Probability distributions

IBIO 851

Sept 20 2016

Suggested reading: **Chapter 4** in **Ecological Models & Data in R** (Bolker)

Quiz answers

- Q1: You are analyzing diameters of oak saplings the year after planting. Your 95% CI is between 6 & 13 cm. What does this mean?
- **A: 95% of the time upon repeated sampling, the interval of 6 to 13 cm will overlap the true population mean**

Quiz answers

- Q2: What are the null & alternative hypotheses for a one sample t-test? What is the linear model?
- **A: Null—No significant difference in means ($\bar{x}$-μ = zero)**
- **Alternative—Significant difference in means ($\bar{x}$-μ $\neq$ zero)**
- **Y(i) = α + β*x(i) + error(i), where error(i) ~ normal(0, σ^2)**

Quiz answers

- Q3: You are testing treatment of CO_2. Your experiment yields a p-value of 0.03. What does this mean?
- **A: Assuming CO_2 treatment had no effect on plant growth, you'd obtain the observed difference or more in 3% of studies due to random sampling error**

Goals for today

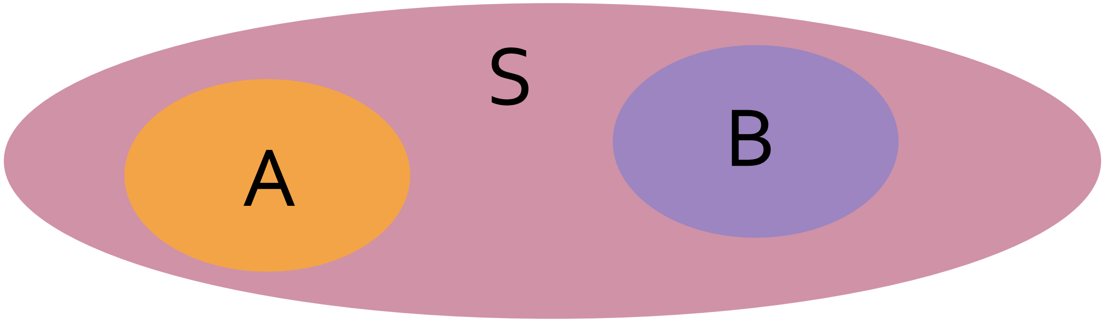
- Re-familiarize ourselves with conditional probabilities and language of probability
- Clarify probability distributions

Probability theory

- ‘Or’ statement: the union ∪
  - For 2 mutually exclusive events, the probability of getting A or B
  - $(A \cup B) = P(A \text{ or } B) = P(A) + P(B)$

Probability theory

- Two independent events
  - Pr(A)=Area of A / Area of S
  - Pr(B)=Area of B / Area of S
  - Pr(A or B)=(Area of A+B) / Area of S = **Pr(A)+Pr(B)**

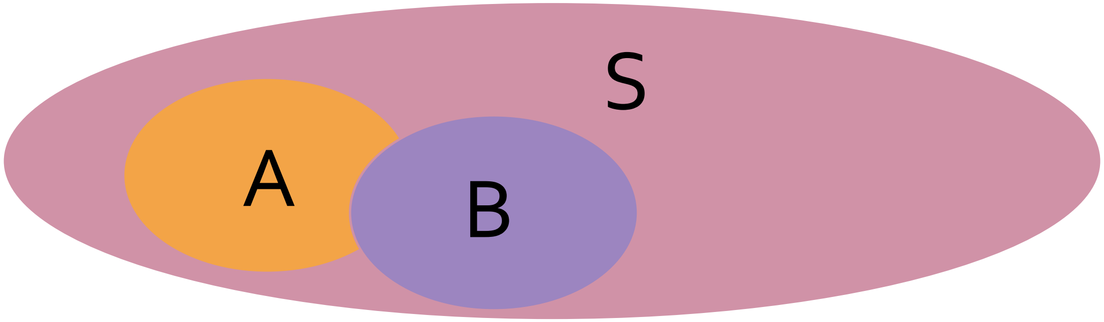


Probability theory

- Probability of a shared event
  - The joint probability of 2 events A & B occurring (assuming independence) is the **product** of their probabilities (intersection)
  - $P(A \& B) = P(A,B) = P(A \cap B) = P(A) * P(B)$
  - $P(A)*P(B) = 4/20 * 3/20 = 0.03$

