Module 1 – Section 2

Distributions for Counts: Bernoulli, Binomial, and Multinomial Random Variables

Outline

- Background: Discrete Random Variables
- Bernoulli and Binomial Distributions
- Multinomial Distribution

Random Variables

- Variable whose value is determined based on random event.
  - Value is unknown before random event occurs.
  - Long-term behavior (distribution) of random variable is knowable.

Random Variables

- Two Types
  - **Discrete** – finite or countably infinite number of possible values
  - Continuous – infinite number of possible values
- Denoted with capital letters (X, **Y**, Z)
  - Z is typically reserved for continuous normal random variable.

Discrete Random Variables

- Distribution – table, formula or graph of values of Y and their probabilities for all possible values of y
- Denoted as $P(Y = y)$

Ex. Discrete Distribution

- The manager of a stockroom in a factory has constructed the following probability distribution for $Y =$ daily cost for a particular tool.

y	\$0	\$100	\$200
$P(Y = y)$	0.1	0.4	0.5

Expected Value

- Average or mean value of the random variable
- Does not have to be a possible value of the variable
- Denoted as $E(Y)$ or μ
- Calculated as

$$\mu = E(Y) = \sum_{y} y * P(Y = y)$$

Ex. Expected Value

y	\$0	\$100	\$200
$P(Y = y)$	0.1	0.4	0.5

$$E(Y) = 0(0.1) + 100(0.4) + 200(0.5) = 140$$

The mean daily cost of this particular tool is \$140.

Variance

- Measures a “mean squared distance” of values of random variable from expected value
- Denoted as $V(Y)$ or σ^2
- Calculated as

$$\sigma^2 = V(Y) = \sum_y \big(y - E(Y)\big)^2 P(Y = y)$$

Standard Deviation

- Variance is in squared units of the random variable
- Calculate standard deviation

$$\sigma = \sqrt{V(Y)}$$

- Standard deviation is in same units as random variable

Ex. Variance

y	\$0	\$100	\$200
$P(Y = y)$	0.1	0.4	0.5

$$
\begin{aligned}
V(Y) &= (0-140)^2(0.1) + (100-140)^2(0.4) + (200-140)^2(0.5) \\
&= 4400
\end{aligned}
$$

The variance of the daily cost of this particular tool is 4400 squared dollars

Ex. Standard Deviation

y	\$0	\$100	\$200
$P(Y = y)$	0.1	0.4	0.5

$$V(Y) = 4400$$

$$\sqrt{V(Y)} = \sqrt{4400} = 66.3325$$

The standard deviation of the daily cost of this particular tool is approximately $66.33.

Connection to Categorical Data

- Observations from Categorical Data = Counts
  - Finite or countably infinite number of possible values
- Discrete Random Variables for Categorical Data
  - Binomial – Two possible categories (outcomes) for categorical variable
    - Special case: Bernoulli Random Variable
  - Multinomial – More than two possible categories (outcomes) for categorical variable

Bernoulli Random Variables

- Random event with only 2 outcomes
- Outcomes
  - Success = Category of Interest
  - Failure = Not in Category of Interest
- Probabilities
  - Success = p
  - Failure = $1 - p$

Ex. Bernoulli Random Variables

- Outcome of flipping a coin
- Developing a disease or not
- Making a free throw or not
- Having Rh+ factor blood or not

Bernoulli Random Variables

- $Y = \begin{cases} 1 & \text{if Success} \\ 0 & \text{if Failure} \end{cases}$

- Distribution of Y

y	0	1
$P(Y = y)$	$1 - p$	p

Bernoulli Random Variables

- $E(Y) = p$

- $V(Y) = p(1 - p)$

Ex. Bernoulli Random Variable

- Outcome of flipping a coin
  - Success = Heads ($Y = 1$)
  - Failure = Tails ($Y = 0$)
  - $p = 0.5$
  - $E(Y) = 0.5$
  - $V(Y) = 0.5(0.5) = 0.25$

Binomial Random Variables

- Random event with 2 outcomes
- Outcomes
  - Success = Category of Interest
  - Failure = Not in Category of Interest
- Probabilities
  - Success = p
  - Failure = $1 - p$

Binomial Random Variables

- Y = number of successes in n independent and identical trials of random event
- Independent – outcome on one trial does not affect outcomes on other trials
- Identical – same probability of success

Outcome of Binomial Random Variable

- General Frequency Table

Outcome	Freq
Success	Y
Failure	$n - Y$
Total	n

Binomial Random Variables

- Possible values of Y are from 0 to n.
- Distribution of Y determined by the parameters n and p

$$P(Y = y) = \binom{n}{y} p^y (1 - p)^{n-y} \quad \text{where } \binom{n}{y} = \frac{n!}{y!(n-y)!}$$

Binomial Random Variables

- Expected Value

$$\mu = E(Y) = np$$

- Variance

$$\sigma^2 = V(Y) = np(1-p)$$

- Standard Deviation

$$\sigma = \sqrt{np(1-p)}$$

Ex. Binomial Random Variable

- Pairs of nesting birds
  - Success = Produce an offspring
  - Failure = Do not produce an offspring
- $p = 0.6$
- $n = 5$ pairs of nesting birds

Outcome	Freq
Offspring	Y
No Offspring	$5 - Y$
Total	5

Ex. Binomial Random Variable

- Find probability that 2 pairs of nesting birds successfully produce an offspring.

$$P(Y = 2) = \binom{5}{2} 0.6^2 0.4^3 = 0.2304$$

Ex. Binomial Random Variable

- Find probability that no pair of nesting birds successfully produce an offspring.

