Welcome to ZOL851: Statistical Methods in Ecology and Evolution

An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.

John Tukey

Your guide: Dr. Ian Dworkin

$$Y \sim N(b_0 + b_1 X, \sigma^2)$$

Why we are here….

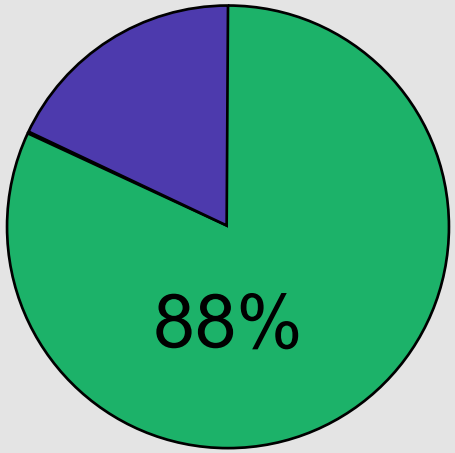
- http://www.youtube.com/watch?v=PbODigCZqL8

Goals for today

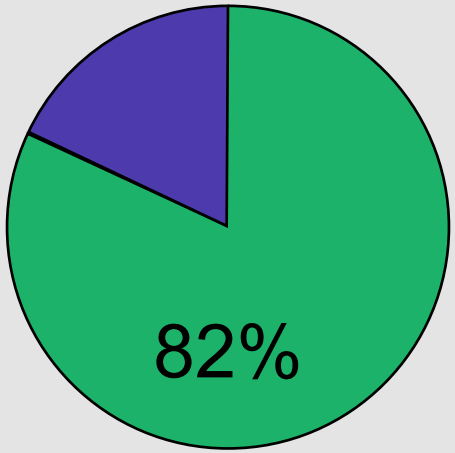
- Introductions.
- Who am I?
- Who are you?
- Goals of the course.
- Syllabus.
- Course content and evaluation
- P-values..
- What is the goal of statistics (discussion).