Probability theory

- Probability of A or B (non independence)?
  - Pr(A or B)= (Area of A + Area of B – Area of AB) / Area of S =
  - Pr(A) + Pr(B) – Pr(A&B) =
  - **Pr(A) + Pr(B) – Pr(A*B)**



Conditional probabilities

- If we are calculating the probability of an event & we have information about the outcome of another event, we should include this information
- These ‘updated’ estimates are called conditional probabilities
  - Written as Pr(A|B)

Conditional probabilities

- $P(A|B) = \frac{P(A,B)}{P(B)}$
- Thus the conditional probability of A given B is equal to the intersection of A and B divided by the probability of B
- This can be rearranged: P(A,B) = P(A|B)*P(B)

Conditional probabilities

- **Example**: Rolling an octahedron dice
- S = {1, 2, 3, 4, 5 ,6, 7, 8}
- A is event of getting an odd number
- B is event of getting at least 7
- P(A) = 4/8 = ½
- P(B) = 2/8 = ¼
- P(A,B) = {7} = 1/8
- $P(A|B) = P(A,B)/P(B) = P(A) = 1/2$

Conditional probabilities

- Sometimes we know the conditional probabilities without any information on the joint probabilities
  - $P(B|A)*P(A)=P(A|B)*P(B)$
- We can rearrange this to get **Bayes Rule**:
  - $P(B|A) = \frac{P(A|B)*P(B)}{P(A)}$

Bayes' Rule

- Example: You are interested in finding out a patient's probability of having lung cancer if he/she is a smoker
- **A** is the event 'patient has lung cancer'
  - 10% of patients entering the clinic have lung cancer: P(A) = 0.10
- **B** is the event that 'patient is a smoker'
  - 5% of clinic's patients are smokers: P(B) = 0.05

Bayes' Rule

- Among those diagnosed with lung cancer, 7% are smokers
  - $P(B|A) = 0.07$
- $P(A|B) = \frac{P(B|A)*P(A)}{P(B)}$
- $P(A|B) = (0.07*0.1)/0.05 = 0.14$
  - If a patient is a smoker, the probability of having lung cancer is 14%

Random variables

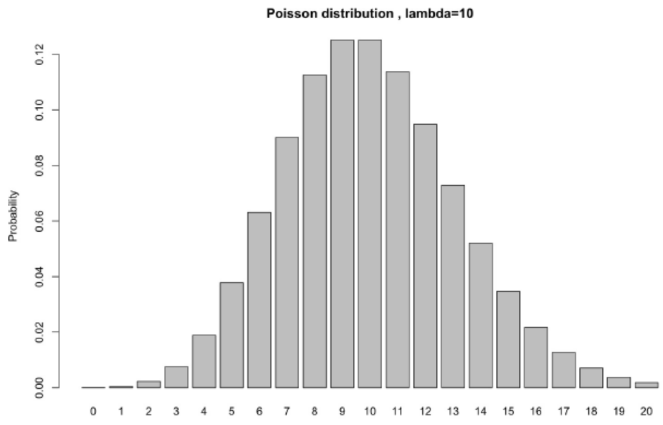
- A random variable X is a variable that assumes numerical values associated with the random outcome of an experiment, where one numerical value is assigned to each sample point
- The distribution of a random variable is the collection of possible outcomes along with their probabilities
  - Discrete or continuous

