

Maximum Likelihood Estimation

CMEE MSc

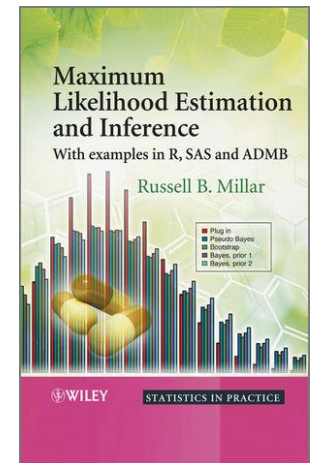
Dr Tin-Yu Hui

<tin-yu.hui11@imperial.ac.uk>

22 Feb 2021

Schedule

- A.M. Lectures
- P.M. Q&A and Practical
- Friday: special topics
- Suggested readings:
 - Millar, *Maximum Likelihood Estimation and Inference*.
 - Crawley, *The R Book*.
 - Hogg & Tanis, *Probability and Statistical Inference*.



Learning outcome

- Define random variables, probability distributions, expectations, and associated concepts
- Understand the principles of Maximum Likelihood Estimation
 - and its relation to other branches of Statistics (e.g. Bayesian)
- Perform hypothesis testing, point and interval estimation under the likelihood framework
- Develop your own likelihood models

- Appreciate Statistics, and start to believe that it is more than a subject 😊

Probability vs Statistics

- A Probabilistic question:
 - Given a fair coin, what is the probability of tossing three heads in a row?
- A Statistical question:
 - I tossed three heads in a row, is the coin fair?

Calculate the chance of occurrence of a certain event, based on some (given) random mechanisms.

Given the observation, what inferences can we make about the underlying mechanism?

- e.g. The Wright-Fisher model
- If the current allele frequency is p_0 , then the allele counts in the next generation due to drift will be binomially distributed with size $2N$ and prob p_0
- In t generations time, the mean allele frequency will not change, but $var(p_t) = p_0(1 - p_0)[1 - \left(1 - \frac{1}{2N}\right)^t]$
- The mean persistence time of an allele is approximately $\bar{t} \approx -4N[p_0 * \log(p_0) + (1 - p_0) * \log(1 - p_0)]$

- If I obtained some temporal changes in allele frequency, what is my best guess for N ?
- Is it normal for a locus with initial frequency p_0 to have remained polymorphic for $> t$ generations under neutrality? Or have there been other forces (e.g. migration or selection) acting on the locus?

Statistical inference

- Point estimation
 - our “best guess”, one-number summary
- Interval estimation
 - e.g. 95% confidence interval
- Hypothesis testing
 - H_0 vs H_1
 - model selection

Day 1

- Random variables
 - discrete and continuous
- Probability mass/density functions
 - cumulative functions
- Expectations, statistical moments, moment-generating functions

A random variable is...

- a variable, and it is random...



A random variable is...

- A variable who takes on its value by chance. A random variable can take on a set of possible values, each with an associated probability.
- To fully characterise a random variable (r.v.) we need to know:
 - all its possible outcomes (domain/support)
 - the probability of hitting each outcome

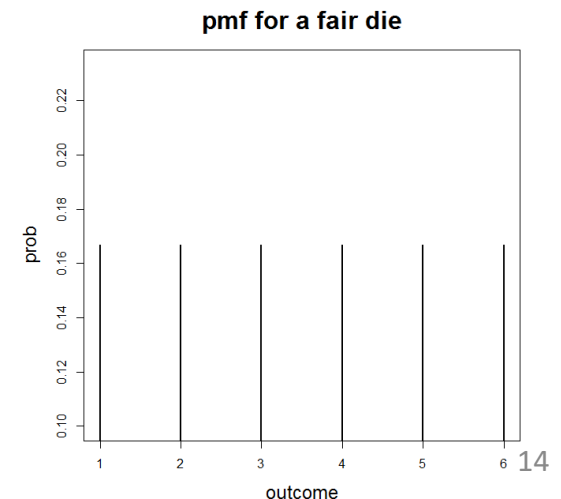
- Let X be the outcome from tossing a fair coin.
 - X is a random variable
 - two possible outcomes: $\{head, tail\}$
 - $\Pr(X = head) = 0.5, \Pr(X = tail) = 0.5$
- Let X be the outcome from rolling a fair die.
 - six possible outcomes: $\{1, 2, 3, 4, 5, 6\}$
 - $\Pr(X = 1) = 1/6, \Pr(X = 2) = 1/6, \dots$
- Let X be tomorrow's temperature.
 - possible outcomes: from -15°C to 35°C
 - how can we quantify the probabilities then...?

Discrete and Continuous r.v.

- A quantity X is called a **discrete** r.v. if 1) it can only take a discrete collection of values, and 2) it is random.
- A quantity X is called a **continuous** r.v. if 1) it can take a whole range of real-numbered values, and 2) it is random.

Probability mass function for discrete r.v.

- A probability **mass** function (or pmf) for a discrete r.v. X is a function that describes the relative probability that X takes each of its possible values.
- Denoted by $f_X(x)$ or $f(x)$.
- pmf is in form of vertical bars



Probability density function for continuous r.v.

- A probability **density** function (or pdf) for a continuous r.v. X is a function that describes the relative probability that X takes each value in the range of possible values.
- The range of possible values (with non-zero probabilities) is called the *support* of r.v. X .

Some common discrete r.v.

Bernoulli r.v.

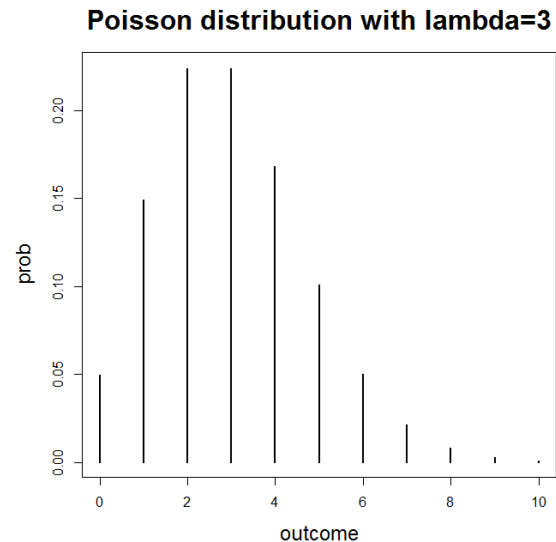
- Binary outcome: success (1) or failure (0).
- One parameter: p , probability of success.
- pmf: $\Pr(X = 1) = p, \Pr(X = 0) = 1 - p$
 - alternative expression: $f_X(x) = p^x(1 - p)^{1-x}$
- $X \sim \text{Bernoulli}(p)$

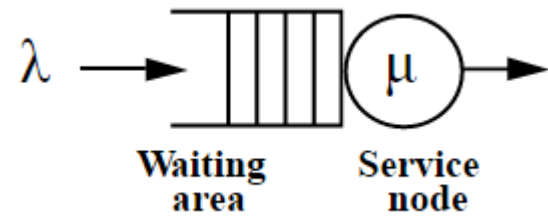
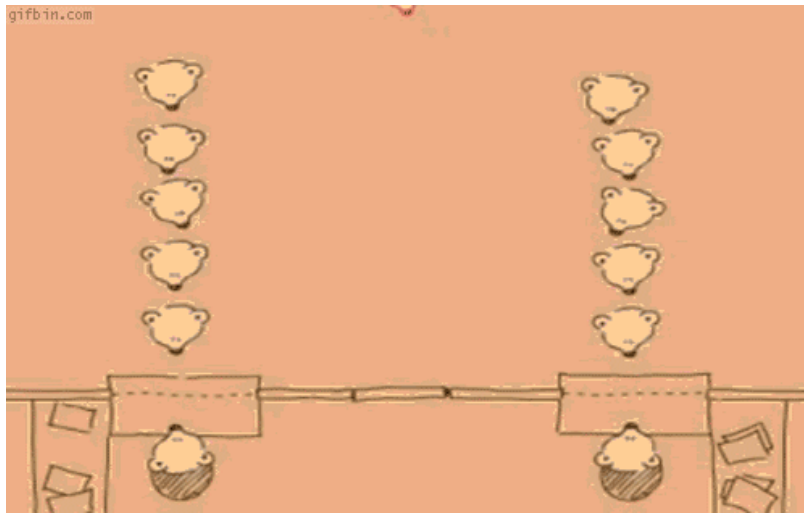
Binomial r.v.

- Sum of n independent and identically distributed (i.i.d.) Bernoulli r.v.
- Takes values on $\{0, 1, 2, \dots, n\}$
- Two parameters:
 - n : Number of independent Bernoulli trials
 - p : Probability of success (inherited from Bernoulli r.v.)
- $f_X(x) = C_x^n p^x (1 - p)^{n-x}$
- $X \sim \text{binomial}(n, p)$ or $X \sim \text{bin}(n, p)$

Poisson r.v.

- Number of events occurring in a fixed interval of time
- Possible outcomes: $\{0, 1, 2, 3, \dots\}$, all non-negative integers
- Rate parameter: $\lambda > 0$
- $f_X(x) = \frac{\lambda^x e^{-\lambda}}{x!}$
- $X \sim \text{Poisson}(\lambda)$





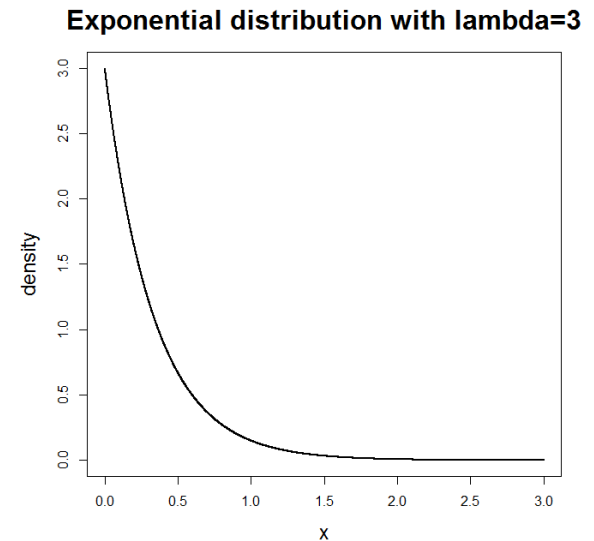
Some common continuous r.v.

Uniform r.v.

- Two parameters: a and b (the lower and upper bound)
- $f_X(x) = \frac{1}{b-a}$
- $X \sim \text{uniform}(a, b)$

Exponential r.v.

- Time between events (remember Poisson?)
- Support: $[0, \infty)$
- λ : the rate parameter, $\lambda > 0$
- $f_X(x) = \lambda e^{-\lambda x}$
- $X \sim \text{Exponential}(\lambda)$

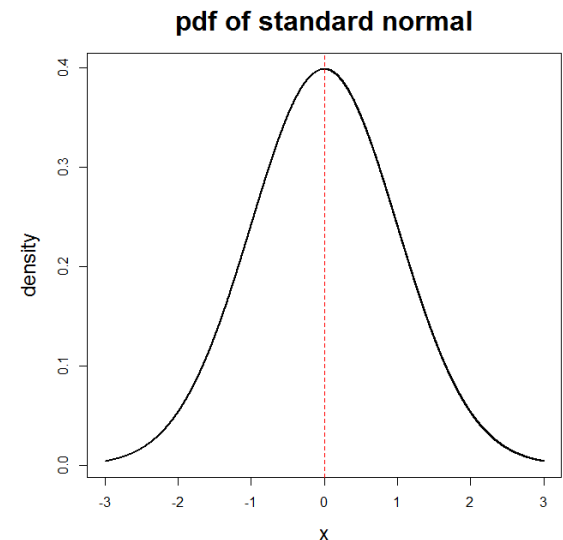


Normal r.v.

- The most famous one (why?)
- Takes values over the real number line
- Two parameters: μ, σ^2

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

- $X \sim N(\mu, \sigma^2)$



Some notations

- We understand that the pdf/pmf can be written as $f_{\mathbf{X}}(x)$ or $f(x)$. The former expression specifies the r.v. of interest through the subscript \mathbf{X} .
 - “the pdf of the r.v. X is f_X ”
- This can avoid confusion when handling more than one r.v., say, $f_{\mathbf{X}}(x)$ and $f_{\mathbf{Y}}(y)$, or $f_{\mathbf{X}_1}(x_1)$ and $f_{\mathbf{X}_2}(x_2)$
- The lowercase (e.g. x or y) inside the round bracket indicates the value at which the pdf/pmf is evaluated.
- Some texts may even state the associated parameter(s) θ while quoting a pdf/pmf. E.g. $f(x; \theta)$ or $f(x|\theta)$

Properties of pmf/pdf

- Always above the horizontal axis
 - probabilities are non-negative
- [Discrete r.v.] Sum of pmf (bars) = 1
- [Continuous r.v.] Area under pdf = 1

Cumulative mass/density function

- $F_X(x) = \Pr(X \leq x)$, hence the name cumulative
- $F(-\infty) = 0$ and $F(\infty) = 1$
- Always non-decreasing
- For discrete r.v.,

$$F_X(x) = \sum_{x_i \leq x} f_X(x_i)$$

- For continuous r.v., $F_X(x)$ is the area under the pdf curve, from $-\infty$ to x :

$$F_X(x) = \Pr(X \leq x) = \int_{-\infty}^x f_X(t) dt$$

- Calculus!

Expectation

Expectation

- Imagine you repeat the same experiment for infinitely many times (e.g. keep tossing a coin, keep drawing r.v. from a distribution), the expectation is the “average” from these repeated experiments.
- Note that the “average” here is the hypothetical average of infinitely many trials. Try not to confuse with the “sample average” that we calculate from data.
 - e.g. Population mean vs Sample mean

Expected value

- $E(X) = \sum_{all\ outcomes} x \cdot f(x)$
- $E(X) = \int_{-\infty}^{+\infty} x \cdot f(x) dx$
- “Average” value weighted according to the probability distribution. Often called the expected value of X .
- $E(X)$ is the population mean or true mean of the r.v. X . It is a measure of central tendency.