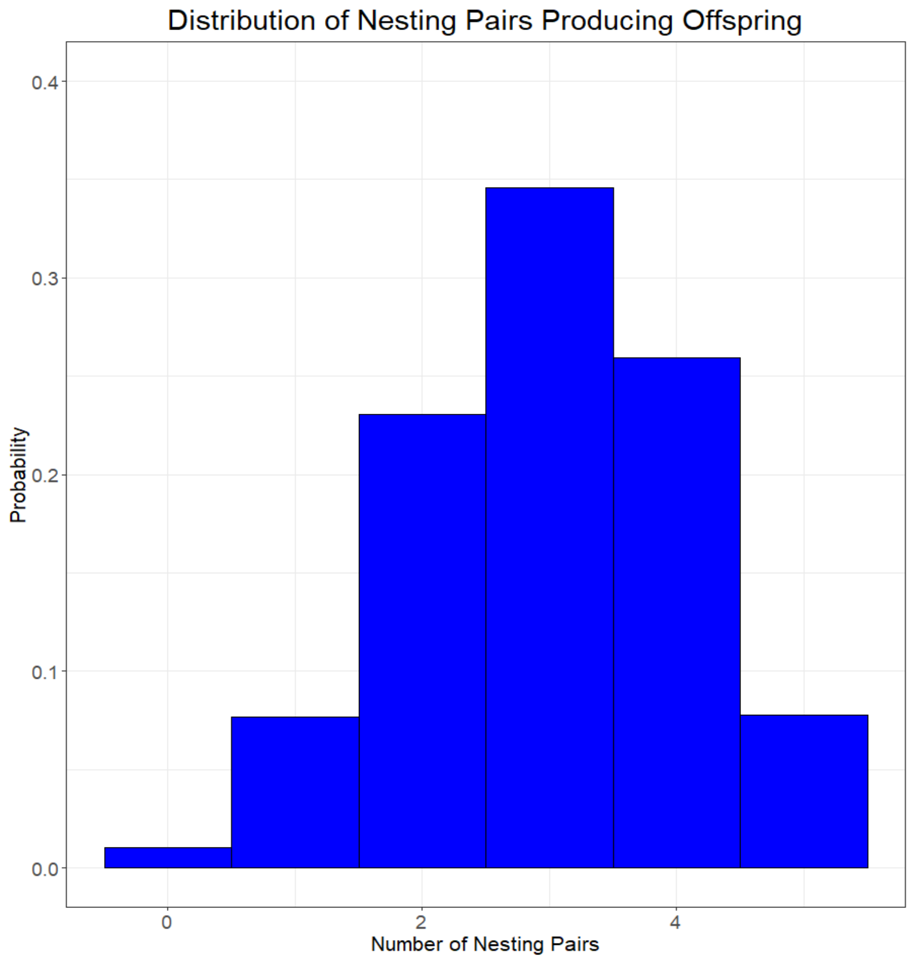
$$P(Y = 0) = \binom{5}{0} 0.6^0 0.4^5 = 0.0102$$

Ex. Binomial Random Variable

- Find probability that at least one pair of nesting birds successfully produce an offspring.

$$\begin{aligned} P(Y \geq 1) &= P(Y=1)+P(Y=2)+P(Y=3)+P(Y=4)+P(Y=5) \\ &= 0.0768+0.2304+0.3456+0.2592+0.0778=0.9898 \\ \\ &= 1-P(Y=0) \\ &= 1-0.0102=0.9898 \end{aligned}$$

Ex. Binomial Random Variable



Ex. Binomial Random Variable

- Mean

$$\mu = E(Y) = np = 5(0.6) = 3$$

  - The mean number of 5 pairs of nesting birds that successfully produce an offspring is 3 pairs.

Ex. Binomial Random Variable

- Variance

$$\sigma^2 = V(Y) = np(1-p) = 5(0.6)(0.4) = 1.2$$

  - The variance of the number of 5 pairs of nesting birds that successfully produce an offspring is 1.2 squared pairs.

