Biost 517 / Biost 514 Applied Biostatistics I / Biostatistics I

Timothy A. Thornton, Ph.D.
Associate Professor of Biostatistics
University of Washington

Lecture 9:

Discrete Random Variables; Binomial Distribution; Normal Approximation to Binomial; Inference for Proportions

Discrete Random Variables

- So far we have largely focused on continuous random variable, and in particular, random variables that follow a normal distribution
- Many variables of interest are discrete. Discrete random variables take values in a finite or countable set
- The probability distribution for a discrete random variable X tells us the possible values that X can take and the probabilities for each of those values
- For discrete random variables, the probability distribution function is often called the **probability mass function**

Probability Distribution of Discrete Random Variables

Denote the possible outcomes of a discrete random variable X as $x_1, x_2, x_3, \ldots$

The **probability distribution** of X is simply a list of the probabilities of each of the possible outcomes.

X	x_1	x_2	x_3	$\ldots$	x_k
Probability	p_1	p_2	p_3	$\ldots$	p_k

The probabilities must satisfy two requirements:

$$0 \leq p_i \leq 1 \qquad \forall i$$

$$p_1 + p_2 + \cdots + p_k = 1$$

Expected Value of Discrete Random Variable

Let X be a discrete random variable with k possible outcomes, $x_1, x_2, \ldots, x_k$.
Let $p_i = P(X = x_i)$.

The **mean** or **expected value** of X is given by

$$\begin{aligned}\mu_X &= E(X) \\ &= p_1 x_1 + p_2 x_2 + \cdots + p_k x_k \\ &= \sum_{i=1}^{k} p_i x_i\end{aligned}$$

Variance and SD of Discrete Random Variable

Let X be a discrete random variable with k possible outcomes, $x_1, x_2, \ldots, x_k$ and corresponding probabilities $p_1, p_2, \ldots, p_k$, respectively.

The **variance** of X is given by

$$\sigma_X^2 = \mathrm{Var}(X) = p_1(x_1 - \mu_X)^2 + \cdots + p_k(x_k - \mu_X)^2$$

$$= \sum_{i=1}^{k} p_i(x_i - \mu_X)^2$$

Variance and SD of Discrete Random Variable

As we saw in the continuous random variable case, another formula for the variance is:

$$\begin{aligned}\sigma_X^2 &= E[(X-\mu_X)^2] \\ &= E(X^2) - \mu_X^2 \\ &= \sum_{i=1}^{k} p_i (x_i)^2 - \mu_X^2\end{aligned}$$

The standard deviation of X is given by

$$\sigma_X = \mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}$$

Example: Rolling 2 Dice

Consider rolling two dice.

$$S = \begin{Bmatrix} (1,1) & \cdots & (1,6) \\ \vdots & \ddots & \vdots \\ (6,1) & \cdots & (6,6) \end{Bmatrix}$$

Let X be the total value of the roll, i.e., the sum of the two dice. Thus,

$$X((a,b)) = a + b, \ \forall (a,b) \in S$$

X, then, is a discrete random variable taking values $1, 2, \ldots, 12$.

What is the **probability distribution** of X?

Example: Rolling 2 Dice

Probability distribution of X is:

$$P(X=1)=\frac{0}{36} \qquad P(X=7)=\frac{6}{36}$$
$$P(X=2)=\frac{1}{36} \qquad P(X=8)=\frac{5}{36}$$
$$P(X=3)=\frac{2}{36} \qquad P(X=9)=\frac{4}{36}$$
$$P(X=4)=\frac{3}{36} \qquad P(X=10)=\frac{3}{36}$$
$$P(X=5)=\frac{4}{36} \qquad P(X=11)=\frac{2}{36}$$
$$P(X=6)=\frac{5}{36} \qquad P(X=12)=\frac{1}{36}$$

Example: Rolling 2 Dice

The **expected value** of X is

$$\begin{aligned}\mu_X &= E(X) \\ &= \sum_{i=1}^{k} p_i x_i \\ &= \frac{1}{36}(2) + \frac{2}{36}(3) + \frac{3}{36}(4) \cdots + \frac{3}{36}(10) + \frac{2}{36}(11) + \frac{1}{36}(12) \\ &= 7\end{aligned}$$

