Section 3.6 Binomial Distributions

Tuesday, September 13, 2016 9:34 AM

Section 3.6
Binomial ...

3.6 Binomial distribution

BERNOULLI TRIALS: Many processes can be envisioned as consisting of a sequence of "trials," where

(i) each trial results in a "success" or a "failure,"

(ii) the trials are **independent**, and

(iii) the probability of "success," denoted by p, $0 < p < 1$, is the same on every trial.

TERMINOLOGY: In a sequence of n Bernoulli trials, denote by Y the number of successes out of n (where n is fixed). We say that Y has a **binomial distribution** with number of trials n and success probability p. Shorthand notation is $Y \sim b(n, p)$.

Example 3.12. Each of the following situations could be conceptualized as a binomial experiment. Are you satisfied with the Bernoulli assumptions in each instance?

(a) We flip a fair coin 10 times and let Y denote the number of tails in 10 flips. Here, $Y \sim b(n = 10, p = 0.5)$.

(b) Forty percent of all plots of land respond to a certain treatment. I have four plots to be treated. If Y is the number of plots that respond to the treatment, then $Y \sim b(n = 4, p = 0.4)$.

(c) In rural Kenya, the prevalence rate for HIV is estimated to be around 8 percent. Let Y denote the number of HIV infecteds in a sample of 740 individuals. Here, $Y \sim b(n = 740, p = 0.08)$.

(d) Parts produced by a certain company do not meet specifications (i.e., are defective) with probability 0.001. Let Y denote the number of defective parts in a package of 40. Then, $Y \sim b(n = 40, p = 0.001)$. □

DERIVATION: We now derive the pmf of a binomial random variable. The support of Y is $R = \{y : y = 0, 1, 2, ..., n\}$. We need to find an expression for $p_Y(y) = P(Y = y)$ for each value of $y \in R$.

QUESTION: In a sequence of n trials, how can we get exactly y successes? Denoting "success" and "failure" by S and F, respectively, one possible sample point might be

$$SSFSFSFFS \cdots FSF.$$

QUESTION: In a sequence of n trials, how can we get exactly y successes? Denoting "success" and "failure" by S and F, respectively, one possible sample point might be

$$SSFSFSFFS \cdots FSF.$$

Because the trials are **independent**, the probability that we get a particular ordering of y successes and $n-y$ failures is $p^y(1-p)^{n-y}$. Furthermore, there are $\binom{n}{y}$ sample points that contain exactly y successes. Thus, we add the term $p^y(1-p)^{n-y}$ a total of $\binom{n}{y}$ times to get $P(Y=y)$. The pmf for Y is, for $0<p<1$,

$$p_Y(y) = \begin{cases} \binom{n}{y}p^y(1-p)^{n-y}, & y = 0, 1, 2, ..., n \\ 0, & \text{otherwise.} \end{cases}$$

Example 3.13. In Example 3.12(b), assume that $Y \sim b(n=4, p=0.4)$. Here are the probability calculations for this binomial model:

$$\begin{aligned}
P(Y=0) &= p_Y(0) = \binom{4}{0}(0.4)^0(1-0.4)^{4-0} = 1 \times (0.4)^0 \times (0.6)^4 = 0.1296 \\
P(Y=1) &= p_Y(1) = \binom{4}{1}(0.4)^1(1-0.4)^{4-1} = 4 \times (0.4)^1 \times (0.6)^3 = 0.3456 \\
P(Y=2) &= p_Y(2) = \binom{4}{2}(0.4)^2(1-0.4)^{4-2} = 6 \times (0.4)^2 \times (0.6)^2 = 0.3456 \\
P(Y=3) &= p_Y(3) = \binom{4}{3}(0.4)^3(1-0.4)^{4-3} = 4 \times (0.4)^3 \times (0.6)^1 = 0.1536 \\
P(Y=4) &= p_Y(4) = \binom{4}{4}(0.4)^4(1-0.4)^{4-4} = 1 \times (0.4)^4 \times (0.6)^0 = 0.0256.
\end{aligned}$$

TI-84.

