# NRES_798_3_201501

Review of probability and distributions

# Outline

- Discrete random variables
  - Bernoulli
  - Binomial
  - Poisson
  - Negative Binomial
- Continuous random variables
  - Uniform distribution
  - Normal distribution
- Central Limit Theorem
- Other Continuous random variables
  - Log-normal distribution
  - Exponential

# Bernoulli Random variable

- Event with only 2 outcomes
- Pitcher plant
  - Event: visit
  - Set: capture, escape
  - Assumes visits independent
- Habitat suitability
  - Random sample of quadrats with species being present or absent.

***Nepenthes sp.***

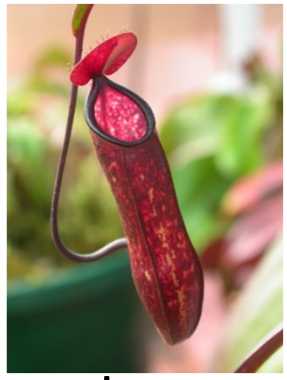
# Bernoulli Random variable

- Bernoulli distribution
  - Probability of success P(X=1)= p
  - Probability of failure P(X=0= 1- p (first axiom)
  - X has a Bernoulli distribution with parameter P

$$X \sim Bernoulli(p)$$

- **Single event** or observation distributed as a Bernoulli random variable

# Binomial distribution



- Pitcher plant
  - Observe 1000 events (visits)
  - Events are independent and identically distributed random variables each with parameter p
  - n = (0,1,0,0,1,0…)
  - Number of captures
  - X = 364 = count of number of successes from n trails
  - Random variable X is a binomial distribution

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

  - Read as: X number of successes in n Bernoulli trials based on p

$$\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

# Binomial probability mass function

$$\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- Probability mass function: the probability that a discreet random variable is exactly equal to some value
- Binomial coefficient: “n choose k”

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

- How many ways can k success be obtained from n trails

# Binomial probability mass function

What is the probability of observing 2 successes in 5 trials if the Bernoulli p = 0.5

Probability of 2 independent successes: 0.5^2 = 0.25

$$\binom{n}{k} p^k (1-p)^{n-k} = \Pr(X = k) = 0.3125$$

Probability of 3 independent failures: $(1\text{-}0.5)^{(5\text{-}2)} = 0.125$

10 = Number of ways 2 successes can result from 5 trails

(1 1 0 0 0) (0 1 1 0 0) (0 0 1 1 0) (0 0 0 1 1)

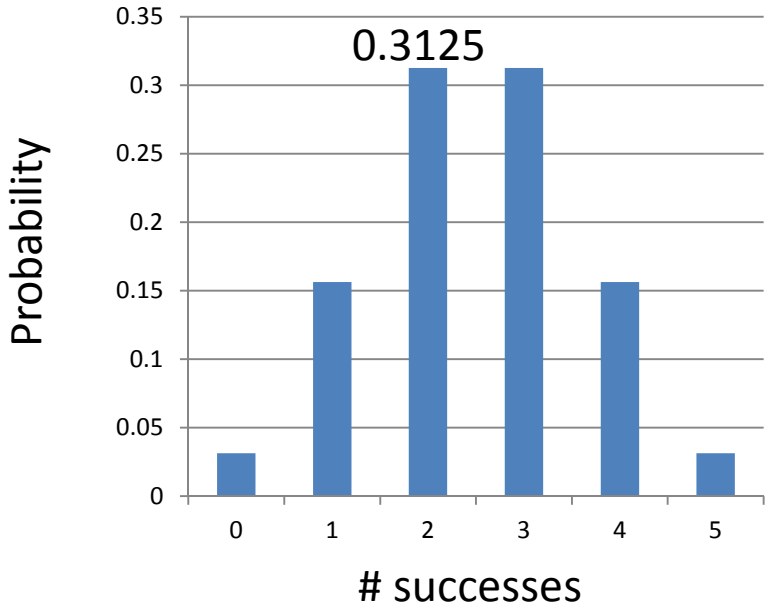
(1 0 1 0 0) (0 1 0 1 0) (0 0 1 0 1)

(1 0 0 1 0) (0 1 0 0 1)

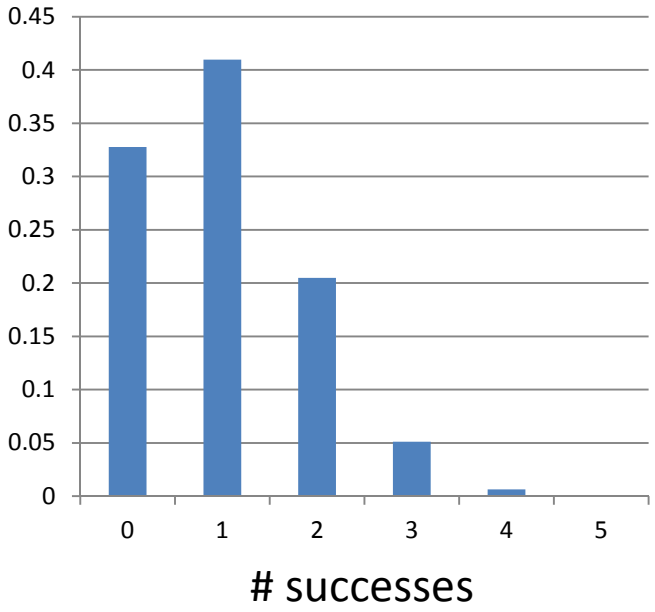
(1 0 0 0 1)

