Background statistical concepts

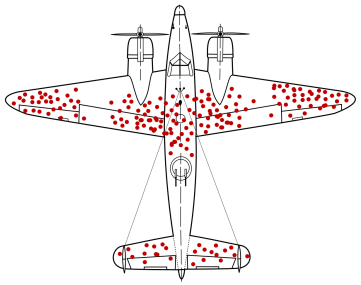
Simon Wood
School of Mathematics, University of Edinburgh, U.K.

Aim of statistics

- Extract information from data for understanding and prediction.
- While…
  1. Avoiding being misled by irreproducible random features of data.
     - measurement errors, sampling variability, patient-to-patient variability etc.
  2. Avoiding biases from systematic selection effects in data.
     - biased and non-random sampling, systematic patterns of missingness, survivorship bias, publication bias (in meta-analysis)
  3. Avoiding our own cognitive biases.
     - availability bias, confirmation bias, seeing patterns in the noise (evolutionary heritage perhaps: better to 'see' three bears that aren't there, than miss one bear that is).

Data bias example: Survivorship[1]

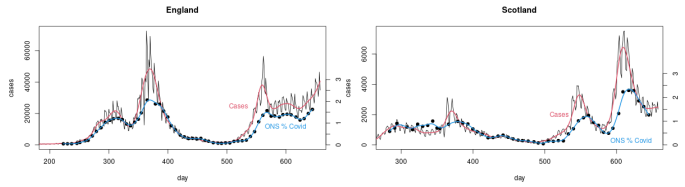
- WW2 analysis of the damage to returning bombers to decide where reinforcement should be added.
- Eventually realized: reinforcement should be added to the areas **not** damaged in the *surviving* bombers.
- Those were the areas most likely hit in the non-returners.



- Similarly when combining evidence from scientific literature…
- Surprising, interesting and 'positive' results much more likely to 'survive' to publication (and more likely to be wrong).
- Especially in 'top' journals demanding high interest/novelty.

[1] Image: Grandjean, McGeddon, Moll, Wikimedia

Data bias examples: biased samples



- Covid 'recorded cases' (black line, red curve) were routinely used to assess the state of the pandemic, as if they were a representative sample of people with Covid.
- They are people who tested positive among those who decided to, or were told to by track and trace, and could get a test.
- The actual prevalence of Covid, measured by randomly sampling UK residents, is shown as black dots and a blue smooth curve.
- Treating cases as representative of prevalence overestimated the severity of each upswing.

Data bias examples: biased design

- A celebrated example is the 1930 Lanarkshire milk trial[2]
- The trial examined relative growth benefits, if any, of daily drinking of raw or pasteurised milk.
- 5000 children each were given 3/4 pint of raw or pasteurized milk daily for 4 months, for comparison with 10000 controls.
- Within schools, allocation to milk or control was by lottery or alphabetic, **but** teachers were given discretion to adjust the groups if they appeared 'unbalanced'.
- The teachers appear to have adjusted by allocating undernourished children to receive milk.
- On average controls were initially larger than the milk receivers by an amount greater than 4 months average growth!
- Data a can of worms! Avoidable by proper randomization.

[2] whose design was first criticised by Student (of t-test fame).

Cognitive biases I

- Availability bias.
  - Concentrate on readily available data (e.g. early deaths from Covid), ignoring difficult to access or delayed data (e.g. early deaths from effects of Covid measures).
  - Similar to selection biases. Excessive weight given to what's visible and accessible.
- Confirmation bias.
  - Look for (notice) data supporting a theory/model, not for data contradicting it.
  - Tendency to require much higher standard of proof for contradiction than for confirmation.

Every time I did this the sun rose next morning!

Cognitive biases II

- The question and answer switch.
  - Use an easy question to provide the answer to a difficult question.
  - e.g. Use *who would be a better drinking buddy?* to answer *who has the better economic policies?*
  - or use *how did google mobility data change around lockdown?* to answer *how did Covid relevant inter-personal contacts change around lockdown?*
- Seeing patterns in the noise.
  - We are so good at spotting patterns that we see them in the arrangement of stars, the sequence of lottery numbers, and almost inevitably in the noise in our data.
  - Whole books are written on these problems: See *Thinking Fast and Slow* (Daniel Kahneman, 2011) for more.

The statistical approach: *learning from random samples*

- Treat data as a random sample from a population, where some fixed property of the population is the information of interest.
- Use the data sample variability to learn about the population.
- *Random samples from populations avoid data selection biases.*
- The population can be concrete or abstract. e.g.
  - The population of UK adults.
  - The population of possible energy yields from replication of a collision experiment under practically identical conditions.