Ex. Binomial Random Variable

- Standard Deviation

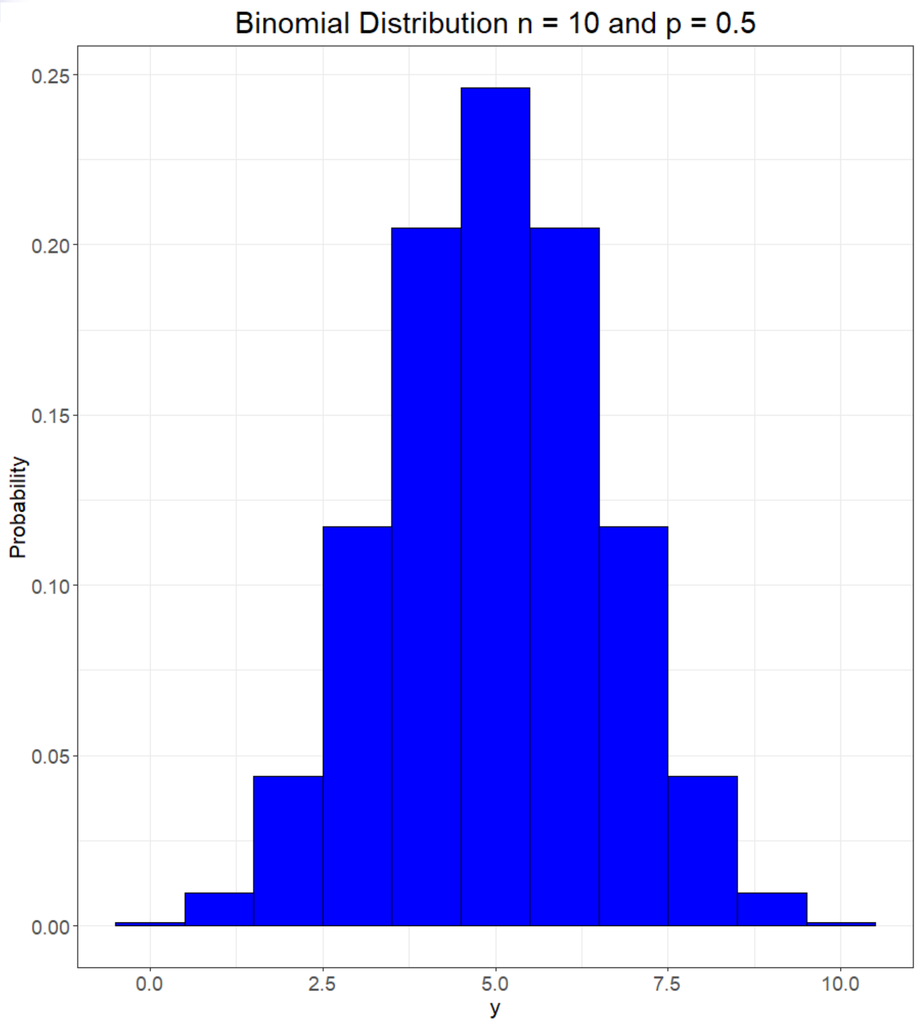
$$\sigma = \sqrt{np(1-p)} = \sqrt{5(0.6)(0.4)} = 1.095$$

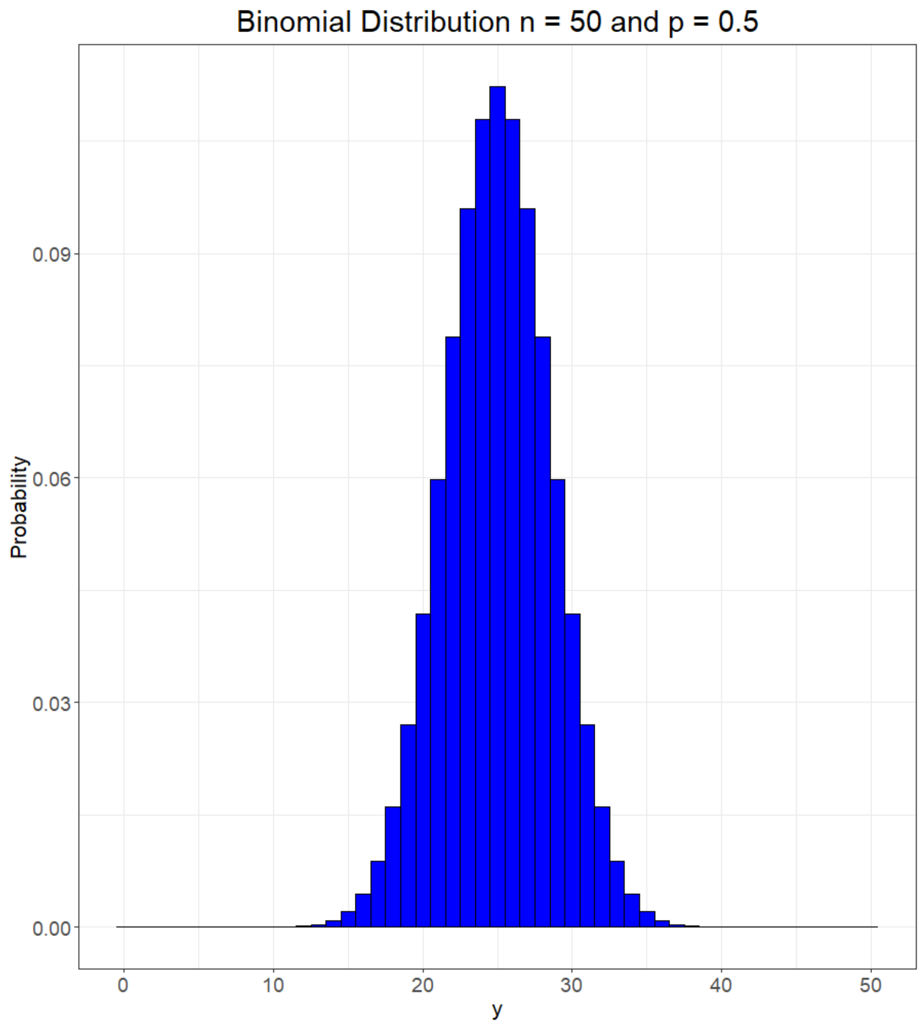
  - The standard deviation of the number of 5 pairs of nesting birds that successfully produce an offspring is 1.095 pairs.

