ROSLIN

Lecture 9: Introduction to Statistical Inference

Osvaldo Anacleto
Genetics and Genomics, Roslin Institute
osvaldo.anacleto@roslin.ed.ac.uk

THE UNIVERSITY *of* EDINBURGH

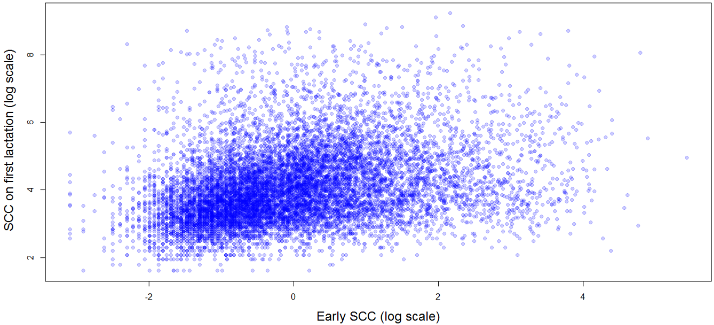
BBSRC
20 Years of Pioneering Great British Bioscience

Overview

- Why do we need statistical models?
- The likelihood function
- Frequentist inference: maximum likelihood estimates and confidence intervals
- Likelihood and frequentist inference for stochastic SIR models

Why do we need models? An example

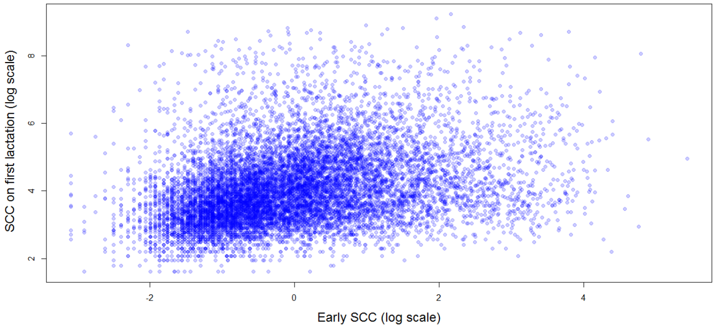
Early vs first day lactation somatic cell counts of young cows
(De Vliegher *et al.*, 2004)



- somatic cell count (SCC) is an indicator of milk quality and cow health
- early indicators of SCC can be useful to guide farm management

Somatic cell count (SCC) study: Research questions

Early vs first day lactation somatic cell counts of young cows



- how early SCC is related to first lactation SCC?
- is early SCC **really** related to first lactation SCC?
- are there any other variables associated with first lactation SCC??

models can provide answers to these questions

Models

Models are devices to **answer questions** and **represent reality**

mathematical models use **equations** to represent relationships

example: 1st lactation SCC $= \alpha + \beta$(early lactation SCC)

- mathematical models represent **assumptions** and **underlying knowledge** about quantities of interest

problem: these models do not deal with the **uncertainty** regarding the phenomenon.
(*is early SCC really related to 1st lactation SCC? are there any other variables associated with 1st lactation SCC?*)

How to deal with the uncertainty underlying the problem?

Statistical models

Statistics deals with **uncertainty** by incorporating **variation** into the model

Sources of variation

- **systematic (deterministic) variation**: this can be based on **knowledge about the system** (example: early lactation SCC)
- **random (stochastic) variation:** this is due to **unknown factors/variables** which might be affecting the response

Statistics uses **probability distributions** to deal with random variation

example: linear regression model: $y = \alpha + \beta x + \epsilon, \;\; \epsilon \sim \mathrm{N}(0, \sigma^2)$
ϵ has a normal distribution with mean 0 and variance σ^2

probability distributions represent data variation

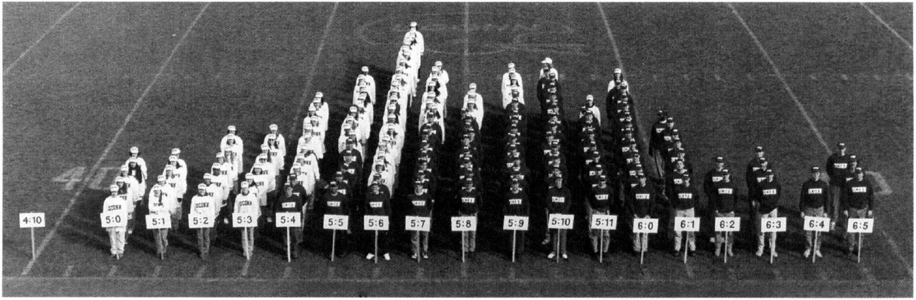
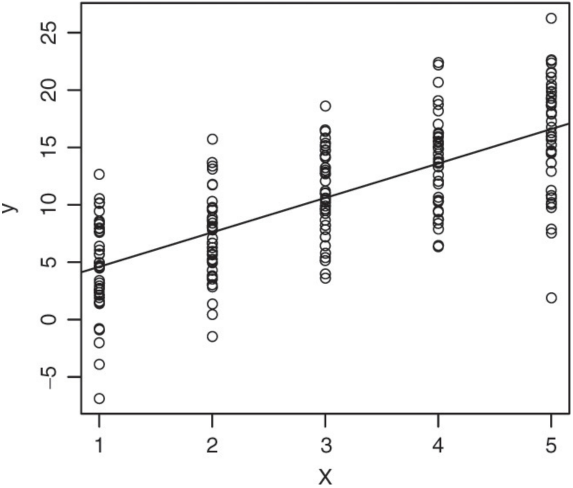