Random variables

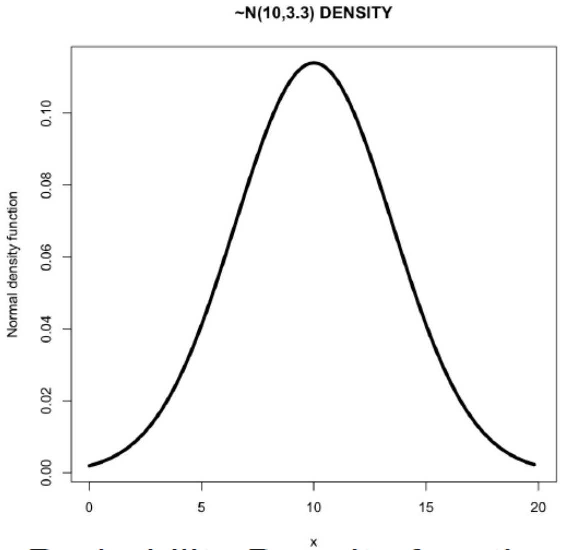
- Examples:
- X= height of an individual in our class?
  - Continuous random variable
- X=number of female students in the class?
  - Discrete random variable
- X=number of tosses required of a coin to obtain heads for the 5^{th} time?
  - Discrete random variable

Probability density vs. mass function

- A PDF is associated with a continuous variable; a PMF is associated with a discrete variable
  - A PDF must be integrated over an interval to yield a probability



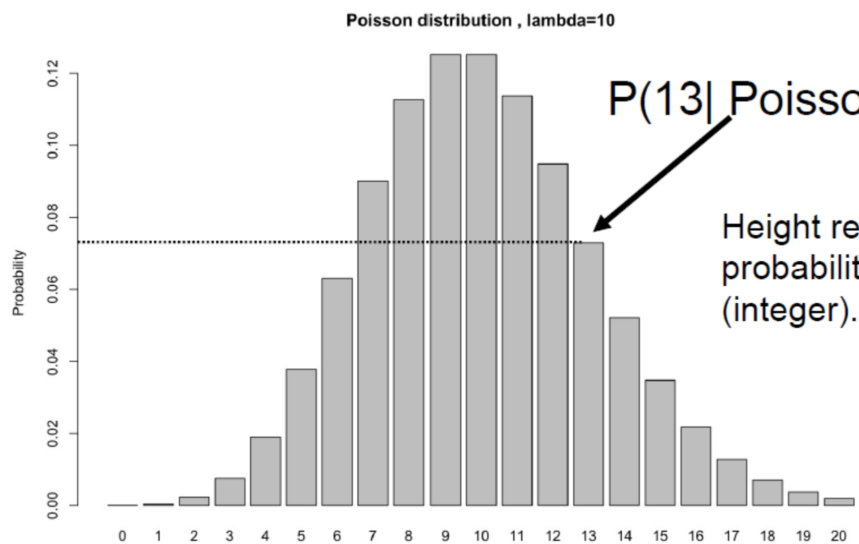
Probability Mass function for a discrete variable.



Probability Density function for a continuous variable.

Probability mass function

Probability Mass function



P(13| Poisson (λ=10)) =0.073

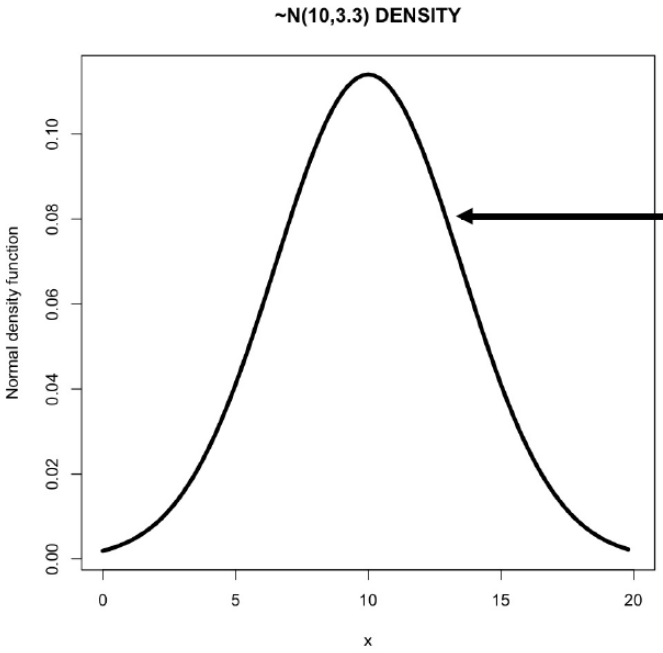
Height represents the probability at that point (integer).

“Area” of the box has no particular meaning.

P(integer) ≥ 0
P(non-integers) = 0.