Binomial Distribution

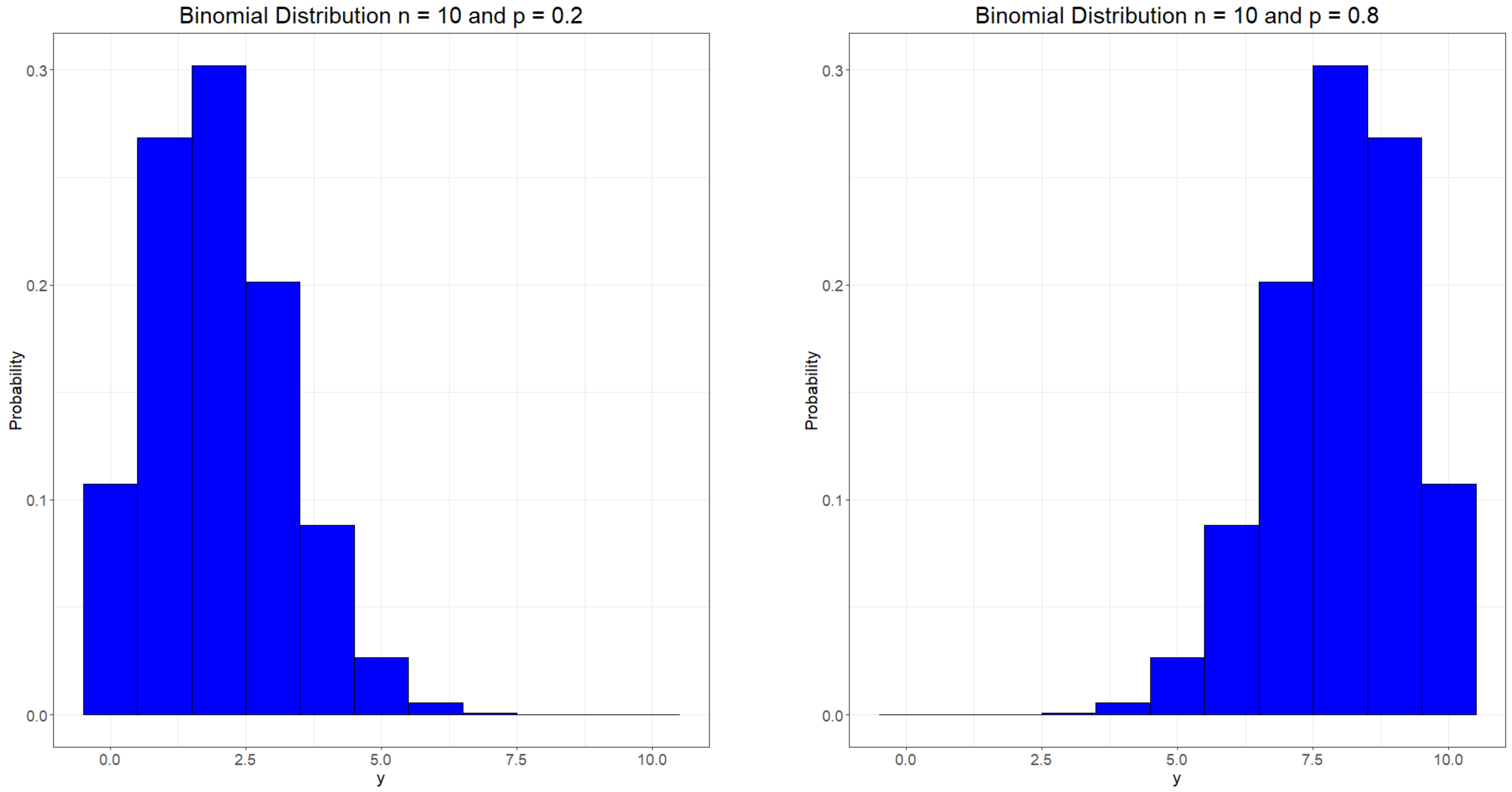
- Depends on values of n and p
- For fixed n, distribution for p and $1 - p$ are mirror images of each other
- If $np \geq 10$ and $n(1 - p) \geq 10$, shape of binomial distribution is close to normal distribution

