NRES_798_4_201501

Probability distributions
and descriptions of data

Normal random variable
Probability density function

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

Normal PDF defined by:

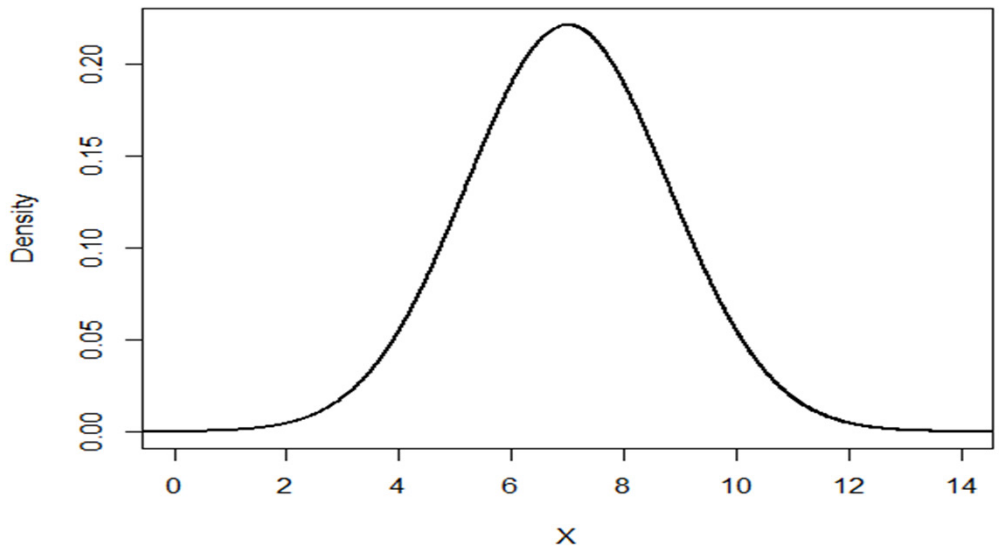
μ = mean

σ = standard deviation (σ^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

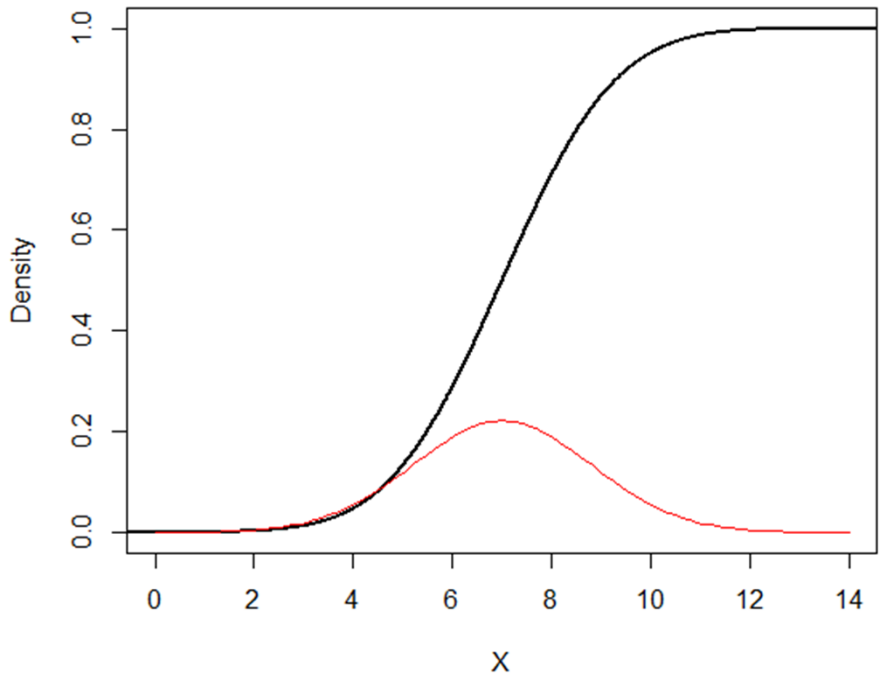



```
# Normal distribution pdf
x <- seq(-2,15,0.01)
snd <- dnorm(x,mean=7,sd=1.8)
plot(x,snd,xlim=c(0,14),
    type="l",lwd=2,
    xlab="X",
    ylab="Density")
```

