ESS 575: Probability Lab 2 - Probability Distributions

Team England

07 September, 2022

Team England:

- Caroline Blommel
- Carolyn Coyle
- Bryn Crosby
- George Woolsey

cblommel@mail.colostate.edu, carolynm@mail.colostate.edu, brcrosby@rams.colostate.edu, george.woolsey@colostate.edu

Question 0

Barn swallows form pair bonds (male/female pairings) in the spring before mating season. Each male/female pair share a nest and care for the offspring of the female. Often a number of the female's offspring were sired by males other than her mate. We are interested in the random variable, the number of offspring sired by the females mate. Suppose previous literature suggests the probability an offspring's father is the female's mate is 0.8. Plot the probability mass function.

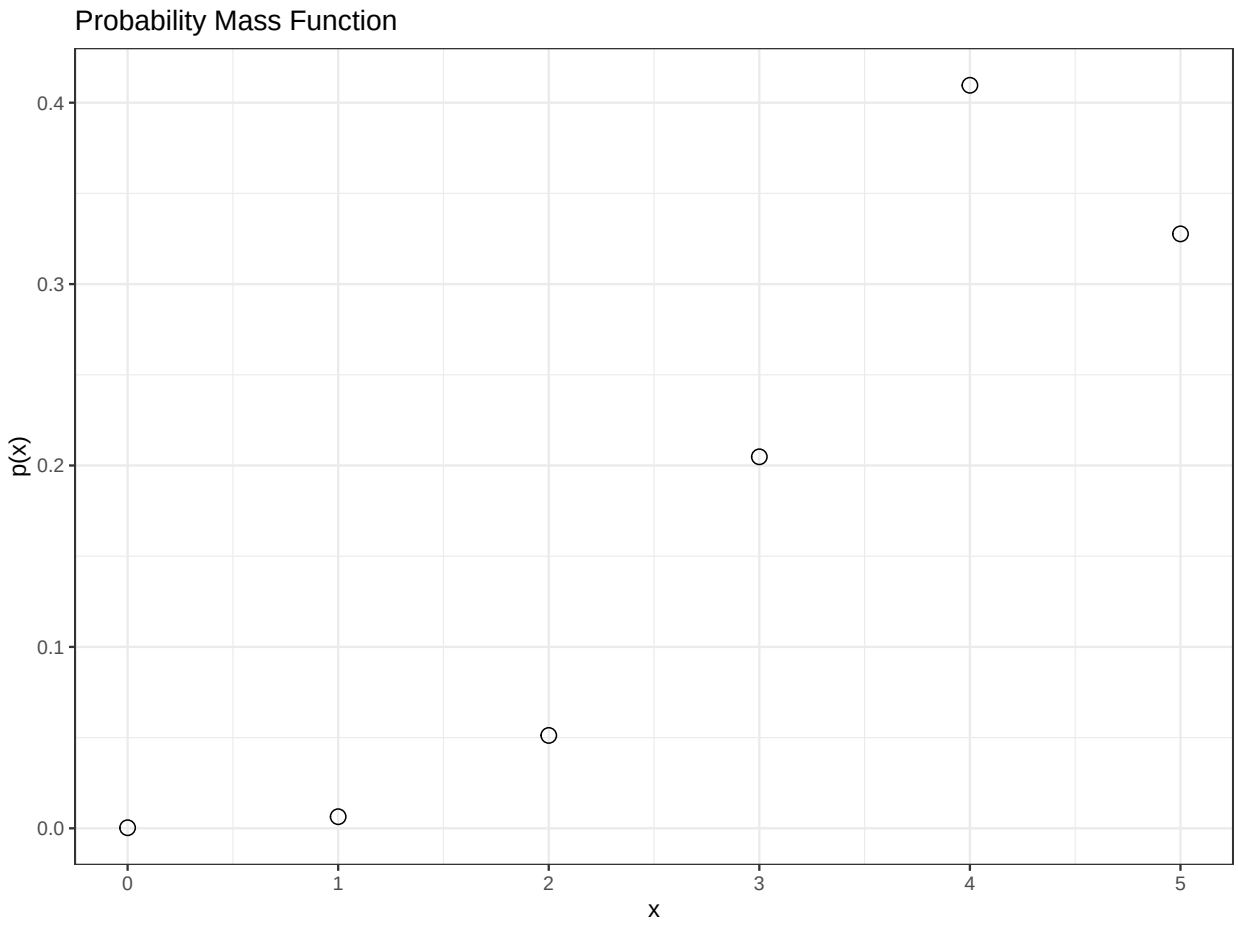
Short Answer

Write a model describing how the data arise.