Probability density function

Probability Density function



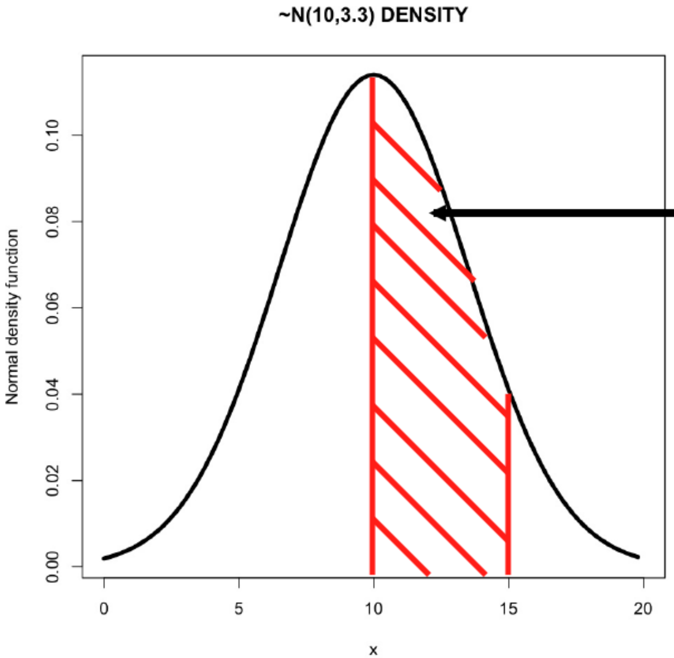
Height at x= 13 is 0.0799

This is not the probability at x=13, but the density.
i.e. f(13) = 0.0799, where f(x) is the normal distribution.

P(x=13| N(mean=10,sd=3.3)) = 0

Probability density function

Probability Density function



We can define the probability in the interval $10 \le x \le 15$

$P(10 \le x \le 15 | N(10,3.3)) = 0.435$

What is likelihood?

- P(data | hypothesis) = P(D|H)
- If we want to know what the probability of our data is, we need to have a context, i.e. hypothesis
- What do we mean by hypothesis?
  - A model!

What is likelihood?

- $Y \sim N(\beta_0 + \beta_1 X, \sigma) = \theta$
  - $\beta_0 + \beta_1 X$ is the linear predictor of the mean
- $P(D|H) = P(Y|\, N[\beta_0 + \beta_1 X, \sigma]\,) = P(Y|\,\theta)$
- Usually we don't know θ
- A natural estimation process is to choose that value of θ that would maximize the probability that we would actually observe Y
  - $L(\theta|Y) = P(Y|\,\theta)$ [discrete RVs]
  - $L(\theta|Y) = f(Y|\,\theta)$ [continuous RVs]

Probability example

- We have a small dataset of plant heights (cm): 7, 4, 3, 7, 7
- A previous plant population we examined looked like plant heights were approx. normally distributed with a mean of 5 cm, SD = 1 cm.
- What is the probability of a plant of height 7 cm coming from that population?

Probability example

- P(7 | ~N(5,1))
- Feed this into the normal distribution PDF:

$$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$\frac{1}{\sqrt{2\pi 1^2}}\exp\left(-\frac{(7-5)^2}{2*1^2}\right)\approx 0.054$$

0.054 is proportional to the probability (not exactly the probability since we're not looking at the interval)

Probability example

- So we figured out the density of a single data point, given an arbitrary hypothesis. Next step?
- Extend this concept to the whole dataset!
- We can use optimization criteria (e.g. maximum likelihood) to find an estimate of ‘best fit’ (given the parameters)

Probability

- Exercises in RStudio: using the pnorm and qnorm commands

Probability distributions

- We do not know the actual form of the distributions of the data
- We use known probability distributions to approximate what we observe &/or predict
- They provide usable frameworks for posing our questions and allowing for most methods of inference

Probability distributions

- For example, even if we do not know the actual distribution, it is clear frequency data is generally going to be better fit by a binomial distribution than a normal distribution
- Why?
  - Binomial is bounded by zero & 1
  - Other distributions (e.g. gamma, Poisson) have lower boundaries