Variation in Ecology & Evolution

Statistics are everywhere in Ecology and Evolution



Ecology



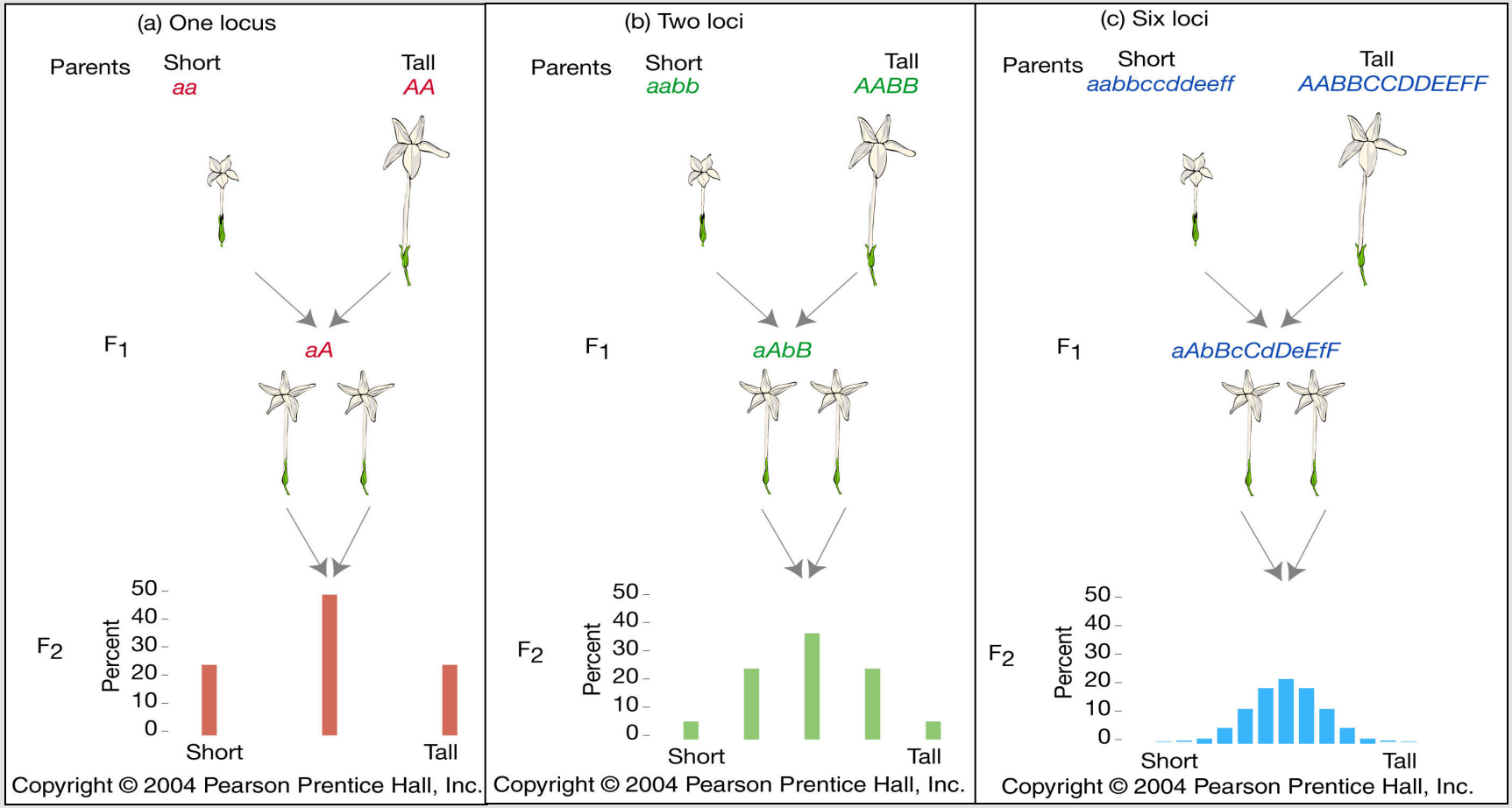
Evolution

- data from July 2005 issues

Sources of Variability

- Measurement error
- Random variation (demographic, environmental…)
- Complexity
  - Deterministic processes can lead to variability

How do we deal with variability and uncertainty?



"We shall term this quantity the variance."

- Fisher, 1918

Variation

Introductions

Ian Dworkin

Office: 056 Giltner Hall

Phone: 432-6733

E-mail: idworkin@msu.edu

(Please include “ZOL851” in subject)

Office hours: Tuesday 12:30-1:30

Evolutionary/Quantitative Geneticist

Dworkin Quantitative & Statistical Biography

Calculus, Algebra No Stats	Quantitative Methods	Generalized Linear Models
B.Sc.	PhD	Post Doc+
Biology	ANOVA ANCOVA Regression Principal Components Analysis MANOVA Logistic Regression Log-linear models	Quantitative Genetics Population Genetics Maximum likelihood GLiM Resampling Monte Carlo Simulations Bayesian & MCMC
Abacus, Excel	Mathematica, Statistica, SAS	SAS, R

Introductions.. Of each other.

- Name
- Department
- Brief outline of research
- Brief statistical background

My Goals

1. Encourage you to think about the way you do science, and the use of models in statistics.
2. Provide an ***overview*** of modern statistical techniques. In particular computationally intensive methods of estimation and inference.
3. Introduce you to a powerful programming environment for statistical analysis and graphing.
4. Facilitate your ability to learn new statistical techniques **on your own**.
5. Give you hands-on experience analyzing **YOUR** data and presenting results relevant to ecology and evolution.

Not my goals:

- Memorize formulae.
- Promote point and shoot stats.
- Survey the multitude of statistical errors to scare you from attempting new or more complex statistical techniques.

Productive failure

- Learning requires productive failure.
- Do not expect to get everything in this course the first time.

- Also see.

Productive Stupidity

Work load - readings

- This class has a well deserved reputation for a heavy work load, in particular for readings.
- In this class we often need to introduce the philosophy behind a statistical approach as well as the approach itself, and how it compares to related ideas.

Work load - programming

- We will be doing a lot of programming in R.
- While you do not need to have any background in using R, or programming in general, it can take a lot of time to learn.
- You need a zen like patience at the beginning. See the slide about productive failure..
- In addition to getting help from me or the TA, I also recommend that you “pair up” with someone with some programming experience (preferably in R).
- Raise hands!

Work load - thinking

- Since my goal is not to promote “point and click” statistical approaches, you will need to spend a great deal of time thinking about how to use certain methods with respect to your data.
- Indeed, one of the most important questions to ask is whether such an approach is useful to address the question at hand.

Work load – Researching the methods that work for you.