# Binomial probability mass function

10 trails
P = 0.5




10 trails
P = 0.2




$$\sum_{i=1}^{n} P(A_i) = 1.0$$

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

- Exact probabilities can be easily calculated
- When p = 0.5, probability distribution symmetric
- Shape of distribution depends on N and P

# Poisson distribution

- Poisson: the number of individuals, arrivals, events, counts, etc., in a given time/space unit of counting effort.
  - Number of seeds/seedling falling in a gap
  - Number of offspring produced in a season (if the number of breeding attempts is not recorded)
  - Number of prey caught per unit time

- Often used when number of counts is small

# Poisson distribution

$$X \sim Poisson(\lambda)$$

- Distribution described by single parameter lamda (λ)
- Lamda is the average number of occurrences in each sample

# Poisson Probability mass function

$$P(x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

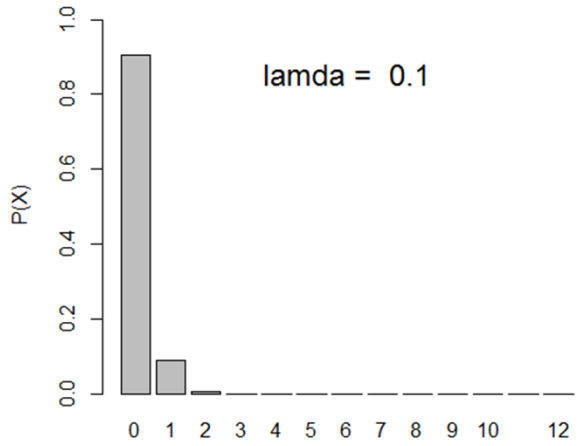
- What is the probability of observing 3 birds in a 625 $m^2$ patch if the average number of birds is hypothesized to be 2
- X = 3, $\lambda = 2$

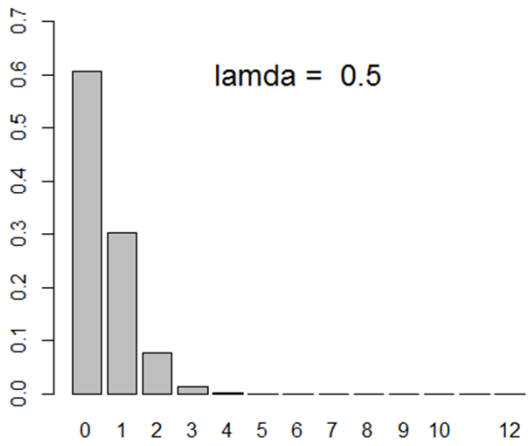
$$P(x) = \frac{\lambda^x}{x!} e^{-\lambda} = \frac{2^3}{3!} e^{-2} = \frac{8}{36} e^{-2} = 0.18$$

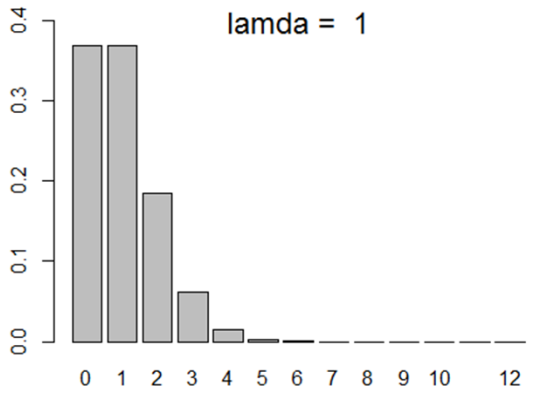
# Poisson

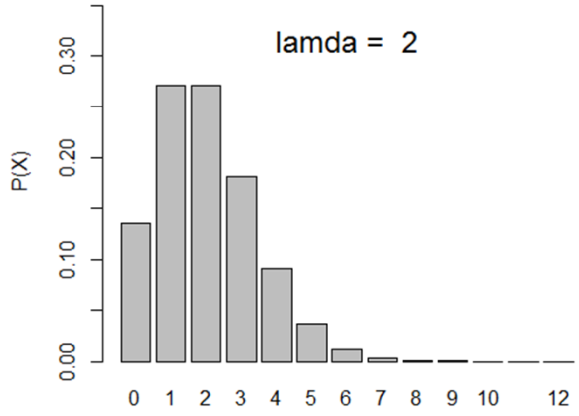
- When $\lambda$ is small ($< 1$) the distribution has a strong "reverse-j" shape
- When the expected number of counts gets large ($\lambda > 10$) the Poisson becomes approximately normal
- $\lambda$ sometimes referred to as a rate parameter because it can describe the frequency of rate events in time
- Poisson has no upper limit (0, unlimited)
- Variance of the Poisson is equal to its mean