The **expected value** of X^2 is

$$\begin{aligned}E(X^2) &= \sum_{i=1}^{k} p_i (x_i)^2 \\ &= \frac{1}{36}(2)^2 + \frac{2}{36}(3)^2 + \frac{3}{36}(4)^2 \cdots + \frac{3}{36}(10)^2 + \frac{2}{36}(11)^2 + \frac{1}{36}(12)^2 \\ &= 54.833\end{aligned}$$

Example: Rolling 2 Dice

The variance of X is:

$$\begin{aligned}\sigma_X^2 &= E(X^2) - \mu_X^2 \\ &= 54.833 - (7)^2 = 5.833\end{aligned}$$

The standard deviation of X is

$$\sigma_X = \mathrm{SD}(X) = \sqrt{5.833} = 2.415$$

Binary Random Variables

- Many variables of interest can take on only two values.
- Clinical and epidemiological studies often generate outcomes that are dichotomous
  - Presence/absence of a condition or characteristic at a particular time
  - Indication of whether a response occurred within a defined period of observations
  - For convenience, often coded as 0 or 1 “indicator” variable
    - Vital Status: “Dead” coded 0= alive 1= dead
    - Prostate Cancer: “In Remission” coded 0=no 1= yes
    - Sex: “Female” coded 0= male 1= female
    - Intervention: “Treatment x” coded 0= control 1= new therapy

Binary Random Variables

- Sometimes continuous variables are dichotomized
  - For scientific reasons (statistically less precise)
    - Systolic Blood pressure: greater than 160 mm Hg (Hypertension)
    - Prostate Specific Antigen (PSA) : normal is less than 4 ng/ml
    - Serum glucose: normal range is less than 120 mg/dl (important for diabetics)

Bernoulli Random Variable

The Bernoulli Distribution

- A binary variable Y must have a Bernoulli probability distribution, where Y takes values of 0 or 1.
- The Bernoulli distribution has a single parameter: $p = P(Y = 1)$ with $0 < p < 1$.
- Note that $P(Y = 0) = 1 - p$.
- The probability mass function for a Bernoulli random variable Y is

$$P(Y = y) = p^y(1 - p)^{1-y}$$

where $y = 1$ or $y = 0$.

Bernoulli Random Variable

The expected value of Bernoulli random variable Y is

$$\begin{aligned}\mu_Y &= E(Y) \\ &= \sum_{y=0}^{1} y p^y (1-p)^{1-y} \\ &= 0(1-p) + 1p \\ &= p\end{aligned}$$

The expected of Y^2 is

$$\begin{aligned}E(Y^2) &= \sum_{y=0}^{1} y^2 p^y (1-p)^{1-y} \\ &= 0^2(1-p) + 1^2 p = p\end{aligned}$$

Bernoulli Random Variable

The variance of Bernoulli random variable Y is:

$$\begin{aligned}\sigma_Y^2 &= E(Y^2) - \mu_Y^2 \\ &= p - p^2 = p(1-p)\end{aligned}$$

The standard deviation of Y is

$$\sigma_X = \mathrm{SD}(X) = \sqrt{p(1-p)}$$

Counting: Permutations

How many ways can we I order n distinct objects? The answer is $n!$

The *factorial* of a non-negative integer n is defined to be $n! = n * (n-1) * (n-2) \cdots * 2 * 1$.

Note that 0! is defined to be 1.

- Example: How many ways can we arrange the letters A,B,C,D?
- Answer: $4! = 4 * 3 * 2 * 1 = 24$

Counting: Combinations

Counting: Combinations

How many ways can we choose r distinct objects from a set of n distinct objects, ignoring order?

$$\binom{n}{r} = \frac{n!}{(n-r)!r!}$$

- Example: How many possible three-topping pizzas can be made with 10 ingredients? In this case, order does not matter.
- Answer:

$$\binom{10}{3} = \frac{10!}{7!3!} = \frac{10 * 9 * 8}{3 * 2 * 1} = 120$$

Binomial Distribution

The binomial distribution is a very common discrete probability distribution that arises in the following situation:

- A fixed number, n, of trials
- The n trials are independent of each other
- Each trial has exactly two outcomes: "success" and "failure"
- The probability of a success, p, is the same for each trial

Binomial Random Variable

Consider n independent Bernoulli random variables $Y_1, Y_2, \ldots, Y_n$ with parameter p.