$P_Y(y) = \text{binompdf}(n, p, y)$

EXERCISE: What is the probability that at least 2 plots respond? at most one? What are $E(Y)$ and $V(Y)$? □

Example 3.14. In a small clinical trial with 20 patients, let Y denote the number of patients that respond to a new skin rash treatment. The physicians assume that a binomial model is appropriate and that $Y \sim b(n=20, p=0.4)$. Under this model, compute (a) $P(Y=5)$, (b) $P(Y \geq 5)$, and (c) $P(Y<10)$.

(a) $P(Y=5) = p_Y(5) = \binom{20}{5}(0.4)^5(0.6)^{20-5} = 0.0746.$

binompdf(20, .4, 5)

(b)

$$P(Y \geq 5) = \sum_{y=5}^{20} P(Y=y) = \sum_{y=5}^{20} \binom{20}{y}(0.4)^y(0.6)^{20-y}.$$

cumulative distribution function

Y : pmf $P_Y(y) = P(Y=y)$

cdf : $F_Y(y) = P(Y \leq y)$

$= 1 - P(Y<5) = 1 - P(Y \leq 4) = 1 - \sum_{y=0}^{4} \binom{20}{y}(.4)^y(.6)^{20-y}$

$= 1 - \text{binomcdf}(20, .4, 4)$

$\text{binomcdf}(n, p, y) = P(Y \leq y)$

(c) $P(Y<10) = P(Y \leq 9) = \text{binomcdf}(20, .4, 9)$

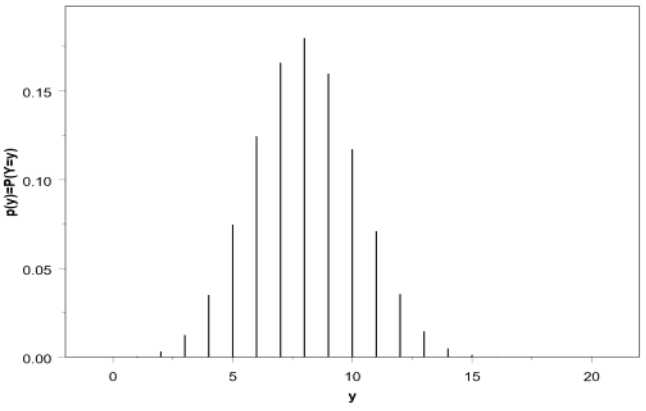

Figure 3.2: *Probability histogram for the number of patients responding to treatment. This represents the* $b(n = 20, p = 0.4)$ *model in Example* 3.14.

This calculation involves using the binomial pmf 16 times and adding the results!

TRICK: Instead of computing the sum $\sum_{y=5}^{20} \binom{20}{y}(0.4)^y(0.6)^{20-y}$ directly, we can write

$$P(Y \geq 5) = 1 - P(Y \leq 4),$$

by the complement rule. We do this because WMS's Appendix III (Table 1, pp 839-841) contains binomial probability calculations of the form

$$P(Y \leq a) = \sum_{y=0}^{a} \binom{n}{y} p^y (1-p)^{n-y},$$

for different n and p. With $n = 20$ and $p = 0.4$, we see from Table 1 that

$$P(Y \leq 4) = 0.051.$$

Thus, $P(Y \geq 5) = 1 - 0.051 = 0.949$.

(c) $P(Y < 10) = P(Y \leq 9) = 0.755$, from Table 1. □

REMARK: The function $P(Y \leq y)$ is called the **cumulative distribution function** of a random variable Y; we'll talk more about this function in the next chapter.