Probability distributions

- The multitude of distributions allows us to choose those that match our data or theoretical expectations in terms of shape, location & scale
- Avoid dangers of under- & overfitting
  - Overfitting can be a problem because models can be too specific & not broadly applicable to other datasets

Probability distributions

- There are 3 types of parameters to specify probability distributions
  - Shape, scale, location
- Some distributions have only 2 (normal) or 1 (Poisson) parameter(s)

Parameters for normal probability distribution

- Location = mean for a normal
- Scale = standard deviation
- Shape = no shape (symmetrical)
- These parameters are NOT the same for all distributions
- **Keep in mind**: it is not the distribution of the whole dataset that will necessarily determine what distribution to use; distribution of the residual variation (once all other parameters are accounted for) is often of interest

Two types of random variables

- A **discrete random variable** can assume a countable number of values
  - Number of steps to the top of the Eiffel Tower
- A **continuous random variable** can assume any value along a given interval of a number line
  - Amount of time a tourist stays at the top of the Eiffel Tower

Probability distributions for discrete random variables

- The **probability distribution** of a discrete random variable is a graph, table or formula that specifies the probability associated with each possible outcome the random variable can assume
  - $p(x) \geq 0$ for all values of x
  - $\Sigma p(x) = 1$

Probability distributions for discrete random variables

- Say a random variable *x* follows this pattern:

$p(x) = (.3)(.7)^{x-1}$

for *x* > zero

- This table gives the probabilities (rounded to two digits) for *x* between 1 and 10

x	*P(x)*
1	.30
2	.21
3	.15
4	.11
5	.07
6	.05
7	.04
8	.02
9	.02
10	.01

Expected values of discrete random variables

- The **mean**, or **expected value**, of a discrete random variable is

$$\mu = E(x) = \sum xp(x).$$

- The **variance** of a discrete random variable is

$$\sigma^2 = E[(x-\mu)^2] = \sum (x-\mu)^2 p(x).$$

- The **standard deviation** of a discrete random variable is

$$\sqrt{\sigma^2} = \sqrt{E[(x-\mu)^2]} = \sqrt{\sum (x-\mu)^2 p(x)}.$$

Expected values of discrete random variables

- In a roulette wheel in a U.S. casino, a $1 bet on "even" wins $1 if the ball falls on an even number
- The odds of winning this bet are 47.37%

$$P(win\,\$1) = .4737$$

$$P(lose\,\$1) = .5263$$

$$\mu = +\$1 \cdot .4737 - \$1 \cdot .5263 = -.0526$$

$$\sigma = .9986$$

???

On average, bettors lose about a nickel for each dollar they put down on a bet like this. (These are the *best* bets for patrons.)

Binomial distribution

- A Binomial random variable
  - n identical trials
  - Two outcomes: **S**uccess or **F**ailure
  - $P(\mathbf{S}) = p$; $P(\mathbf{F}) = q = 1 - p$
  - Trials are independent
  - x is the number of **S**uccesses in n trials

Binomial distribution

- A Binomial random variable
  - n identical trials
  - Two outcomes: **S**uccess or **F**ailure
  - P(**S**) = p; P(**F**) = $q = 1 - p$
  - Trials are independent
  - x is the number of **S**'s in n trials

Flip a coin 3 times

Outcomes are Heads or Tails

P(H) = .5; P(F) = 1-.5 = .5

A head on flip i doesn't change P(H) on flip $i + 1$

Binomial distribution

Results of 3 flips	Probability	Combined	Summary
HHH	$(p)(p)(p)$	p^3	$(1)p^3q^0$
HHT	$(p)(p)(q)$	p^2q	
HTH	$(p)(q)(p)$	p^2q	$(3)p^2q^1$
THH	$(q)(p)(p)$	p^2q	
HTT	$(p)(q)(q)$	pq^2	
THT	$(q)(p)(q)$	pq^2	$(3)p^1q^2$
TTH	$(q)(q)(p)$	pq^2	
TTT	$(q)(q)(q)$	q^3	$(1)p^0q^3$