$$y_i \sim \text{binomial}(5, 0.8) \tag{}$$

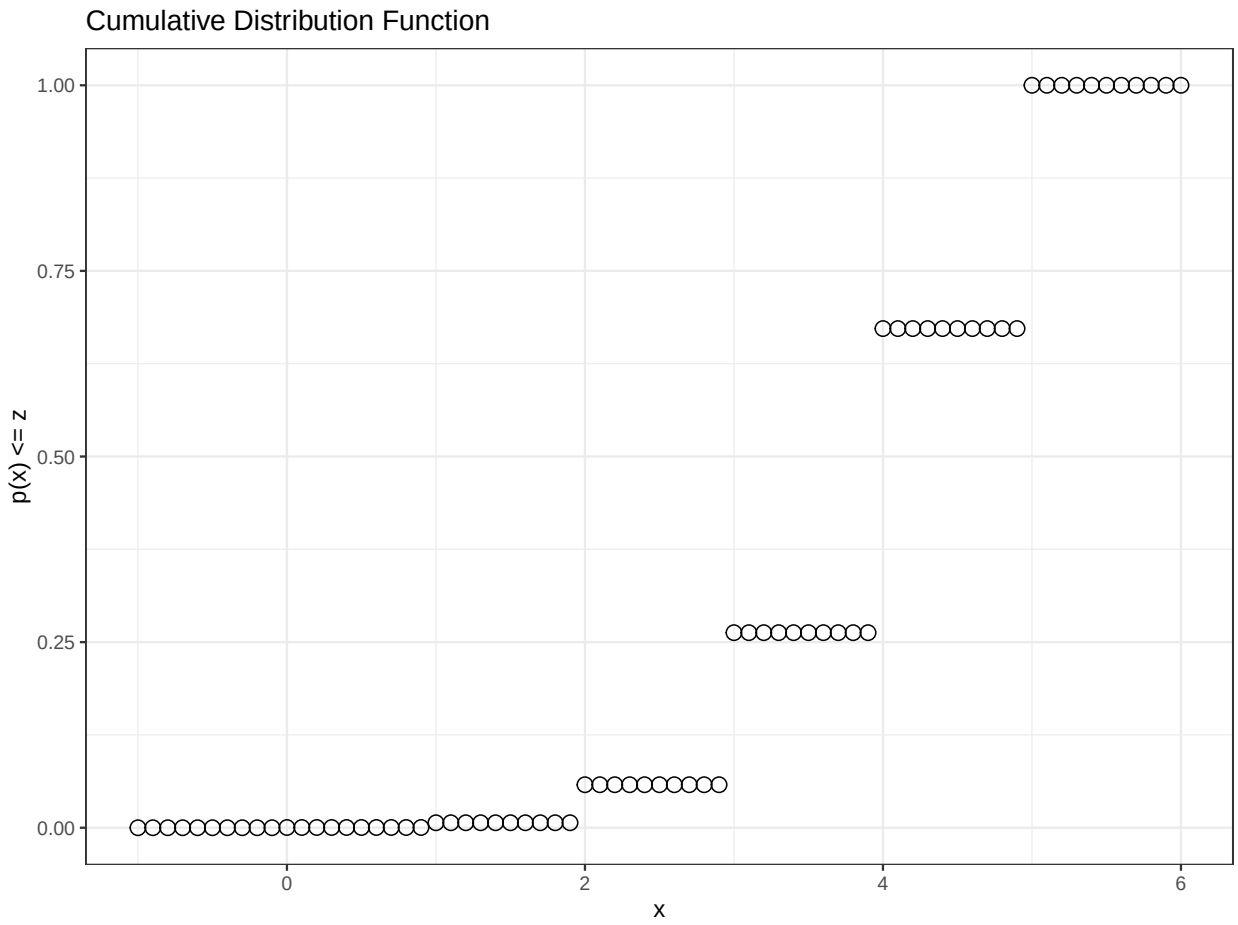
Plot the probability mass function.

```
p <- 0.8
x <- seq(0, 5, 1)
y <- dbinom(x = x, size = 5, prob = p)
ggplot(data.frame(x,y), aes(x = x, y = y)) +
  geom_point(shape = 1, size = 3) +
  labs(title = "Probability Mass Function") +
  xlab("x") +
  ylab("p(x)") +
  theme_bw()
```



