Statistical Methods for Plant Biology

PBIO 3150/5150

Anirudh V. S. Ruhil

February 2, 2016

The Voinovich School of Leadership and Public Affairs

Table of Contents

The Binomial Distribution

The Binomial Distribution

- Many phenomena can be dichotomized ... category A or B?
- The Binomial Distribution characterizes the distribution of such phenomena, with the category of interest being tagged as *success* and the other category tagged as *failure*
- The distribution is premised on some assumptions:
  1. The number of trials (n) is fixed
  2. Each trial is independent of all other trials
  3. The probability of observing a success (p) does not vary across trials
- Mathematically, then, the probability of observing X successes in n trials is given by

$$P\left[X \text{ successes}\right] = \binom{n}{X} p^X \left(1-p\right)^{n-X}$$

$$\text{where } \binom{n}{x} = \frac{n!}{X!(n-X)!} \text{ and}$$

$$n! = n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1$$

Understanding the Binomial Distribution

If I toss a coin 2 times, what is the probability of getting exactly 1 head? Let $X = 1$. We know for unbiased coins $p(Heads) = 0.50$. We are also conducting $n = 2$ independent trials.
How many outcomes are likely in 2 independent trials? We know this to be $(2)^2 = 4$ ... these are $[HH, HT, TH, TT]$. In how many ways can we get 1 Head out of 2 tosses? ... $[HT, TH]$. So the probability of getting exactly 1 Head in 2 tosses is $\frac{2}{4} = 0.5$

$$\begin{aligned} P[X \text{ Successes}] &= \binom{n}{X} p^X (1-p)^{n-X} \\ \therefore P[1 \text{ Success}] &= \binom{2}{1} (0.50)^1 (1-0.50)^{2-1} \\ &= \binom{2}{1} (0.50)^1 (0.50)^1 \\ \binom{2}{1} &= \frac{2 \times 1}{(1)(1)} = 2 \\ \therefore, P[1 \text{ Success}] &= (2) \times (0.5) \times (0.5) = 0.50 \end{aligned}$$

If I toss a coin 3 times, what is the probability of getting exactly 1 head? Let $X = 1$. We know for unbiased coins $p(Heads) = 0.50$. We are also conducting $n = 3$ independent trials.
How many outcomes are likely in 3 independent trials? We know this to be $(2)^3 = 8$ ... these are $[HHH, HHT, HTH, HTT, TTT, TTH, THT, THH]$. In how many ways can we get 1 Head out of 3 tosses? ... $[HTT, THT, TTH]$. So the probability of getting exactly 1 Head in 3 tosses is $\frac{3}{8} = 0.375$

$$
\begin{aligned}
P\left[X \text{ Successes}\right] &= \binom{n}{X} p^X \left(1-p\right)^{n-X} \\
\therefore P\left[1 \text{ Success}\right] &= \binom{3}{1} (0.50)^1 \left(1-0.50\right)^{3-1} \\
&= \binom{3}{1} (0.50)^1 (0.50)^2 \\
\binom{3}{1} &= \frac{3 \times 2 \times 1}{(1)(2 \times 1)} = 3 \\
\therefore, P\left[1 \text{ Success}\right] &= (3) \times (0.5) \times (0.25) = 0.375
\end{aligned}
$$

The Wasp Example

- A random sample of 5 wasps are gathered. What is the probability that exactly 3 of these wasps will be male?
- Let $X =$ A wasp is a male; $p =$ probability the wasp is male
- Now, assume we know that the probability of randomly picking a male wasp (p) is 0.20

$$P[X \text{ successes}] = \binom{n}{X} p^X (1-p)^{n-X}$$

$$\therefore P[3 \text{ Males}] = \binom{5}{3} (0.20)^3 (0.80)^2$$

$$\binom{5}{3} = \frac{5!}{3!(2)!} = \frac{5 \times 4 \times 3 \times 2 \times 1}{(3 \times 2 \times 1)(2 \times 1)} = \frac{120}{12} = 10$$

$$\therefore P[3 \text{ Males}] = (10)(0.20)^3(0.80)^2 = (10)(0.008)(0.64) = 0.0512$$

Right-Handed Toads Revisited

- We had a random sample of 18 toads with the probability of a right-handed toad being $p = 0.50$. What is the probability that in such a sample we would observe exactly 9 right-handed toads?

$$P[9 \text{ Right-Handed Toads}] = \binom{18}{9}(0.50)^9(0.50)^9$$

$$= \frac{18!}{9!(9!)} \times (0.50)^9 \times (0.50)^9 = 0.1854706$$

$$P[0 \text{ Right-Handed Toads}] = \binom{18}{0}(0.50)^0(0.50)^{18}$$

$$= \frac{18!}{0!(18!)} \times (0.50)^0 \times (0.50)^18 = 3.814697e-06 = 0.00000381$$

Left-Handed Flowers Revisited

- Assume we sampled 27 mud plantains from a population of which 25% are believed to have left-handed flowers (*success*).
- What is the probability of ending up with exactly 6 left-handed flowers in our random sample?