Normal random variable cumulative probability distribution

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

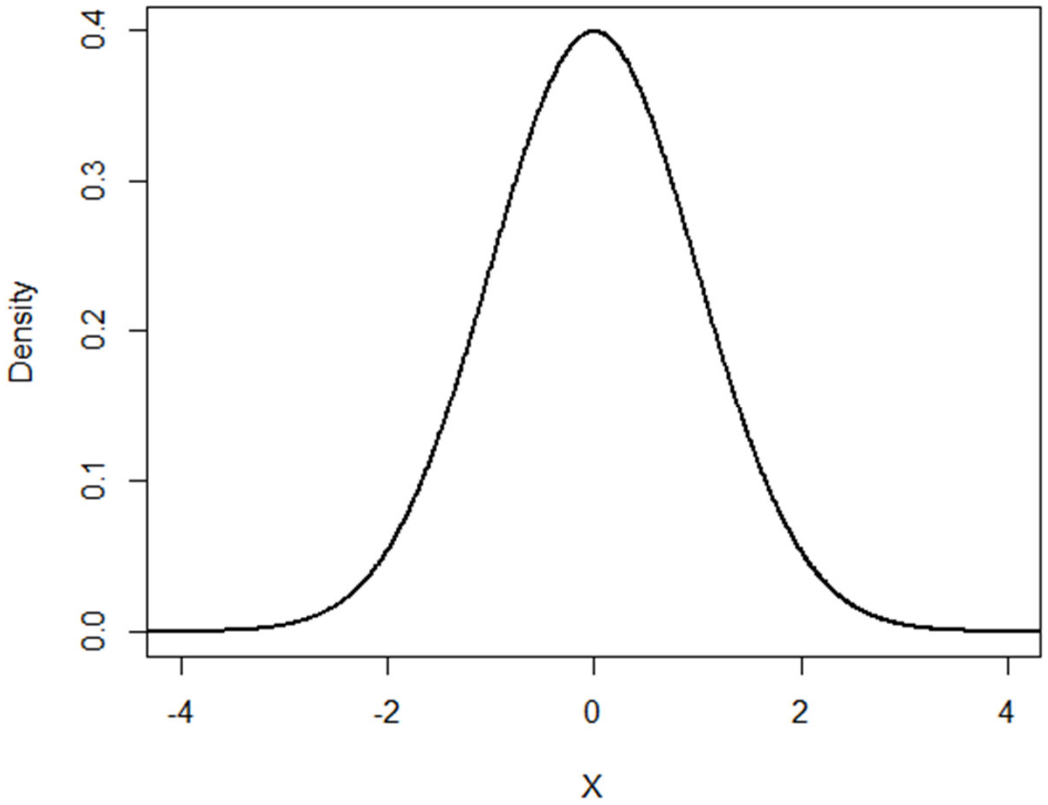
- Integral of PDF (no analytical solution)




```
# 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

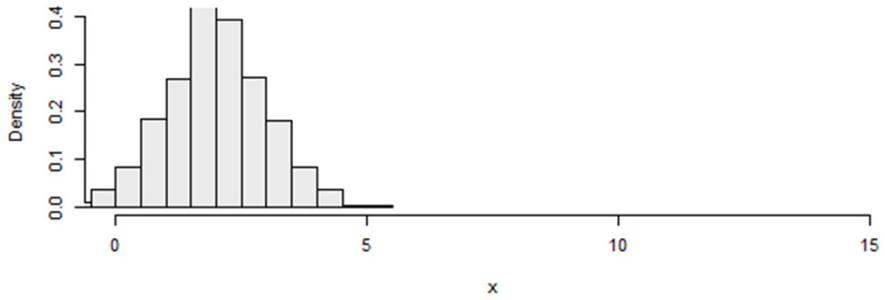
1. Normal distributions can be added and the result is a normal distribution
2. Normal distributions can be transformed with shift and change of scale operations
   - and a normal distribution is retained