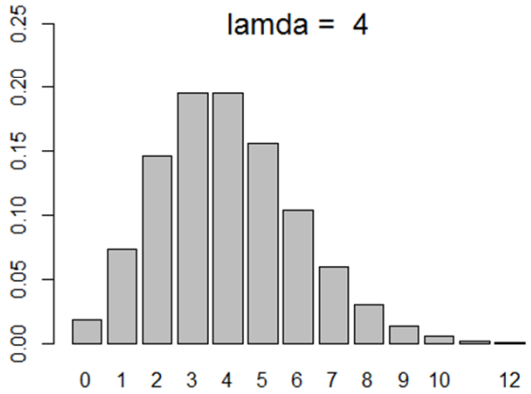
# Poisson distribution with different λ




```
# Poisson distribution
x <- c(0:12)
lamda <- 8 # 0.1, 0.5, 1,2,3,8
p <- dpois(x,lamda)
barplot(p,axes = TRUE,
        names.arg = x,
        ylim=c(0,max(p)+0.1),
        ylab = "P(X)"
        )
mtext(paste("lamda = ",lamda),side=3,
      outer=FALSE,line=-3,cex=1.5)
```

# Negative binomial distribution

- The negative binomial counts the number of failures before a predetermined number of success occurs
  - Remember the binomial is number of successes in a fixed number of trials
  - Discrete, similar to the Poisson, but variance can be larger than its mean (can be over dispersed, which can valuable for ecological data)

- In ecology it is sometimes used because it is a good phenomenological description of a clustered distribution with no upper limit, and more variance than the Poisson.

# Negative binomial distribution

$$X \sim NB(r; p)$$

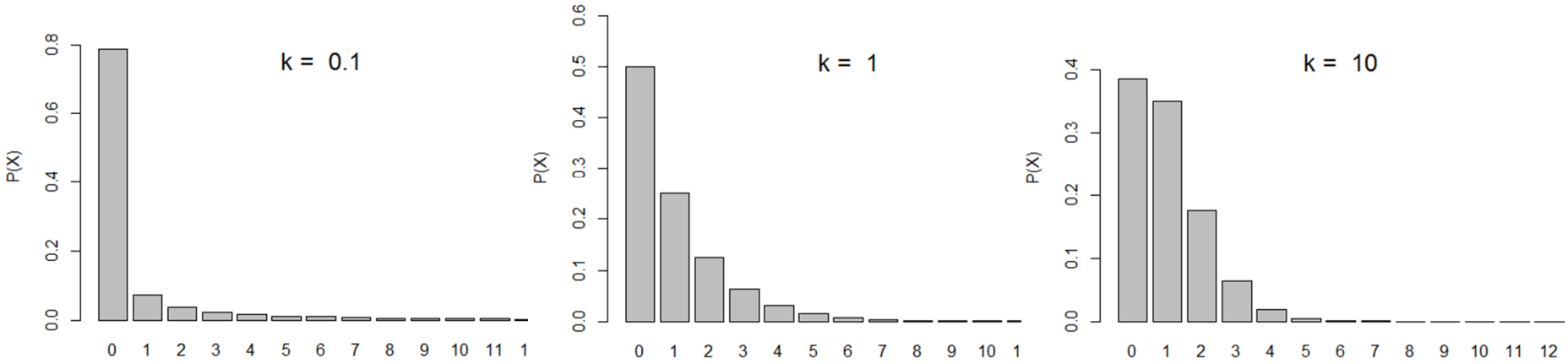
- p is the probability of success per trail (Bernoulli random variable)
- r is the predefined number of successes that need to be observed

# Negative binomial probability mass function

$$\Pr(X = k) = \left( \frac{(k + r\ - 1)!}{k!\,(r - 1)!} \right) p^k (1 - p)^r$$

- For k > 10 the NB resembles the Poisson
- k often < 1 when used in ecological applications

# Examples of negative binomial distributions




```
# Negative binomial distribution
X <- c(0:12)
k <- 10
#nbp <- dnbinom(x,size=1,prob = 0.2)
nbp <- dnbinom(x,size=k,mu=1)
barplot(nbp,axes = TRUE,
        names.arg = x,
        ylim=c(0,max(nbp)+0.1),
        ylab = "P(X)"
        )
mtext(paste("k = ",k),side=3,
      outer=FALSE,line=-3,cex=1.5)
```

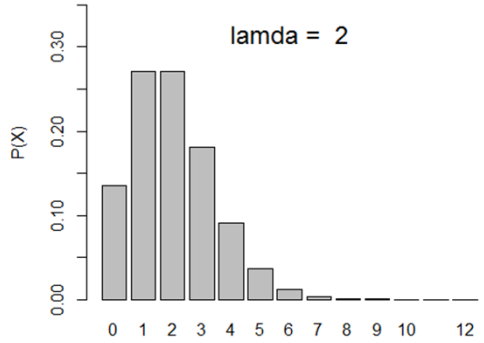
# Discrete vs. continuous random variables