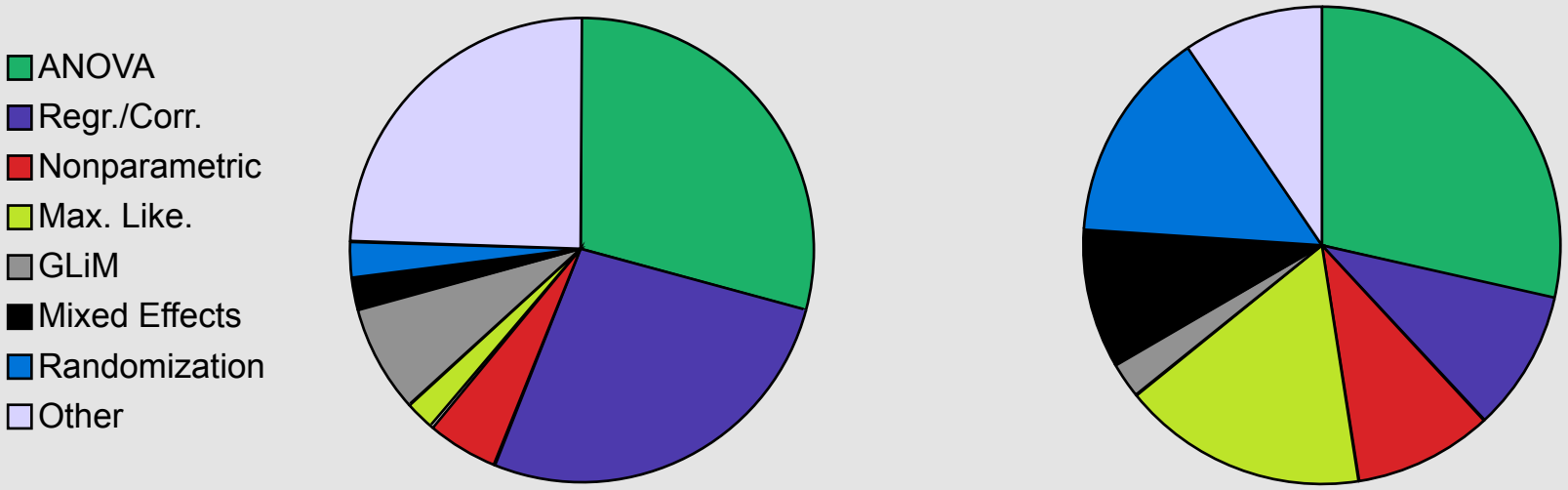
- Since you will be working with your data (or a dataset that is otherwise important or interesting to you), it will be up to you to figure out what approaches you ultimately should be using for your final paper.
- I will happily offer guidance, and feel free to discuss this members of the class.
- It is not appropriate for your approaches to be conceived of or models built by your PI or statistical consultants.

What are your goals?

Many ways of approximating the same result

- Everything is statistics that is valuable is an approximation!!!!
- We will spend some time comparing different methods for estimation & inference (LSE, MLE,Monte Carlo, Bayesian, Resampling).
- Sometimes they give similar results, sometimes not. We will discuss when each has its advantages, and downsides (yes, they are all valuable).

Which statistics are being done in Ecology and Evolution?



Ecology

Evolution

- data from July 2005 issues

Statistical Concepts to be Covered

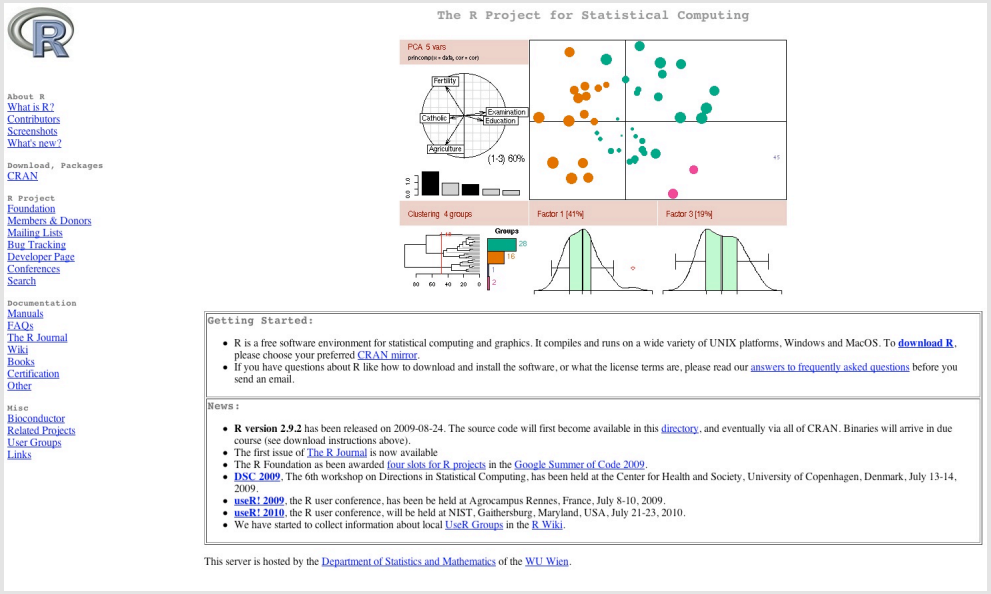
- Philosophical basis and approaches to statistics
- Measurement Theory
- Estimation of Effect sizes.
- A field guide to probability
- Monte Carlo simulation
- Maximum Likelihood Estimation
- Bayesian estimation
- Resampling methods
- GLM: ANOVA & Regression reviewed
- Model Selection
- Generalized Linear Models (GLiM)
- Mixed Effects Models
- Process Models

Concepts not covered.