3. Any normal distribution can be transformed into the standard normal distribution through shift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

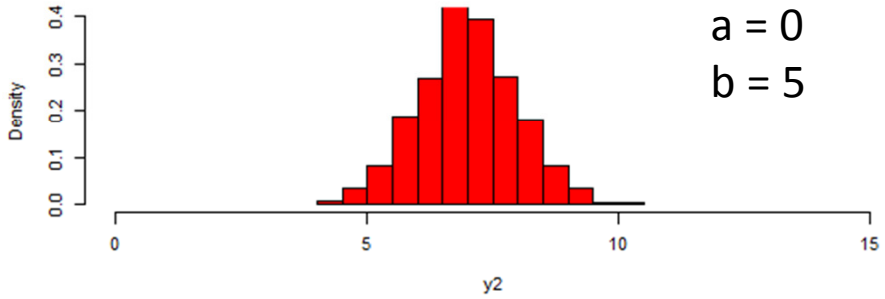
Normal distribution transformations

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

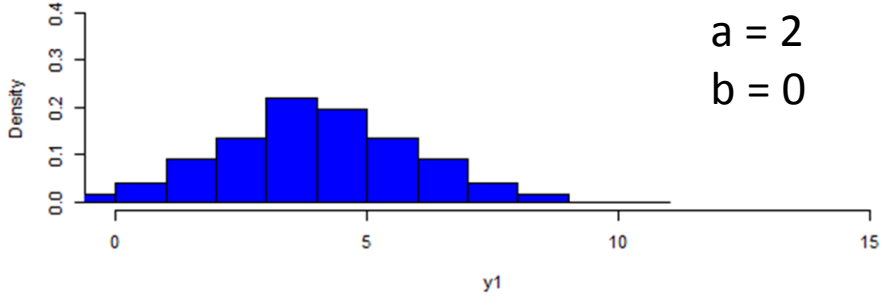
Y = aX + b

- 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”.

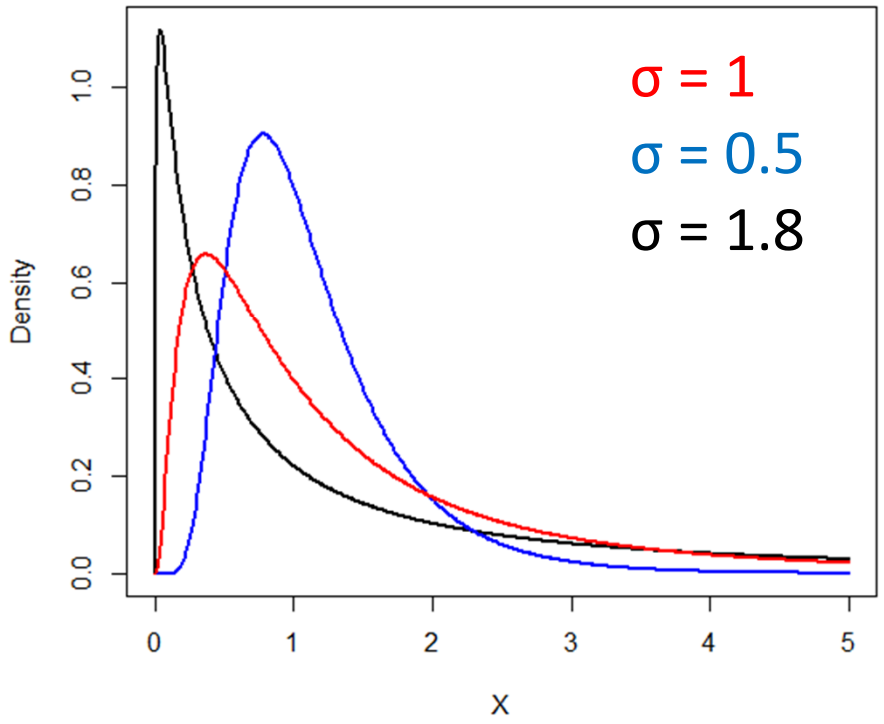
Central Limit Theorem

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

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

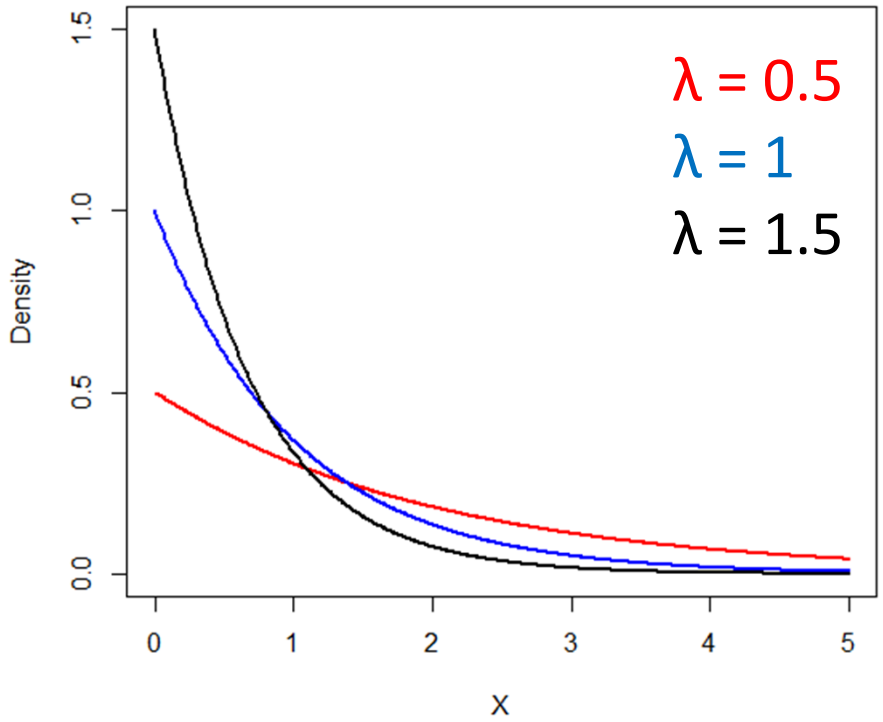