Variance

- $$\begin{aligned} \text{Var}(X) &= E \left[(X - E(X))^2 \right] \\ &= E(X^2) - [E(X)]^2 \end{aligned}$$
- The variance is the expected squared distance of the r.v. X from its population mean
- $$E(X^2) = \int_{-\infty}^{+\infty} x^2 f(x) dx$$
 - if X is an r.v., then X^2 is also an r.v.
 - expected value of the transformed r.v. X^2
- Variance is a measure of dispersion

Example

- $X \sim \text{Bernoulli}(p)$, two possible outcomes: $\{0, 1\}$

$$\begin{aligned} E(X) &= \sum x f(x) \\ &= 0 * (1 - p) + 1 * p \\ &= p \end{aligned}$$

$$\begin{aligned} E(X^2) &= \sum x^2 f(x) \\ &= 0^2 * (1 - p) + 1^2 * p \\ &= p \end{aligned}$$

$$\begin{aligned} \text{Var}(X) &= E(X^2) - [E(X)]^2 \\ &= p - p^2 \\ &= p(1 - p) \end{aligned}$$

More on expectation

- $E(X^n) = \int_{-\infty}^{+\infty} x^n f_X(x) dx$
- $E[g(X)] = \int_{-\infty}^{+\infty} g(x) f_X(x) dx$
 - for any real function g
- $E(X + Y) = E(X) + E(Y)$ for any r.v. X and Y (linearity)
- BTW, functions / transformations of r.v. are also r.v.

Statistical moments

$E(X)$: central tendency, mean

$E(X^2)$: dispersion, variance

$E(X^3)$: skewness

$E(X^4)$: kurtosis

- The n^{th} moment of a r.v. X is $E(X^n)$

Moment generating function

- Moment generating function (mgf) $M_X(t)$ can also be used to characterise a r.v.
 - t is a dummy variable, X in the subscript indicates the r.v. of interest
- The mgf “generates” statistical moments through its derivatives at $t = 0$:
- n^{th} moment of $X = E(X^n) = \frac{d^n M_X(t)}{dt^n} \Big|_{t=0}$

Yesterday we...

- Learned about some common discrete and continuous r.v.
- Plotted some pmf/pdf in R
- Calculated statistical moments of r.v.
 - moment generating function

Day 2

- Multivariate random variable
 - correlation and covariance
 - nuisance variables and marginalisation
- Independence
- Likelihood function (finally!)

Multivariate r.v.

- Sometimes events happen at the same time, or interact with each other. For example,
 - allele frequencies at different loci (genetic linkage)
 - population sizes of species within a dynamical / eco system
 - wind speed and rainfall in the same region
 - different traits of an individual
 - stock prices
- The **joint** pmf/pdf is multi-dimensional
- For bivariate case, the joint distribution of the two r.v. X and Y is often denoted as $f_{XY}(x, y)$
- $f_{XY}(x, y)$ looks like a landscape, 3D plot

Bivariate normal r.v.

- Support: \mathbf{R}^2 (two-dimensional real number plane)
- pdf: $f_{XY}(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(\frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right)$
- Parameters: mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2)$ and variance-covariance matrix $\boldsymbol{\Sigma}$
- $\begin{pmatrix} X \\ Y \end{pmatrix} \sim MVN\left(\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}\right)$
- <http://socr.ucla.edu/htmls/HTML5/BivariateNormal/>

Marginal distribution

- Given $f_{XY}(x, y)$ the joint pdf. Sometimes we are interested in only one of them (X , say)
 - i.e. we would like to obtain the marginal pdf of X
 - without referencing to the values of Y
- $f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x, y) d\mathbf{y}$
 - integrate (marginalise) out the uninterested r.v. Y
 - there will be no Y in $f_X(x)$
 - $f_X(x) = \sum_{\text{all } \mathbf{y}} f_{XY}(x, y)$ for discrete case
- Similarly, the marginal pdf of Y is $f_Y(y) = \int_{-\infty}^{+\infty} f_{XY}(x, y) d\mathbf{x}$

Conditional distribution

- If the value of Y is known (i.e. $Y = y$), this may give extra information on another r.v. X
- This conditional distribution of X is

$$f_{X|Y}(x|y) = \frac{f_{XY}(x, y)}{f_Y(y)}$$

- $X|Y$ reads as ‘ X given Y ’
- $f_{X|Y}(x|y)$ is a slice of the joint pdf at $Y = y$
- $f_{Y|X}(y|x)f_X(x) = f_{XY}(x, y) = f_{X|Y}(x|y)f_Y(y)$
 - “Joint = conditional \times marginal”
 - very useful in Bayesian and MCMC

Nuisance variables

- Say, Y is a r.v. with pdf $f_{Y|U}(y|u)$, but U is also a r.v. following another pdf $f_U(u|\theta)$
 - θ is a parameter
 - the same principle applies to hierarchical models/r.v.
- Ultimately, we would like to know $f_Y(y|\theta)$, the density of Y given the parameter θ , while U is just an intermediate (latent, nuisance) r.v.
- $f_Y(y|\theta) = \int f_{Y|U}(y|u)f_U(u|\theta)du$
 - law of total probability
 - sum/integrate across all possible values of U
 - U is marginalised

Covariance and Correlation

- Describe the linear association between two r.v.
- $cov(X, Y) = E[XY] - E[X]E[Y]$
 - $E[XY] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xy f_{XY}(x, y) dx dy$
 - “product moment”
- $corr(X, Y) = \frac{cov(X, Y)}{\sqrt{var(X)var(Y)}}$
 - bounded between -1 and +1

Independence

- Two events are **independent** if the occurrence of one does not affect the occurrence of another.
 - i.e. gives no extra information
- Perhaps the strongest assumption in statistics (we cannot actually test for independence).
- If X and Y are independent then $\text{corr}(X, Y) = 0$
- But $\text{corr}(X, Y) = 0$ **DOES NOT** imply independence!

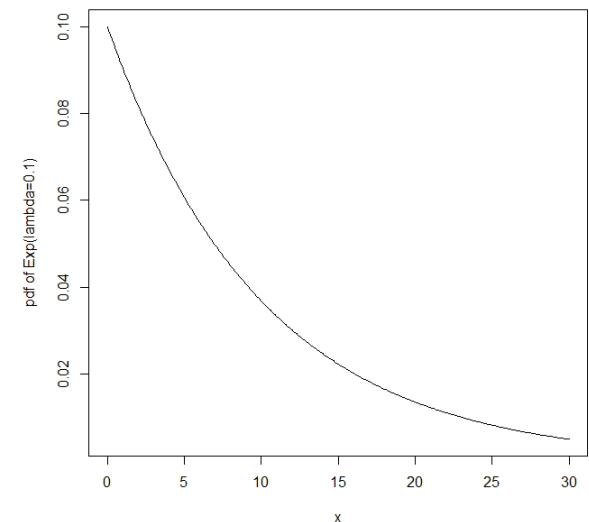
- Two r.v. X and Y are independent if and only if their joint probability density/mass function is the product of their marginal distributions:

$$f_{XY}(x, y) = f_X(x)f_Y(y)$$

- Remember our definition to independence: “The outcome of X provides no extra information about Y ”

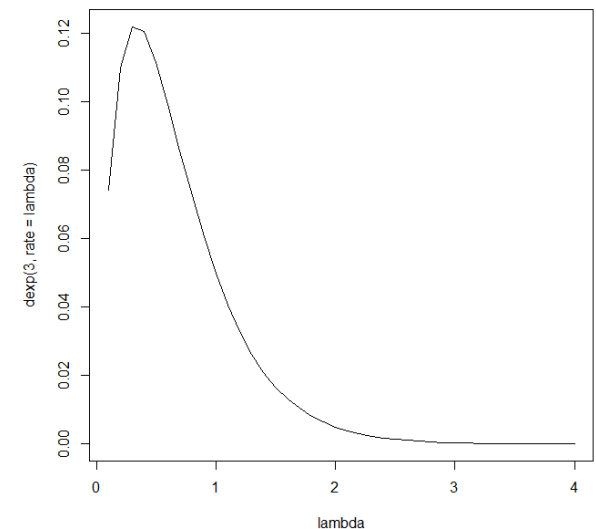
-PAUSE-

- So far we have been predicting outcomes, calculating the associated probabilities, expectations etc. of an r.v.
 - given a (known and fixed) parameter value
- For example, let X be the waiting time for a bus which is exponentially distributed with rate $\lambda = 0.1$
 - X is a r.v.
- $f(x; \lambda) = \lambda e^{-\lambda x} = 0.1 * e^{-0.1x}$
- f is expressed as a function of x



-PAUSE-

- Alternatively, if we have waited 3 units of time before getting on a bus
 - i.e. given some data, x is observed
 - the parameter is fixed but unknown
- $f(x; \lambda) = \lambda e^{-\lambda x} = \lambda e^{-3\lambda}$
- Express f as a function of λ
- Inference on the parameter λ
- Statistics!



Maximum Likelihood Estimation

- Likelihood is the central idea of statistics
- Invented (?) by Sir Ronald Fisher
- “One of the greatest ideas of the 20th century and probably one of the greatest of civilization” – Dr Dan Reuman, a co-founder of this MSc course.

Maximum Likelihood Estimation

- Maximum Likelihood estimation (MLE) is a method to estimate parameters of a statistical model
- When the method is applied to a **dataset** with a statistical **model**, MLE provides estimates for the associated **parameters**.
- “The parameter values that make the observed dataset most *probable*.”

- The likelihood function $L(\underline{\theta})$ is used to quantify how “likely” the parameter values are. The symbol $\underline{\theta}$ denotes a vector of parameters.
 - also $\underline{x} = \{x_1, x_2, \dots, x_n\}$, a vector of observations
- $L(\underline{\theta}|\underline{x}) = f(x_1, \dots, x_n|\underline{\theta})$ by definition
 - “the likelihood function is the **joint density** of \underline{x} ”
- Further, if \underline{x} are independent samples then the joint density of \underline{x} is the product of their individual densities:

$$L(\underline{\theta}|\underline{x}) = f(x_1, \dots, x_n|\underline{\theta}) = \prod_{i=1}^n f_{X_i}(x_i|\underline{\theta})$$
- Once \underline{x} is observed, $L(\underline{\theta}|\underline{x})$ **becomes a function of $\underline{\theta}$ only**

- For each set of \underline{x} (fixed) and given a model, let $\underline{\hat{\theta}}$ be a parameter value at which $L(\underline{\theta}|\underline{x})$ attains its maximum. $\underline{\hat{\theta}}$ is the maximum likelihood estimate for the observed data \underline{x} .
- Maximising the log-likelihood function is equivalent to maximising the likelihood function.

- Treat the parameters as unknowns (a bit counter-intuitive)
- The triplets:
 - Model
 - Parameters
 - Data

Example 1: Coin tossing


- If we flip 10 coins, independently, and observe 7 heads and 3 tails
- If we define p as $Prob(head)$, what is the MLE for p ?
- Each coin toss is a Bernoulli trial, and the joint density of 10 independent coin tosses is *binomial*(n, p).
- Let Y be the number of heads out of 10 tosses
$$f(Y = y) = C_y^{10} p^y (1 - p)^{10-y}$$

- Now, put $y = 7$ as this is what we observed

$$f(Y = 7) = C_7^{10} p^7 (1 - p)^{10-7}$$

- And this is our likelihood function

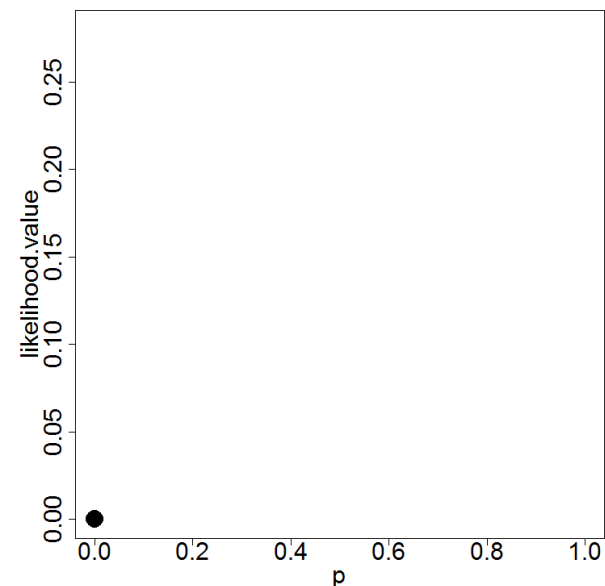
$$L(p) = f(Y = 7) = C_7^{10} p^7 (1 - p)^3$$



The likelihood function depends on p only after observing the data.