- Discrete random variables
  - Presence vs. absence
  - Count data, integers, e.g. 1,4,7
  - E.g. # offspring, # prey captured, # species

- Continuous random variables
  - Can have values within an interval
  - Real numbers, e.g. 1.74, 14.9
  - E.g. spine length, N concentration in soil, body mass, pesticide concentration in fish tissue

# Discrete random variables

- Exact probability mass function can be calculated for each count expectation
- E.g. probability of observing a count of 2 = 0.72




- With continuous random variables we can not identify all the possible events or outcomes
- For continuous random variables a specific probability can not be directly calculated for each measured value
- E.g. probability of  observing a body mass of 67.34 kg
- Use probability density functions
- Constructed by getting the probability that a measurement occurs within a sub interval (e.g. p(67.3 < x <67.4))

# Calculating probability distributions of continuous variables

Probability of observing a wing length of 14.86cm?

0 cm 20 cm

- Assume max wing length = 20cm
- Assume any wing length between 0 and 20cm is equally likely to occur.

10 cm

$0 < X \leq 10$ $10 < X \leq 20$

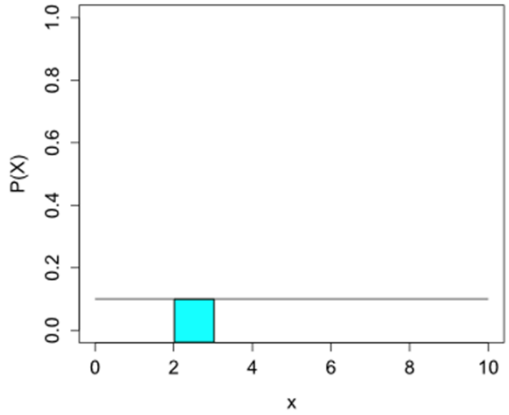
P = 0.5 P = 0.5

Ho: Wing length x is between
0 and 10cm

Discrete intervals have defined probabilities

# Uniform random variable

- Assume a closed unit interval (real number bounded by: $0 \leq x \leq 10$)
- This interval can be divided into sub-intervals
- Within a closed interval the probability of a value occurring within a subinterval can be defined (interval = 0:1)
- In a continuous sample space, all the probabilities of events must still sum to 1 (first axiom)




```
# Uniform distribution PDF
x <- seq(0,10,0.1)
updf <- dunif(x,min=0,max=10)
plot(x,updf,ylim=c(0,1),
    type="l",xlim=c(0,10),
    ylab="P(X)")
```

- Probability density function (PDF)
- Probability of x occurring in interval I is given by the area under the curve
- Total area under the curved described by a PDF = 1

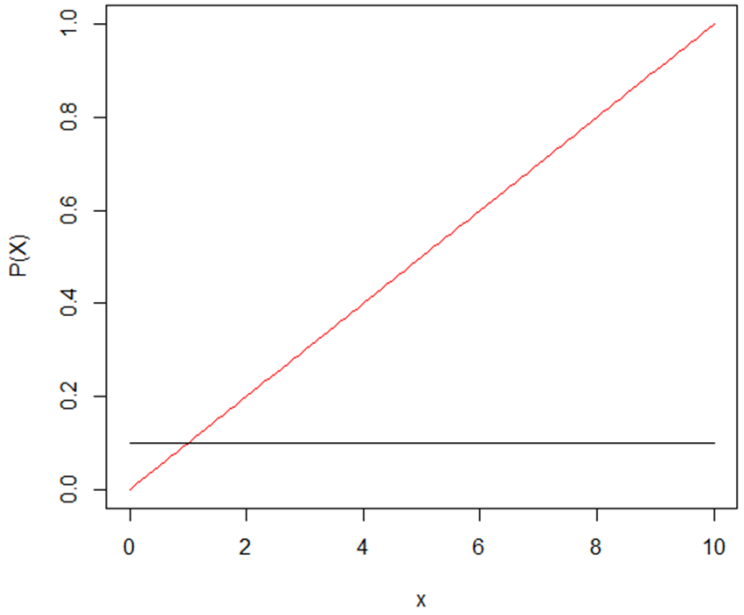
$$f(x) = \begin{cases} 1/10, 0 \leq x \leq 10 \\ 0 \; otherwise \end{cases}$$

- What are ecological examples of a uniform random variable?