Plot the cumulative distribution function for values between -1 and 6 in steps of .1 using pbinom. (Don't worry about the unfilled circles bit or making a fancy plot. Just get the concepts.)

```
p <- 0.8
x <- seq(-1, 6, 0.1)
y <- pbinom(q = x, size = 5, prob = p)
ggplot(data.frame(x,y), aes(x = x, y = y)) +
  geom_point(shape = 1, size = 3) +
  labs(title = "Cumulative Distribution Function") +
  xlab("x") +
  ylab("p(x) <= z") +
  theme_bw()
```



Write a mathematical expression for the the probability that there are fewer than four offspring sired by the female's mate and compute the probability using dbinom and pbinom.

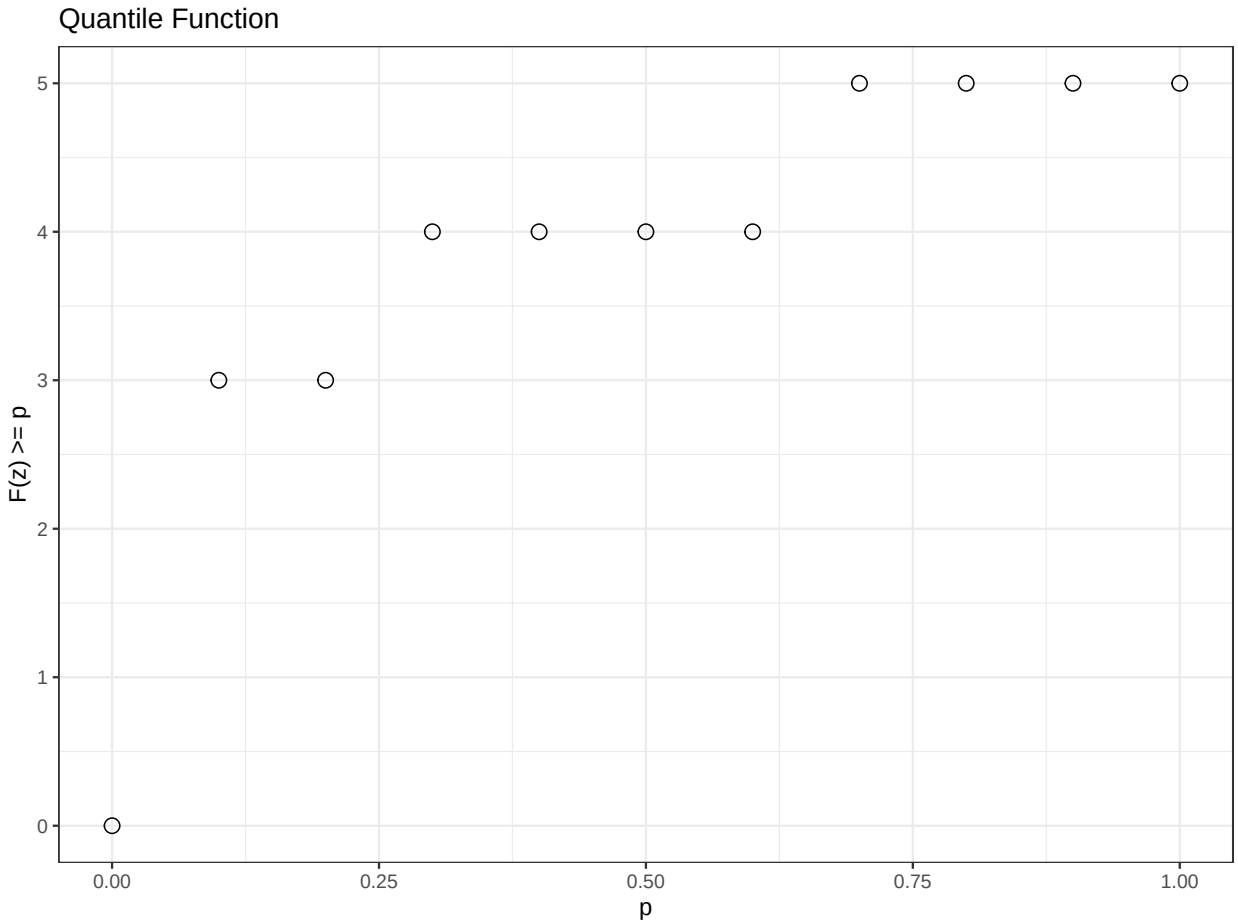
```
p <- 0.8
x <- seq(0, 3, 1)

sum_dbinom <- sum(dbinom(x = x, size = 5, prob = p))
p_binom <- pbinom(q = 3, size = 5, prob = p)
```