Binomial distribution

- Binomial distribution specification
  - p = P(**S**) on a single trial
  - $q = 1 - p$
  - n = number of trials
  - x = number of successes

$$P(x) = \binom{n}{x} p^x q^{n-x}$$

Binomial distribution

The number of ways of getting the desired results

The probability of getting the required number of successes

$$P(x)=\binom{n}{x}p^{x}q^{n-x}$$

The probability of getting the required number of failures

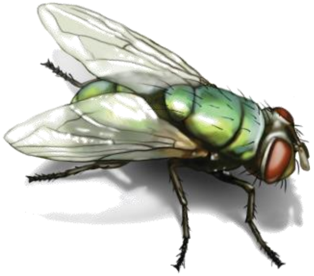
Binomial distribution

- A binomial random variable has

Mean	$\mu = np$
Variance	$\sigma^2 = npq$
Standard Deviation	$\sigma = \sqrt{npq}$

Binomial distribution: example

- You set up a series of enclosures. Within each, you place 25 flies & a pre-determined set of predators
- You want to know what the distribution (across enclosures) of flies getting predated is, according to a pre-determined probability of success for a given predator species



Binomial distribution: example

- Set this up as a binomial problem
- **N** = 25 (total # of individuals in each enclosure; or # of ‘trials’)
- **P** = probability of successful predation ‘trial’ (i.e. the coin toss)
- **X** = # of trials of successful predation

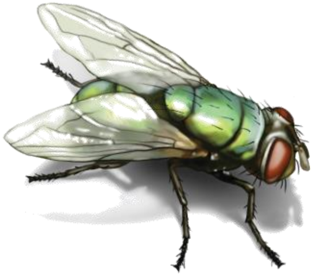
Binomial distribution: example

Binomial probability distribution $$P(x) = \binom{n}{x} p^x q^{n-x}$$

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

- You can think of 'n choose x' in 2 ways:

1. A normalizing constant so that probabilities sum to 1.
2. Number of different combinations to allow for x 'successful' predation events out of N total



Binomial distribution: example

- If predator A had a per 'trial' probability of successfully eating a prey item of 0.2, **what would be the probability of exactly 10 flies (out of the 25) being consumed in a single enclosure?**
- $P(x=10|bi(N=25,p=0.2))=$ **???**

Binomial distribution: example

$$P(x) = \binom{n}{x} p^x q^{n-x}$$

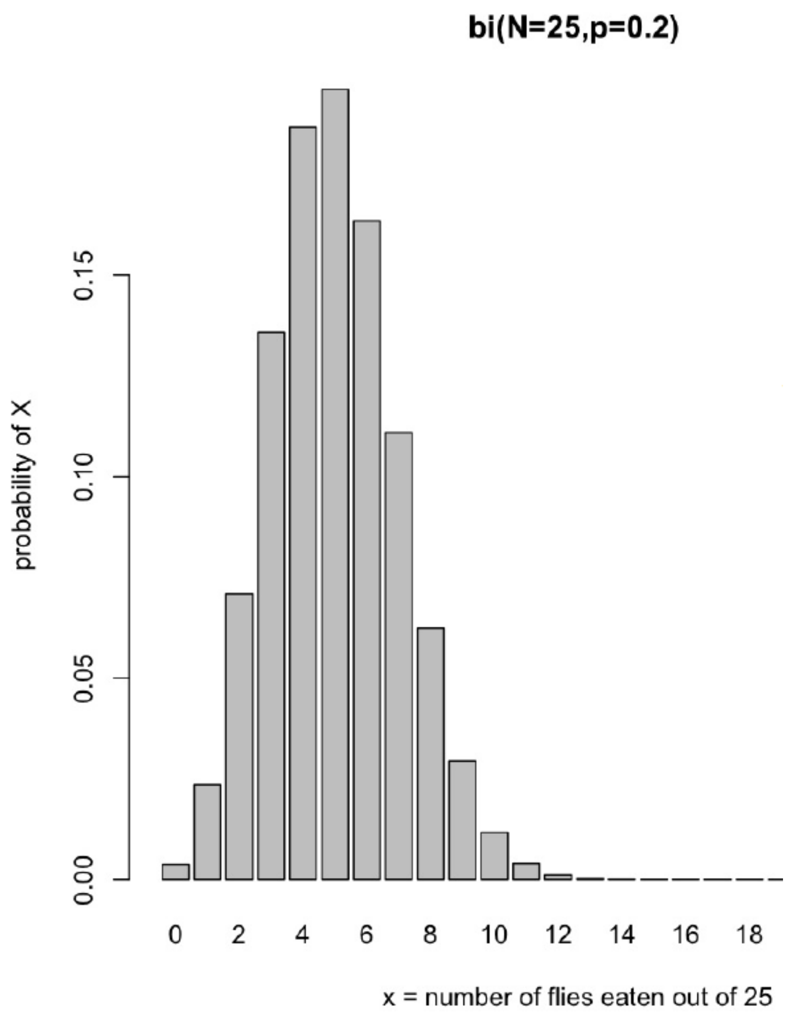
$$= \binom{25}{10} (.2^{10})(.8^{25-10})$$