# Uniform variable cumulative distribution function

- Probability that a random variable X is less than or equal to a value y.

$$F(y) = P(X < y)$$

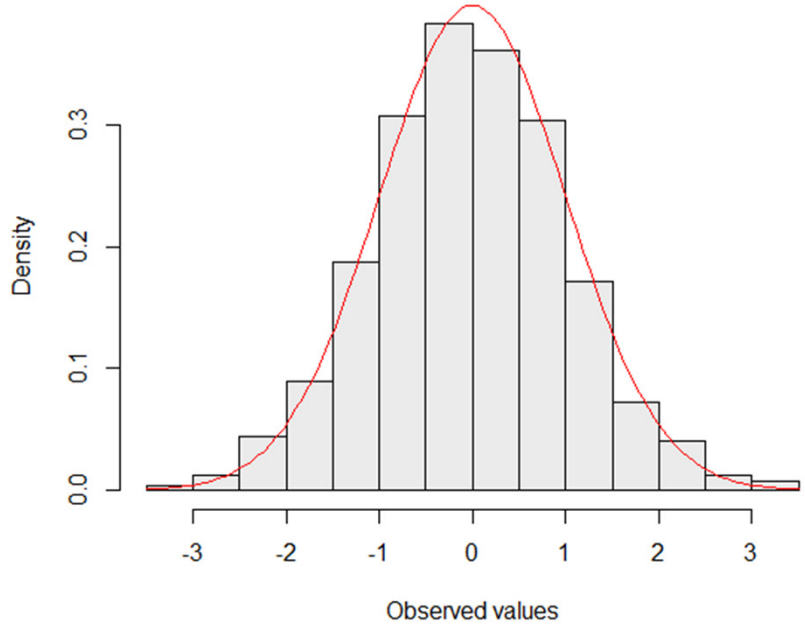



```
# Uniform distribution CDF
x <- seq(0,10,0.1)
ucdf <- punif(x,0,10)
plot(x,ucdf,ylim=c(0,1),
    type="l",xlim=c(0,10),
    ylab="P(X)",col="red")
points(x,updf,type="l",
     col="black")
```

- CDF is the area under the PDF for x < y

# Normal random variables (Gaussian)

- Observations clustered around a central value
- Long tails (infinite in the PDF)
- Distribution is approximately symmetrical




```
# Normal distribution
rN <- rnorm(1000,mean=0,sd =1)
hist(rN,main = "",
    xlab="Observed values",
    col="grey92",prob=TRUE)
curve(dnorm(x,mean=0,sd=1),
    add=TRUE,col="red")
```

- Assumptions of normal distributions are the basis of many statistical tests
  - Regression, ANOVA
- Ecological examples of normal distributions?

# Normal variable
# Probability density function

$$X \sim N(\mu, \sigma)$$

Normal PDF defined by:

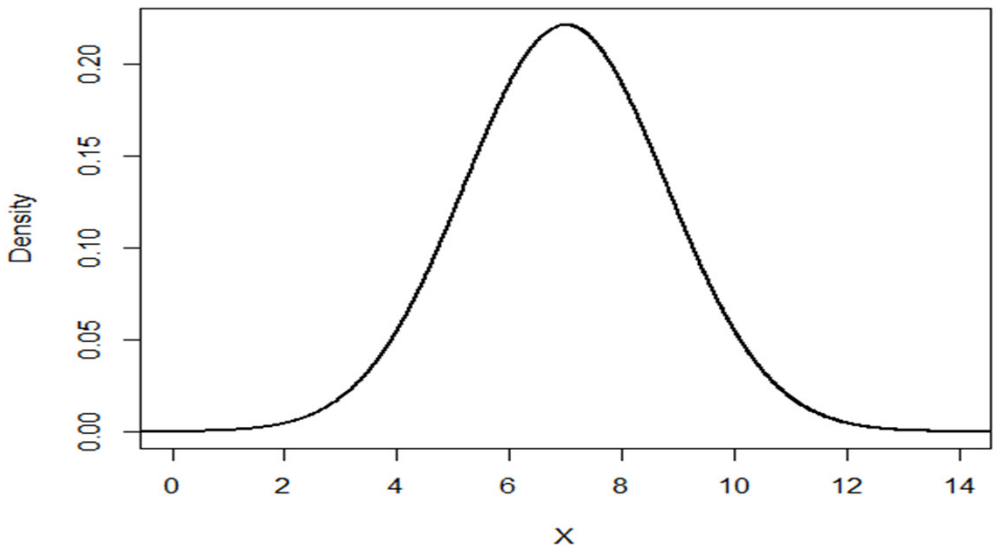
μ = mean

σ = standard deviation  ($\sigma^2$ = variance)