```
# Log normal distribution
x <- seq(0,5,0.01)
ln1 <- dlnorm(x,meanlog =0, sdlog =1)
ln2 <- dlnorm(x,meanlog =0, sdlog =0.5)
ln3 <- dlnorm(x,meanlog =0, sdlog =1.8)
plot(x,ln3,col="black",type="l",
    xlab="X",ylab="Density",lwd=2)
points(x,ln2,col="blue",type="l",lwd=2)
points(x,ln1,col="red",type="l",lwd=2)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

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

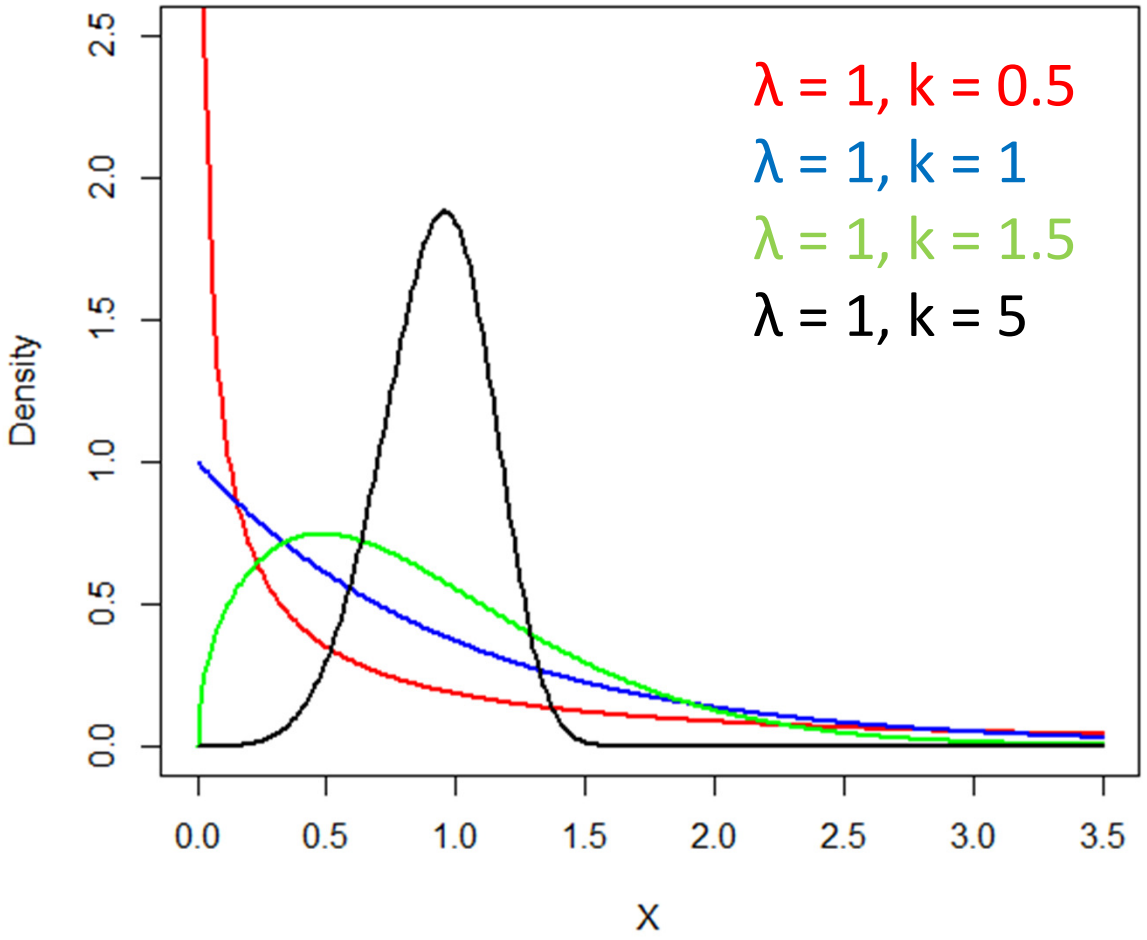
Exponential distribution (negative exponential distribution)

- 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)
```

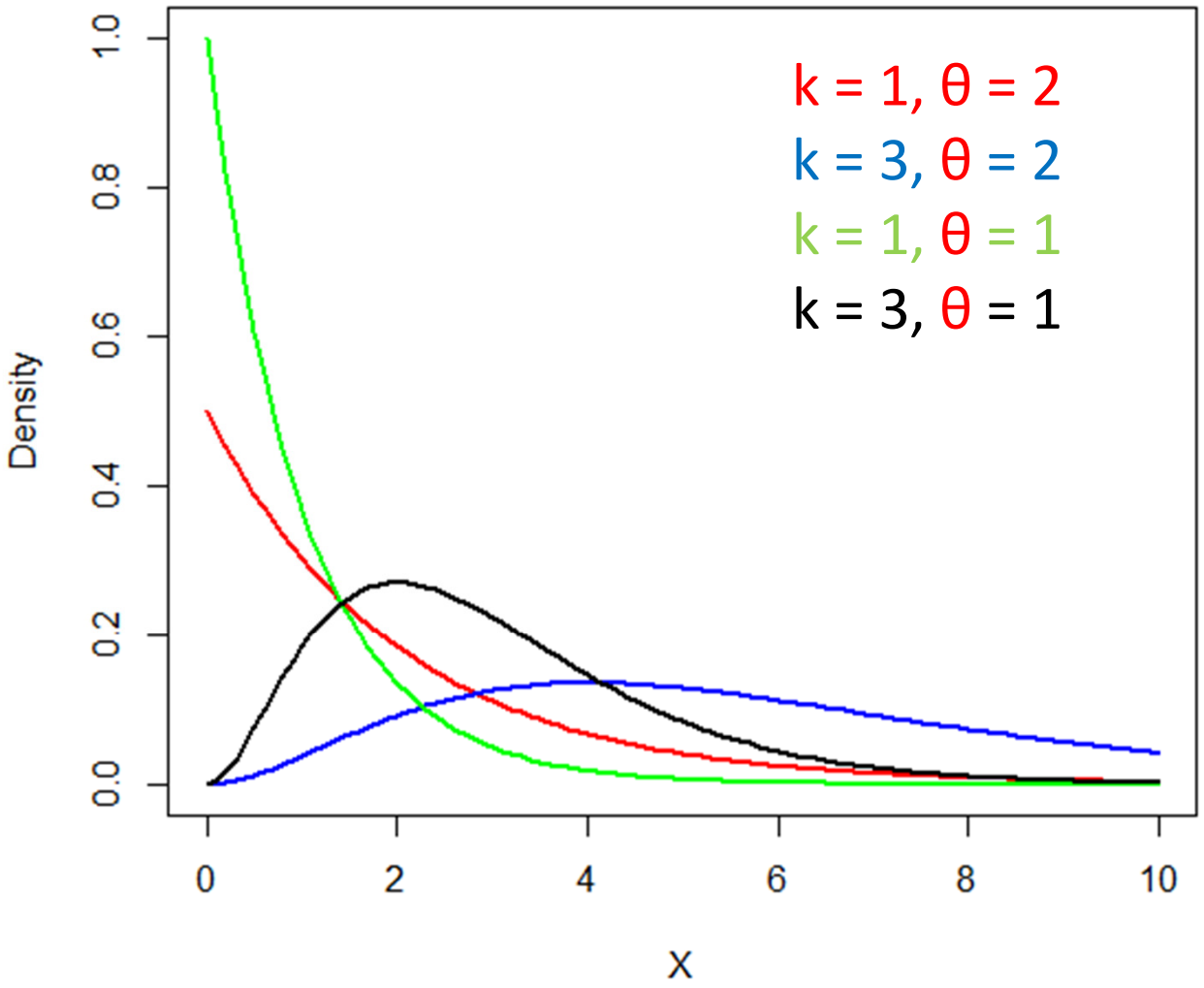
Weibull distribution