- For each value of p , there is a corresponding value of the likelihood function $L(p)$

p	$L(p)$
0	$C_7^{10} 0^7 (1 - 0)^3 = 0$
0.1	$C_7^{10} 0.1^7 0.9^3 = 8.748 \times 10^{-6}$
0.2	$C_7^{10} 0.2^7 0.8^3 = 0.000786$
0.3	$C_7^{10} 0.3^7 0.7^3 = 0.0090$
0.4	$C_7^{10} 0.4^7 0.6^3 = 0.0424$
\vdots	\vdots



Some R code

```
# WRITE DOWN THE LIKELIHOOD FUNCTION
binomial.likelihood<-function(p) {
  choose(10,7)*p^7*(1-p)^3
}

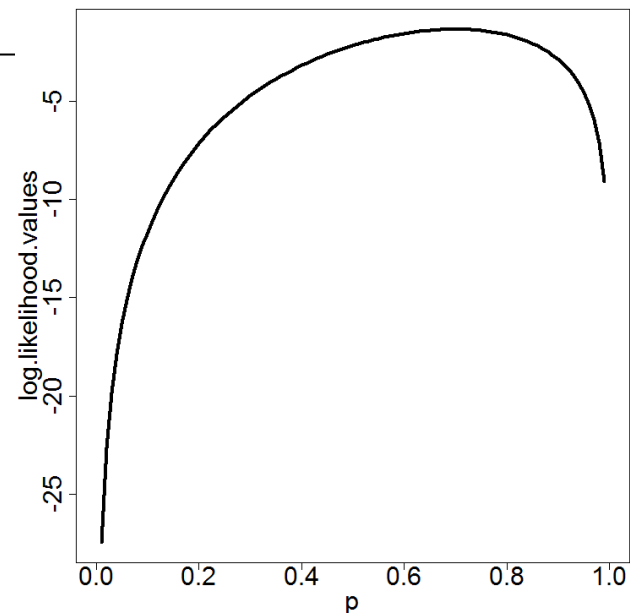
# LET US CALCULATE THE LIKELIHOOD VALUE AT p=0.1
binomial.likelihood(p=0.1)
      # YOU GOT SOMETHING AROUND 8.748e-06, RIGHT?

# PLOT THE LIKELIHOOD FUNCTION FOR A RANGE OF p
p<-seq(0,1,0.01)
likelihood.values<-binomial.likelihood(p)
plot(p, likelihood.values, type='l')
```

```
# MORE OFTEN WE STUDY THE LOG-LIKELIHOOD
# WE CAN REUSE THE FUNCTION WE'VE JUST WRITTEN
log.binomial.likelihood<-function(p) {
  log(binomial.likelihood(p=p))
}

# PLOT THE LOG-LIKELIHOOD
p<-seq(0,1,0.01)
log.likelihood.values<-log.binomial.likelihood(p)
plot(p, log.likelihood.values, type='l')
```

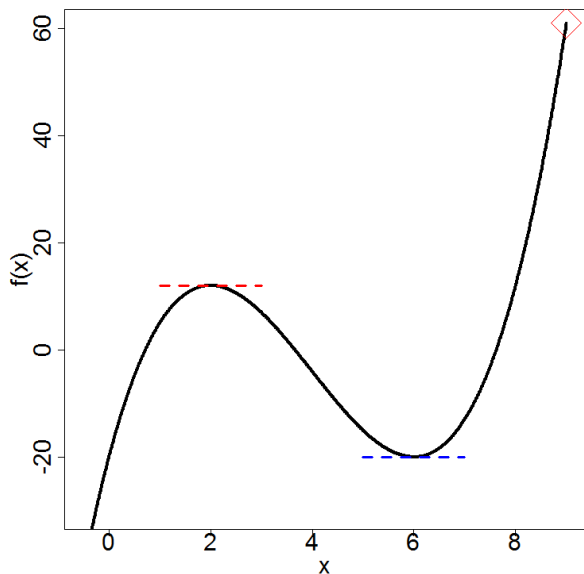
We can see that both the likelihood and log-likelihood function are maximised when p is around 0.7



- Remember in day 1 we made some **probabilistic** statements: If we toss 10 fair coins, independently, then the probability of getting 7 heads out of 10 tosses is around 0.117.
- Today we make some **statistical inferences**: If we observed 7 heads out of 10 tosses, what can we say about the coin? Is the coin loaded?

Maximisation: some mathematical considerations

- To optimise a function we need to calculate its derivatives. Other conditions (e.g. the boundaries or saddle points) need to be examined as well.
- Some proficiency in calculus is certainly required, and things can be complicated if we have multiple parameters (multivariate calculus).



- In many cases, because of the complexity of the model, or high dimensionality of the parameters (or both!), MLE cannot be solved explicitly.
- More often, log-likelihood functions are maximised numerically via computer
 - `optim()` or `optimize()` in R

```
optimize(binomial.likelihood, interval=c(0,1), maximum=TRUE)
```

```
$maximum  
[1] 0.6999843  
  
$objective  
[1] 0.2668279
```

Solve MLE analytically

In general, if we obtain y heads out of n tosses, the likelihood function is

$$L(p) = f(y|p) = C_y^n p^y (1 - p)^{n-y}$$

and the log-likelihood is

$$l(p) = \ln(L(p)) = \ln(C_y^n) + y \ln p + (n - y) \ln(1 - p)$$

Differentiate $l(p)$ w.r.t. p

$$\frac{\partial}{\partial p} l(p) = 0 + y \left(\frac{1}{p} \right) + (n - y) \left(\frac{-1}{1 - p} \right)$$

Then find $p = \hat{p}$ such that $\frac{\partial}{\partial p} l(p)|_{p=\hat{p}} = 0$

$$\frac{y}{\hat{p}} + (n - y) \left(\frac{-1}{1 - \hat{p}} \right) = 0$$

$$\frac{y}{\hat{p}} = \frac{n - y}{1 - \hat{p}}$$

$$\dots$$
$$\hat{p} = \frac{y}{n}$$

Example 2: i.i.d. normal samples

- X_1, X_2, \dots, X_n are i.i.d. random samples from $N(\mu, 1)$. Variance is known but we need to estimate μ , the population mean.

- Parameter of interest: μ

- $L(\mu) = f(x_1, x_2, \dots, x_n) = \prod_{i=1}^n f(x_i)$

$$= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i - \mu)^2}{2}\right)$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2\right)$$

Because of
independence!

x_i are the observed
samples, fixed.

μ is the only quantity to be
estimated.

The log-likelihood is

$$l(\mu) = \text{constant} - \frac{1}{2} \left(\sum_{i=1}^n (x_i - \mu)^2 \right)$$

Differentiate the log-likelihood wr.t. μ

$$\frac{\partial l}{\partial \mu} = 0 - \frac{1}{2} \left[-2 \sum_{i=1}^n (x_i - \mu) \right]$$

Does not depend on μ

Find $\mu = \hat{\mu}$ such that the derivative is zero. i.e.

$$\sum_{i=1}^n (x_i - \hat{\mu}) = 0$$

$$\sum_{i=1}^n x_i - n\hat{\mu} = 0$$

$$\hat{\mu} = \frac{\sum_{i=1}^n x_i}{n}$$

MLE suggests that the arithmetic average of our samples is an estimate for μ .

Example 3: normal samples with unknown variance

- X_1, X_2, \dots, X_n are i.i.d. random samples from $N(\mu, \sigma^2)$. Both μ, σ^2 are unknown.
- Parameters of interest: μ, σ^2 (bivariate parameter space)
- Similar to the previous example, the likelihood function is $L(\mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma}\right)^2\right)$

$$l(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2$$

We need to find $\frac{\partial l}{\partial \mu}$ and $\frac{\partial l}{\partial \sigma^2}$ (exercise)

The remaining question is to find $(\hat{\mu}, \hat{\sigma}^2)$ such that $\frac{\partial l}{\partial \mu} = 0$ and $\frac{\partial l}{\partial \sigma^2} = 0$ simultaneously. (more exercise!)

You are getting there!

Example 4: Linear regression

- The model

$$y_i = a + bx_i + \varepsilon_i$$

with i.i.d. normally distributed error term $\varepsilon_i \sim N(0, \sigma^2)$, $i = 1, 2, \dots, n$, where n is the number of data points

- Data: $\begin{cases} \underline{x} & \text{independent (explanatory) variable} \\ \underline{y} & \text{response} \end{cases}$
- Parameters: $\underline{\theta} = \begin{cases} a & \text{intercept} \\ b & \text{slope} \\ \sigma^2 & \text{variance} \end{cases}$

- [Perspective 1] The distribution of the responses y_i
- $y_i \sim N(a + bx_i, \sigma^2)$, independently
 - but with a different mean
- $L(\underline{\theta}) = f(y_1, y_2, \dots, y_n | \underline{\theta}) = f_{Y_1}(y_1 | \underline{\theta}) f_{Y_2}(y_2 | \underline{\theta}) \dots f_{Y_n}(y_n | \underline{\theta}) = \prod_{i=1}^n f_{Y_i}(y_i | \underline{\theta})$

- [Perspective 2] The distribution of the error terms
- Let us rearrange the model such that ε_i is the subject: $\varepsilon_i = y_i - a - bx_i$
 - also note that ε_i are i.i.d. $N(0, \sigma^2)$
- $L(\underline{\theta}) = f(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n | \underline{\theta}) =$
 $f(\varepsilon_1 | \underline{\theta}) f(\varepsilon_2 | \underline{\theta}) \dots f(\varepsilon_n | \underline{\theta}) = \prod_{i=1}^n f(\varepsilon_i | \underline{\theta})$

- And the log-likelihood becomes

$$l(\underline{\theta}) = \sum_{i=1}^n \ln(f(\varepsilon_i | \underline{\theta}))$$

Can you see why we prefer log-likelihood to the original likelihood?

- We can find a set of $(\hat{a}, \hat{b}, \widehat{\sigma^2})$ such that the likelihood function is maximised
- The remaining challenges are to 1) write down the log-likelihood function in R, and 2) to maximise it
- Q6 of today's practical

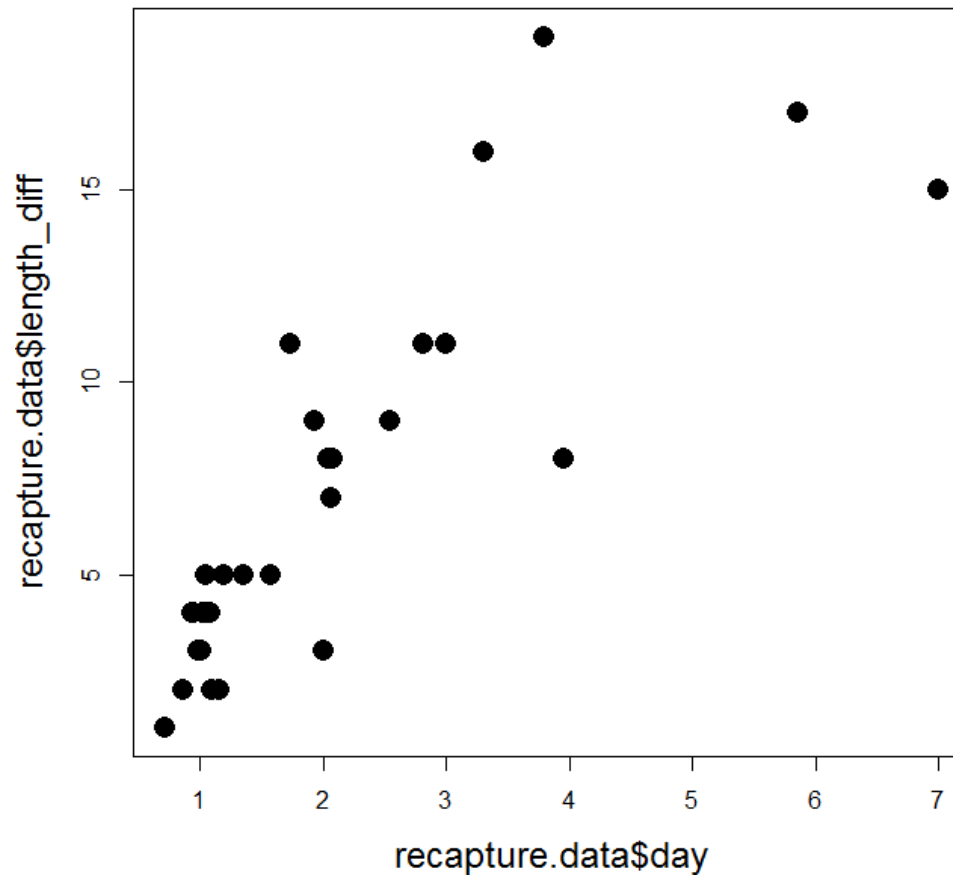
Rabbit example

- We had tagged and released some rabbits, and then some 29 of them were recaptured
- `day` measures the days a rabbit had spent before being recaptured (Explanatory variable)
- `diff_length` is the grow in body length in between (Response)
- What is the relationship between these two variables?



```
# READ IN DATASET
recapture.data<-read.csv('recapture.csv', header=T)

# SCATTERPLOT
plot(recapture.data$day, recapture.data$length_diff)
```



[Perspective 1] Log-likelihood

```
# THE LOG-LIKELIHOOD FOR THE LINEAR REGRESSION
# PARAMETERS HAVE TO BE INPUT AS A VECTOR
regression.log.likelihood<-function(parm, dat)
{
# DEFINE THE PARAMETERS parm
# WE HAVE THREE PARAMETERS: a, b, sigma. BE CAREFUL OF THE ORDER
a<-parm[1]
b<-parm[2]
sigma<-parm[3]

# DEFINE THE DATA dat
# FIRST COLUMN IS x, SECOND COLUMN IS y
x<-dat[,1]
y<-dat[,2]

# MODEL ON y
# EACH y[i] IS NORMALLY AND INDEPENDENTLY DISTRIBUTED. WITH MEAN a+b*x[i]
# AND A COMMON VARIANCE sigma^2. VECTORISED CODE
density<-dnorm(y, mean=a+b*x, sd=sigma, log=T)

# THE LOG-LIKELIHOOD IS THE SUM OF INDIVIDUAL LOG-DENSITY
return(sum(density))
}
```

