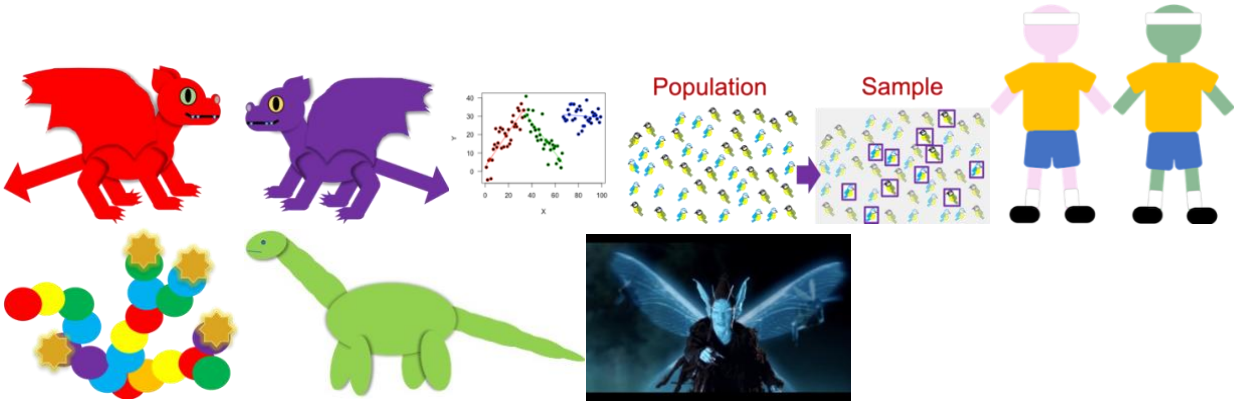
Summary of ST2304 2021

By Emily G Simmonds



This PDF contains a broad summary of the course ST2304.

A: Core concepts/skills

These are concepts, ideas, and skills that spread across the whole of the course.

They are:

- **Maximum likelihood estimation.** This was introduced in Module 1.

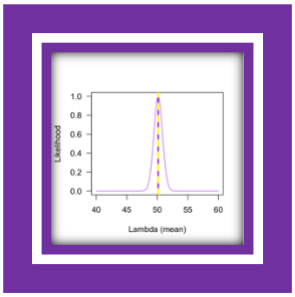
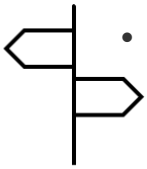


Figure 1: Likelihood for a lambda (mean of Poisson distribution)



- **Choosing an appropriate model.** This is a skill that is used throughout the course. To achieve this skill, you need to be able to look at the data you have and work out which of the models you know (or can find out about via Google) will capture how those data were generated. An example of this, would be choosing a Poisson GLM to model counts of the abundance of turtles as a result of sea surface temperatures. If instead you had data on the height of giraffes that you assume is influenced by the habitat they live in, this would use either a GLM with a Gaussian distribution for the error or a linear model.

- **Interpreting statistical estimation.** This skill fits closely to maximum likelihood estimation. It is being able to interpret the meaning of estimates of parameters in terms of the original question that was being asked.

- **Uncertainty in statistical estimation.** Following on from being able to interpret estimates of parameter values, we have also looked quantifying uncertainty in those estimates. This is an essential skill to be able interpret the results of a statistical analysis correctly.

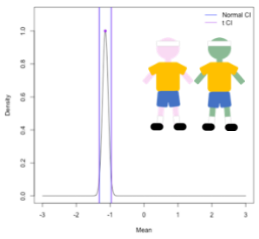


Figure 2: Example of confidence intervals for t- and normal distributions

- **Model checking (including Model selection).** All models covered in this course have assumptions. We should always check whether the data we give these models meet the assumptions required for the method to work correctly. We also use model selection to see which explanatory variables we should include.

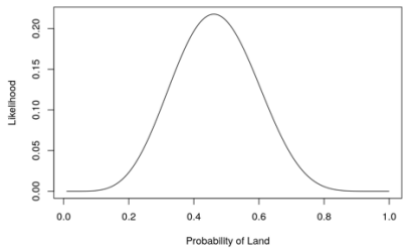
- **Finding the 'best' model.** This focusses on how you can choose which explanatory variables to include, called model selection.

B: Model types

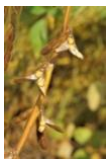
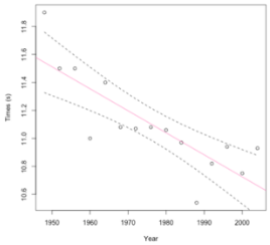
There are several different types of statistical model that we have used in this course. These can be considered tools in your modelling toolkit. You can apply the core concepts list above to each of these models. How this is done is detailed in section C.