Let X be the sum of the n independent Bernoulli random variables, i.e., $X = \sum_{i=1}^{n} Y_i$.

Then the probability distribution of X is a **binomial distribution** with parameters n and p:

$$X \sim B(n, p)$$

The binomial probability distribution function is:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad k = 0, 1, 2, \ldots, n$$

Binomial Random Variable

If $X \sim B(n, p)$,

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad k = 0, 1, 2, \ldots, n$$

Explanation for this formula?

- Since the trials are independent, any particular sequence of 0's and 1's with k 1's and $(n - k)$ 0's has probability $p^k(1 - p)^{n-k}$ of happening.
- There are $\binom{n}{k}$ distinct sequences with k 1's and $(n - k)$ 0's.

Example: Binomial Distribution

Joe reads that 1 out of 4 eggs contains salmonella bacteria. After that, Joe decides to never uses more than 3 eggs in cooking. Assume eggs are infected with salmonella independently of each other. What is the probability that Joe selects 3 uninfected eggs?

- Let X be number of infected eggs. Then X follows a binomial distribution with $n = 3$ and $p = \frac{1}{4}$:

$$P(X = k) = \frac{3!}{(3-k)!k!} \times \left(\frac{1}{4}\right)^k \left(\frac{3}{4}\right)^{3-k}$$

- So the probability that Joe selects 3 uninfected eggs is $P(X = 0)$ where

$$P(X = 0) = \frac{3!}{(3!)0!} \times \left(\frac{1}{4}\right)^0 \left(\frac{3}{4}\right)^3 = \frac{27}{64}$$

- So probability that Joe selects 3 uninfected eggs is $\frac{27}{64} \approx 0.42$.

Mean and Variance of Binomial

$$E(X) = np$$
$$\text{Var}(X) = np(1-p)$$
$$\text{SD}(X) = \sqrt{np(1-p)}$$

Can easily derive this based on Bernoulli random variables.

$$E(X) = E\left(\sum_{i=1}^{n} Y_i\right) = \sum_{i=1}^{n} E\left(Y_i\right) = \sum_{i=1}^{n} p = np$$

$$Var(X) = Var\left(\sum_{i=1}^{n} Y_i\right) = \sum_{i=1}^{n} Var\left(Y_i\right)$$
$$= \sum_{i=1}^{n} p(1-p) = np(1-p)$$

Note that $Var\left(\sum_{i=1}^{n} Y_i\right) = \sum_{i=1}^{n} Var\left(Y_i\right)$ because the Bernoulli Y_i's are independent.

Proportion of Successes

For a binomial random variable X, the proportion of "successes" in n trials is $\frac{X}{n}$.

$$E\left(\frac{X}{n}\right) = p$$

$$\text{Var}\left(\frac{X}{n}\right) = \frac{p(1-p)}{n}$$

$$\text{SD}\left(\frac{X}{n}\right) = \sqrt{\frac{p(1-p)}{n}}$$

Example 2: Binomial Distribution

Suppose a random sample of 1500 African Americans adults are selected for a study of diabetes. Assume that the prevalence of diabetes is 12% in African American adults

- What is the expected number of people in the sample with diabetes?

$$E(X) = np = 1500(0.12) = 180$$

- What is the probability that the sample contains 170 or less individuals with diabetes?

$$P(X \leq 170) = \sum_{j=0}^{170} P(X = j)$$

$$= \sum_{j=0}^{170} \binom{1500}{j} (0.12)^j (0.88)^{1500-j}$$

That's pretty ugly. Is there an easier way?