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

μ = 7

σ = 1.8

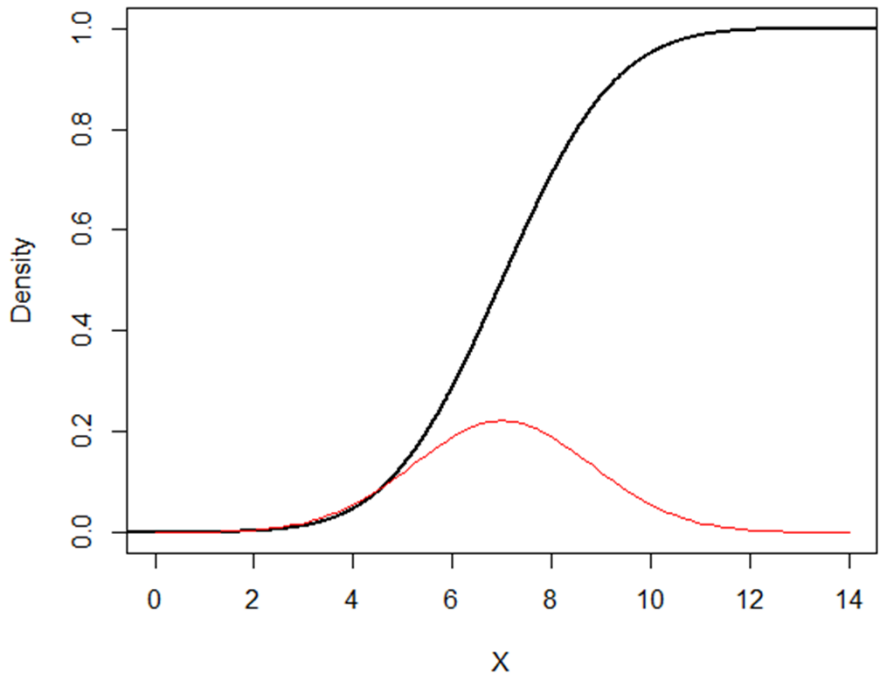



```
# Normal distribution pdf
x <- seq(-2,15,0.01)
snd <- dnorm(x,mean=7,sd=1.8)
plot(x,snd,xlim=c(0,14),
    type="l",lwd=2,
    xlab="X",
    ylab="Density")
```

# Normal variable cumulative probability distribution

$$F(X) = \int_{-\infty}^{X} f(x)dx$$

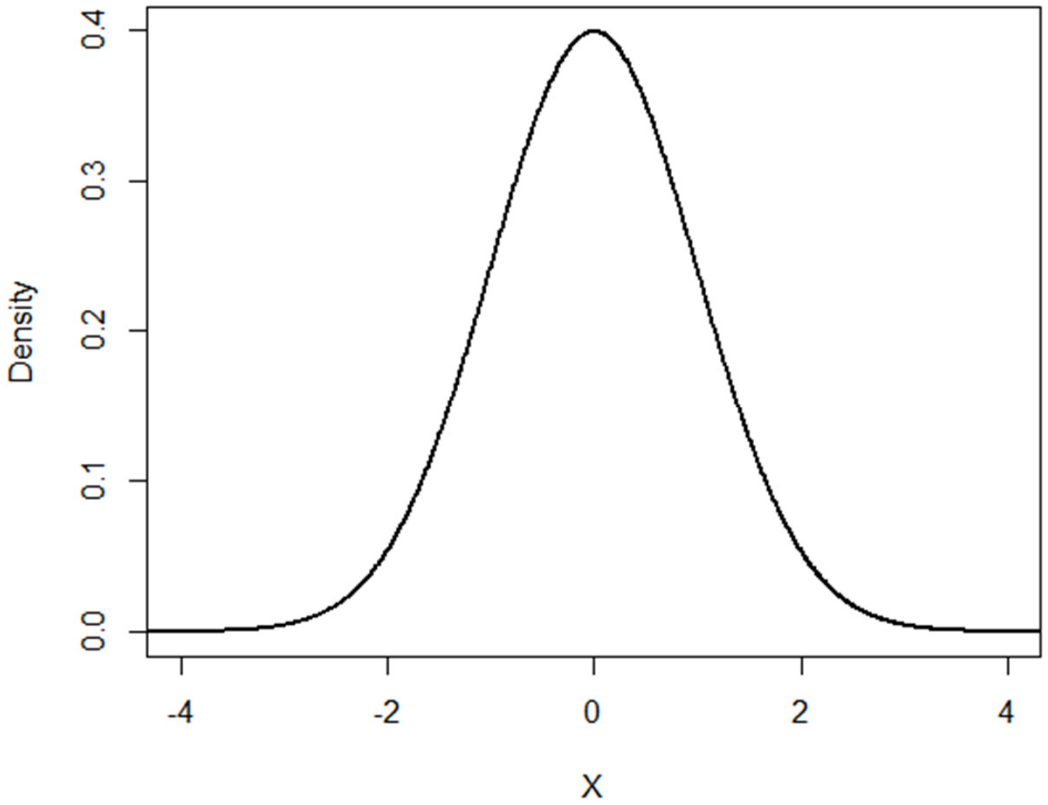
- Integral of PDF (no analytical solution)




```
# Normal distribution cdf
x <- seq(-2,15,0.01)
snd  <- dnorm(x,mean=7,sd=1.8)
scnd <- pnorm(x,mean=7,sd=1.8)
plot(x,scnd,xlim=c(0,14),
    type="l",lwd=2,
    xlab="X",
    ylab="Density")
points(x,snd,type="l",col="red")
```

# Normal distribution examples

- Standard normal distribution 0,1




```
# Standard Normal distribution pdf
x <- seq(-5,5,0.01)
snd <- dnorm(x,mean=0,sd=1)
plot(x,snd,xlim=c(-4,4),
    type="l",lwd=2,
    xlab="X",
    ylab="Density")
```