Binomial Distribution

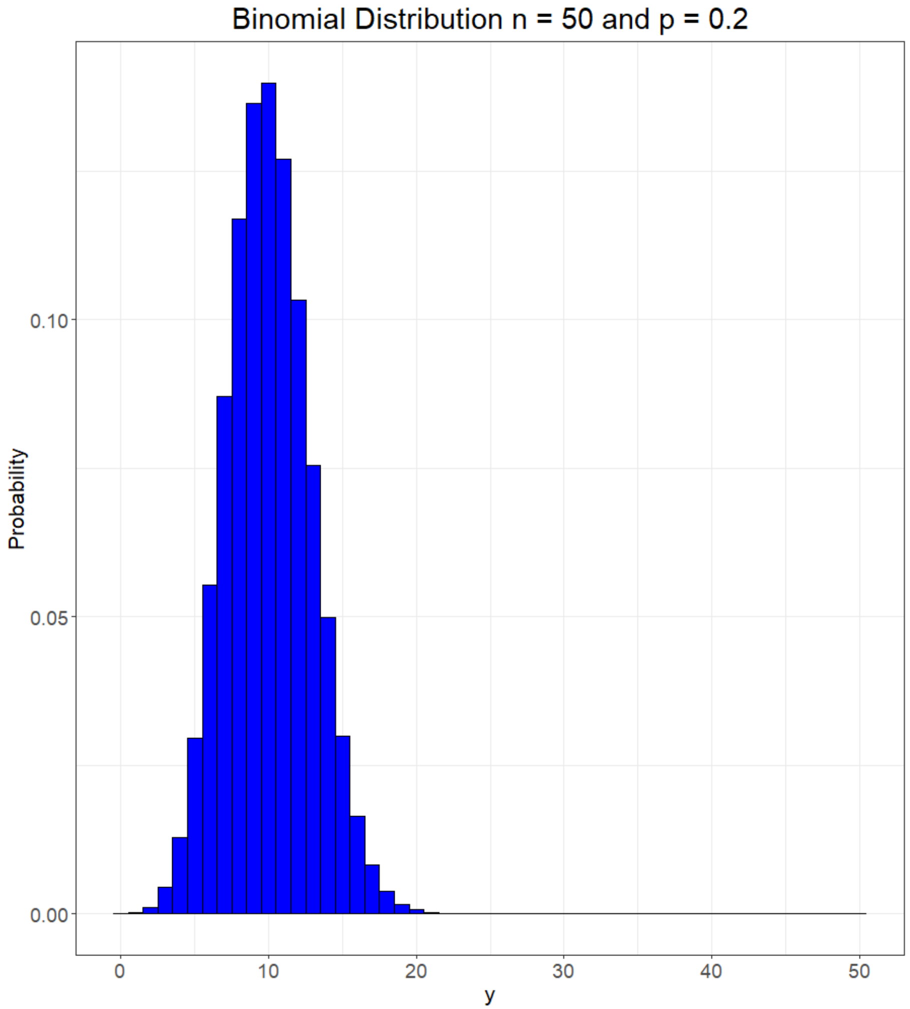


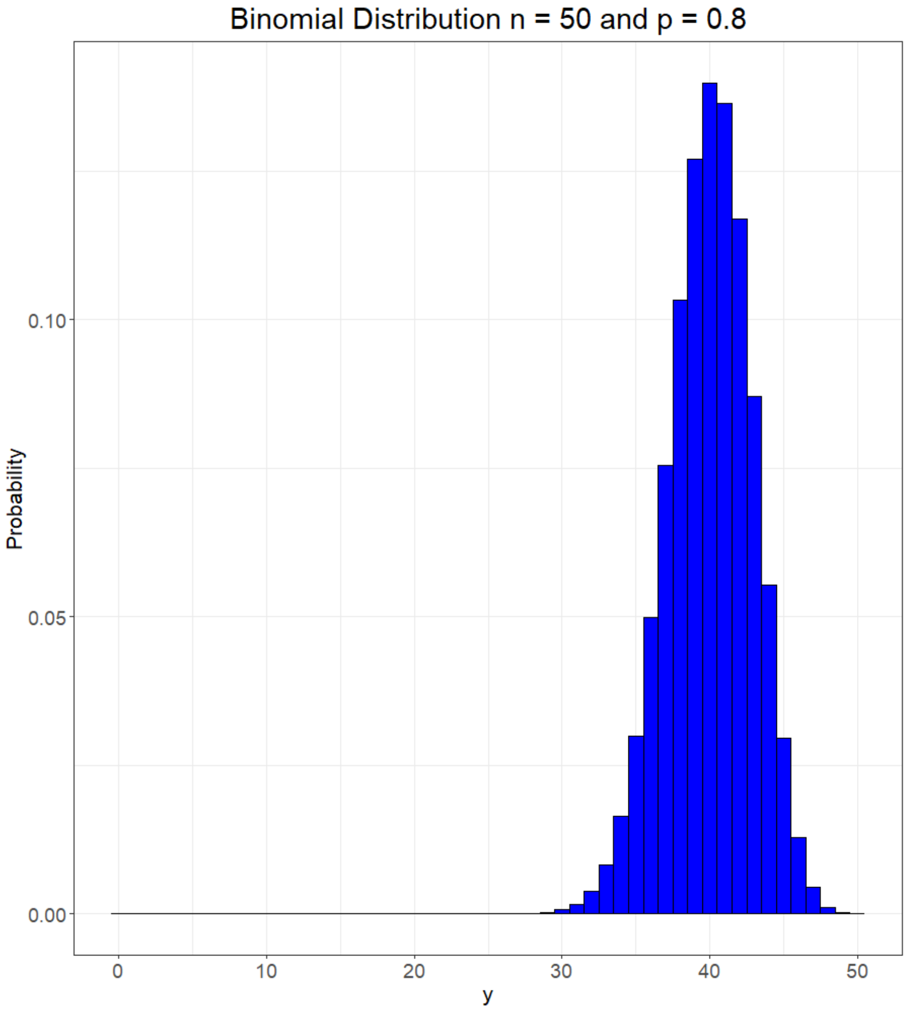


Binomial Distribution

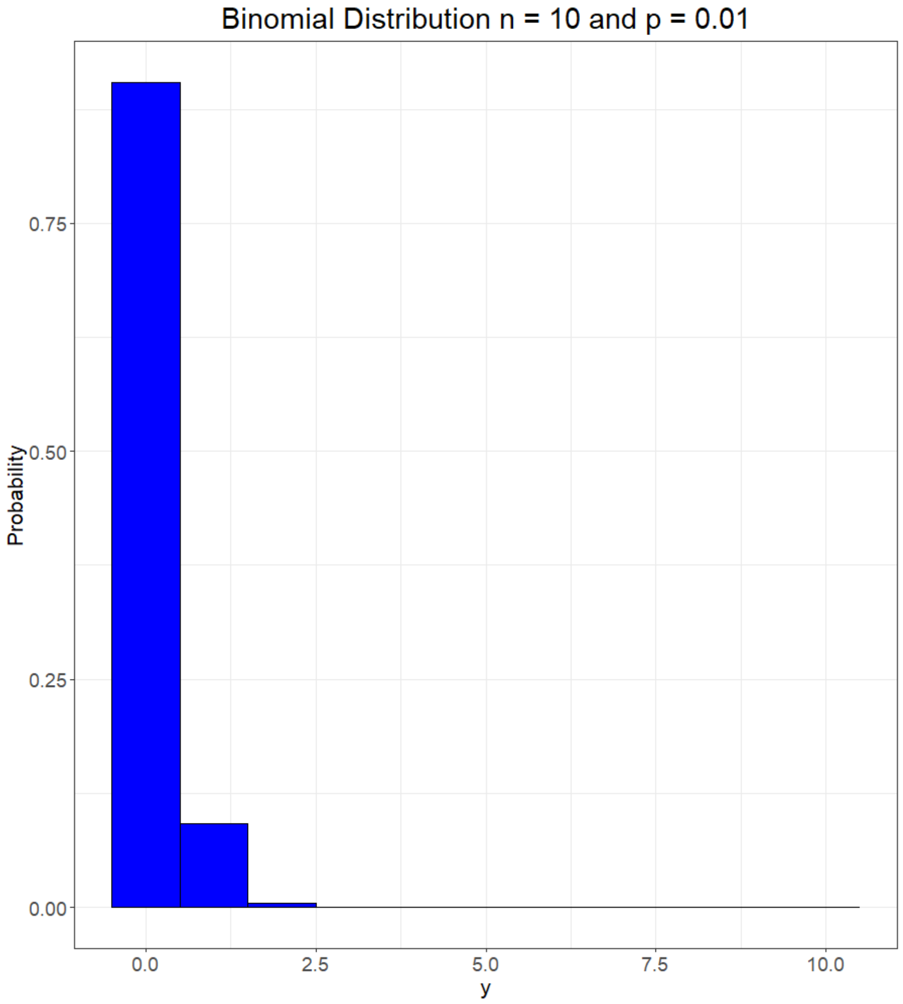