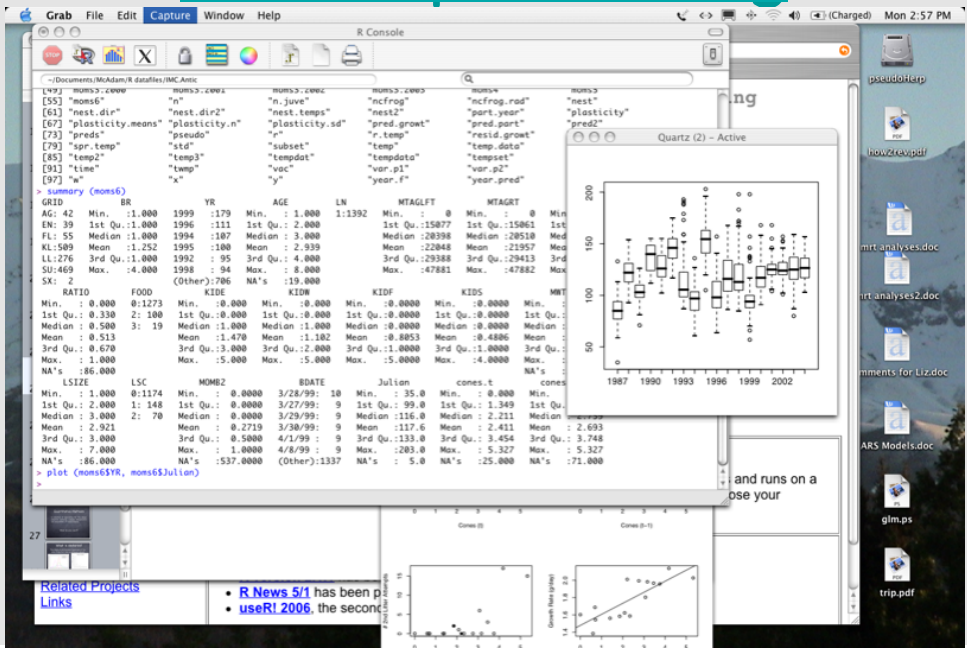
- Sadly there are many important statistical topics and methods that biologists use regularly that we will not cover.

- An important goal is to give you the confidence to go and learn those you need.

- Powerful statistical package
- Multi-platform
- Ongoing development
- On-line manuals, support, newsletter
- Free



www.r-project.org



R

Why R rocks: a sermon from a reformed SAS programmer

Please install R 2.15.1 on Your computer

R style guide for writing your code.

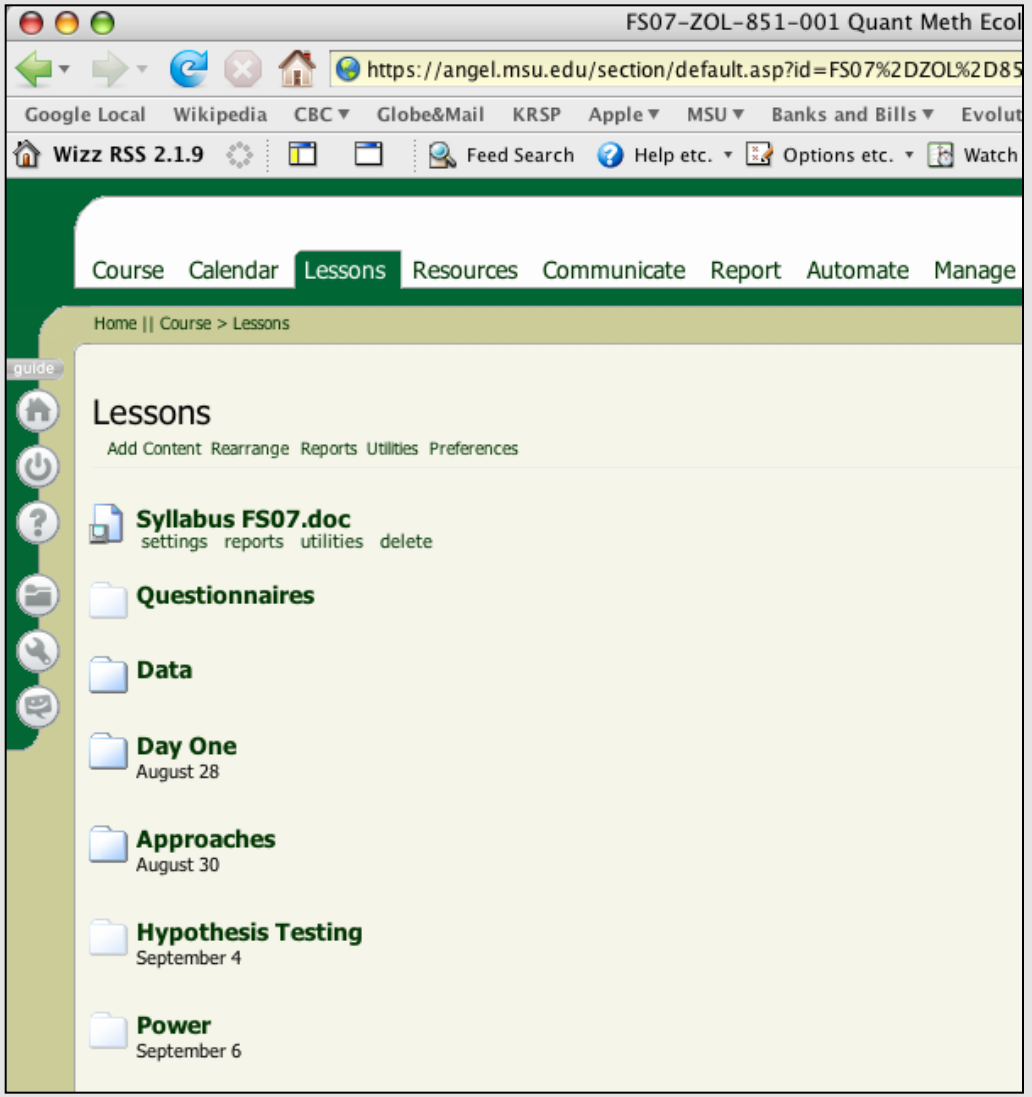
- This year we will be using a “style guide” for our programming, so that all of us are following the same basic conventions for naming of files, variables, spacing, etc…