# Properties of normal distributions

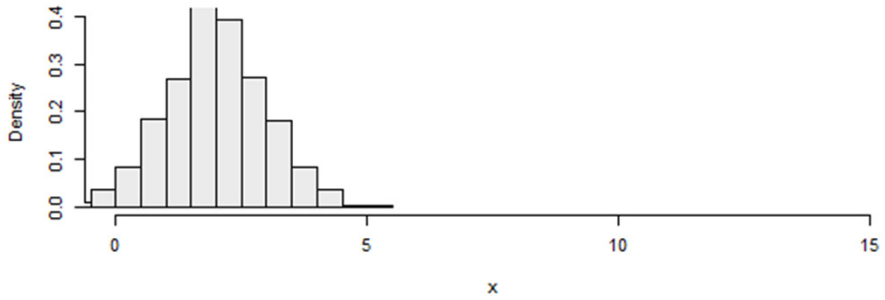
1. Normal distributions can be added and the result is a normal distribution
2. Normal distributions can be transformed with shift and change of scale operations
   - and a normal distribution is retained
3. Any normal distribution can be transformed into the standard normal distribution through shift and change of scale operations

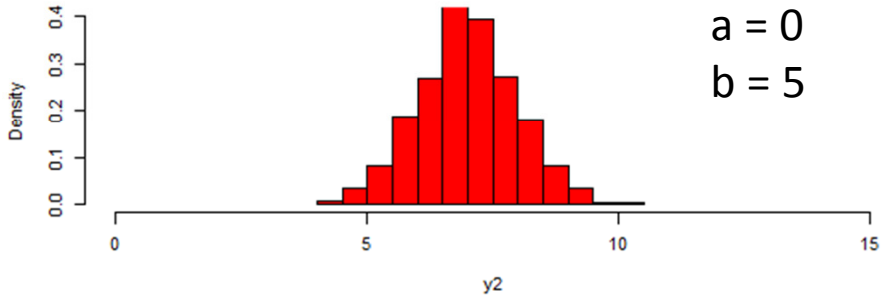
# Normal distribution transformations

$$X \sim N(\mu, \sigma)$$

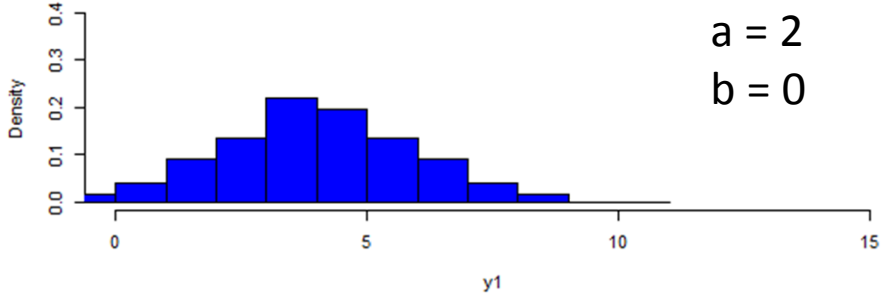
Y = aX + b

- Shift: a = 1, b != 0
  - Move random variable over b units
- Scale: a != 1, b = 0
  - One unit of X becomes a units of Y




a = 0
b = 5




a = 2
b = 0

```
# Transforme Normal distributions pdf
par(mfrow=c(3,1))
x <- rnorm(1000,2,1)
hist(x,xlim=c(0,15),ylim=c(0,0.4),
    prob=TRUE,col="gray92",
    main="")
a = 2; b = 5
y1 <- a*x + 0  # scale
y2 <- 1*x + b  # shift
hist(y2,xlim=c(0,15),ylim=c(0,0.4),
    prob=TRUE,col="red",
    main="")
hist(y1,xlim=c(0,15),ylim=c(0,0.4),
    prob=TRUE,col="blue",
    main="")
```

# Central limit theorem

- Ubiquity of Normal Distribution due to the Central Limit Theorem

- Most classical statistics are premised on a normal distribution due to the central limit theorem (ANOVA, regression, etc.)

# Central Limit Theorem

- Central Limit Theorem says that if you **add** a “large” number of independent samples from the same distribution (binomial, Poisson, gamma etc.) , the distribution of the sums will be approximately normal
  - Standardizing the resulting distribution will produce a new random variable that is close to one that has a standard normal distribution

- “Large” varies between distributions and conditions but can be reasonably small (>5)
- The central limit theorem does not mean that “all samples with large numbers are normal”.