Binomial Distribution and CLT

- We know from the **Central Limit Theorem (CLT)** that the sampling distribution for a sum (or mean) of random variables will be approximately normally distributed for large enough n.
- A binomial random X is the sum of the n independent Bernoulli random variables:

$$X = \sum_{i=1}^{n} Y_i$$

- Similarly $\frac{X}{n}$ is the mean of n independent Bernoulli random variables:

$$\frac{X}{n} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

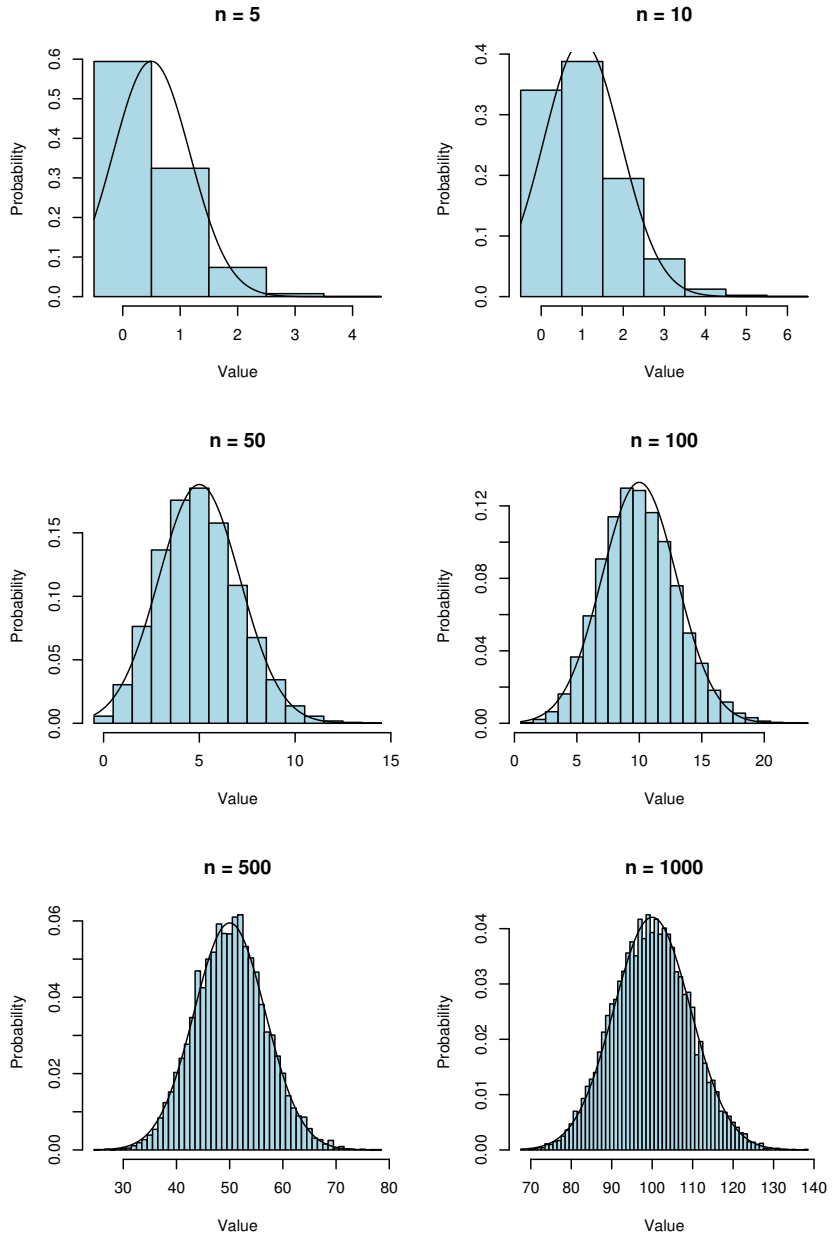
Binomial Distribution and CLT

From the Central Limit Theorem, as n gets larger, the Binomial distribution looks increasingly like the Normal distribution.

Recall that we defined **sampling distribution** of a statistic to be the distribution of the statistic in all possible samples of the same size from the same population.

If we repeatedly drew samples of size n and calculated X, the number of successes, we ascertain that the **sampling distribution** of X is approximately normally distributed.