```
#Weibull distribution
x <- seq(0,3.5,0.01)
#0.5,1,1.5,5
w1 <-dweibull(x,shape=0.5)
w2 <-dweibull(x,shape=1)
w3 <-dweibull(x,shape=1.5)
w4 <-dweibull(x,shape=5)
plot(x,w1,col="red",type="l",
    ylim=c(0,2.5),
    xlab="X",ylab="Density",lwd=2)
points(x,w2,col="blue",type="l",lwd=2)
points(x,w3,col="green",type="l",lwd=2)
points(x,w4,col="black",type="l",lwd=2)
```

- Often used in survival analysis (time to death of organism)

Gama distribution




```
#Gamma distribution
x <- seq(0,10,0.1)
g1 <- dgamma(x,shape=1,scale=2)
g2 <- dgamma(x,shape=3,scale = 2)
g3 <- dgamma(x,shape=1, scale = 1)
g4 <- dgamma(x,shape=3,scale = 1)
plot(x,g1,col="red",type="l",
    ylim=c(0,1),
    xlab="X",ylab="Density",lwd=2)
points(x,g2,col="blue",type="l",lwd=2)
points(x,g3,col="green",type="l",lwd=2)
points(x,g4,col="black",type="l",lwd=2)
```

Characterizing distributions

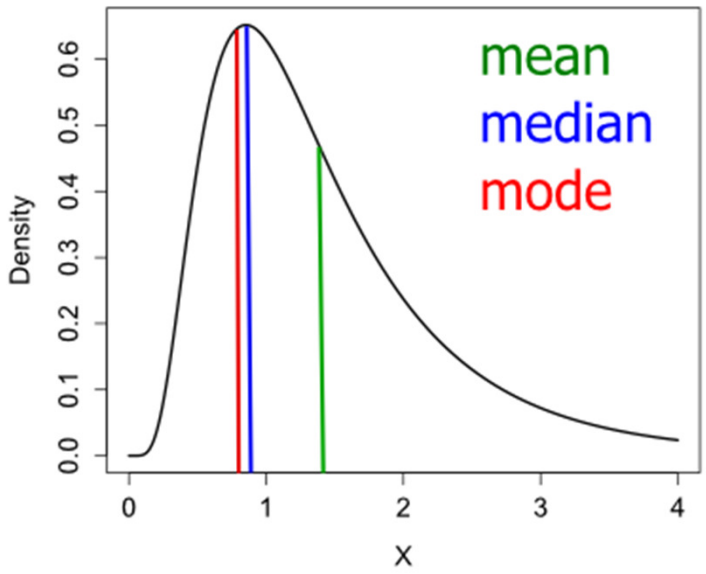
- Location and spread
  - Location: Mean, median, mode
  - Spread: variance, standard deviation, standard error
- Distribution characteristics
  - Central moments
    - Arithmetic mean (first moment)
    - Variance (second moment)
    - Skewness (third moment)
    - Kurtosis (fourth moment)

Location: quick refresher

- Mean (arithmetic)
  - Arithmetic average of a distribution (set of values)

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

  - Unbiased estimator of μ if:
    - Randomly selected individuals
    - Samples independent
    - Samples drawn from a larger population described by a normal random variable
- Median
  - Value separating the higher half of a data set from the lower half