  . . . the key is to identify what population your data can be treated as sampling, and what property of that population is of interest.
- Models of how the data were randomly sampled from the population allow us to
  - *infer* properties of the population from the data.
  - *avoid* over-interpreting random patterns in the data.

The statistical roles of randomness

1. By modelling the component of data that would change from replicate to replicate in a random unexplained way, we avoid over-interpretation of 'noise' and can characterize the reliability of the information gained from the data.
2. By designing data collection to randomly sample from the population of interest we can avoid data selection biases.
3. In *experiments* on non-identical experimental units (people, guinea pigs, 5-year old crash helmets) where we manipulate one 'treatment' variable to find its effect on a 'response' variable, systematic association between the unit characteristics and treatment is avoided by *randomizing* units to treatment levels.

Causality and caution

- *Randomized experiments* can show that a treatment variable and nothing else *caused* the changes seen in the response, because all other unit properties are forced to vary only randomly and independently across the treatment levels.
- For data not from a randomized experiment – *observational data*[3] – we would need to be able to allow for the effect of every possible variable also influencing the response before we could conclude anything causal about the treatment's effect.
- Usually we don't know what these variables are, let alone have measurements for them.
- This makes *causal inference* difficult with observational data.
- Unless we measure every variable relevant to the response, and have a very good model relating the response to these variables, great care is then needed in drawing causal conclusions!

[3] still a random sample hopefully!

Statistical regression models

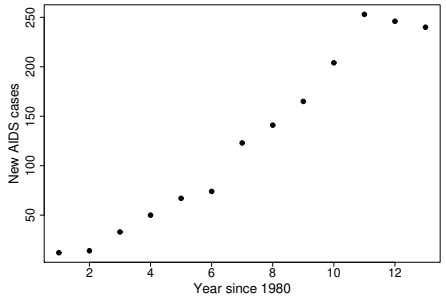
- n observations, y_i, of a *response variable* of primary interest.
- With each y_i is a *covariate* vector $\mathbf{x}_i$ that influences its value.
- $y_i, \mathbf{x}_i$ *may* be sampled randomly from the joint distribution of $y, \mathbf{x}$.
- But we only *require* that each y_i is sampled randomly from the sub-population for which the covariates take the corresponding observed value $\mathbf{x}_i$.
- We create a model relating y_i to the $\mathbf{x}_i$, and use it for all such sub-populations.
- In particular we model the distribution of y *given* $\mathbf{x}$, not the more complicated joint distribution of y *and* $\mathbf{x}$.
- Crucially, $y_i|\mathbf{x}_i$ can often be modelled as independent of $y_j|\mathbf{x}_j$ for all $i \neq j$. This simplification is untrue for y_i and y_j 'marginally'[4].

[4] i.e. without conditioning.

Regression model general structure

- Regression models specify some mathematical form for the relationship between the statistical distribution of the response and the covariates, in the population.
- This mathematical expression contains some unknown parameters, whose values provide interesting information about the population.
- We learn about the parameters from the sample of data.

Simple regression model example: Poisson GLM



- $\texttt{cases}_i \sim \text{Poi}(\mu_i)$ where $\log(\mu_i) = \theta_0 + \theta_1 \texttt{year}_i + \theta_2 \texttt{year}_i^2$.
- μ_i represents the underlying case rate in the population over time.
- The actual number of cases seen in a year is assumed to be a Poisson random variable with mean μ_i.
- The parameters $\boldsymbol{\theta}$ control the change in μ_i over time.
- Need to estimate $\boldsymbol{\theta}$ from the data.

Basic inference methods: Maximum Likelihood

- A regression model specifies $\pi(y_i|\mathbf{x}_i, \boldsymbol{\theta})$, the p.d.f. of $y_i|\mathbf{x}_i$. Given conditional independence the p.d.f. of $\mathbf{y}$ given $\mathbf{x}$ is

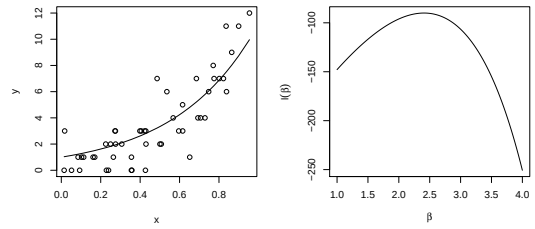
$$\prod_{i=1}^{n} \pi(y_i|\mathbf{x}_i, \boldsymbol{\theta})$$

- Plug the observed $\mathbf{y}$ values into the joint p.d.f. take logs and consider it as a function of $\boldsymbol{\theta}$

$$l(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \pi(y_i|\mathbf{x}_i, \boldsymbol{\theta})$$

  – the *log likelihood*. $\boldsymbol{\theta}$ values are more *likely* to be correct, the higher probability they ascribe to the observed data.
- So $\hat{\boldsymbol{\theta}} = \text{argmax}_\theta l(\boldsymbol{\theta})$ is the *maximum likelihood estimate* of $\boldsymbol{\theta}$.