[Perspective 2] Log-likelihood

```
# THE LOG-LIKELIHOOD FOR THE LINEAR REGRESSION
# PARAMETERS HAVE TO BE INPUT AS A VECTOR
regression.log.likelihood<-function(parm, dat)
{
# DEFINE THE PARAMETERS parm
# WE HAVE THREE PARAMETERS: a, b, sigma. BE CAREFUL OF THE ORDER
a<-parm[1]
b<-parm[2]
sigma<-parm[3]

# DEFINE THE DATA dat
# FIRST COLUMN IS x, SECOND COLUMN IS y
x<-dat[,1]
y<-dat[,2]

# MODEL ON THE ERROR TERMS. VECTORISED CODE
error.term<-(y-a-b*x)
# error.term[i] ARE IID NORMAL, WITH MEAN 0 AND A COMMON VARIANCE sigma^2
density<-dnorm(error.term, mean=0, sd=sigma, log=T)

# THE LOG-LIKELIHOOD IS THE SUM OF INDIVIDUAL LOG-DENSITY
return(sum(density))
}
```



```
# JUST TO SEE WHAT THE LOG-LIKELIHOOD VALUE IS WHEN a=1, b=1, and sigma=1
# YOU MAY TRY ANY DIFFERENT VALUES
regression.log.likelihood(c(1,1,1), dat=recapture.data)
```

```
[1] -452.6903
```

```
# TO OPIMISE THE LOG-LIKELIHOOD FUNCTION IN R
# optimize() IS ONE-DIMENSIONAL,
# optim() GENERALISES TO MULTI-DIMENSIONAL CASES
optim(par=c(1,1,1), regression.log.likelihood, method='L-BFGS-B',
      lower=c(-1000,-1000,0.0001), upper=c(1000,1000,10000),
      control=list(fnscale=-1), dat=recapture.data, hessian=T)
```

```
$par
[1] 1.527870 2.676240 2.678428
```

```
$value
[1] -69.72089
```

```
$counts
function gradient
      40      40
```

```
$convergence
[1] 0
```

```
$message
[1] "CONVERGENCE: REL_REDUCTION_OF_F <= FACTR*EPSMCH"
```

par=c(1,1,1)	Initial values for the parameters
log.likelihood.regression	The function to be optimised
method='L-BFGS-B'	Optimisation algorithm
lower=c(-1000,-1000,0.0001)	Lower bound of parameter space
upper=c(1000,1000,10000)	Upper bound of parameter space
control=list((fnscale=-1))	fnscale=-1 means to maximise

Notes on using `optim()`

- Parameters are input as a vector. Order does matter.
- Initial parameter values are set by the first argument `par=`
- Choice of optimisation `method` can be tricky even for advanced users. See R help for details.
- The method `L-BFGS-B` requires a box-like `upper` and `lower` bound for parameter values. Nothing to specify for `Nelder-Mead`
- If you wish to maximise a function, set `fnscale=-1` in your `control` list. The default is to minimise. You can put multiple control parameters (such as tolerance) in the `control` list.
- The Hessian matrix provide information about the variance-covariance structure of your parameter estimates (more on this later).
- Try multiple sets of initial parameters to ensure they all converge to the global maximum.

- “Stumble around” the parameter space towards the best parameters, just like a drunkard trying to stumble home (the best place).
- Not every step is in the right direction, and it takes some time to go home.
- Ideal if the drunkard finds his place.
- But he may get stuck at the local maximum (not the most comfortable place, but, still..., okay..., at a tube station?)

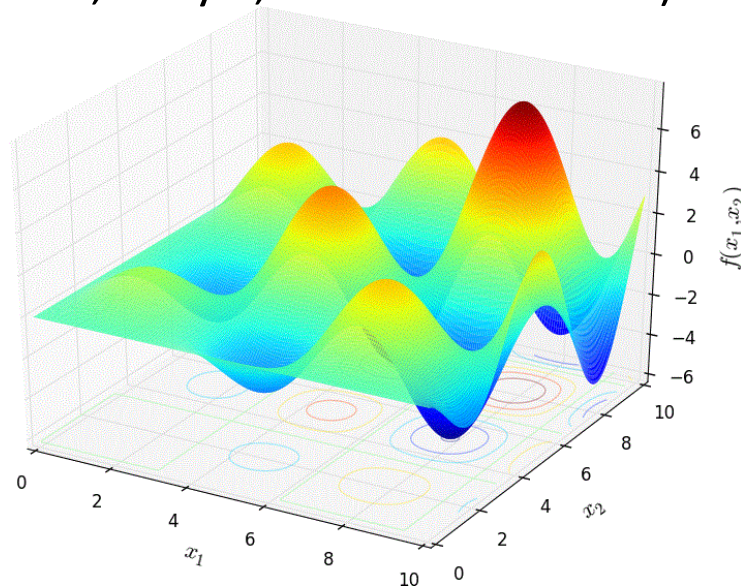


Photo credit: Dan Reuman

- Of course you can perform the same analysis with `lm()`

```
# REGRESSION WITH THE BUILT-IN lm()
m<-lm(length_diff~day, data=recapture.data)
summary(m)

> summary(m)

Call:
lm(formula = length_diff ~ day, data = recapture.data)

Residuals:
    Min       1Q   Median       3Q      Max
-5.2499 -1.2226 -0.1297  0.9099  7.3179

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.5279     0.8833   1.730   0.0951 .
day           2.6762     0.3464   7.725 2.62e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.776 on 27 degrees of freedom
Multiple R-squared:  0.6885,    Adjusted R-squared:  0.677
F-statistic: 59.67 on 1 and 27 DF,  p-value: 2.622e-08
```

```
n<-nrow(recapture.data)
sqrt(var(m$residual)*(n-1)/n)
```

```
[1] 2.6784281
```

2009

$$Y = \beta X + \epsilon$$

STATISTICS

2019

$$Y = \beta X + \epsilon$$

MACHINE LEARNING

~~10~~ 10 YEARS CHALLENGE

Yesterday we...

- Constructed our first likelihood function
 - data, model, and parameters of interest
- Maximised likelihood functions by differentiation
- Maximised likelihood functions in R using `optim()` and `optimize()`

Day 3

- Properties of ML estimator
- Logistic regression example
- Likelihood-Ratio test

Properties of ML Estimator

- Asymptotically unbiased
 - on average we are hitting the target
 - $E[\hat{\theta}] \rightarrow \theta$ when $n \rightarrow \infty$
- Low variance (efficient)
 - better use of data
 - narrower confidence interval compared to other estimators



- Consistent: ML estimator converges in probability to the true parameter when $n \rightarrow \infty$
- Asymptotically normal
 - ML estimator is asymptotically distributed as normal with mean equals the true parameter value
 - remember Central Limit Theorem?
 - construction of confidence intervals (more on this later)

- Invariant principle
 - if $\hat{\theta}$ is the ML estimator for θ , then $g(\hat{\theta})$ is the ML estimator for $g(\theta)$

Example: Logistic regression

- Binary responses: dead or alive, yes or no, success or failure...
- Explanatory variable x is often called a risk factor (affect the risk/probability of “bad” outcome)
- Very common in public health/ medicine/ biology/ classification

#	State	Average cholesterol
1	Dead	5.0
2	Alive	4.4
3	Alive	3.4
4	Dead	3.7
5	Alive	3.6
6	Dead	4.7
...

- We need to find r.v. with binary outcomes to model the response variable y_i
- Bernoulli r.v.! Logistic regression assumes each response variable y_i follows a Bernoulli distribution
- Each individual will have its own p_i , which is a function of the risk factor x_i
 - x_i is the risk factor
 - $a + bx_i$ is the linear predictor
- $y_i \sim \text{Bernoulli}(p_i)$, where $p_i = \eta^{-1}(a + bx_i)$, a and b are our parameters.
- What is η^{-1} ?

- In logistic regression, $\eta^{-1}(a + bx_i) = \frac{e^{a+bx_i}}{1+e^{a+bx_i}}$
- η^{-1} is called “expit” transformation. The inverse of “logit” transformation
- $\eta^{-1}(a + bx_i)$ is bounded between 0 and 1 (remember, p_i is the probability of success), regardless of the values of $a + bx_i$
- Let us construct the likelihood function

- Two parameters: a and b

$$L(a, b) = \prod_{i=1}^n f(y_i) = \prod_{i=1}^n [p_i^{y_i} (1 - p_i)^{1-y_i}]$$

$$= \prod_{i=1}^n [\text{expit}(a + bx_i)^{y_i} (1 - \text{expit}(a + bx_i))^{1-y_i}]$$

- Take to log of the likelihood function

$$l(a, b) = \sum_{i=1}^n \{y_i \ln[\text{expit}(a + bx_i)] + (1 - y_i) \ln[1 - \text{expit}(a + bx_i)]\}$$

- It becomes a function of a and b only (with known y_i and x_i). We can maximise the log-likelihood function w.r.t. a and b .

Non-standard regression

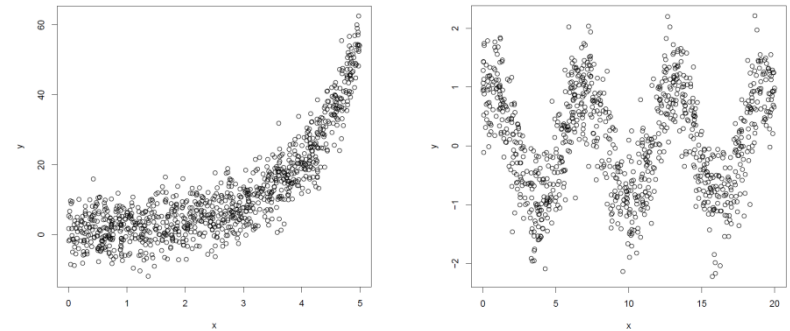
- Learning MLE means you can build your own statistical models
- Especially for non-standard cases where no “instant meals” are available

- $y_i = \exp(mx_i + b) + \epsilon_i$

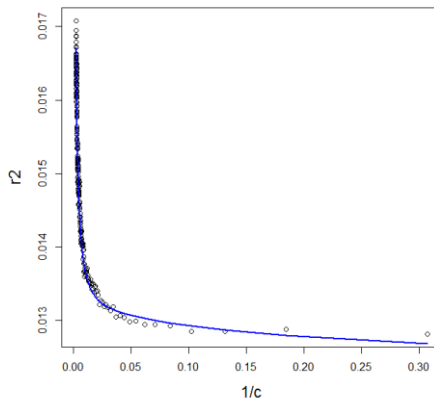
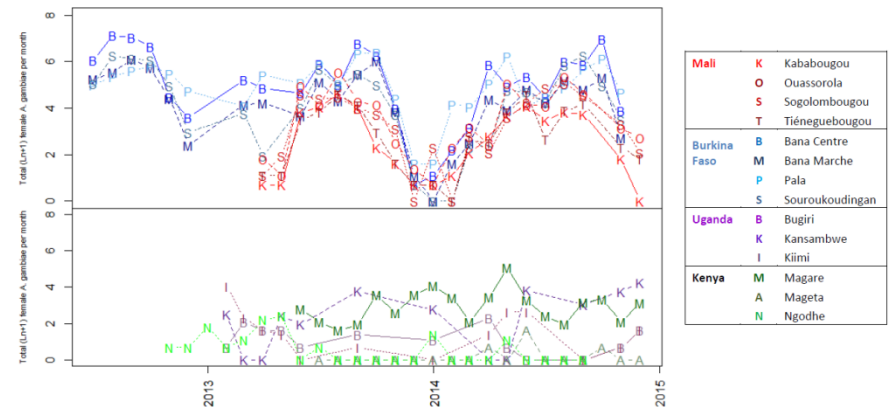
- Over-dispersed data

- Seasonal data

- State-space model



All PSC Time series together - Ln+1



$$\begin{array}{ccccccc}
 p_0 & \rightarrow & p_1 & \rightarrow & p_2 & \rightarrow & \dots & \rightarrow & p_{t-1} & \rightarrow & p_t \\
 \downarrow & & \downarrow & & \downarrow & & & & \downarrow & & \downarrow \\
 x_0 & & x_1 & & x_2 & & \dots & & x_{t-1} & & x_t
 \end{array}$$

Likelihood-Ratio Test

- Hypothesis testing
- Let $M1$ and $M2$ be two models, and that $M1$ is **nested in** $M2$. If $M2$ has $d2$ parameters and $M1$ has $d1$ parameters ($d2 > d1$), then $D = 2 * (\ln(L2) - \ln(L1))$ follows approximately a chi-square distribution with $(d2 - d1)$ degrees of freedom.
 - D is the LRT statistic
- The procedure is as follows:
 - Fit $M1$ to the data, record the **maximised** log-likelihood value $\ln(L1)$
 - Fit $M2$ to the data, record the **maximised** log-likelihood value $\ln(L2)$
 - Compute the likelihood-ratio statistic $D = 2 * (\ln(L2) - \ln(L1))$
 - Look up χ^2_{d2-d1} table for critical value. Accept $M1$ as the simplified model if D is smaller than the critical value

- Rationale:
 - The larger the log-likelihood value the better fit the model
 - M2 fits the data better with more parameters, thus yields a larger maximised log-likelihood value
 - M1 is the simplified model who has less explanatory power than M2 and therefore a smaller maximised log-likelihood value
 - D measures the difference in ‘explanatory power’
 - If the parameters dropped by M1 are unimportant, then there will only be a small decrease in explanatory power, hence a small D statistic
 - Dropping unimportant terms means we tend to accept M1 as the simplified model

Linear regression: test for intercept

- In yesterday's `recapture.csv`, we may think (biologically) that the intercept should be zero, because if a rabbit falls back to the trap “within zero days”, then there should be no difference in its body length
- We let M1 be a linear regression model without an intercept i.e. $y_i = bx_i + \varepsilon_i$ (Two parameters)
- We let M2 be the full linear regression model we fitted yesterday i.e. $y_i = a + bx_i + \varepsilon_i$ (Three parameters)
- Clearly M1 is a special case of M2 with $a = 0$. We say M1 is nested in M2.