- Mode
  - The value (observation) that occurs most often in a data set
  - For symmetric distribution mean ≈ median ≈ mode
  - Median and mode useful when data doesn't conform to a standard distribution



Spread: quick refresher

- Variance
  - Measure of how far observed values from a random variable differ from the expected E(X) value
  - PFD variance

$$Var(x) = \sigma^2 = \int (x-\mu)^2 f(x) dx$$

  - Estimated population variance

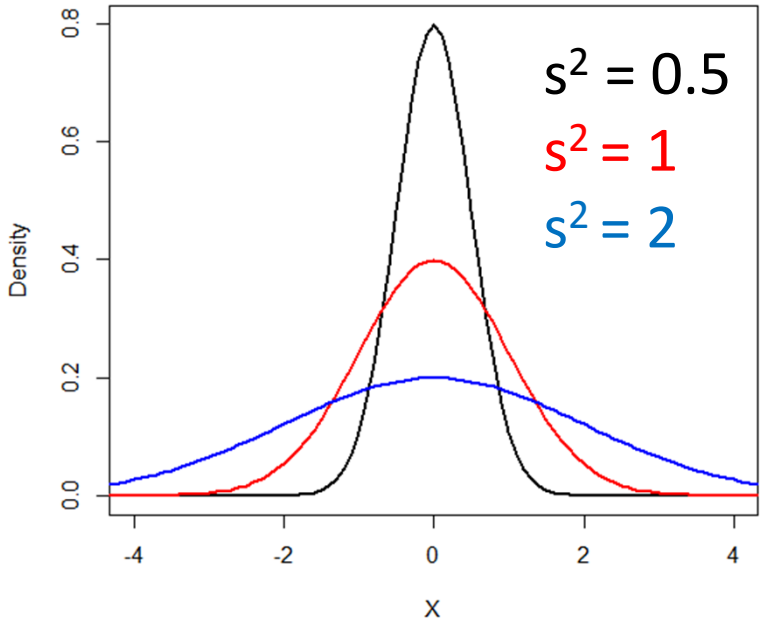
$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

- Standard deviation

$$s = \sqrt{s^2}$$

- Standard error

$$s_{\bar{Y}} = \frac{s}{\sqrt{n}}$$



- Sum of squares
  - Regression, ANOVA
- Degrees of freedom
  - # independent data points that we can use for estimation

Central moments

- A **moment** is quantitative measure of the shape of a set of points.
- Moments about a random variables mean are **central moments**
- Evaluating moments is one of the easiest ways to characterize and distinguish probability distributions

- Arithmetic mean: first moment
- Variance: second central moment
- Skewness: third central moment
- Kurtosis: fourth central moment

Skewness (third central moment)

- Central moment (general)

$$CM = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^r$$

r = 1 arithmetic mean

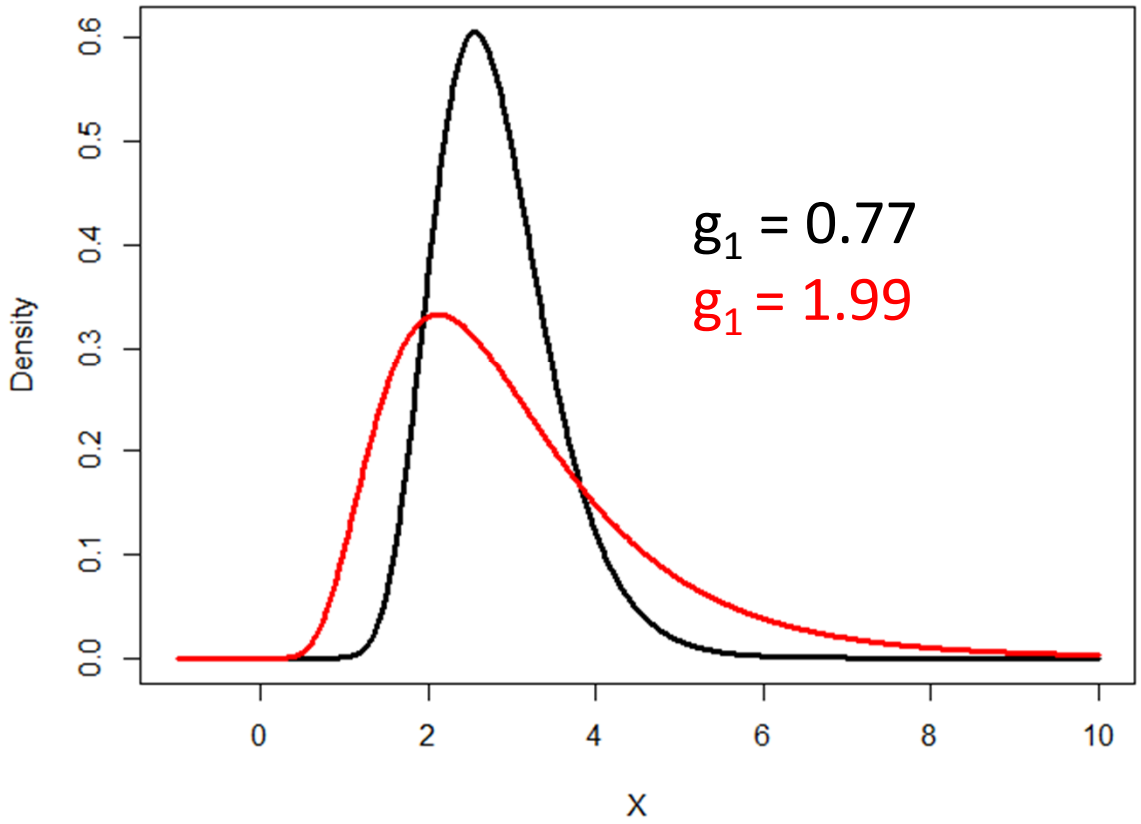
r = 2 variance

- Third central moment

$$g_1 = \frac{1}{ns^3}\sum_{i=1}^{n}(Y_i - \bar{Y})^3$$

- How the sample differs in **shape** from a symmetrical distribution
  - $g_1 > 0$ is right skewed
  - $g_1 < 0$ is left skewed

Skewness



Positive skew
Right skewed
mean > mode