The probability that there are fewer than four offspring sired by the female's mate: **26.3%**

Plot the quantile function over the range of zero to 1 in steps of .1 using qbinom.

```
p <- 0.8
x <- seq(0, 1, 0.1)
y <- qbinom(p = x, size = 5, prob = p)
ggplot(data.frame(x,y), aes(x = x, y = y)) +
  geom_point(shape = 1, size = 3) +
  labs(title = "Quantile Function") +
  xlab("p") +
  ylab("F(z) >= p") +
  theme_bw()
```



Compute the smallest number of offspring with a probability equal to or exceeding .3

```
p <- 0.8
x <- 0.3
q_binom <- qbinom(p = x, size = 5, prob = p)
```

The smallest number of offspring with a probability equal to or exceeding .3 is: **4**

Question 1

We commonly represent the following general framework for linking models to data:

$$[y_i \mid g(\theta, x_i), \sigma^2] \tag{1}$$

which represents the probability of obtaining the observation y_i given that our model predicts the mean of a distribution $g(\theta, x_i)$ with variance σ^2. Assume we have count data. What distribution would be a logical choice to model these data? Write out a model for the data.

Short Answer

The Poisson distribution is the classic distribution for (integer) counts; e.g., things in a plot, things that pass a point, etc.). The Poisson distribution has 1 parameter: λ = mean = variance; that is, the first and

second central moments are equal.

$$y_i \sim \text{Poisson}(g(\theta, x_i)) \quad ()$$

```
pois <- rpois(n = 100000, lambda = 50)
pois[1:10]
```

```
##  [1] 50 48 40 45 52 41 48 45 57 55
```

Question 2

We commonly represent the following general framework for linking models to data:

- The mass of carbon in above ground biomass in square m plot.
- The number of seals on a haul-out beach in the gulf of AK.
- Presence or absence of an invasive species in forest patches.
- The probability that a white male will vote republican in a presidential election.
- The number of individuals in four mutually exclusive income categories.
- The number of diseased individuals in a sample of 100.
- The political party affiliation (democrat, republican, independent) of a voter.

Short Answer

Table 1: Question 2 Answers

Random Variable	Distribution	Support
The mass of carbon in above ground biomass in square m plot.	lognormal or gamma	non-negative real numbers
The number of seals on a haul-out beach in the gulf of AK.	Poisson or negative binomial	counts
Presence or absence of an invasive species in forest patches.	Bernoulli	0 or 1
The probability that a white male will vote republican in a presidential election.	beta	0 to 1 real numbers
The number of individuals in four mutually exclusive income categories.	multinomial	counts in >2 categories
The number of diseased individuals in a sample of 100.	binomial	counts in 2 categories
The political party affiliation (democrat, republican, independent) of a voter.	multinomial	counts in >2 categories

Question 3

Find the mean, variance, and 95% quantiles of 10000 random draws from a Poisson distribution with $\lambda = 33$.

```r
pois <- rpois(n = 10000, lambda = 33)
pois[1:10]
```

```
##  [1] 35 31 38 36 30 22 47 29 23 31
```