Log-likelihood function for M1

```
# THE LOG-LIKELIHOOD FUNCTION FOR M1 WITHOUT AN INTERCEPT
regression.no.intercept.log.likelihood<-function(parm, dat)
{
# DEFINE THE PARAMETERS
# NO INTERCEPT THIS TIME
?????
?????

# DEFINE THE DATA
# SAME AS BEFORE
x<-dat[,1]
y<-dat[,2]

# DEFINE THE ERROR TERM, NO INTERCEPT HERE
error.term<-?????

# REMEMBER THE NORMAL pdf?
density<-dnorm(error.term, mean=0, sd=sigma, log=T)

# LOG-LIKELIHOOD IS THE SUM OF DENSITIES
return(sum(density))
}
```

Log-likelihood function for M1

```
# THE LOG-LIKELIHOOD FUNCTION FOR M1 WITHOUT AN INTERCEPT
regression.no.intercept.log.likelihood<-function(parm, dat)
{
# DEFINE THE PARAMETERS
# NO INTERCEPT THIS TIME
b<-parm[1]
sigma<-parm[2]

# DEFINE THE DATA
# SAME AS BEFORE
x<-dat[,1]
y<-dat[,2]

# DEFINE THE ERROR TERM, NO INTERCEPT HERE
error.term<- (y-b*x)

# REMEMBER THE NORMAL pdf?
density<-dnorm(error.term, mean=0, sd=sigma, log=T)

# LOG-LIKELIHOOD IS THE SUM OF THE DENSITIES
return(sum(density))
}
```

Performing likelihood-ratio test

```
# PERFORMING LIKELIHOOD-RATIO TEST
M1<-optim(par=c(1,1), regression.no.intercept.log.likelihood,
          dat=recapture.data, method='L-BFGS-B',
          lower=c(-1000,0.0001), upper=c(1000,10000),
          control=list(fnscale=-1), hessian=T)
M2<-optim(par=c(1,1,1), regression.log.likelihood,
          dat=recapture.data, method='L-BFGS-B',
          lower=c(-1000,-1000,0.0001), upper=c(1000,1000,10000),
          control=list(fnscale=-1), hessian=T)

# THE TEST STATISTIC D
D<-2*(M2$value-M1$value)
D

[1] 3.047676
```

```
# CRITICAL VALUE
qchisq(0.95, df=1)

[1] 3.841459
```

We accept the hypothesis that the intercept is zero at $\alpha = 0.05$ (Same conclusion is drawn from `lm()` using anova table)

Model selection

- *AIC* is a tool to determine which of two models is better by weighting the improved fit of more complex models against their larger number of parameters.
- $AIC = -2l(\hat{\theta}) + 2K$, where $l(\hat{\theta})$ is the maximised log-likelihood and K is the number of parameters in the model
- Find the model with the lowest AIC value

Exercise: Non-constant variance regression

- In `recapture.csv`, we observe that the variance of the response is increasing with `day`. (Why?)
- Can we incorporate non-constant variance in our regression?
- Not sure about how we can do it with `lm`. Transformation of variables may help, but it is relatively simple MLE.
- How about $\varepsilon_i \sim N(0, x_i^2 \sigma^2)$? The standard deviation of the error terms increases linearly with the number of days?

Log-likelihood function: non-constant variance

```
# THE LOG-LIKELIHOOD FUNCTION FOR M1 WITHOUT AN INTERCEPT
regression.non.constant.var.log.likelihood<-function(parm, dat)
{
# DEFINE THE PARAMETERS
# NO CHANGE FROM M1
b<-parm[1]
sigma<-parm[2]

# DEFINE THE DATA
# SAME AS BEFORE
x<-dat[,1]
y<-dat[,2]

# DEFINE THE ERROR TERM, NO INTERCEPT HERE
error.term<-(y-b*x)

# REMEMBER THE NORMAL pdf
density<-dnorm(error.term, mean=0, sd=x*sigma, log=T)

# THE LOG-LIKELIHOOD IS THE SUM OF INDIVIDUAL DENSITIES
return(sum(density))
}
```

```
# MAXIMISE THE LOG-LIKELIHOOD
# HOW ABOUT CALLING IT M4?
M4<-optim(par=c(1,1), regression.non.constant.var.log.likelihood,
          dat=recapture.data, method='L-BFGS-B',
          lower=c(-1000,0.0001), upper=c(1000,10000),
          control=list(fnscale=-1))
```

M4

```
> M4
$par
[1] 3.483407 1.149874

$value
[1] -60.62583

$counts
function gradient
      25      25

$convergence
[1] 0

$message
[1] "CONVERGENCE: REL_REDUCTION_OF_F <= FACTR*EPSMCH"
```

This afternoon...

- Free 😊

Yesterday we...

- Discussed the theoretical guarantees of MLE
 - Why do we prefer MLE to other estimators?
- Fitted a logistic regression via MLE
- Introduced Likelihood-Ratio test
 - hypothesis testing for nested models

Day 4

- Interval estimation
 - Confidence interval
 - Joint Confidence region
 - Profile likelihood
 - Normal approximation

Confidence interval estimation

- We are now able to find point estimates by maximising the log-likelihood function
- Usually confidence intervals (C.I.) are also required while quoting them
- There are many ways to calculate C.I., some of which can be directly obtained from the log-likelihood function

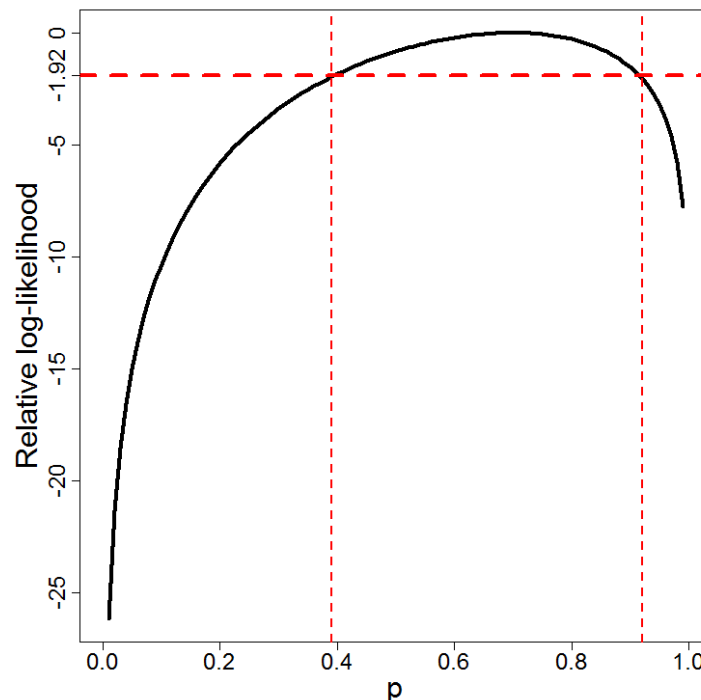
1661 and 4365. The published estimate using MLNE (Wang 2001) was 2169 (C.I. = 1221–5744), while the estimate from the *F*-statistic (Waples 1989) was 2247 (C.I. = 1127–8370). The complete result can be found in table 2 of Cuveliers *et al.* (2011, p. 3561). We found that all three estimates mostly

- Let us revisit the coin tossing example. We performed a likelihood-ratio test to test for $H_0: p = 0.5$ (fair coin hypothesis), with $n = 10$ and $y = 7$
- For the simplified model M1, there is no free parameter as p is fixed to 0.5
 - $l(0.5) = -2.14398$
- The full model M2 has one free parameter p , $0 \leq p \leq 1$, and l is maximised at $p = 0.7$
 - $l(0.7) = -1.32115$
- The critical value of this test is $\chi^2_{0.95, df=1} = 3.84$
- $D = 2 * [-1.32115 - (-2.14398)] = 1.64566 < 3.84$
- According to LRT, $H_0: p = 0.5$ is not rejected

C.I example: coin tossing

- Instead of performing a LRT, or multiple LRTs against different values of p , we can find a range of p , such that D remains within the “acceptance region”.
 - a collection of p such that $D < 3.84$
- Equivalently, we can find a collection of p such that the log-likelihood descends by no more than 1.92 units from the maximum
 - $D = 2 * (\ln(L2) - \ln(L1)) < 3.84$
 - $\ln(L2) - \ln(L1) < 1.92$

- In most cases, if we want to find the 95% C.I. for a single parameter, we look at the range of parameter values such that the log-likelihood is within 1.92 units from its maximum
- Rule of thumb: -1.92, or -2



- If we observe 7 heads out of 10 tosses, the 95% C.I. for p is $[0.39, 0.92]$.
- Since 0.5 lies inside the 95% C.I., we do not reject the “fair coin” hypothesis.

C.I. example: linear regression

- Back to our `recapture.csv`, M1 has two parameters: b, σ
- Plot the log-likelihood surface against b and σ
- Bivariate function \rightarrow 3D plot
- 3D plot in R using `persp()`

```

# DEFINE THE RANGE OF PARAMETERS TO BE PLOTTED
b<-seq(2, 4, 0.1)
sigma<-seq(2, 5, 0.1)

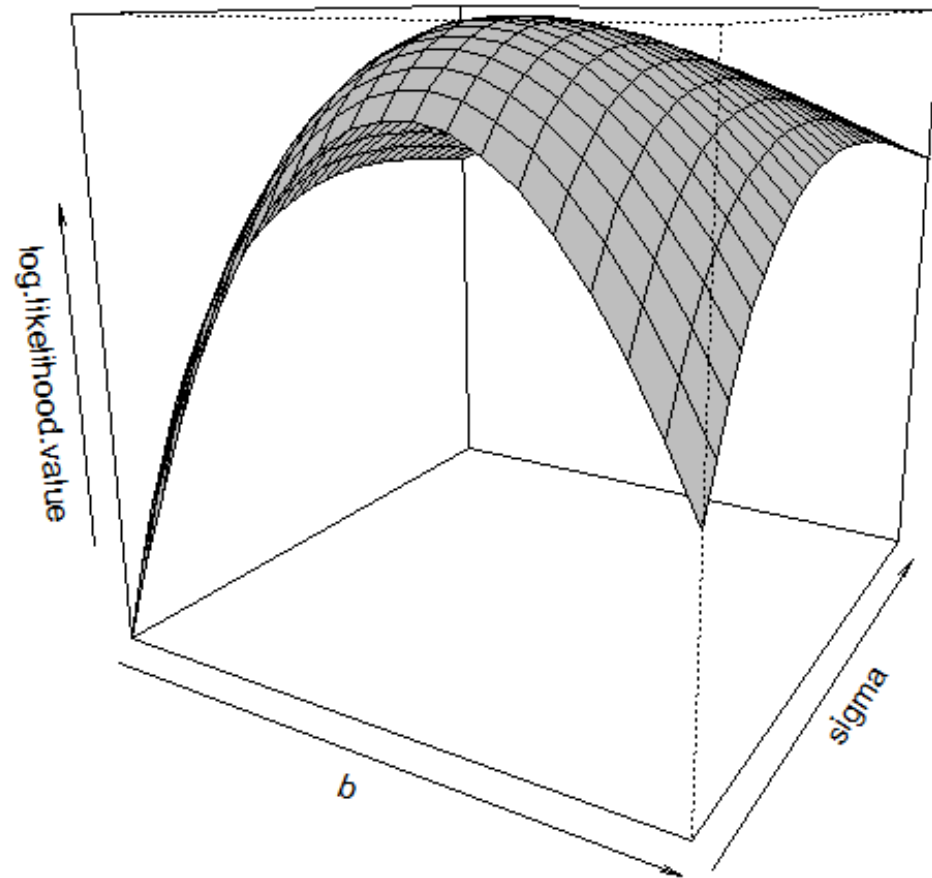
# THE LOG-LIKELIHOOD VALUE IS STORED IN A MATRIX
log.likelihood.value<-matrix(nr=length(b), nc=length(sigma))

# COMPUTE THE LOG-LIKELIHOOD VALUE FOR EACH PAIR OF PARAMETERS
for (i in 1:length(b))
{
  for (j in 1:length(sigma))
  {
    log.likelihood.value[i,j]<-
    regression.no.intercept.log.likelihood(parm=c(b[i],sigma[j]),
    dat=recapture.data)
  }
}

# WE ARE INTERESTED IN KNOWING THE RELATIVE LOG-LIKELIHOOD VALUE
# RELATIVE TO THE PEAK (MAXIMUM)
rel.log.likelihood.value<-log.likelihood.value-M1$value

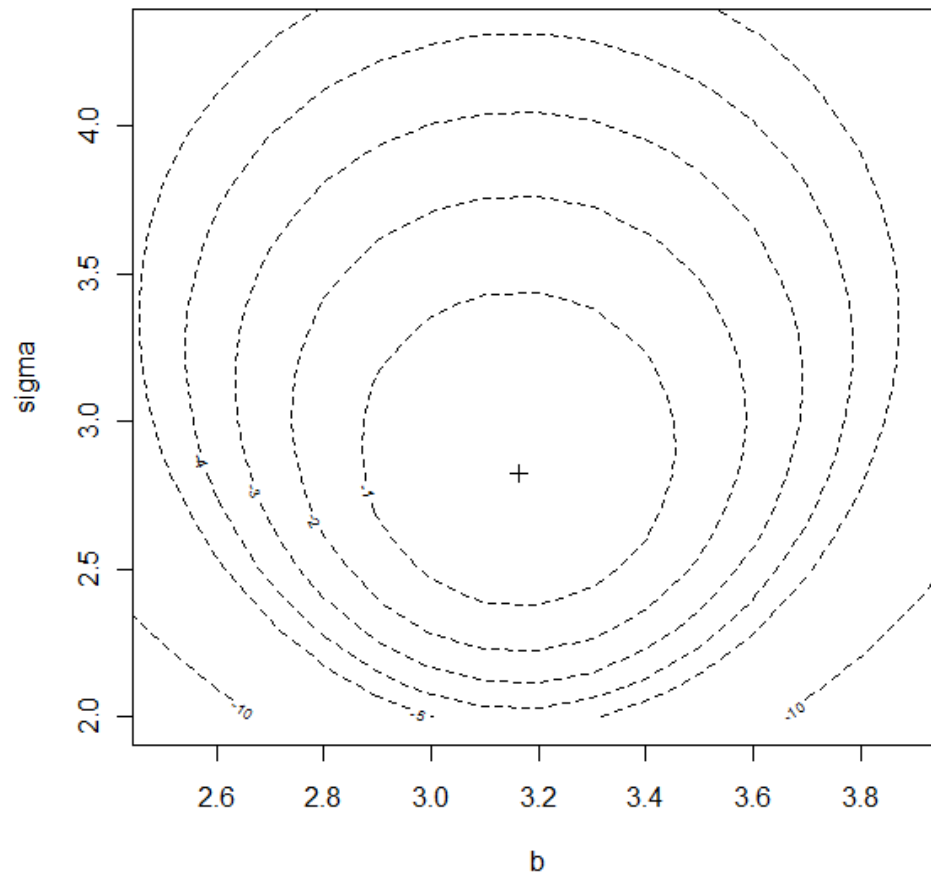
# FUNCTION FOR 3D PLOT
persp(b, sigma, rel.log.likelihood.value, theta=30, phi=20,
      xlab='b', ylab='sigma', zlab='rel.log.likelihood.value',
      col='grey')