```
#Skewness
#install.packages("moments") # package e1071
#library(moments)
x <- seq(-1,10,0.01)
g1 <- dlnorm(x,1,0.25)
g2 <- dlnorm(x,1,0.5)
plot(x,g1,col="black",type="l",
    xlab="X",ylab="Density",lwd=3)
points(x,g2,type="l",col="red",lwd=3)
skewness(rlnorm(1000,1,0.25))
skewness(rlnorm(1000,1,0.5))
```

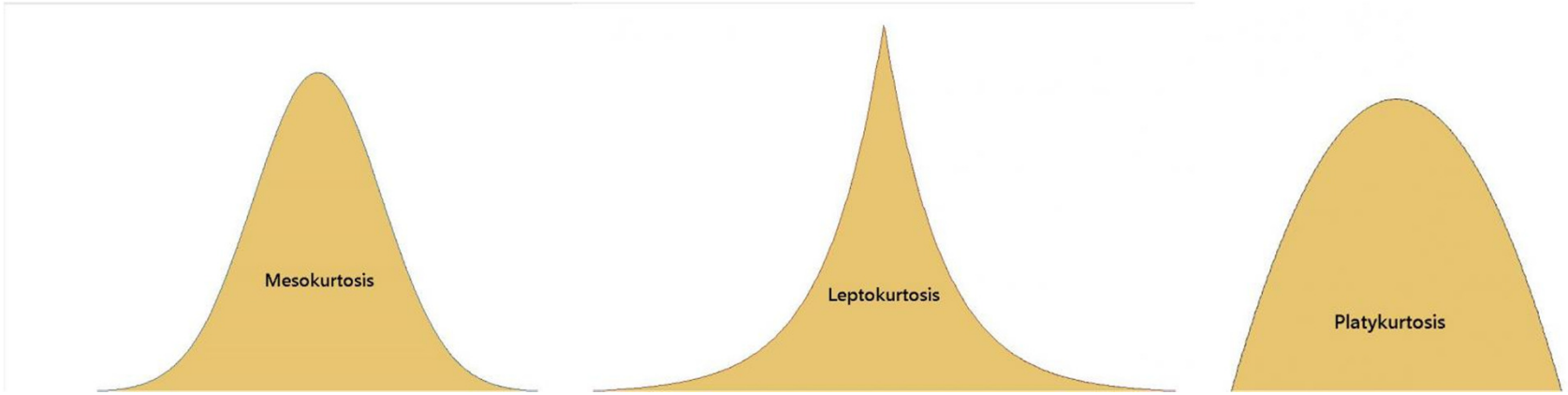
Kurtosis

- Fourth central moment

$$g_2 = \left[\frac{1}{ns^4}\sum_{i=1}^{n}(Y_i - \bar{Y})^4\right] - 3$$

- Represents the extent to which data is distributed in the tails vs. the center of the distribution
  - $g_2 < 0$ is leptokurtic, more probability in the tails
  - $g_2 > 0$ is platykurtic, less probability in the tails

Kurtosis



$g_2 = 0$

$g_2 > 0$

Fat tails

$g_2 < 0$

Skinny tails

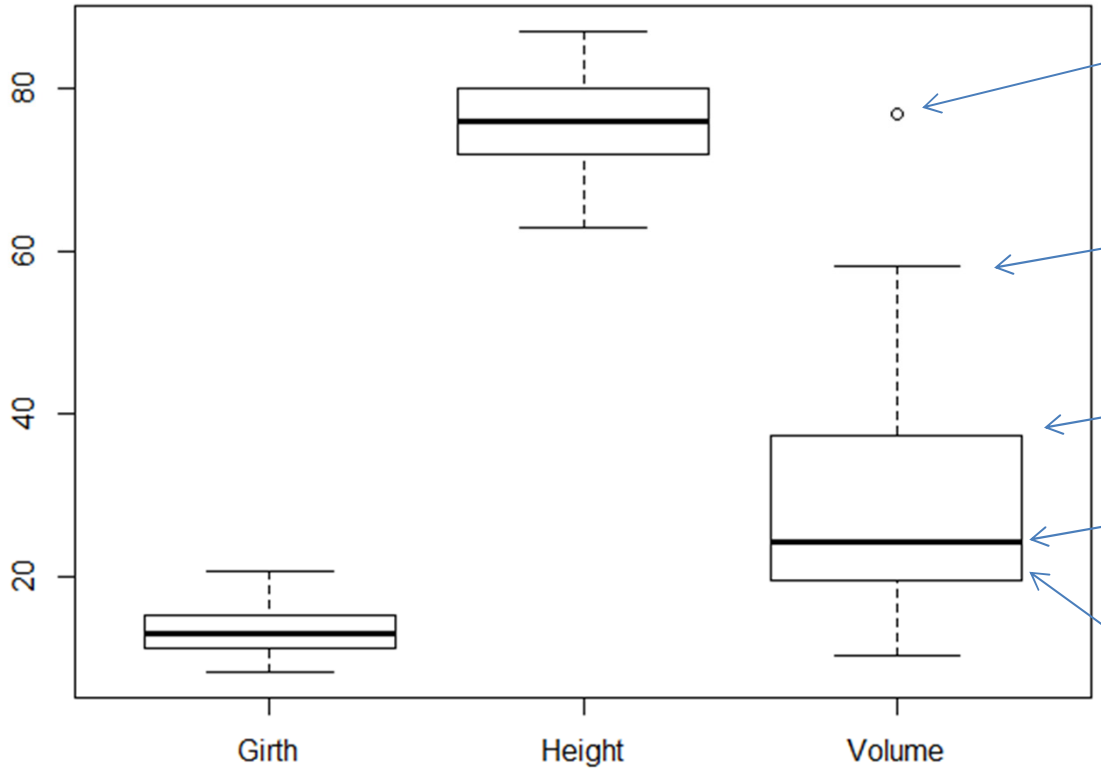
Quantiles

- Another measure of spread
- Point where a defined % of measured data has a smaller value
- Median (50^{th} percentile)
- Upper and lower quartiles (25^{th} and 75^{th} percentiles)
- Upper and lower deciles (10^{th} and 90^{th} percentiles)
- Provides easily accessible information about a distribution
- More meaningful way to describe data that is asymmetric or contains a large number of extreme values

Quantiles

- Box plot