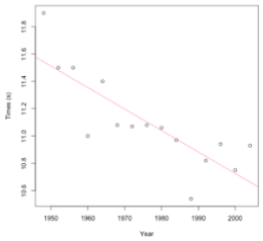
The models are:

- Distributions (Module 1, Module 2, Module 3, Exercise 1, Exercise 2)

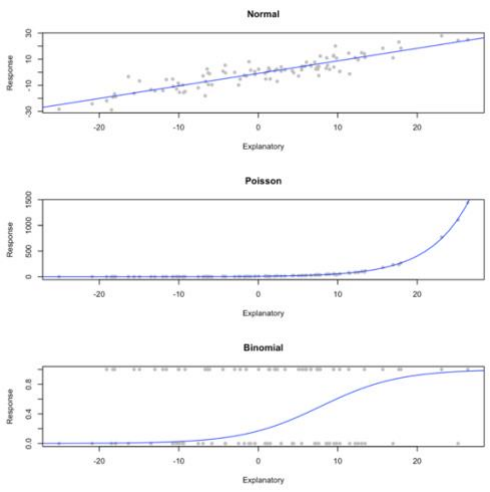


- Linear models (Module 3, Exercise 2, Module 4, Exercise 3, Module 5, Exercise 4, Module 6, Exercise 5, Module 7, Exercise 6,Module 8, Exercise 7, Module 9, Exercise 8)



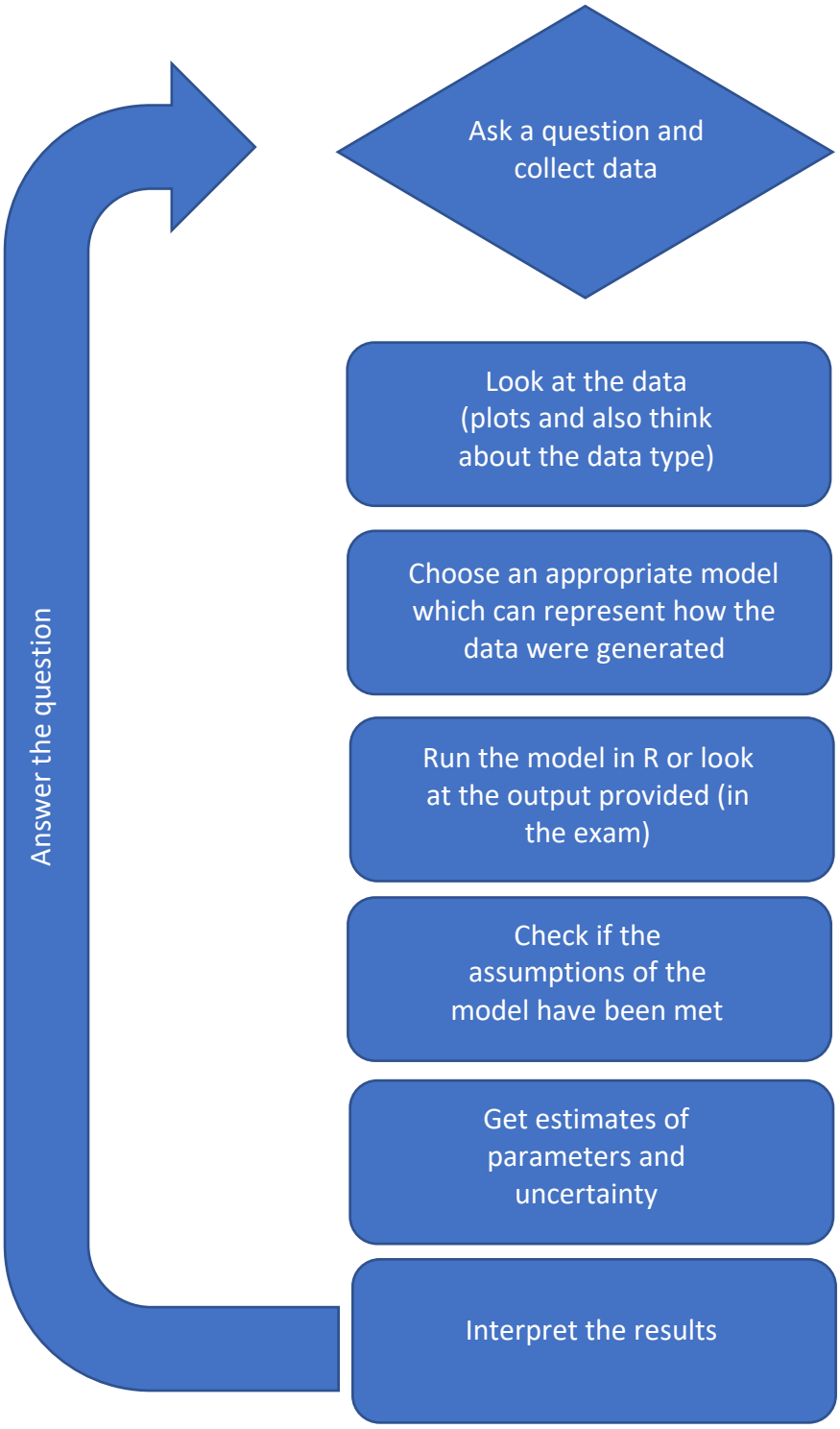

- Generalised linear models (Module 10, Exercise 9, Module 11, Exercise 10, Module 12, Exercise 11)