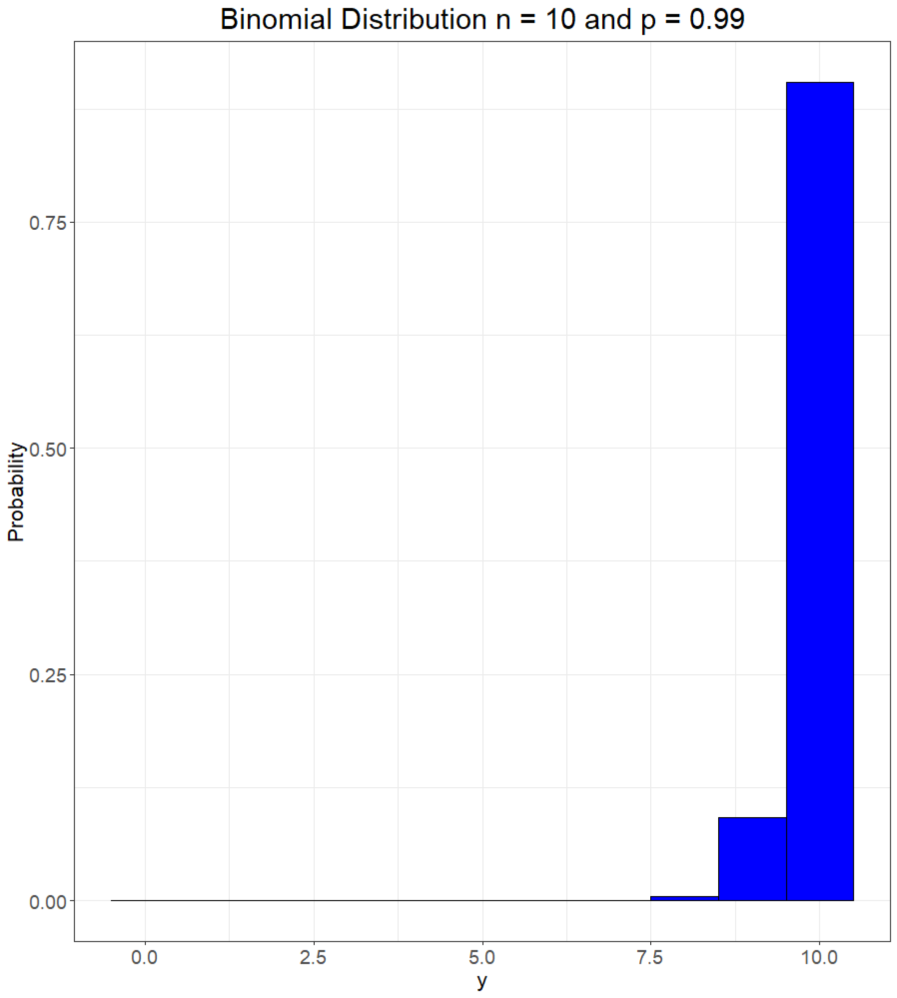
Binomial Distribution



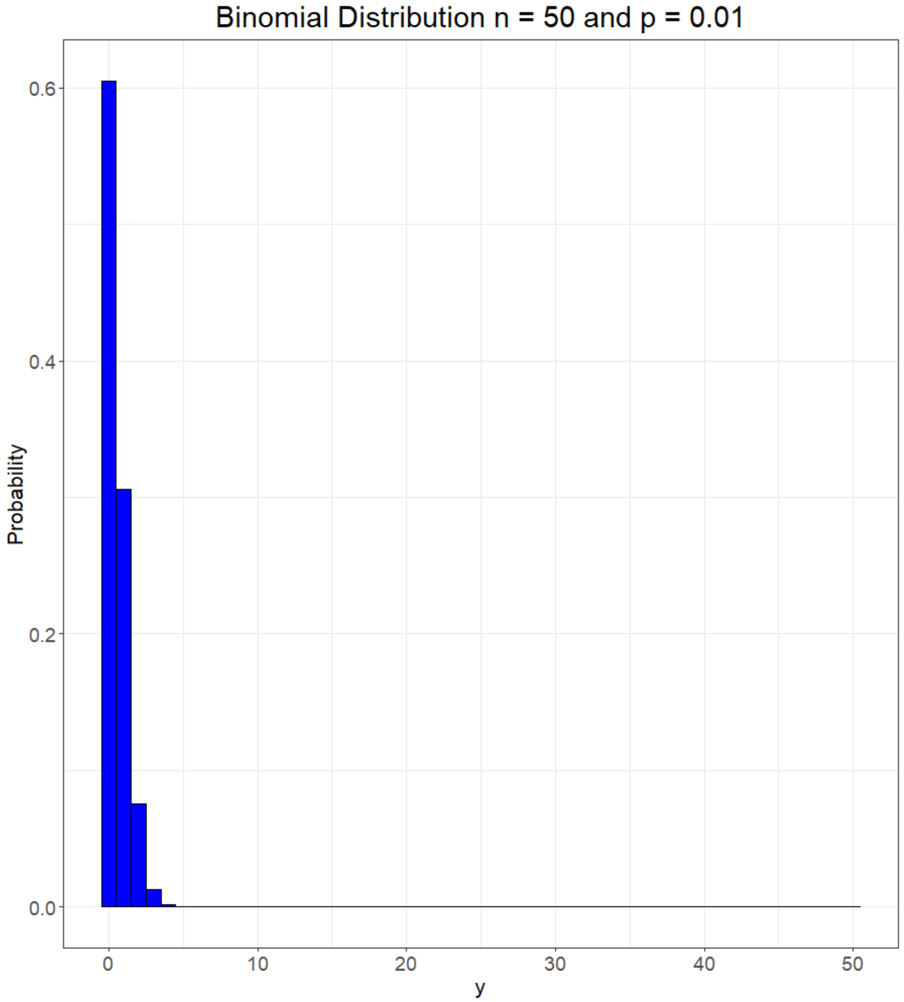


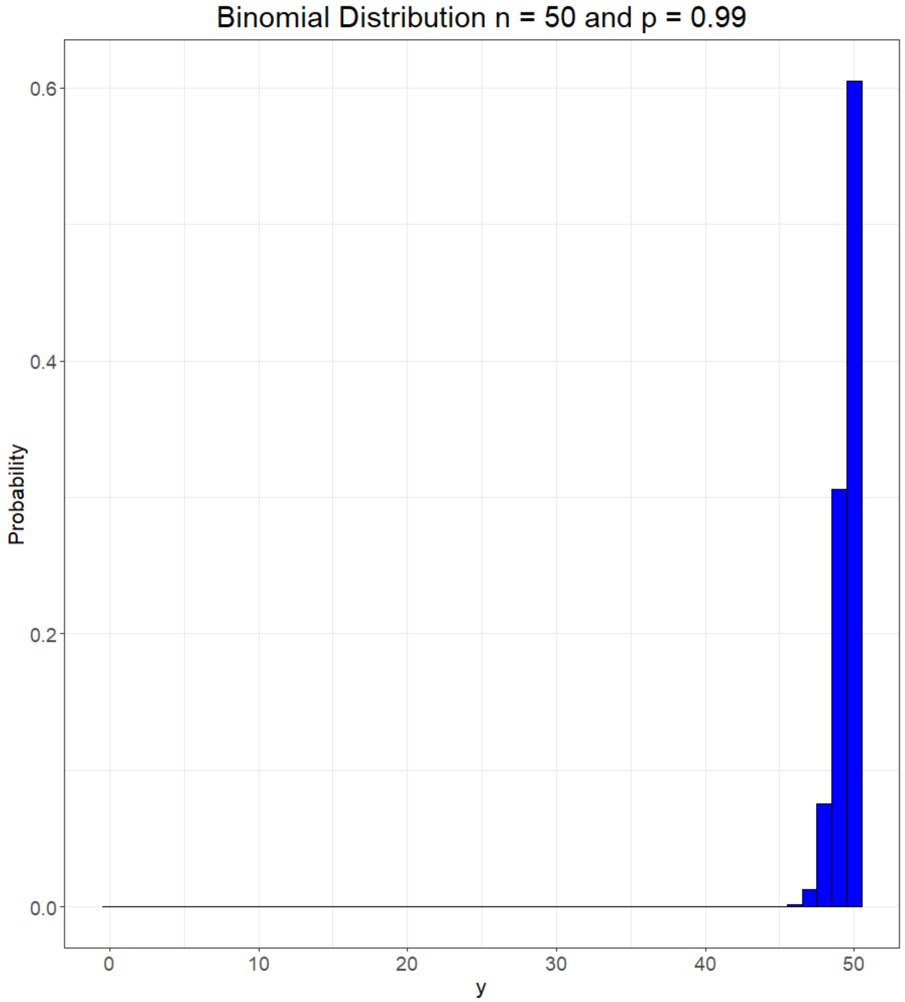
Binomial Distribution





Binomial Distribution





Multinomial Random Variables

- Random event with $J > 2$ outcomes
- Probability of each Outcome = p_j
- $\sum_{j=1}^{J} p_j = 1$