```



How about a contour plot?

```
# CONTOUR PLOT
contour(b, sigma, rel.log.likelihood.value, xlab='b',
        ylab='sigma',
        xlim=c(2.5, 3.9), ylim=c(2.0, 4.3),
        levels=c(-1:-5, -10), cex=2)
# DRAW A CROSS TO INDICATE THE MAXIMUM
points(M1$par[1], M1$par[2], pch=3)
```

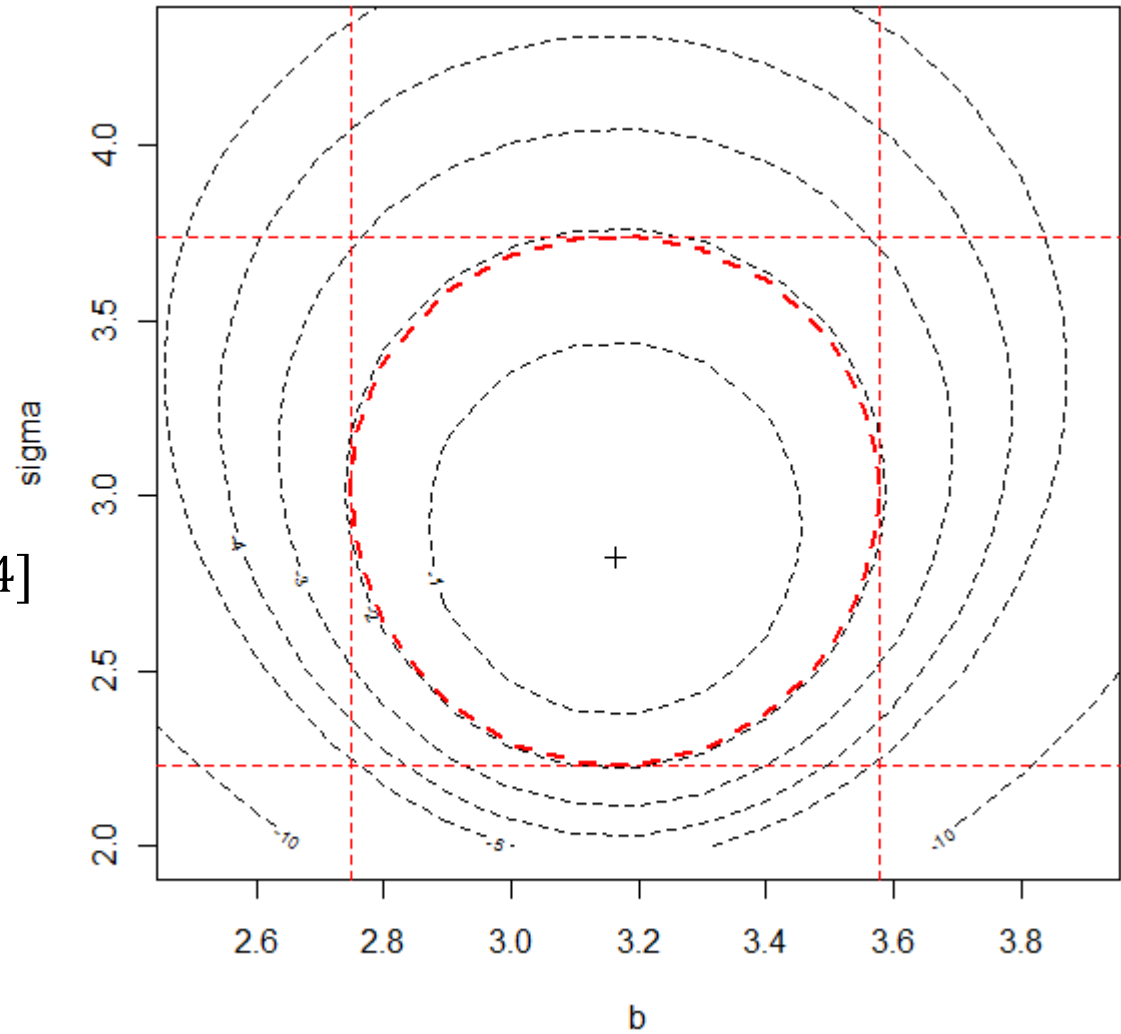


We can, again, draw the -1.92 line (circle) on the contour map

```
contour.line<-contourLines(b, sigma,
rel.log.likelihood.value, levels=-1.92)[[1]]
lines(contour.line$x, contour.line$y, col='red',
      lty=2, lwd=2)
```

IF WE STUDY ONE
PARAMETER AT A TIME

95% C.I. for σ is [2.23, 3.74]



95% C.I. for b is [2.75, 3.57]

Joint confidence interval (region)

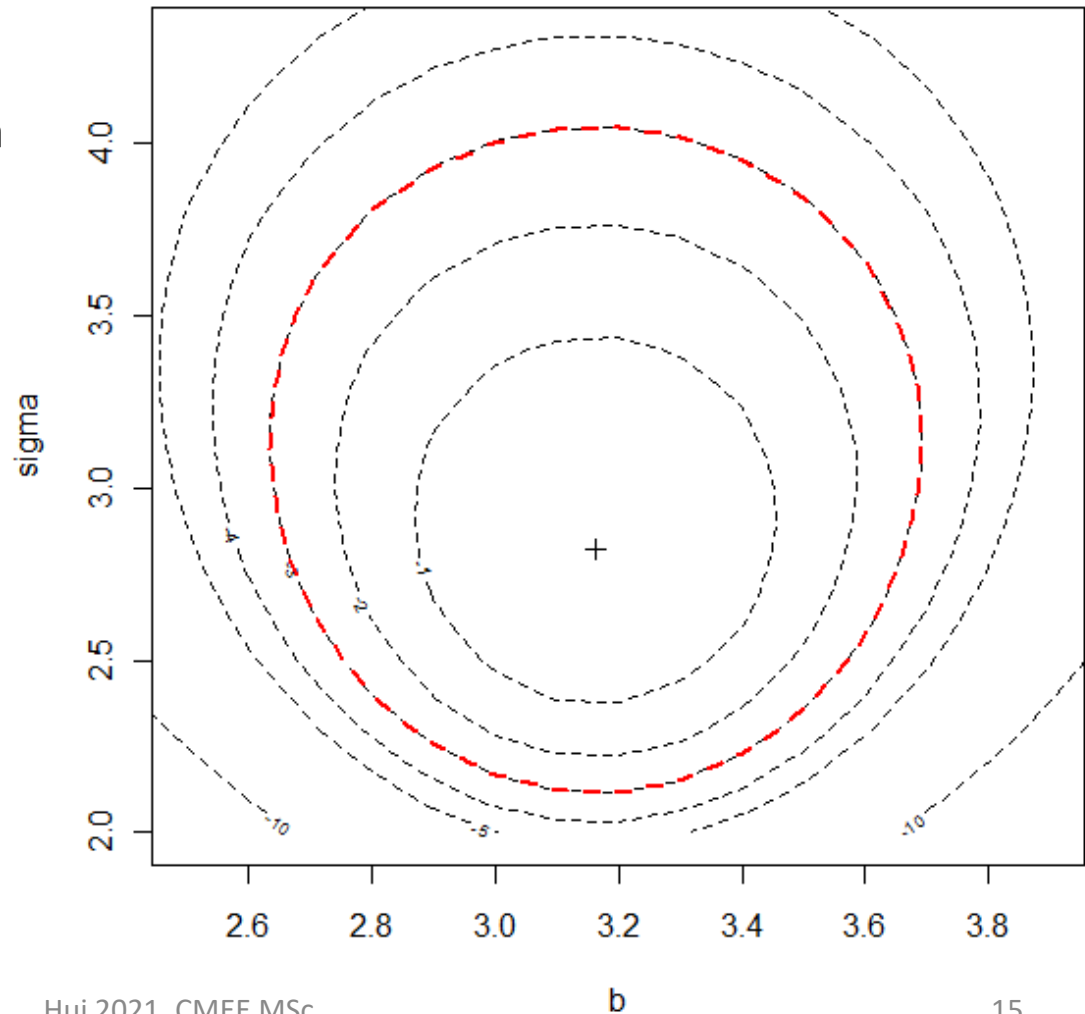
- We know the 95% C.I. for b , and the 95% C.I. for σ
- But it does not mean we know the **joint** 95% confidence **region** for the pair (b, σ)
- We need to consider the correlation between the two ML estimators
- Multiple comparisons?

Joint confidence interval (region)

- The general rule: the 95% joint C.I. (region) for k parameters is the collection of parameter values for which the log-likelihood decreases by no more than half of $\chi^2_{0.95, df=k}$ from its maximum.
- 95% C.I. for one parameter: $0.5 * \chi^2_{0.95, df=1} = 1.92$
- Joint 95% C.I. for two parameters: $0.5 * \chi^2_{0.95, df=2} = 2.99$

- On the contour plot, we can circle the region where the log-likelihood value is 2.99 units below the maximum.

The joint 95% confidence region for (b, σ) are all the points within the red dotted circle



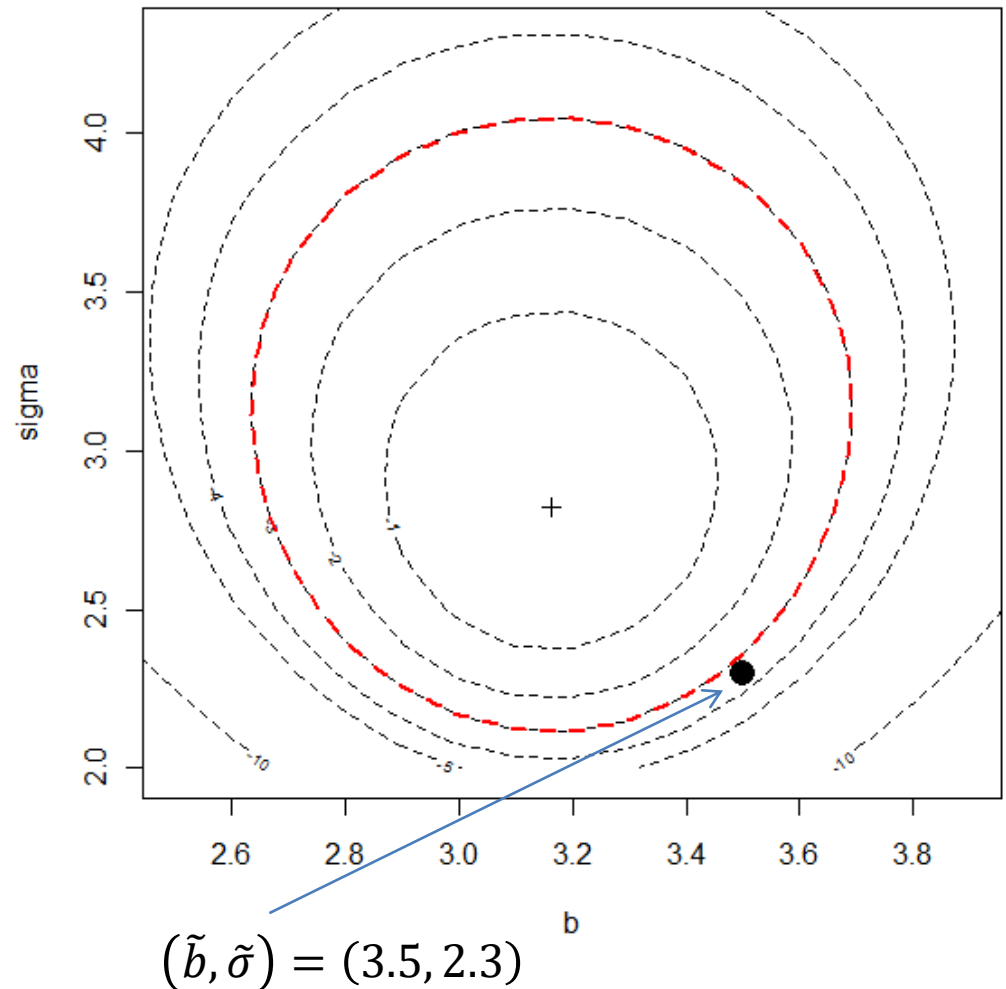
- Consider a set of values $(\tilde{b}, \tilde{\sigma}) = (3.5, 2.3)$

$\tilde{b} = 3.5$ alone is within the 95% C.I. for b

$\tilde{\sigma} = 2.3$ alone is also within the 95% C.I. for σ

But it is possible for the pair $(\tilde{b}, \tilde{\sigma}) = (3.5, 2.3)$ to lie outside the joint 95% C.I.