Course Materials

- Lecture Notes (PowerPoint slides)
  - "Lecture1_ID_2011.ppt" old .Posted before lecture
  - "Lecture1_ID_V1.ppt" Posted before lecture
  - "Lecture1_ID_final.ppt" posted after lecture
- R data and code
  - "Lecture 1 script" posted prior to lecture
- Readings
  - 3 primary texts + some additional books (available online)
  - Relevant readings from the primary literature
  - Not required unless otherwise stated

ANGEL

www.angel.msu.edu



The texts for the course

- Bolker, B. 2008. Ecological Models and Data in R.
- Gelman,A. & Hill, J. 2007. Data Analysis using regression and multilevel/ hierarchical models.
- Dalgaard. 2009. Introductory Statistics with R (available online at MSU library).

Homework Assignments

- Reinforce concepts discussed in class
- Encourage you to think about **YOUR** approach to science
- Give you hands-on experience with analysis of ecological/evolutionary data
- Provide you with experience using a powerful statistical package (R)

P value's and homeworks

- We will talk in some depth in this course about the merits of using "p values" as a tool for inference, model assessment and selection.

I hope you will come to appreciate them as a tool, as they have a role, but a very limited one!!! Thus…

No P-values in your assignment**

- OMG!!! Are you crazy Ian?!
- How can you have a statistics course without p-values?

P-values tend to be a crutch, and a broken one at that.

- Many researchers incorrectly use p-values as an all in one tool for inference & model assessment (among others).
- Instead in this class we will focus on approaches that are more *generally* appropriate (in particular using confidence intervals for inference and other tools we will learn for model assessment).

When can I use p-values?

- There may times when p-values are appropriate to report in your assignments and final paper.
- However if you do use them, I will require a reason about why you are using them as opposed to other tools. (i.e. why are they the most appropriate tool).

**When can I use p-values?

- Also some of our discussion of resampling methods, simulations and power analysis will involve p-values…

When beginning an analysis

Start from the biological question first

What question are you trying to answer?
What kind of data do you need to collect to answer that question?
Observational data? Controlled experiments?
 Are you just looking for patterns?
 Estimating parameters for a model?
 Predicting future observations?
 Testing between explicit hypotheses?

What kind of analyses should you be doing based on the above?
How do you relate the results back to the question?

Goals for an analysis

- When performing an analysis for assignments make sure the goal of performing such an analysis are clear.

- (exploratory analysis, estimating parameters for a model, for making predictions, testing between explicit hypotheses…).