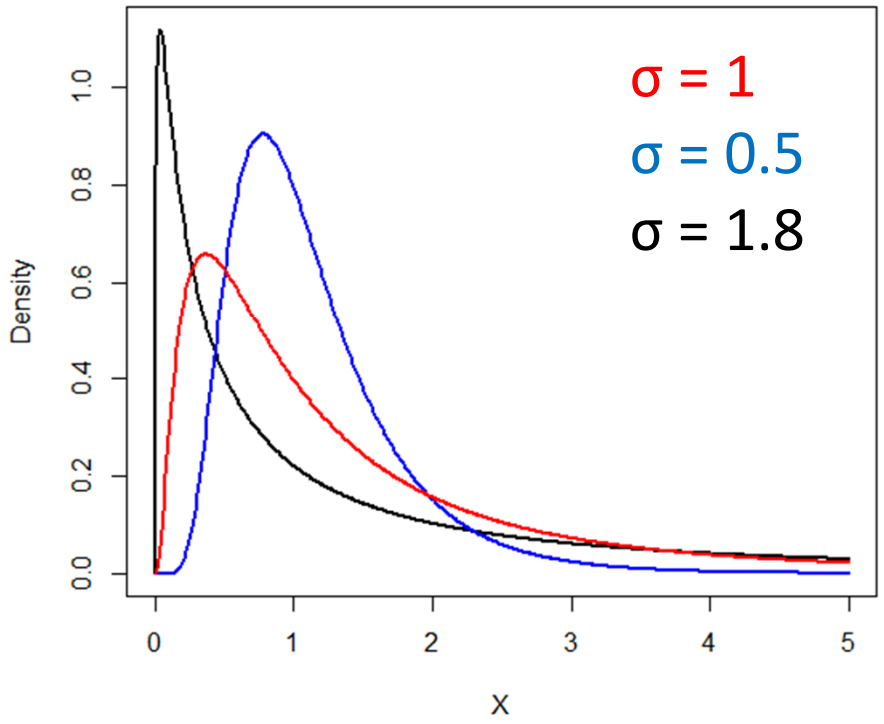
# Central Limit Theorem

Allows us to use statistics that are premised on a normal distribution even though the underlying (root) random variables may themselves not be normally distributed!

- # prey caught by individual represents a Poisson random distribution
- Sample the # of prey caught by multiple individuals in two populations
- Distribution of observed # prey caught in the two populations will approach a normal distribution

# Log-normal distribution

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}$$

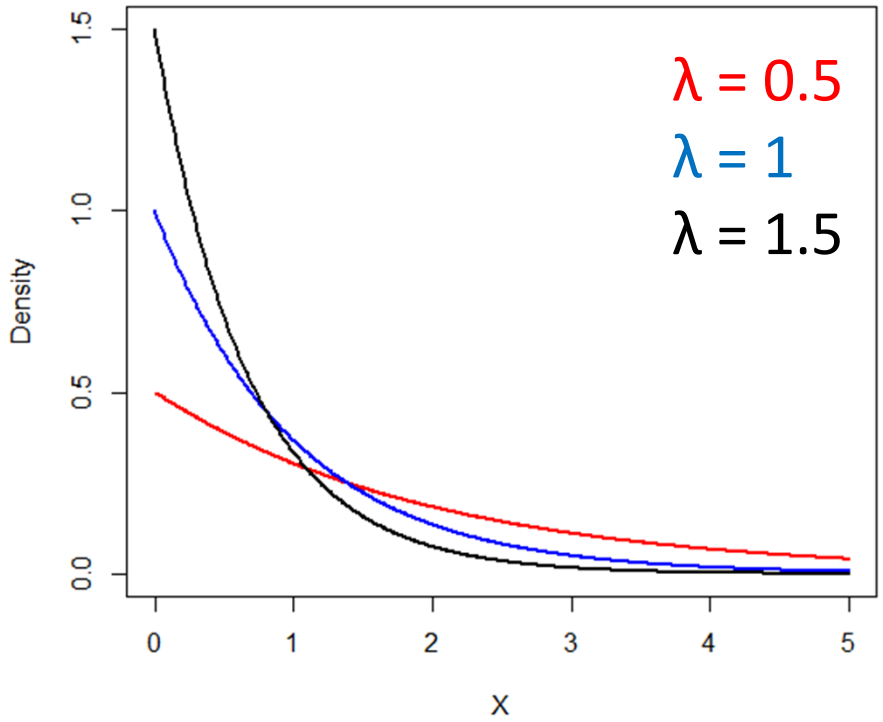



```
# Log normal distribution
x <- seq(0,5,0.01)
ln1 <- dlnorm(x,meanlog =0, sdlog =1)
ln2 <- dlnorm(x,meanlog =0, sdlog =0.5)
ln3 <- dlnorm(x,meanlog =0, sdlog =1.8)
plot(x,ln3,col="black",type="l",
    xlab="X",ylab="Density",lwd=2)
points(x,ln2,col="blue",type="l",lwd=2)
points(x,ln1,col="red",type="l",lwd=2)
```

- Why is Log-normal often observed in biological systems?

# Exponential distribution (negative exponential distribution)

- Describes time between events in a Poisson process




```
# Exponential distribution
x <- seq(0,5,0.01)
e1 <- dexp(x, rate = 0.5)
e2 <- dexp(x, rate = 1)
e3 <- dexp(x, rate = 1.5)
plot(x,e1,col="red",type="l",
    ylim=c(0,1.5),
    xlab="X",ylab="Density",lwd=2)
points(x,e2,col="blue",type="l",lwd=2)
points(x,e3,col="black",type="l",lwd=2)
```

# Characterizing distributions

- Location and spread
  - Location: Mean, median, mode
  - Spread: variance, standard deviation,
- Distribution characteristics
  - Central moments
    - Arithmetic mean (first moment)
    - Variance (second moment)
    - Skewness (third moment)
    - Kurtosis (fourth moment)