Multiple comparisons!

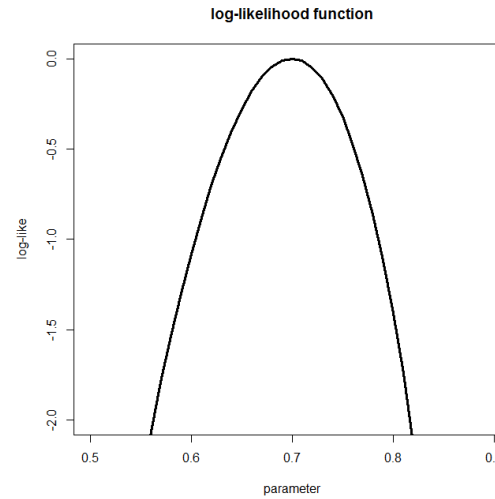
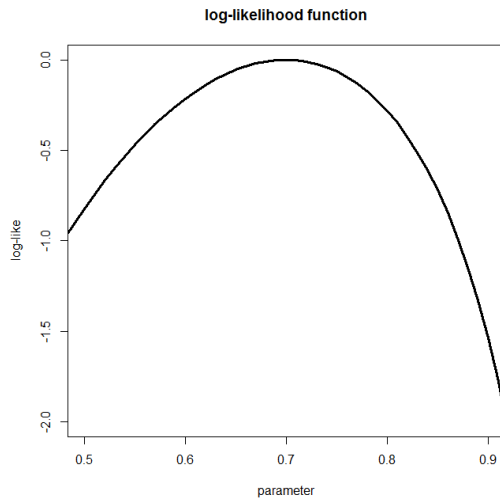


Profile likelihood

- We wish to focus on a subset of parameter(s)
- Partition the parameters into two subsets: $\underline{\theta} = (\underline{\theta}_1, \underline{\theta}_2)$, and aim to obtain the C.I. for $\underline{\theta}_1$ only
- We can perform profiling, partial maximisation of the original log-likelihood along $\underline{\theta}_1$
- $l^*(\underline{\theta}_1) = \max_{\underline{\theta}_2} l(\underline{\theta}_1, \underline{\theta}_2; \underline{x})$
 - Fix $\underline{\theta}_1$, then vary $\underline{\theta}_2$ such that the log-likelihood is (partially) maximised
 - Record down the maximised log-likelihood, and this is your $l^*(\underline{\theta}_1)$
 - Repeat this for a range of $\underline{\theta}_1$, then you get the profile log-likelihood function for $\underline{\theta}_1$
- The LRT statistic (and also C.I.) can be calculated using this profile log-likelihood

CI: Approximate normality of MLE

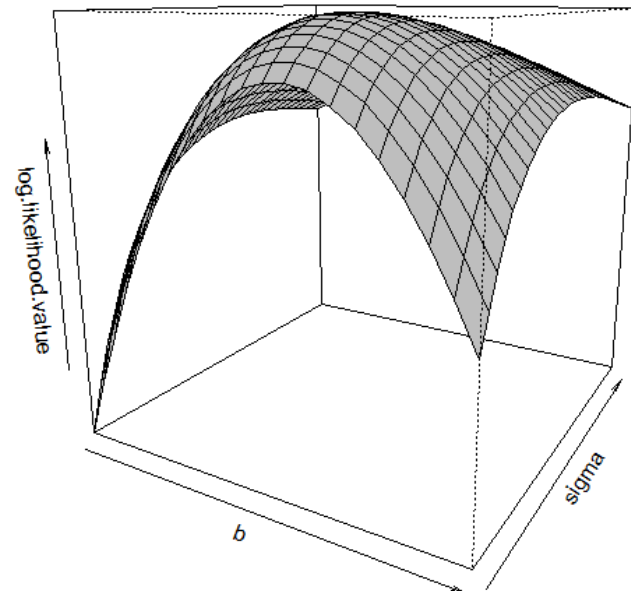
- One key property of ML estimators is asymptotic normality for reasonably large n
- For one-parameter case θ , the 95% C.I. for θ is approximately $\hat{\theta} \pm 1.96 \sqrt{\text{var}(\hat{\theta})}$, where the magical number 1.96 comes from the 2.5- and 97.5-percentile of a standard normal distribution
- But what is $\text{var}(\hat{\theta})$?
- The curvature of the log-likelihood function



The rate of change of slope at the peak!

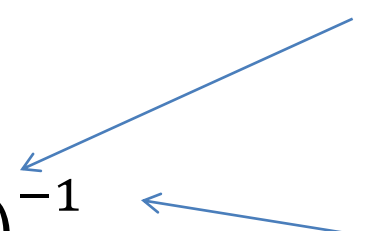
- Look at the two log-likelihood curves above: the right one is “steeper” around its peak
- Steepness = curvature = rate of change of the slope = the **second derivative** of the log-likelihood function
- more concave downwards -> narrower CI -> smaller variance
- $$var(\hat{\theta}) \approx -\frac{1}{l''(\hat{\theta})}$$
 - the second derivative of the log-likelihood function, evaluated at $\hat{\theta}$

- For multiple parameters, the ML estimators (asymptotically) follow a multivariate normal distribution.
- $V(\hat{\theta})$ is now a variance-covariance matrix



Univariate case:

$$\text{var}(\theta) \approx -\frac{1}{l''(\hat{\theta})}$$

- Empirically $V(\hat{\theta}) \approx -H(\hat{\theta})^{-1}$ 
- $H(\hat{\theta})$ is called the Hessian matrix, the second derivative of the log-likelihood function evaluated at its peak $\hat{\theta}$
- [OR] $-H(\hat{\theta})$ is called the *observed Fisher information matrix*.
- It measures the amount of information contained towards the parameters.
- $H(\hat{\theta})$ is readily available in `optim()`

Matrix inverse!

• Back to the rabbit data

```
# optim() FOR TWO-DIMENSIONAL PARAMETER SPACE, b AND sigma
# WITH HESSIAN MATRIX

result<-optim(par=c(1,1), regression.no.intercept.log.likelihood,
             method='L-BFGS-B',
             lower=c(-1000,0.0001), upper=c(1000,10000),
             control=list(fnscale=-1), dat=recapture.data, hessian=T)
# OBTAIN THE HESSIAN MATRIX
result$hessian
```

```
> result$hessian
           [,1]      [,2]
[1,] -2.365675e+01 -2.486900e-07
[2,] -2.486900e-07 -7.278254e+00
```

```
# THE VARIANCE-COVARIANCE MATRIX IS THE NEGATIVE OF
# THE INVERSE OF THE HESSIAN MATRIX.
# BY solve() FUNCTION
var.cov.matrix<-(-1)*solve(result$hessian)
var.cov.matrix
```

```
> var.cov.matrix
           [,1]      [,2]
[1,]  4.227123e-02 -1.444362e-09
[2,] -1.444362e-09  1.373956e-01
```

This is the variance-covariance structure of $(\hat{b}, \hat{\sigma})$

```
> var.cov.matrix
           [,1]      [,2]
[1,]  4.227123e-02 -1.444362e-09
[2,] -1.444362e-09  1.373956e-01
```

- $(\hat{b}, \hat{\sigma})$ follows (approximately) a bivariate normal distribution
- For example, 95% C.I. for b alone is $3.1629 \pm 1.96\sqrt{0.04227}$
- We can also apply multivariate testing to test for $H_0: (b, \sigma) = (b_0, \sigma_0)$
 - multivariate version of z-test
 - multivariate analysis (beyond the scope of this course)

Note on confidence interval

- “In Author’s experience, the Wald (i.e. normality) and likelihood method can give quite different results when used to test joint hypotheses... The likelihood method can require more effort to compute, but is generally preferred.” (Millar, 2011)

Day 5 selected topics

- Gamma MLE and moment estimator
- Memoryless r.v.
- Markov Chain
- Monte-Carlo integration
- Random-effect model
- Examples in population genetics

- Although MLE is “the best” sometimes it is hard to find
 - no closed-form solution (e.g. gamma MLE), and often can only be evaluated numerically
 - computational issue due to dimensionality (to many parameters)
 - the presence of nuisance variables (e.g. mixed effect model)

Memoryless r.v.?

- I arrive at a bus stop. Nobody is there. Let T be the waiting time before the arrival of the next bus.
 $T \sim \text{Exponential}(\lambda)$
- $f_T(t) = \lambda e^{-\lambda t}$ is the pdf
- $F_T(t) = \Pr(T \leq t) = 1 - e^{-\lambda t}$ is the cdf
- I will be late if a bus does not arrive in the next k minutes. The probability that I will be late for work is
- $\Pr(T > k) = 1 - \Pr(T \leq k) = 1 - F_{T(k)} = e^{-\lambda k}$

- Next day, I arrive at the same bus stop. “I have been waiting here for s minutes!” a person says to me.
- I am in the same situation, that if I do not get on a bus in the next k minutes then I will be in trouble.
- $\Pr(T > s + k | T > s)$ is the probability of being late
 - extra information about the r.v. T is given by the person
 - s minutes has passed, so it is known that $T > s$
- $$\Pr(T > s + k | T > s) = \frac{\Pr(T > s + k \text{ \& } T > s)}{\Pr(T > s)} = \frac{\Pr(T > s + k)}{\Pr(T > s)} =$$

$$\frac{e^{-\lambda(s+k)}}{e^{-\lambda s}} = e^{-\lambda k}$$
- WHAT???

- Memoryless: $\Pr(X > m + n | X > m) = \Pr(X > n)$
 - no extra information given
- Another example is the discrete geometric distribution
 - the number of Bernoulli trials required before getting the first success

Markov chain

- A random process, series of r.v., $X(t), X(t + 1), X(t + 2), \dots$
- There are several “states” (possible outcomes) that each $X(j)$ can take on
 - states can be discrete or continuous
- Transits from one state to another by chance over time
- The transition probabilities depend only on the current state
- The transition probability can be represented in a matrix form called Markov matrix
- Time-homogeneous Markov chain: A special case of Markov chain whose transition probabilities remain the same over time

- The four states of PonPon the rabbit



Sleeping (S)



Playing (P)



Eating (E)



Grooming (G)

- The Markov matrix the four states is

	S	P	E	G
S	0.5	0.1	0.2	0.2
P	0	0.4	0.3	0.3
E	0.3	0.1	0.6	0
G	0.25	0.25	0.25	0.25

JC69 substitution model

- Nucleotide substitution
 - mutation
 - four states: A, C, T, G

$$P = \begin{pmatrix} \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} \end{pmatrix}$$

- Other substitution models include
 - Kimura 1980, Felsenstein 1981 etc

Wright-Fisher model as a Markov chain

- For diploids, if the effective population size is N , then the possible number of alleles (states) are $\{0, 1, 2, \dots, 2N\}$.
- Assume there are two alleles: A and B
- The dimension of the Markov matrix is $(2N + 1) * (2N + 1)$
- Genetic drift changes the allele frequency over generations
- If the frequency of allele A is $k/2N$ now, then the number of allele A in the next generation follows $\text{binomial}(2N, k/2N)$

- For instance, for $N = 2$, there are five states: $\{0, 1, 2, 3, 4\}$ representing the number of allele A .
- The $\{i, j\}^{th}$ element of the transition matrix is the probability from state i to state j .

Jump to state j

```

> WF(2)
allele 0 1.00000000 0.000000 0.00000000 0.000000 0.00000000
      1 0.31640625 0.421875 0.2109375 0.046875 0.00390625
      2 0.06250000 0.250000 0.3750000 0.250000 0.06250000
      3 0.00390625 0.046875 0.2109375 0.421875 0.31640625
      4 0.00000000 0.000000 0.0000000 0.000000 1.00000000
allele  0         1         2         3         4

```

from state i

Example

- Given the Wright-Fisher transition matrix of $N = 2$. Let $X(t)$ be the number of allele A at time t .

```
> WF(2)
allele 0 1.00000000 0.000000 0.00000000 0.000000 0.00000000
      1 0.31640625 0.421875 0.2109375 0.046875 0.00390625
      2 0.06250000 0.250000 0.3750000 0.250000 0.06250000
      3 0.00390625 0.046875 0.2109375 0.421875 0.31640625
      4 0.00000000 0.000000 0.0000000 0.000000 1.00000000
allele 0 1 2 3 4
```

What is $\Pr(X(t + 1) = 3 | X(t) = 2)$?

What is $\Pr(X(t + 1) = 3 | X(t) = 0)$?

What is $\Pr(X(\textcolor{red}{t} + \textcolor{red}{2}) = 3 | X(t) = 2)$?

Some properties of Markov matrix

- Non-negative (the elements are probabilities, of course)
- Row sum to one
- If a Markov matrix \mathbf{M} is time-homogeneous, then the transition probabilities for T steps ahead is \mathbf{M}^T
- We can analyse the long-run behaviour of \mathbf{M} . Some Markov chains have limiting distributions, $\lim_{T \rightarrow \infty} \mathbf{M}^T$ exists.
- Some even have stationary distributions π , where $\pi \mathbf{M} = \pi$
- There are some states which you cannot leave once you have entered. They are called the **absorbing states**. For example, the first row and the last row of the Wright-Fisher model (Why?)