Final Project

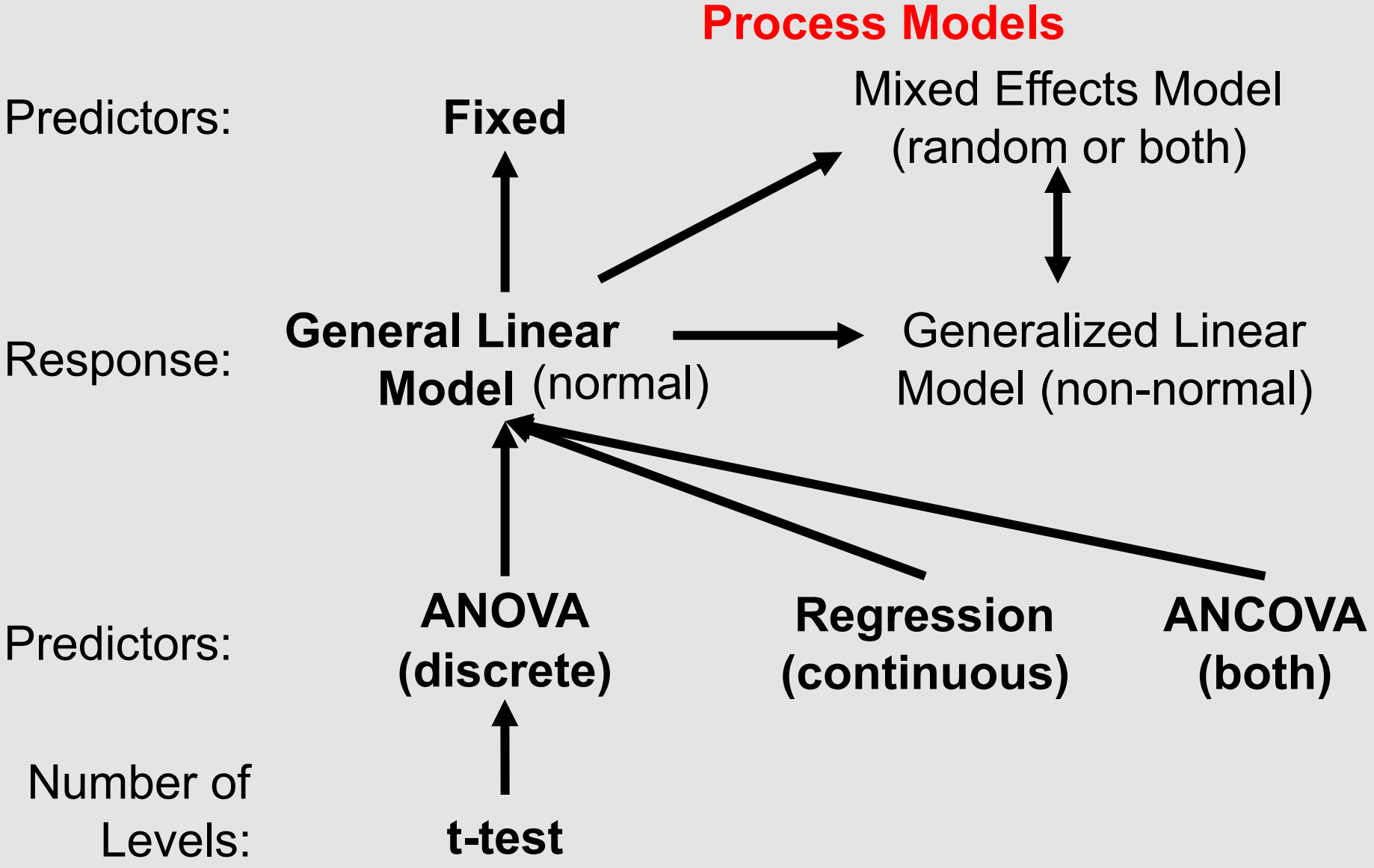
- You will analyze a dataset of your own using R and techniques learned in this course.

- Requires a dataset with:
  - At least 50 records
  - A continuous response variable
  - A discrete response variable as well if possible (0 or 1; Counts)
  - At least 3 covariates (continuous and categorical)