Binomial Distribution and CLT



Binomial Distribution and CLT

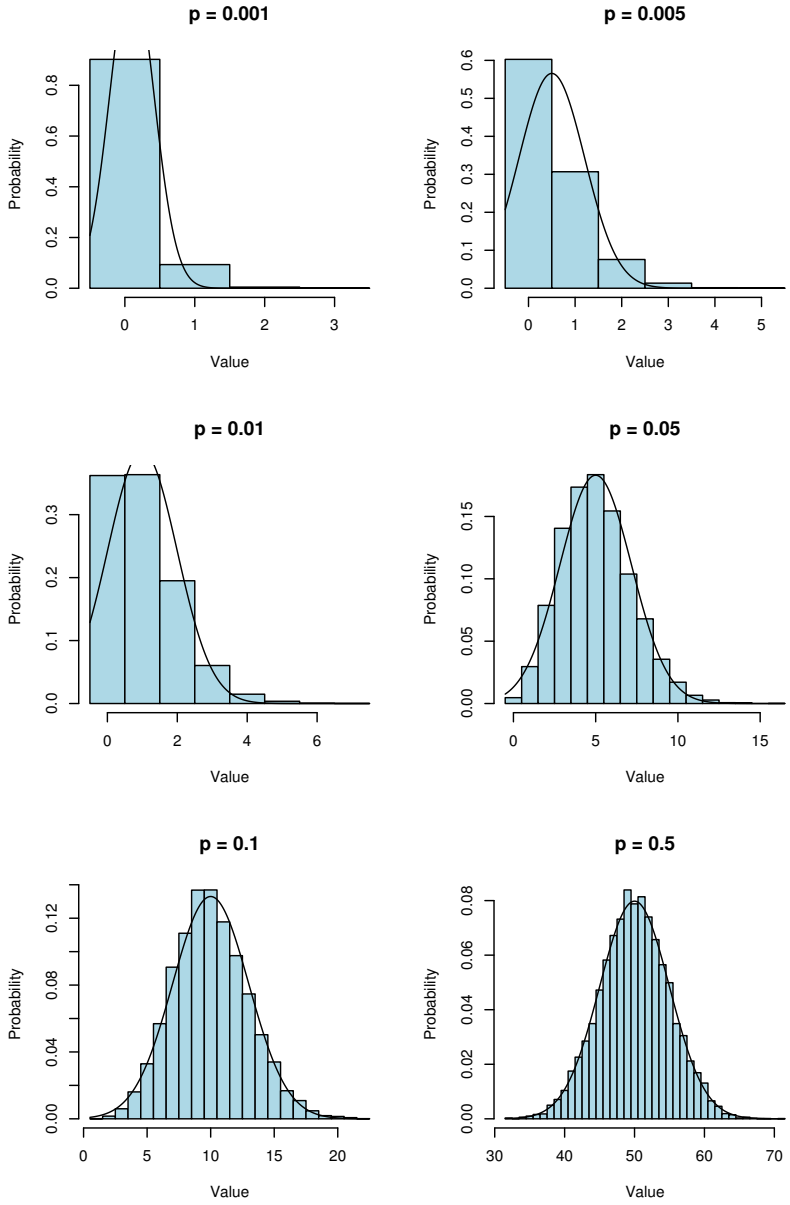
When is the normal approximation for the binomial distribution appropriate?

The farther p is from $\frac{1}{2}$, the larger n needs to be for the approximation to work. Thus, as a rule of thumb, only use the approximation if

$$np \geq 10 \qquad \text{and} \qquad n(1-p) \geq 10$$

Binomial Distribution and CLT

Behavior of the approximation as a function of p, for n=100



Normal Approximation to the Binomial

- Let $X \sim B(n, p)$, having mean $E[X] = np$ and standard deviation $\mathrm{SD}(X) = \sqrt{np(1-p)}$.
- The approximating normal distribution for the binomial has the same mean and standard deviation as the underlying binomial distribution.
- For large enough n, the binomial distribution can be approximated by a normal distribution such that

