
	.45	.55	.55	.55	= .075
	.55	.45	.55	.55	= .075
	.55	.55	.45	.55	= .075
	There are three ways to get three wins: You could win states A,B,C or B,C,D, or A,C,D, or A,B,D				**= .300**
4 wins	.55	.55	.55	.55	**= .092**

Probability Distribution:

P(x=0) = .041
P(x=1) = .200
P(x=2) = .366
P(x=3) = .300
P(x=4) = .092



Expected value = mean of distribution.

$$E(X) = \mu = \sum_{i=1}^{n}(x_i * p(x_i))$$

$$\text{Variance } (X) = \sigma^2 = \sum_{i=1}^{n}((x_i - \mu)^2 * p(x_i))$$

Note: μ and σ^2 = population mean and variance (as opposed to a sample). Why?

$E(X)= 0*.041 + 1*.200 + 2*.366 + 3*.300 + 4*.092 =$ **2.2**

$\sigma^2 = (0-2.2)^2*(.041) + (1-2.2)^2*(.200)+(2-2.2)^2*(.366)+(3-2.2)^2*(.300)+(4-2.2)^2*(.092)$

$= .198 + .288 + .015 + .192 + .298 =$ **.991**

$\sigma = \sqrt{\sigma^2} =$ **.995**

So, for strategy #1, the expected number of states won is 2.2 states, with a standard deviation of 0.995.*

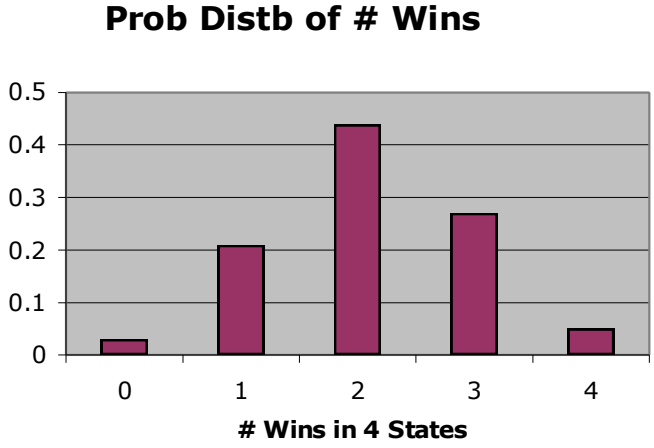
*Note: you cannot actually win 2.2 states—you can't win two-tenths of a state. But that is still the expected value. Just like with averages, the expected value does not have to be one of the possible outcomes. Do not let the name confuse you.

Strategy 2

	State A	State B	State C	State D	
Outcome	**P(w) = .75**	**P(w) = .75**	**P(w) = .30**	**P(w) = .30**	**Probability**
0 wins	.25	.25	.70	.70	**= .031**
1 win	.75	.25	.70	.70	= .092
	.25	.75	.70	.70	= .092
	.25	.25	.30	.70	= .013
	.25	.25	.70	.30	= .013
					= 0.21
2 wins	.75	.75	.70	.70	= .276
	.25	.75	.30	.70	= .039
	.25	.25	.30	.30	= .006
	.75	.25	.70	.30	= .039
	.75	.25	.30	.70	= .039
	.25	.75	.70	.30	= .039
					= .438
3 wins	.75	.75	.30	.70	= .118
	.25	.75	.30	.30	= .017
	.75	.25	.30	.30	= .017
	.75	.75	.70	.30	= .118
					= .270
4 wins	.75	.75	.30	.30	**= .051**

Probability Distribution:

P(0) = .031
P(1) = .210
P(2) = .438
P(3) = .270
P(4) = .051



E(X)= 0*.031 + 1*.210 + 2*.438 + 3*.270 + 4*.051 = **2.1**
σ^2 = .137 + .254 + .004 + .219 + .184 = **.798**
$\sigma = \sqrt{\sigma^2}$ = **.893**

So, using Strategy 2, we expect 2.1 wins with a std dev of 0.893.
Using Strategy 1, we expected 2.2 wins with a std dev of .995.

Which strategy would you prefer? That really depends on your goal. Are you comfortable with a slightly higher expected value but a greater variation? If so, you'd want Strategy 1. If you prefer a slightly lower expected value with slightly less variation, you'd prefer Strategy 2.

Do we really need to go through this laborious process each time? Fortunately, there's a shortcut for scenarios like these.

The Binomial Distribution

The binomial distribution describes the probability distribution of events that can only be described as "success" or "failure." (These don't mean "good" and "bad.")

P(*s* successes in *n* trials) = $\frac{n!}{(n-s)!*s!} * (\pi^s(1-\pi)^{n-s})$

Where 1) π = probability of a success
2) *n*! = *n*(n-1)*(n-2)*(n-3) … *(2)*(1)* (Example: 4! = 4*3*2*1 = 24.)

This equation can be used to create a binomial distribution only when

- there is a binary outcome;
- the outcomes are independent of one another; and
- the underlying probability of outcome is uniform (for instance, p =.55 in each state).

Look at bit more carefully at this equation:

$$\frac{n!}{(n-s)!*s!} * (\pi^s(1-\pi)^{n-s})$$

The second part of the equation is what we calculated above. For instance, for three wins in four states with p(w) = .55, we had $.55^3(1-.55)^1 = .075$.

The first part is the total number of ways that the independent events can be ordered. 4! / [(4-3)! * 3!] = 4*3*2*1 / [1*3*2*1] = 4.

So, the binomial tells us that the probability of three wins in four states = .075*4 = .300

When these conditions are met (and only then), we can use the following estimates:
E(X) = n*p
σ^2 = n*p(1-p)

Go back and calculate these for Strategy 1:
E(X) = 4*.55 = 2.2
σ^2 = 4*(.55)*(1-.55) = .99