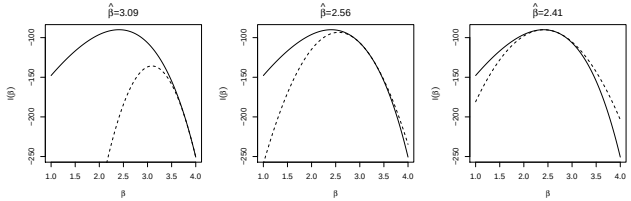
Simple simulated one parameter likelihood example



- Left: data + expected value curve for model $y_i \sim \text{Poi}\{\exp(\beta x_i)\}$.
- Right: corresponding $l(\beta)$ function. $\hat{\boldsymbol{\beta}} \simeq 2.4$.
- Poisson p.d.f of y_i is: $\exp(\beta x_i)^{y_i} \exp\{-\exp(\beta x_i)\}/y_i!$
- So log-likelihood function is

$$l(\beta) = \sum_{i=1}^{n} y_i \beta x_i - \exp(\beta x_i) - \log y_i!$$

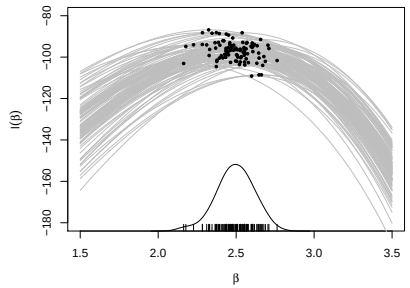
Maximizing log likelihoods by Newton's method



- Log likelihoods can be maximized numerically by Newton's method. Iterate...
  1. Find the quadratic matching the first and second derivatives of $l(\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$ at current $\hat{\boldsymbol{\theta}}$ guess.
  2. Maximize the quadratic to get an updated $\hat{\boldsymbol{\theta}}$ estimate.
- To guarantee convergence, perturb Hessian[5] to be negative def. if it's not and step half (repeatedly) if log likelihood not increased.
- Note: log likelihood of a Gaussian is exactly quadratic - this is successive Gaussian approximation, improving with iteration.

[5] second derivative matrix

Sampling distribution of MLE



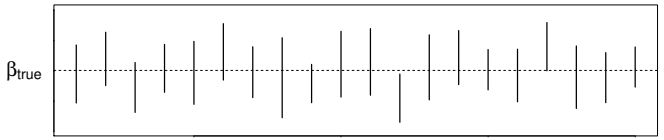
- How would the MLE vary under repeated replication of the data sampling process?
- The figure illustrates how the likelihood curves of the simple simulated example vary under replication.
- This variability in the likelihood function leads to variability in the MLE (black dots maxima, black ticks MLEs).

Theoretical sampling distribution of MLE

- Actually replication impractical, but if $\hat{\boldsymbol{\mathcal{I}}} = -\partial^2 l/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\mathsf{T}$ and $n \to \infty$ we have theoretical result

$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, \hat{\boldsymbol{\mathcal{I}}}^{-1}).$$

- Can also substitute $\mathbb{E}(\hat{\boldsymbol{\mathcal{I}}})$ for $\hat{\boldsymbol{\mathcal{I}}}$.
- Can immediately use this to summarize uncertainty in $\boldsymbol{\theta}$ by constructing *confidence intervals*.
- A confidence interval is a random interval having a specified probability (e.g. 0.95) of containing the true parameter value, over imagined replication of the data sampling process. e.g.



Hypothesis testing: comparing nested models

- Consider testing whether a simplified model could be adequate for our data.
- Express the simplification as a *null hypothesis* placing r restrictions on $\boldsymbol{\theta}$. Say, $H_0 : R(\boldsymbol{\theta}) = \mathbf{0}$.
- If H_0 is true, $\hat{\boldsymbol{\theta}}_r$ is the MLE given $R(\boldsymbol{\theta}) = \mathbf{0}$ and $n \to \infty$[6]

$$2\{l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_r)\} \sim \chi^2_r$$

  but if H_0 is untrue the LHS will be too large for χ^2_r.
- Use result to compute *p-value* (prob. a χ^2_r r.v. $\geq$ observed LHS)
  - Very roughly: what's the chance of getting these data if H_0 is true.
  - Exactly: if H_0 is true, how probable are data with at least this low a ratio of probability under H_0 to probability without restrictions.