$$= 3268760(.000000102)(.0352)$$

$$= .0117$$

Not a very high probability of eating 10 flies!

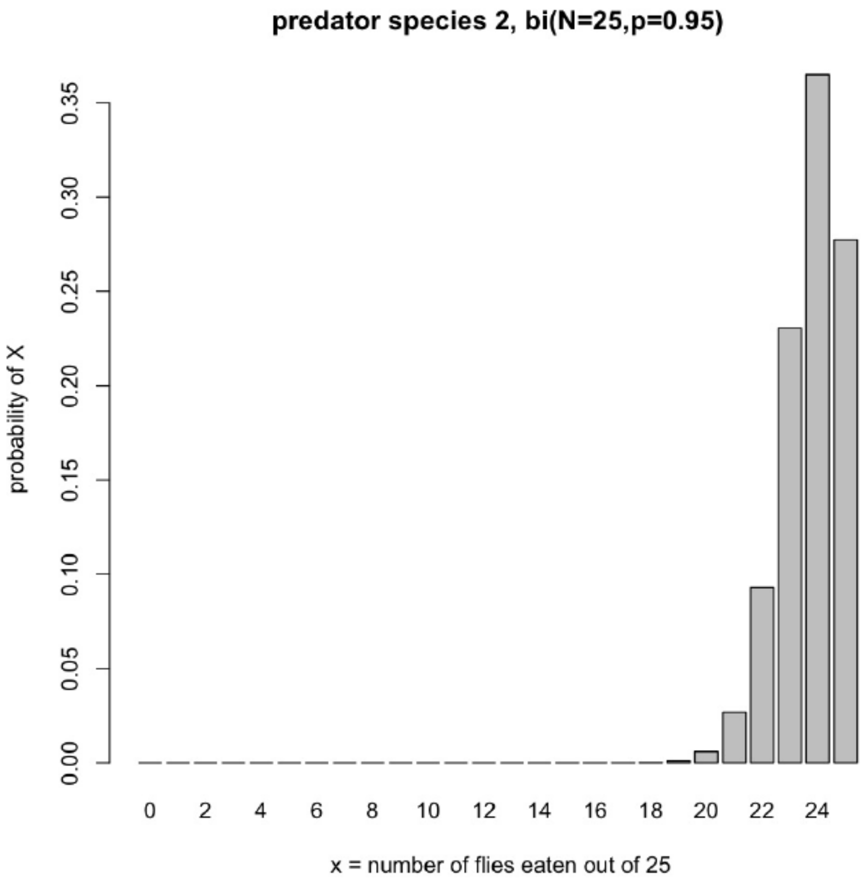
Binomial distribution: example



This is the expected distribution if we set up many replicate enclosures with 25 flies and this predator

Binomial distribution: example

- What about a ‘hungrier’ predator whose probability of successfully eating a prey item = 0.95?



Binomial distribution

- So if you're modeling using the binomial distribution, which parameter are you estimating?
  - N, the number of trials?
  - P, the proportion of successes?
  - X, the number of successful trials?
- You're modeling X!

Binomial distribution

- Other examples:
- Number of surviving individuals out of an initial sample
- Number of infested/affected animals in a sample
- Number of a particular haplotype in a larger population

Binomial distribution exercises

- Let’s go to RStudio