Multinomial Random Variables

- Y_j = number of observations in j^{th} outcome in n independent and identical trials of random event
- Independent – outcome on one trial does not affect outcomes on other trials
- Identical – same probabilities for outcomes

Multinomial Random Variable

- General Frequency Table

Outcome	Freq
Cat 1	Y_1
Cat 2	Y_2
Cat 3	Y_3
$\vdots$	$\vdots$
$\vdots$	$\vdots$
Cat J	Y_J
Total	n

Multinomial Random Variables

- Possible values for each Y_j are from 0 to n.
- Distribution

$$p(y_1, y_2, \ldots, y_J) = \frac{n!}{y_1!\, y_2! \cdots y_J!} p_1^{y_1} p_2^{y_2} \cdots p_J^{y_J}$$

where $\sum_{j=1}^{J} p_j = 1$ and $\sum_{j=1}^{J} Y_j = n$

Multinomial Random Variables

- Means

$$\mu_j = E(Y_j) = np_j \text{ for all } j = 1, 2, \dots, J$$

Multinomial Random Variables

- Variances

$$\sigma_j^2 = V\left(Y_j\right) = np_j(1 - p_j) \text{ for all } j = 1, 2, \dots, J$$

- Standard Deviations

$$\sigma_j = \sqrt{np_j(1 - p_j)} \text{ for all } j = 1, 2, \dots, J$$

Covariance of Two Random Variables

- Joint variability of two random variables with respect to their means.
  - Positive covariance = values of the two variables are most often smaller or larger than their means at the same time.
  - Negative covariance = Values below the mean of one variable are most often associated with values above the mean of the other variable (and vice versa).
- Sign is direction of linear relationship, magnitude is not interpretable.

Correlation of Two Random Variables

- Joint variability of two standardized* random variables
  - Parameter value is called ρ.
  - Sign is same as covariance = direction of linear relationship
- Magnitude is strength of linear relationship

Equal to 0 – none	From 0.7 to 0.9 – strong
From 0 to 0.3 – weak	From 0.9 to 1 – very strong
From 0.3 to 0.7 – moderate	Equal to 1 – perfect

*(value – mean)/std. dev.

Multinomial Random Variables

- Covariances

$$\mathrm{Cov}(Y_j, Y_k) = -np_j p_k \text{ for } j \neq k$$

- Correlations

$$\rho(Y_j, Y_k) = -\sqrt{\frac{p_j p_k}{(1-p_j)(1-p_k)}} \text{ for } j \neq k$$

Ex. Multinomial Random Variable

- Distribution of Colors of Milk Chocolate M&Ms*
  - Blue – 24%
  - Orange – 20%
  - Yellow – 14%
  - Red – 13%
  - Green – 16%
  - Brown – 13%

*from company's website – June 2008

Outcome	Freq
Blue	Y_1
Orange	Y_2
Yellow	Y_3
Red	Y_4
Green	Y_5
Brown	Y_6
Total	n

Ex. Multinomial Random Variable

- What is the probability of obtaining a handful of M&Ms with 2 blue, 2 orange, 1 yellow, 1 red, 1 green, and 1 brown?

$$p(2,2,1,1,1,1) = \frac{8!}{2!\,2!\,1!\,1!\,1!\,1!} 0.24^2 0.2^2 0.14^1 0.13^1 0.16^1 0.13^1$$

$$= 0.0088$$

Ex. Multinomial Random Variable

- What is the expected number of orange M&Ms in a small bag with 50 M&Ms?

$$\mu_2 = E(Y_2) = np_2 = 50(0.2) = 10$$

  - The mean number of orange M&Ms in a small bag of 50 M&Ms is 10 M&Ms.

Ex. Multinomial Random Variable

- What is the variance of the number of orange M&Ms in a small bag with 50 M&Ms?

$$\sigma_2^2 = V(Y_2) = np_2(1 - p_2) = 50(0.2)(0.8) = 8$$

  - The variance of the number of orange M&Ms in a small bag with 50 M&Ms is 8 squared M&Ms.

Ex. Multinomial Random Variable

- What is the standard deviation of the number of orange M&Ms in a small bag with 50 M&Ms?

$$\sigma_2 = \sqrt{np_2(1-p_2)} = \sqrt{50(0.2)(0.8)} = 2.83$$

  - The standard deviation of the number of orange M&Ms in a small bag with 50 M&Ms is approximately 2.83 M&Ms.

Ex. Multinomial Random Variable

- What is the covariance of the number of orange and red M&Ms in a small bag with 50 M&Ms?

$$\mathrm{Cov}(Y_2, Y_4) = -np_2p_4 = -50(0.2)(0.13) = -1.3$$

  - A covariance of -1.3 tells me there is a negative relationship between the number of orange and red M&Ms in a small bag with 50 M&Ms. This means an above average number of orange M&Ms is associated with a below average number of red M&Ms and vice versa.

Ex. Multinomial Random Variable

- What is the correlation between the number of orange and red M&Ms in a small bag with 50 M&Ms?

$$\rho(Y_2, Y_4) = -\sqrt{\frac{p_2 p_4}{(1-p_2)(1-p_4)}} = -\sqrt{\frac{0.2(0.13)}{(0.8)(0.87)}} = -0.1933$$

  - A correlation of -0.1933 tells me there is a negative weak linear relationship between the number of orange and red M&Ms in a small bag with 50 M&Ms.