$$X \dot{\sim} N\left(\mu_X = np, \sigma_X = \sqrt{np(1-p)}\right)$$

Normal Approximation for Sample Proportions

- If X is the number of "success" in n trials, then the proportion of "success" is $\frac{X}{n}$.
- Since $\frac{X}{n}$ in an unbiased estimator of the population proportion p, i.e., $E\left(\frac{X}{n}\right) = p$, the proportion of "success" in a sample is often denoted as $\hat{p}$
- The approximating normal distribution for $\hat{p} = \frac{X}{n}$ is

$$\hat{p} \dot{\sim} N\left(\mu = p, \sigma = \sqrt{\frac{p(1-p)}{n}}\right)$$

Example 2 cont.: Normal Approximation to Binomial

Recall the problem we set out to solve:

$$P(X \leq 170), \text{ where } X \sim B(1500, 0.12)$$

How do we calculate this using the Normal approximation?

If we were to draw a histogram of the $B(1500, 0.12)$ distribution with bins of width one, $P(X \leq 170)$ would be represented by the total area of the bins spanning

$$(-0.5, 0.5], (0.5, 1.5], \ldots, (169.5, 170.5]$$

Thus, using the approximating Normal distribution $W \sim N(180, 12.59)$, we calculate

$$P(X \leq 170) \approx P(W \leq 170.5) = 0.2253$$

For reference, the *exact* Binomial probability is 0.2265, so the approximation is apparently pretty good.

The Continuity Correction

Note that the Binomial distribution is discrete while the approximating Normal distribution is continuous. A **continuity correction** is often used to refine the approximation by accounting for this.

In general, if W is used for the approximating distribution of the discrete distribution of X, we make the following adjustments:

$$P(X \leq x) \approx P(W \leq x + 0.5)$$
$$P(X < x) = P(X \leq x - 1) \approx P(W \leq x - 0.5)$$
$$P(X \geq x) \approx P(W \geq x - 0.5)$$
$$P(X > x) = P(X \geq x + 1) \approx P(W \geq x + 0.5)$$

Inference for a population proportion

- Inference about a population proportion p is often of interest. How do we obtain inference about p from a sample?
- We know that $\hat{p} = \frac{X}{n}$ is an unbiased estimate of the unknown true population proportion p and the the approximating normal distribution for $\hat{p}$ is

$$\hat{p} \dot{\sim} N\left(\mu = p, \sigma = \sqrt{\frac{p(1-p)}{n}}\right)$$

where the above normal approximation can be used when n is sufficiently large – i.e. if $np \geq 10$ and $n(1-p) \geq 10$

Confidence Interval for population proportion

- An approximate $(1-\alpha)$ CI for the population proportion p is given by

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where as before, z^* is chosen so that $P(Z > z^*) = \alpha/2$ for $Z \sim N(0,1)$.

z^* is the **critical value**, selected so that a standard Normal density has area $(1-\alpha)$ between $-z^*$ and z^*. The quantity $z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ is the **margin error**.

Inference for population proportion: Hypothesis Testing

Suppose we want to test whether $H_0 : p = p_0$ for some fixed value p_0

The null hypothesis is $p = p_0$, and under this hypothesis,

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \dot{\sim} N(0,1)$$

Notice that we are using a different value for the SD of $\hat{p}$ than was used for the CI. Since H_0 specifies a true value for p, the SD of $\hat{p}$ under H_0 is given by

$$\sqrt{\frac{p_0(1-p_0)}{n}}$$

The p-values for this test are:

- $H_a : p > p_0 \qquad P(Z \geq z)$
- $H_a : p < p_0 \qquad P(Z \leq z)$
- $H_a : p \neq p_0 \qquad 2P(Z \geq |z|)$

for $Z \sim N(0,1)$.

Choosing an Appropriate Sample Size

Suppose we want a certain margin of error m for a $(1-\alpha)$ CI. What sample size should we use? If we knew p, we could solve for n in the margin of error formula, and get

$$n = \left(\frac{z^*}{m}\right)^2 p(1-p)$$

Of course, we don't know p – and we can't even estimate it, because we haven't collected the data yet! What to do?

- Make a guess p^*. If the guess is good, the margin of error will be close to what we want.
- Be conservative: choose $p^* = 0.5$. This will yield the largest n of any p^*.