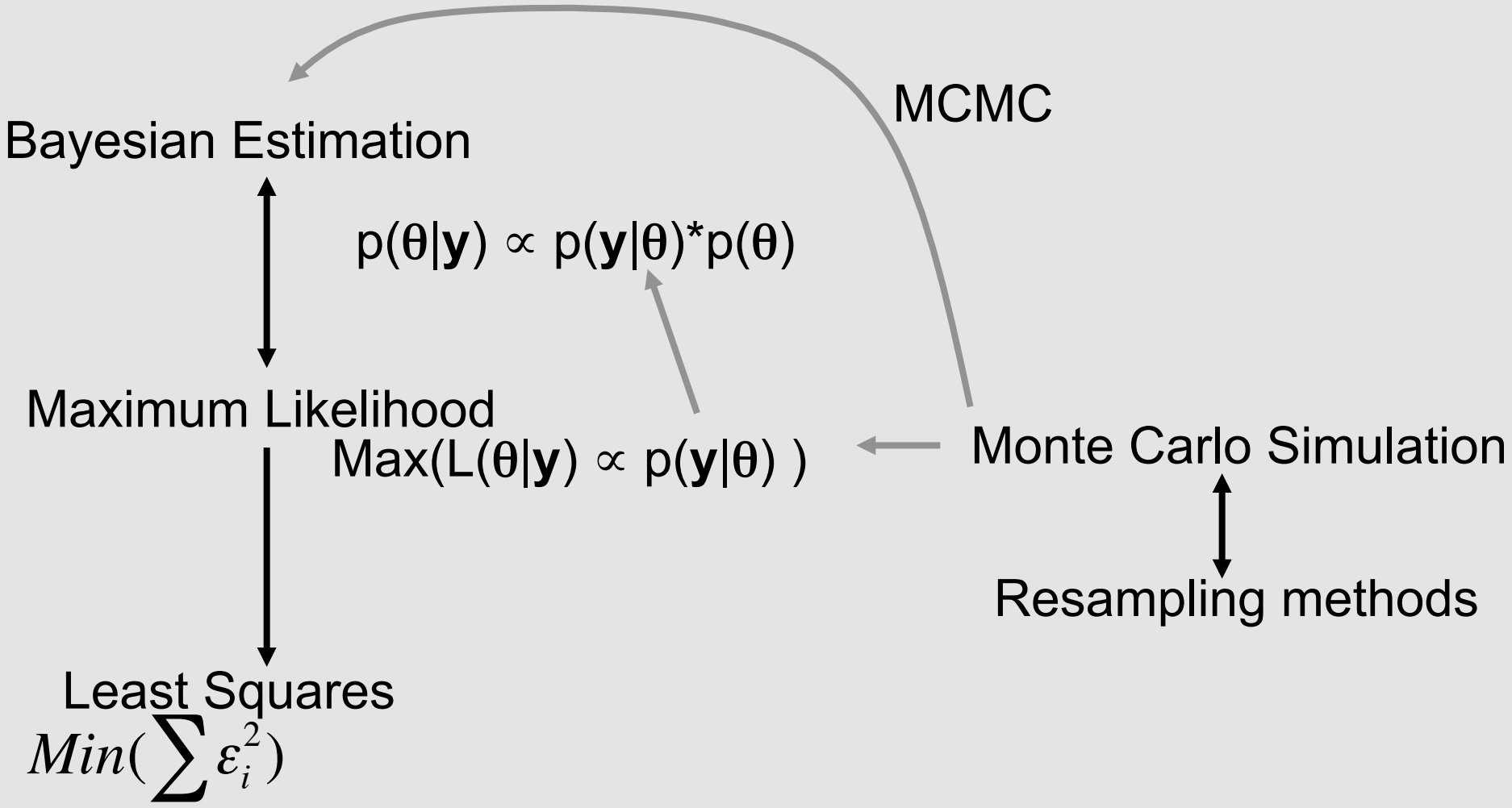
Your tasks for Tuesday…

1. Think about sources of variation in your own data. Why are you here?
2. Start to search out a database for your final project.
3. Read the papers on ANGEL