	S	P	E	G
S	0.5	0.1	0.2	0.2
P	0	0.4	0.3	0.3
E	0.3	0.1	0.6	0
G	0.25	0.25	0.25	0.25

 R Console

```
> m
      [,1] [,2] [,3] [,4]
[1,] 0.50 0.10 0.20 0.20
[2,] 0.00 0.40 0.30 0.30
[3,] 0.30 0.10 0.60 0.00
[4,] 0.25 0.25 0.25 0.25

> m%*%m%*%m%*%m%*%m%*%m%*%m%*%m%*%m%*%m%*%m%*%m%*%m%*%m%*%m%*%m%*%m%
      [,1] [,2] [,3] [,4]
[1,] 0.3 0.175 0.375 0.15
[2,] 0.3 0.175 0.375 0.15
[3,] 0.3 0.175 0.375 0.15
[4,] 0.3 0.175 0.375 0.15

> |
```

Stationary distribution $\pi = (0.3, 0.175, 0.375, 0.15)$

- For WF model M^{30} looks like this:

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	1.0000000	0.0000000e+00	0.0000000e+00	0.0000000e+00	0.0000000
[2,]	0.7499203	5.102345e-05	5.740139e-05	5.102345e-05	0.2499203
[3,]	0.4998937	6.803127e-05	7.653518e-05	6.803127e-05	0.4998937
[4,]	0.2499203	5.102345e-05	5.740139e-05	5.102345e-05	0.7499203
[5,]	0.0000000	0.0000000e+00	0.0000000e+00	0.0000000e+00	1.0000000

- According to the WF model, all alleles go fixed/extinct in 30 generations if $N = 2$
- Genetic drift reduces genetic variation!

Other Markov processes

- Moran model
 - alternative to Wright-Fisher model
 - allows overlapping generations
- Birth and Death process
 - continuous-time Markov process
 - birth: $\text{state}+1$; death: $\text{state}-1$
 - infinitely many states

Monte Carlo integration

- To evaluate the following integral: $I = \int_0^1 \sqrt{1-x^2} dx$
- Calculate by hand... ☹️
- Numerical methods such as quadrature rule
 - counting areas of rectangles / trapeziums
 - `integrate()` in R

```
integrate(function(x) {sqrt(1-x^2)}, lower=0, upper=1)
```

Monte Carlo integration

- $I = \int_0^1 \sqrt{1 - x^2} dx$
- Sample $\{x_1, x_2, \dots, x_n\}$ from *uniform*(0, 1) distribution
- Compute $I_n = \frac{1}{n} \sum_{i=1}^n \sqrt{1 - x_i^2}$
- I_n is an approximation to the integral I , if n is reasonably large

Justification

- $I = \int_a^b g(x) dx = \int_a^b \frac{g(x)}{f(x)} f(x) dx = E_X\left[\frac{g(X)}{f(X)}\right]$
 - where X is a r.v. with pdf f and support (a, b) . The integral becomes the expectation of the transformed r.v. $\frac{g(X)}{f(X)}$.
- Remember, expectation is the population mean, the average of infinitely many trials, which can be “replicated” by computer
- Draw $\{x_1, \dots, x_n\}$ from f , $I_n = \frac{1}{n} \sum_{i=1}^n \frac{g(x_i)}{f(x_i)}$ is a good approximation to I for sufficiently large n

- $\int_0^{\infty} x e^{-2x} dx$

- $\int_0^{\infty} x e^{-2x} dx$
- Let $X \sim \text{Exponential}(\lambda = 1)$, $f_X(x) = e^{-x}$

$$\begin{aligned} & \int_0^{\infty} \frac{x e^{-2x}}{e^{-x}} e^{-x} dx \\ &= \int_0^{\infty} x e^{-x} e^{-x} dx \\ &= E_X[X e^{-X}] \end{aligned}$$

- Sample $\{x_1, x_2, \dots, x_n\}$ from *Exponential*(1) distribution for some large n
- $I_n = \frac{1}{n} \sum_{i=1}^n x_i e^{-x_i}$ is an approximation to the original integral

```
x<-rexp(1e6, 1)
mean(x*exp(-x))
```

- Stochastic simulation -> no deterministic answers (unlike `integrate()` in R)
 - “random” answers
- Able to work with $\pm\infty$ bounds
 - as long as the chosen r.v. is defined in those bounds (e.g. normal)

Exercise

- Evaluate $\int_{-\infty}^{+\infty} e^{-x^2} dx$ via MC integration
 - which r.v. should you use? Those with $\pm\infty$ bounds perhaps?
 - some choices of r.v. are better than the others

- The intrinsic variance of MC integration is unavoidable
 - slow convergence
 - variance $\propto \frac{1}{n}$ as each draw is independent
 - need large n
- Multivariate integrals -> Multivariate distributions
 - requires samples from a multivariate distribution
 - requires even more sampling points
- There are ways to reduce variance (beyond this course)
 - e.g. importance sampling
- It is also possible to use dependent samples
 - Markov Chain Monte Carlo (MCMC)

MCMC integration

- $\int_a^b g(x)dx \approx \frac{1}{n} \sum \frac{g(x_i)}{f(x_i)}$
- In pure MC integration we assume $\{x_1, \dots, x_n\}$ are independent
- In fact the above approximation still holds for correlated $\{x_1, \dots, x_n\}$
- Sometimes it is easier to generate a series of dependent $\{x_1, \dots, x_n\}$
- One can (smartly) construct a Markov chain, whose stationary distribution is f
 - Gibbs sampling
 - Metropolis-Hastings (MH) algorithm

Example 1.1: Estimating haplotype frequencies and LD

- Two-locus, two-allele setting
- Four haplotypes: AB, Ab, aB, ab
 - with true haplotype frequencies $p_{AB}, p_{Ab}, p_{aB}, p_{ab}$
 - sum of four haplotype frequencies = 1

	B	b	
A	p_{AB}	p_{Ab}	
a	p_{aB}	p_{ab}	
			1

- If haplotypic data is obtained then the MLE is

- $\widehat{p}_{AB} = \frac{\text{\# of } AB \text{ haplotype observed}}{\text{total haplotype sample size}}$
- same for all four haplotype frequencies

- $$\widehat{r^2} = \frac{(\widehat{p}_{AB}\widehat{p}_{ab} - \widehat{p}_{Ab}\widehat{p}_{aB})^2}{\widehat{p}_{A\cdot}(1 - \widehat{p}_{A\cdot})\widehat{p}_{B\cdot}(1 - \widehat{p}_{B\cdot})}$$

- according to the invariant principle, $\widehat{r^2}$ is also the MLE for r^2 , the standardised LD coefficient
- heavily biased for small sample size

Table 1 Expected genotypic frequencies under HWE

	<i>BB</i>	<i>Bb</i>	<i>bb</i>
<i>AA</i>	$f_1 = p_{AB}^2$	$f_2 = 2p_{AB}p_{Ab}$	$f_3 = p_{Ab}^2$
<i>Aa</i>	$f_4 = 2p_{AB}p_{aB}$	$f_5 = 2(p_{AB}p_{ab} + p_{Ab}p_{aB})$	$f_6 = 2p_{Ab}p_{ab}$
<i>aa</i>	$f_7 = p_{aB}^2$	$f_8 = 2p_{aB}p_{ab}$	$f_9 = p_{ab}^2$

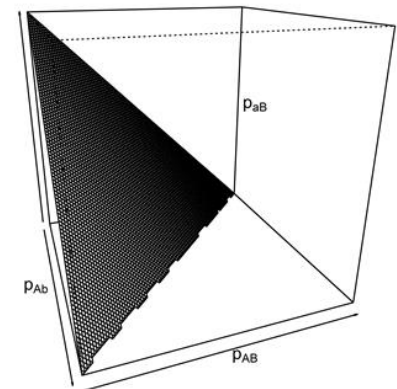
The expected frequency of genotypes given the haplotype frequencies under HWE [2]. All the expected frequencies f_1, f_2, \dots, f_9 add up to one

- Sometimes only genotypic information is obtained.
- Nine genotypes, each has an expected frequency under HWE.
- The genotype counts $\{n_1, n_2, \dots, n_9\}$ are assumed to follow a multinomial distribution with size n and expected frequencies $\{f_1, f_2, \dots, f_9\}$
- The log-likelihood function is

$$l(p_{AB}, p_{Ab}, p_{aB}, p_{ab}) = \text{constant} + \sum_{i=1}^9 n_i \log(f_i)$$

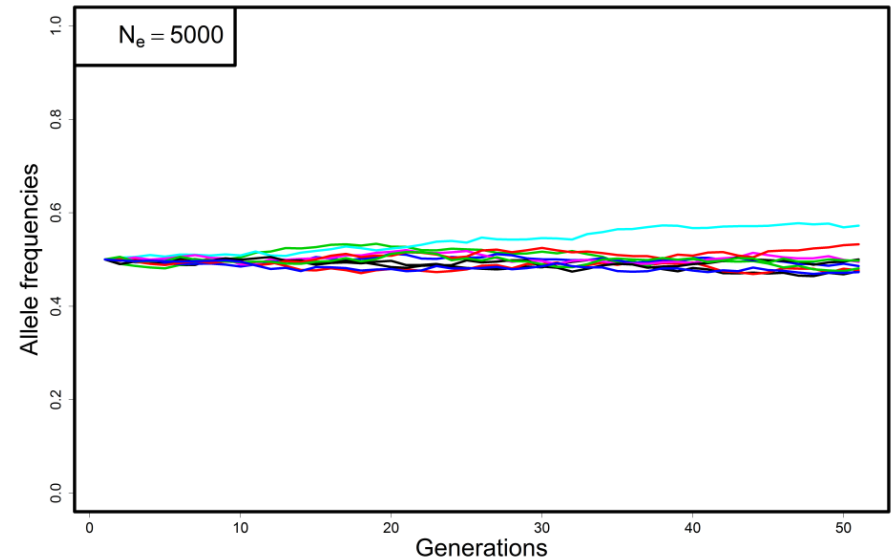
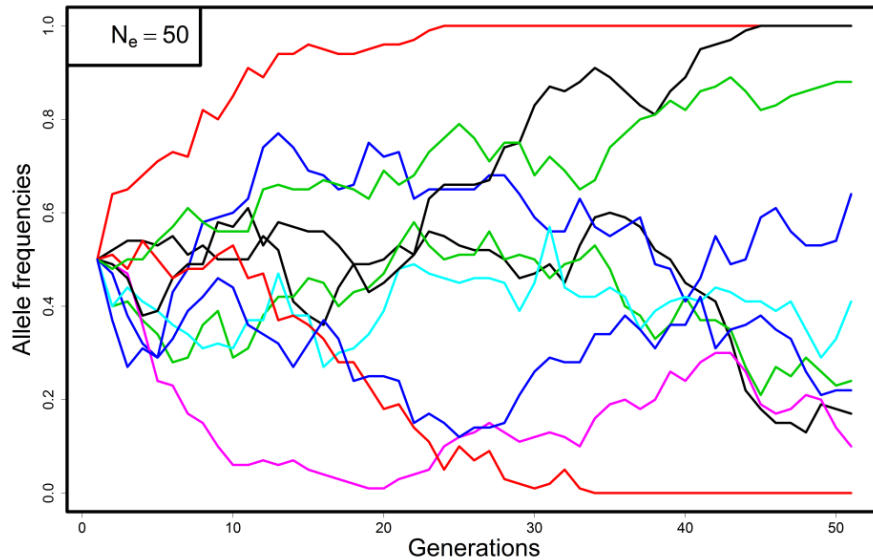
- In theory, we can maximise the above log-likelihood to find the MLE for the four haplotype frequencies, but...

- There are challenges:
 - no closed-form solution
 - The four haplotype frequencies add up to 1; the fourth frequency is redundant. Hence the parameter space is a tetrahedron
 - sometimes there are multiple points with zero gradients (saddle / local points)?



- Possible solutions:
 - Expectation-Maximisation (EM) algorithm (Excoffier & Slatkin 1995)
 - some kind of transformation before maximising the log-likelihood? into a cube? (Hui and Burt, 2020)

Example 2: Drift and population size



- Model: Wright-Fisher model (the Markov matrix)
- Parameter of interest: N , the effective population size
- Data: The allele frequencies at two time points, across many unlinked loci.
- So why not MLE???

Transition prob, from WF matrix

- $L(N) = \sum_{all\ p_t} \sum_{all\ p_0} f(x_t|p_t) f(p_t|p_0, N) f(x_0|p_0)$

Sampling at time t

Sampling at time 0

- The two sampling $f(x_t|p_t)$ and $f(x_0|p_0)$ are modelled by binomial distributions, independently
- The transition probabilities $f(p_t|p_0, N)$ can be obtained from the WF matrix M^t

- “The likelihood of observing x_0 and x_t , given the true allele frequencies p_0 and p_t , and effective population size N ”
 - then sum these likelihood values over all possible p_0 and p_t
 - state-space model
 - from 0 to $2N$
- Williamson & Slatkin (1999); Hui & Burt (2015)

Summary - MLE

- Day 1: Common r.v. and their pmf/pdf. Expectation. Moment generation functions.
- Day 2: Multivariate r.v., independence. Define likelihood functions. The triplet: model, data, parameters. Maximisation via differentiation and `optim()`.
- Day 3: Properties of MLE. Likelihood-ratio test. Logistic regression.
- Day 4: C.I. by log-likelihood. C.I. by approximate normality. Joint confidence region. Profile likelihood.
- Day 5: Examples and more examples

Beyond this course

- R functions and packages that help implement MLE
 - `mle()`, `confint()`
 - `{stats4}`
- Alternative optimisation routines
 - `nlm()`, `nlminb()`
 - `{optimx}`, `{lbfgsb3}`, `{BB}`, ... etc
- Require some 'grammatical' changes
- Multivariate testing

(Possible) solutions

- Approximation to the integrals (e.g. Laplace approximation), EM algorithm
- Statistical sampling (e.g. Monte Carlo, MCMC)
- Approximate Bayesian Computation (ABC)
- More Statistics!

MLE is...

- Not just a method, but THE method
- A collection of methods that share a common belief towards how “the best parameters” should be
- Many canned software and functions make use of the results from MLE (with or without acknowledging it)