  ... so low p-value casts doubt on H_0.

[6] $R(\boldsymbol{\theta})$ must not restrict $\boldsymbol{\theta}$ to edge of feasible parameter space, l must be 'regular'.

Comparing models by prediction performance

- Perhaps we don't have nested models, or don't want to specially favour simplicity.
- Can compare models by their ability to predict new (replicate) y_i data *not used in fitting*.
- Favour the model that would ascribe the highest probability to such new replicate data.
- Idea leads theoretically to choosing model with lowest

$$\text{AIC} = -2l(\hat{\boldsymbol{\theta}}) + 2\text{dim}(\boldsymbol{\theta})$$

- Or use a brute force estimate. Let $\hat{\boldsymbol{\theta}}^{[-i]}$ be MLE on omission of $y_i, \mathbf{x}_i$ from fit. Maximize leave one out cross validation criterion

$$\text{OCV} = \sum_{i=1}^{n} \log \pi(y_i|\mathbf{x}_i, \hat{\boldsymbol{\theta}}^{[-i]}).$$

Basic inference methods: Bayesian

- MLE: $\boldsymbol{\theta}$ are fixed constants to estimate. $\hat{\boldsymbol{\theta}}$ variability over theoretical replication of data sampling characterizes uncertainty.
- Bayesian approach: use probability distributions to model our uncertainty about $\boldsymbol{\theta}$ values, treating $\boldsymbol{\theta}$ as random variables.
- A Bayesian model describes the pre-data uncertainty about parameters using a *prior* distribution, $\pi(\boldsymbol{\theta})$, say.
- The sampling model describes how the data have been sampled from the population given $\boldsymbol{\theta}$ values. It provides the p.d.f. $\pi(\mathbf{y}|\boldsymbol{\theta})$.
- $\pi(\boldsymbol{\theta})$ is then updated given the observed $\mathbf{y}$ using the fact that $\pi(\mathbf{y},\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|\mathbf{y})\pi(\mathbf{y}) = \pi(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ implying *Bayes rule*

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \pi(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})/\pi(\mathbf{y}).$$

- Plug in observed $\mathbf{y}$. We have posterior $\propto$ likelihood $\times$ prior.

Using Bayes

- Given $\pi(\boldsymbol{\theta}|\mathbf{y})$ we can obtain *credible intervals*: fixed intervals containing the *random* parameter with specified probability.
- Despite the switching of fixed and random, confidence and credible intervals frequently converge as $n \to \infty$.
- As $n \to \infty$ the likelihood's impact on the posterior usually dominates the prior, so the choice of prior becomes unimportant.
- Hence we can often use 'uninformative' priors if unsure.
- But we can't escape the impact of $\pi(\boldsymbol{\theta})$ choice if trying to compute the relative posterior probability of models, for model selection. This makes model selection tricky.
- As for MLE, use of $\pi(\boldsymbol{\theta}|\mathbf{y})$ usually requires numerical methods.
  1. One approach is to use stochastic simulation methods to simulate draws from $\pi(\boldsymbol{\theta}|\mathbf{y})$.
  2. Other approaches make judicious use of Gaussian approximations, for approximate posterior calculations.

One technical reminder: covariance matrices

- A covariance matrix for $\mathbf{y}$ is a matrix with variances of $\mathbf{y}$ on leading diagonal and covariances on off diagonals.
- i.e. if $\text{cov}(\mathbf{y}) = \mathbf{V}$ then $V_{ij} = \text{covariance}(y_i, y_j)$. Equivalently $\mathbf{V} = \mathbb{E}[\{\mathbf{y} - \mathbb{E}(\mathbf{y})\}\{\mathbf{y} - \mathbb{E}(\mathbf{y})\}^{\mathsf{T}}]$.
- From these basic definitions it is easy to show that if $\mathbf{y} = \mathbf{A}\mathbf{x}$ and $\mathbf{x}$ has covariance matrix $\mathbf{V}_x$ then the covariance matrix of $\mathbf{y}$ is

$$\mathbf{V}_y = \mathbf{A}\mathbf{V}_x\mathbf{A}^{\mathsf{T}}$$

...this result gets used quite a bit.

Enough background. Let's move on to smooth models, GAMs etc...