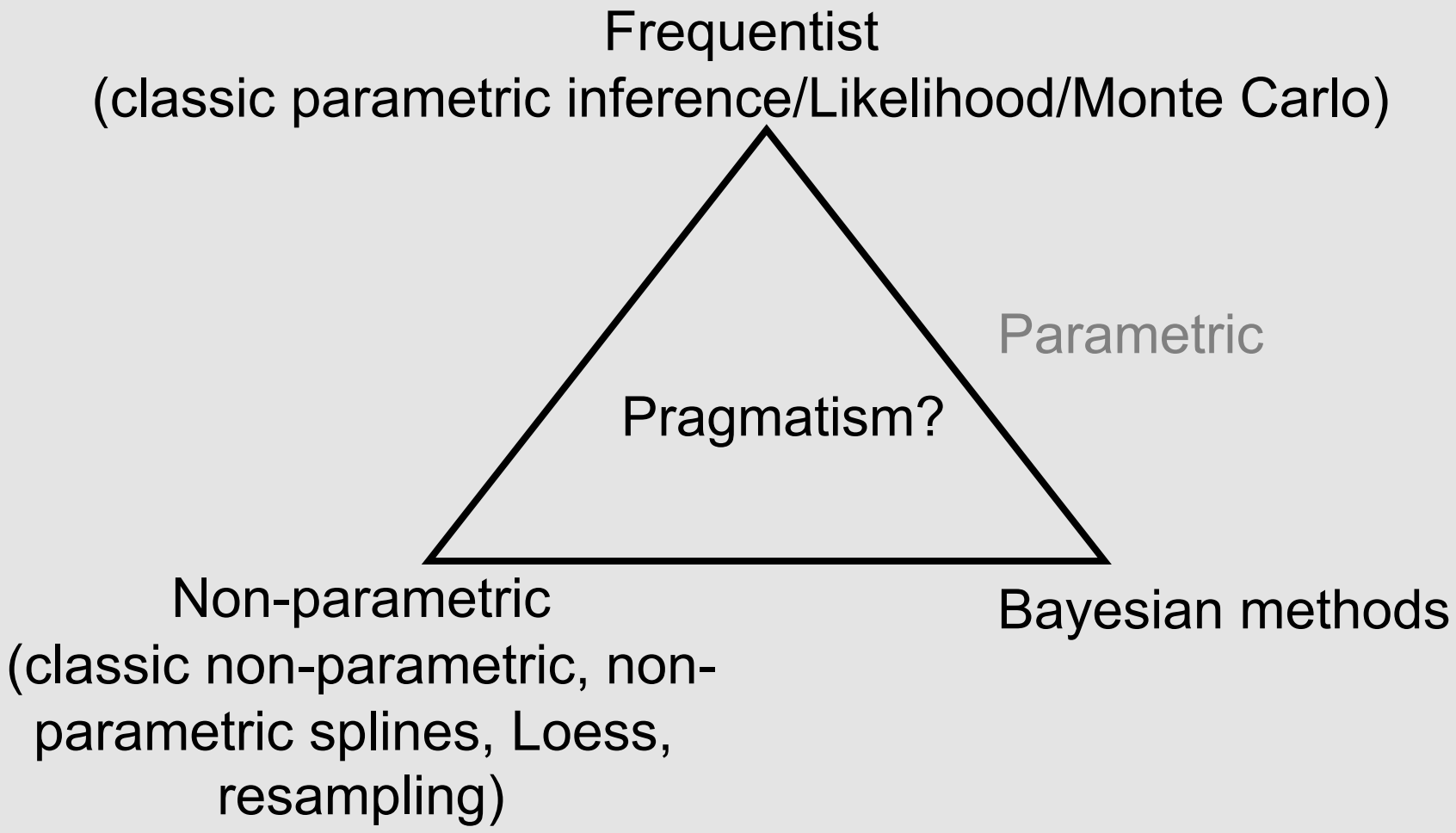
Continuity of Statistical Approaches



Relationship between Estimation methods



The philosophical axes of inference



Thinking about models in statistics

In statistics, most models have a “deterministic” & a “random” component (there are purely random models, and Bayesian models are all “random” in a sense).

The lowly regression model

$y_i = b_0 + b_1 x_i + e_i$

What is “fixed” what is random?

$Y \sim N(\mu, \sigma^2)$

$Y \sim N(b_0 + b_1 X , \sigma^2)$

Thinking about models in statistics

In statistics, most models have a “deterministic” & a “random” component.

Usually we are estimating both the random and “fixed” parameters (either as point estimates, intervals, or random variables in the case of Bayesian methods).

Thinking about models in statistics

• In statistics it is not enough to build a model and estimate parameters. We also need **explicit** information on the **uncertainty** in our estimates.

• We will talk a fair bit about measures of uncertainty (Standard errors, confidence intervals), and different ways of estimating them.

•**Parameter estimates + measures of uncertainty are far more useful than any P value!!!!**

Below I describe a few options with respect to least squares estimation, and inference. Please choose the answer that seems correct.
LSE= least squares estimation

1. Both LSE, and inferences based on LSE are parametric procedures.
2. Both LSE, and inferences based on LSE are non-parametric procedures.
3. LSE is parametric, but inferences based on LSE are often non-parametric.
4. LSE is non-parametric, but inferences based on LSE are often parametric.

It will help to answer the following questions first:
What is the criteria for LSE?
What do I mean by parametric? Non-parametric?

Please choose the answer that best describes what a p-value is:

1. P(D|H0) - probability of observing the data given the null hypothesis is true.
2. P(D|H1) - probability of observing the data given the alternative hypothesis is true.
3. P(H0|D) - probability of the null hypothesis being true given the data.
4. P(H1|D) - probability of the alternative hypothesis being true given the data.

It will help to answer the following questions first:
What is a null hypothesis? What is an alternative hypothesis?
What do we mean by P(x|y)?

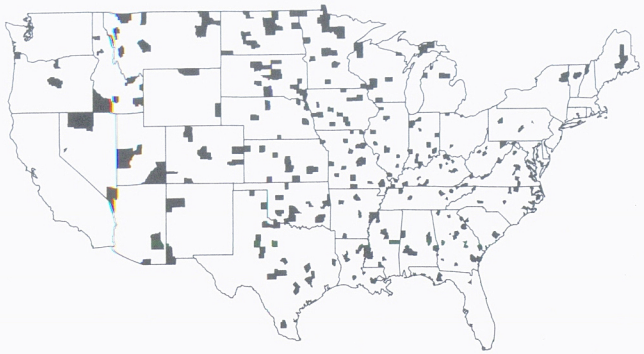
Which of the following are linear models?

1. $y = \mu + \beta x + e$
2. $y = \mu + \beta_1 x + \beta_2 x^2 + e$
3. $y = \mu + x^{\beta} + e$

What defines a linear model? Linear with respect to what?

Cancer incidence maps

Highest kidney cancer death rates



Lowest kidney cancer death rates