Short Answer

The mean of this example Poisson distribution is: **33.0**

The variance of this example Poisson distribution is: **33.4**

The 95% quantile of this example Poisson distribution is: **22.0, 45.0**

Question 4

Simulate `one` observation of survey data with five categories on a Likert scale, i.e. strongly disagree to strongly agree. Assume a sample of `80` respondents and the following probabilities:

a) Strongly disagree = 0.07
b) Disagree = .13
c) Neither agree nor disagree = .15
d) Agree = .23
e) Strongly agree = .42

```r
p <- c(0.07, 0.13, 0.15, 0.23, 0.42)
N <- 80
n <- 1
rmultinom(n = n, size = N, prob = p)
```

```
##      [,1]
## [1,]    5
## [2,]   11
## [3,]   10
## [4,]   22
## [5,]   32
```

Question 5

The average above ground biomass in a grazing allotment of sagebrush grassland is `103 g/m^2`, with a standard deviation of `23`. You clip a `1 m^2` plot. Write out the model for the probability density of the data point. What is the probability density of an observation of `94` assuming the data are normally distributed? Is there a problem using normal distribution? What is the probability that your plot will contain between `90 gm` and `110 gm` of biomass? (For lab report)

Answer

$$y \sim \mathsf{normal}(103, 23^2) \qquad ()$$

```r
x <- 94
mean <- 103
sd <- 23
d_norm <- dnorm(x = x, mean = mean, sd = sd)
scales::percent(d_norm, accuracy = .1)
```

```
## [1] "1.6%"
```

The probability density of an observation of 94 assuming the data are normally distributed is: **1.6%**

Is there a problem using normal distribution?

The normal distrubtion supports random variables from $-\infty$ to $+\infty$ and it is not possible to have negative biomass in a sample area.

```r
samp_low <- 90
samp_high <- 110
p_norm <- pnorm(q = c(samp_low, samp_high), mean = mean, sd = sd)

scales::percent(p_norm[2] - p_norm[1], accuracy = .1)
```

```
## [1] "33.4%"
```

The probability that the plot will contain between 90 gm and 110 gm of biomass is: **33.4%**

Question 6

The prevalence of a disease in a population is the proportion of the population that is infected with the disease. The prevalence of chronic wasting disease in male mule deer on winter range near Fort Collins, CO is `12 percent`. A sample of `24` male deer included `4` infected individuals. Write out a model that represents how the data arise. What is the probability of obtaining these data conditional on the given prevalence (`p=0.12`)? (For lab report)

Answer

$$y \sim \text{binomial}(24, 0.12) \qquad ()$$

```r
x <- 4
n <- 24
p <- 0.12
d_binom <- dbinom(x = x, size = n, prob = p)
```

The probability of obtaining these data conditional on the given prevalence (p=0.12) is: **17.1%**

Question 7

Researchers know that the true proportion of related age-sex classifications for elk in Rocky Mountain National Park are: Adult females (`p = 0.56`), Yearling males (`p = 0.06`), Bulls (`p = 0.16`), and Calves (`p = 0.22`). What is the probability of obtaining the classification data conditional on the known sex-age population proportions given the following counts?

a) Adult females (count = 65)
b) Yearling males (count = 4)
c) Bulls (count = 25)
d) Calves (count = 26)

Answer

```
p <- c(0.56, 0.06, 0.16, 0.22)
n <- c(65, 4, 25, 26)

d_multinom <- dmultinom(x = n, prob = p)
```

The probability of obtaining the classification data conditional on the known sex-age population proportions given the following counts is: **0.030%**

Question 8

Nitrogen fixation by free-living bacteria occurs at a rate of `1.9 g/N/ha/yr` with a standard deviation σ of `1.4`. What is the lowest fixation rate that exceeds `2.5 percent` of the distribution? Use a normal distribution for this problem, but discuss why this might not be a good choice.

Answer

```
p <- 0.025
mean <- 1.9
sd <- 1.4

q_norm <- qnorm(p = p, mean = mean, sd = sd)
```

The lowest fixation rate that exceeds 2.5% of the distribution is: **-84.4%**

Is there a problem using normal distribution?

The normal distrubtion supports random variables from $-\infty$ to $+\infty$ and it is not possible to have negative nitrogen fixation.