RECALL: The **binomial expansion** of $(a+b)^n$ is given by

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k.$$

CURIOSITY: Is the binomial pmf a **valid** pmf? Clearly $p_Y(y) > 0$ for all y. To check that the pmf sums to one, consider the binomial expansion

$$[(1-p)+p]^n = \sum_{y=0}^{n} \binom{n}{y} p^y (1-p)^{n-y}.$$

The LHS clearly equals 1, and the RHS is the $b(n,p)$ pmf. Thus, $p_Y(y)$ is valid. $\square$

BINOMIAL MGF: Suppose that $Y \sim b(n,p)$. The mgf of Y is given by

$$m_Y(t) = E(e^{tY}) = \sum_{y=0}^{n} e^{ty} \binom{n}{y} p^y (1-p)^{n-y} = \sum_{y=0}^{n} \binom{n}{y} (pe^t)^y (1-p)^{n-y} = (q+pe^t)^n,$$

where $q = 1-p$. The last step follows from noting that $\sum_{y=0}^{n} \binom{n}{y} (pe^t)^y (1-p)^{n-y}$ is the binomial expansion of $(q+pe^t)^n$. $\square$

MEAN AND VARIANCE: We want to compute $E(Y)$ and $V(Y)$ where $Y \sim b(n,p)$. We will use the mgf. Taking the derivative of $m_Y(t)$ with respect t, we get

$$m_Y'(t) \equiv \frac{d}{dt} m_Y(t) = \frac{d}{dt}(q+pe^t)^n = n(q+pe^t)^{n-1} pe^t.$$

Thus,

$$E(Y) = \frac{d}{dt} m_Y(t)\Big|_{t=0} = n(q+pe^0)^{n-1} pe^0 = n(q+p)^{n-1} p = np,$$

since $q+p=1$. Now, we need to find the second moment. By using the product rule for derivatives, we have

$$\frac{d^2}{dt^2} m_Y(t) = \frac{d}{dt} \underbrace{n(q+pe^t)^{n-1} pe^t}_{m_Y'(t)} = n(n-1)(q+pe^t)^{n-2}(pe^t)^2 + n(q+pe^t)^{n-1} pe^t.$$

Thus,

$$E(Y^2) = \frac{d^2}{dt^2} m_Y(t)\Big|_{t=0} = n(n-1)(q+pe^0)^{n-2}(pe^0)^2 + n(q+pe^0)^{n-1} pe^0 = n(n-1)p^2 + np.$$

Appealing to the variance computing formula, we have

$$V(Y) = E(Y^2) - [E(Y)]^2 = n(n-1)p^2 + np - (np)^2 = np(1-p).$$

NOTE: WMS derive the binomial mean and variance using a different approach (not using the mgf). See pp 107-108. □

Example 3.15. Artichokes are a marine climate vegetable and thrive in the cooler coastal climates. Most will grow in a wide range of soils, but produce best on a deep, fertile, well-drained soil. Suppose that 15 artichoke seeds are planted in identical soils and temperatures, and let Y denote the number of seeds that germinate. If 60 percent of all seeds germinate (on average) and we assume a $b(15, 0.6)$ probability model for Y, the mean number of seeds that will germinate is

$$E(Y) = \mu = np = 15(0.6) = 9. \text{ (seeds)}$$

The variance of Y is

$$V(Y) = \sigma^2 = np(1-p) = 15(0.6)(0.4) = 3.6 \text{ (seeds)}^2.$$

The standard deviation of Y is $\sigma = \sqrt{3.6} \approx 1.9$ seeds. □

BERNOULLI DISTRIBUTION: In the $b(n, p)$ family, when $n = 1$, the binomial pmf reduces to

$$p_Y(y) = \begin{cases} p^y(1-p)^{1-y}, & y = 0, 1 \\ 0, & \text{otherwise.} \end{cases}$$

This is called the **Bernoulli distribution**. Shorthand notation is $Y \sim b(1, p)$ or $Y \sim \text{Bern}(p)$.

$P_Y(y) = \begin{cases} 1-p \\ p \end{cases}$

$y=0$

$y=1$ Bernoulli

$E[Y] = 0\times(1-p)+1\times p$

$= p$

3.7 Geometric distribution

TERMINOLOGY: Envision an experiment where Bernoulli trials are observed. If Y denotes the trial on which the first success occurs, then Y is said to follow a **geometric distribution** with parameter p, where p is the probability of success on any one trial.