C: How to match core concepts to the different models

Concept	Distribution	Linear Model	Generalised Linear Model
How to choose a model?	Which distribution best characterises the data?	Is the response data normally distributed? Do you think there is a causal relationship between y and x? Is that relationship linear? If yes to all, use LM	Is the response data normally distributed? Do you think there is a causal relationship between y and x? Is that relationship linear? If no, yes, no, use a GLM (can also use for yes, yes, yes but it is no different to an LM).
Parameter estimates	Parameters of the distribution	alpha (intercept) and betas (differences in group mean or slope)	alpha (intercept) and betas (differences in group mean or slope) on the link scale
Uncertainty	Standard errors and confidence interval	Standard errors and confidence interval	Standard errors and confidence interval
Model checking	Do the characteristics of the data match those of the distribution	Use diagnostic plots of the residuals	Use diagnostic plots of the SCALED residuals
Model selection	•	anova (confirmatory) or AIC/BIC (exploratory)	analysis of deviance (confirmatory) or AIC/BIC (exploratory)
Interpretation	Use CIs and the estimate to say something about the distribution that the data came from	Use CIs and the estimate to say something about how x (or xs) influence y	Use CIs and the estimate to say something about how x (or xs) influence y, you will need to take the inverse of the link function to interpret on the original scale

D: General workflow for modelling



E: Objectives from each module (what you should be able to do at the end of each module)

Overall learning outcomes of the course:

- To be able to fit appropriate linear models to data (including regression and ANOVA)
- To be able to assess the fit of the linear model
- To be able to compare models and decide which is 'best'
- To be able to appropriately fit GLMs: (including logistic regression and log-linear models)
- To be able to interpret a GLM
- To be able to use maximum likelihood estimation principles to interpret results of statistical models

Module	Exercise	By the end you should be able to:
1	-	Run R through RStudio and use simple commands (such as plot(), exp(), read.csv()) Explain what: objects, assigning, functions, packages are Import data and scripts
2	-	Explain the difference between likelihood and probability Plot a likelihood curve for a Binomial distribution and find the maximum likelihood estimate of the probability Explain how the likelihood can be used to find the 'best' parameters
3	1	Explain what the maximum likelihood estimate for one

		parameter is, including confidence intervals Calculate the confidence intervals of a maximum likelihood estimate (both Poisson and Binomial distributions), using the normal approximation Interpret and communicate the results of maximum likelihood estimation of a single parameter, including the uncertainty
4	2	Recognise a normal distribution and know what its parameters are Calculate the maximum likelihood estimates of the parameters of a normal distribution and construct confidence intervals Interpret the results of the maximum likelihood estimate and the confidence intervals (the example this week is in a t-test) Make a decision based on your results Recognise a linear model formula
5	3	Explain what a linear model is and how regression fits into this category Determine when a linear regression is an appropriate model for data

		Fit a linear regression model using the lm() function in R Interpret the estimates from the linear regression, including confidence intervals, in a statistical and biological context (i.e. back in original units - women get 0.016 seconds faster at 100m per year)
6	4	Explain the difference between a good and not so good model Explain what an R^2 is and how it relates to how good our model is Check whether the model fits the data, or if it suffers from: non-linearity, heteroscedasticity, or outliers Improve the model to overcome the above deficiencies
7	5	Write out a multiple regression model in R and as an equation Choose when a multiple regression is appropriate Write a model in matrix form Fit a polynomial model in R (e.g. with x^2) Interpret the outputs of these models in a biological context

8	6	Determine when a variable is continuous or categorical (a factor) Fit a simple ANOVA model in R (as an LM) Write the ANOVA model in matrix form Interpret the output of the ANOVA in a biological context
9	7	Fit an interaction in a linear model and interpret the result Explain why an interaction term might be needed Fit a linear model with one continuous and one categorical variable and interpret the result
10	8	Explain why model selection is needed Use the AIC to compare models Compare hypotheses with F tests/ANOVA
11	9	Describe what the parts of a GLM are (distribution, linear predictor, link function) Determine when to use a GLM rather than just a linear model Fit a GLM in R

		Extract and interpret the parameters of the GLM
12	10	Determine when to use a binomial/logistic regression Fit a logistic regression in R Interpret the output of the regression in a biological context Name and explain some of the common link functions in glm()
13	11	Determine when to use a log-linear/Poisson regression Explain how this differs from a simple linear regression/linear model Fit a Poisson regression in R Extract the parameters from this model Interpret the parameters of a log-linear model