```
# Quantiles
data(trees)
boxplot(trees)
```



Outlier

Max

3rd quartile

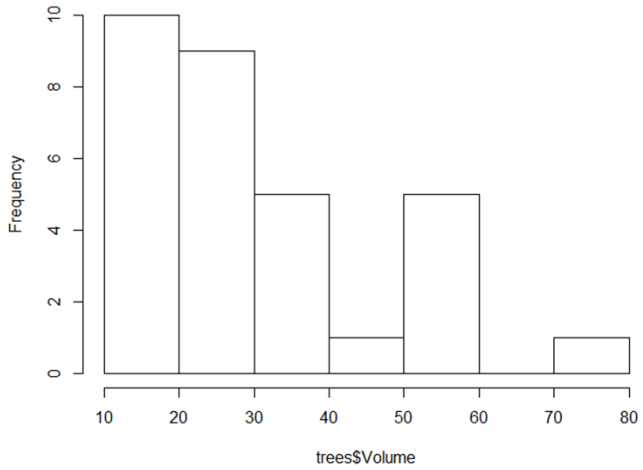
Median

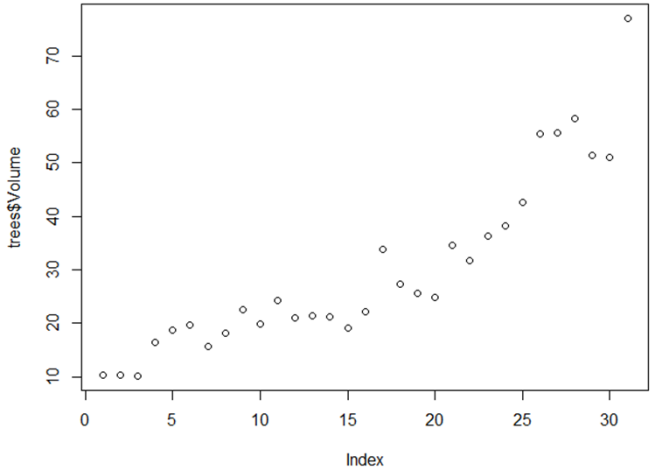
1st quartile
25% quantile

Outliers

- Data points that are distant from other observations
- Hawkins 1980: “An outlier is an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism”

- “flag” for potential problem
- Error or true variation
  - Recording error: coding error, measurement error
    - Remove or fix
  - True variation: distribution characteristic or other biological process





Data exploration (raw)

- Dependent and independent variables

- Plot all data
  - Individual variables
  - Scatter plots (relationships between variables)
- Histograms
- Boxplot
- Mean, median, mode
- Skew and Kurtosis

Data distribution problems

- Transformations?
  - What transformation best?
  - How should the results be interpreted?
  - Ideally statistical model relates directly to ecological process (i.e. measuring and understanding real ecological parameters)

- Different distribution used for statistical test
  - E.g. count data seems to be Poisson distributed
  - Use Poisson regression (log-linear regression model)
  - Generalized linear models

  - E.g. Zero-inflated negative binomial (species count data)

Data exploration

- What type of distribution am I expecting?
- What type of random variable(s) would I expect from the ecological processes I am interested in?
- What other forms of variation (uncertainty) are included in my data (can these be accounted for)?
- Does the statistical hypothesis that I am able to test, correspond to the ecological hypothesis (model) that I want to test?