$$P\left[X \text{ successes}\right] = \binom{n}{X} p^X \left(1-p\right)^{n-X}$$

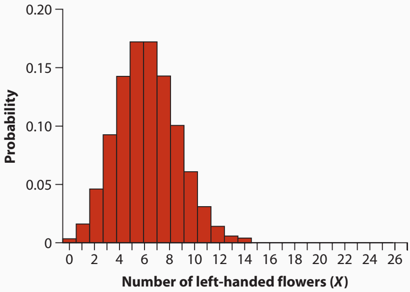
$$\therefore P\left[6 \text{ left-handed flowers}\right] = \binom{27}{6} (0.25)^6 (0.75)^{21}$$

$$\binom{27}{6} = \frac{27 \times 26 \times 25 \times \cdots \times 2 \times 1}{(6 \times 5 \times \cdots \times 2 \times 1)(21 \times 20 \times \cdots \times 2 \times 1)} = 296,010$$

$$\therefore P\left[6 \text{ left-handed flowers}\right] = (296,010)(0.25)^6 (0.75)^{21} = 0.1719$$

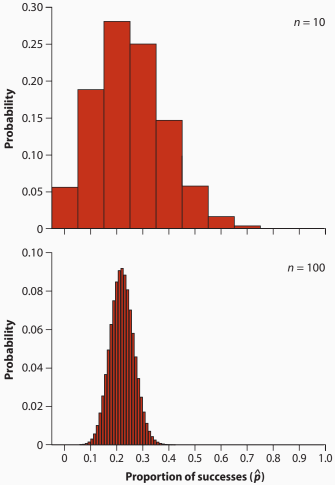
Calculating the Probability of $X = [0, 1, 2, \cdots, 27]$

X	$P(X)$	X	$P(X)$
0	0.000413	10	0.060530
1	0.003836	11	0.031185
2	0.016541	12	0.013945
3	0.045789	13	0.005339
4	0.091652	14	0.001798
5	0.140660	15	0.000514
6	0.171824	16	0.000132
7	0.171711	17	0.000029
8	0.143449	18	0.000006
9	0.100646	19	0.000001



Sampling Distribution of the Proportion

- $\hat{p} = \dfrac{X}{n}$
- We know that if we drew all possible samples of size n and calculated $\hat{p}$ in each such sample we would find the average $\hat{p}$ of all these samples to equal p ... i.e., $Mean[\hat{p}] = p$
- But what is the standard deviation of the sampling distribution ... i.e., the *standard error* of $\hat{p}$?
- $\sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$
- Again, notice n in the denominator; as $n \to \infty$, $\sigma_{\hat{p}} \to 0$ ... the Law of Large Numbers



Testing a Proportion: The Binomial Test

Testing a Proportion: The Binomial Test

- Given a dichotomous (success/failure) outcome of interest
- H_0: The relative frequency of successes in the population is p_0
  H_A: The relative frequency of successes in the population is not p_0

  OR

  H_0: The relative frequency of successes in the population is $\leq p_0$
  H_A: The relative frequency of successes in the population is $> p_0$

  OR

  H_0: The relative frequency of successes in the population is $\geq p_0$
  H_A: The relative frequency of successes in the population is $< p_0$
- ... we use the binomial test to decide whether or not to reject H_0

Sex and the X

- Wang et al.'s (2001) study of 25 genes involved in sperm formation found 10 (40%) on the X chromosome
- If genes for sperm formation occur randomly across the genome then only 6.1% should be on the X chromosome because the X chromosome contains 6.1 of the genes in the genome
- Do the data, then, suggest that spermatogenesis genes occur preferentially on the X chromosome?
- Setup the Hypotheses:
  H_0: The probability that a spermatogensis gene falls on the X chromosome is $p = 0.061$
  H_A: The probability that a spermatogensis gene falls on the X chromosome is $p \neq 0.061$
- Construct the test statistic:
  If H_0 is true then what is the probability of seeing 10 on the X chromosome, *by chance alone*?

$$P[X \text{ successes}] = \binom{n}{X} p^X (1-p)^{n-X}$$

$$P\left[10 \text{ successes}\right] = \binom{25}{10}(0.061)^{10}(0.939)^{15}$$

$$\binom{25}{10} = \frac{25 \times 24 \times \cdots \times 2 \times 1}{(10 \times 9 \times \cdots \times 2 \times 1)(15 \times 14 \times \cdots \times 2 \times 1)} = 3{,}268{,}760$$

$$\therefore P\left[10 \text{ successes}\right] = (3{,}268{,}760)(0.061)^{10}(0.939)^{15}$$

$$= (3{,}268{,}760)(0.0000000000007133)(0.3890307083879447)$$

$$= 0.0000009071211000$$

Calculating the two-tailed P-value yields 1.98×10^{-6}

- Notice how small a probability this is ... Thus it cannot be chance but instead that H_0 is not true
- If H_0 is not true, then what might be true? Well, the most we can say is that about 40% $\left(\hat{p} = \frac{10}{25}\right)$ of the spermatogenesis gene is located on the mouse X chromosome

Standard Errors and Confidence Intervals

- Earlier we said $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$
- But we rarely know p and must, instead, rely on $\hat{p}$ ...
- ... Yielding: $SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}$
- We can also calculate confidence intervals for proportions ... (text recommends the Agresti-Coull method)
  1. Calculate $p^{'} = \frac{X+2}{n+4}$
  2. CI is then given by: $p^{'} - z\sqrt{\frac{p^{'}\left(1-p^{'}\right)}{n+4}} < p < p^{'} + z\sqrt{\frac{p^{'}\left(1-p^{'}\right)}{n+4}}$
- Default in practice is the Wald method[1]: $p^{'} - z\left(SE_{p'}\right) < p < p^{'} + z\left(SE_{p'}\right)$
- Recall what the confidence interval is telling us (WHAT?)

[1]Wald inaccurate when (i) n is small or (ii) p is close to 0 or 1