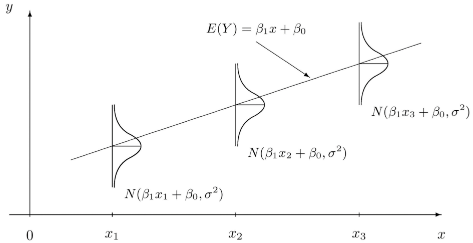
Figure 7. Living histogram of 143 student heights at University of Connecticut.

- What probability distribution gives the best fit to these data?
- Assumptions must be considered and evaluated based on available data

Variation in a variable might depend on the variation in another variable



Regression Analysis idea



y
$E(Y) = \beta_1 x + \beta_0$
$N(\beta_1 x_3 + \beta_0, \sigma^2)$
$N(\beta_1 x_2 + \beta_0, \sigma^2)$
$N(\beta_1 x_1 + \beta_0, \sigma^2)$
0
x_1
x_2
x_3
x

linear regression model:

$$y = \beta_0 + \beta_1 x + \epsilon,$$
$$\epsilon \sim \mathrm{N}(0, \sigma^2)$$

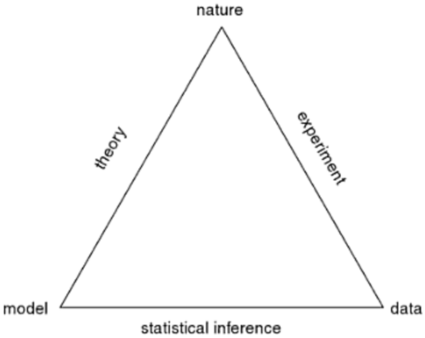
Distribution of the response variable Y given (fixed) x

$$Y|x \sim \mathrm{N}(\beta_0 + \beta_1 x, \sigma^2)$$

task: estimate parameters β_0, β_1 based on a sample of the population: **Statistical Inference**

Parameter inference is one of the goals of Statistics

Scientific Method and Statistics



important statistical tasks

- design of experiments
- **inference**
- prediction/forecasting

statistical conclusions must be translated back into biology

Statistical inference

Statistical inference is the process of reaching **conclusions** from **data**

- data are always limited: usually a sample and/or limited experiments
- information may be limited even when dealing with large datasets (ex. gene expression data)
- different data provide different answers

any **statistical conclusion** involves a degree of **uncertainty**

Statistical inference tasks

- point estimation
- interval estimation
- hypothesis testing (ex. p-values, GWAS)

Point estimation

- Statistics uses **probability** to deal with random variation
- a probability distribution is assumed for a random variable of interest
- probability distributions are functions of **unknown parameters**

point estimation idea:
given the **available data** and assuming an **underlying model** for the variation observed (probability distribution) what single value is plausible for the indexing parameter?

a point estimate is the best guess about the parameter of a distribution

Maximum likelihood estimates (MLE)

Key idea (Fisher, 1922):
ML estimates maximize the probability of having observed the available data (e.g best explain the data)

Example: suppose an artificial data set with 3 **independent** observations:
$x_1 = 3$, $x_2 = 4$ and $x_3 = 8$

Assume these data come from random variable X which follows a geometric distribution:

$$P(X = x; \theta) = (1 - \theta)^{x-1}\theta, \;\; x = 1, 2, 3...$$

The geometric distribution depends on an unknown parameter θ which can be estimated using x_1, x_2 and x_3.

- In the example we assume that X_1, X_2, X_3 are **independent** and follow the **same distribution**

- If that's the case, $X_1, X_2, \ldots, X_N$ **independent** and **identically distributed (i.i.d)** random variables

- The assumption of i.i.d random variables is frequently considered in Statistics, **but it is not suitable for infectious disease data**

Example (cont'd)

assuming i.i.d random variables (random sample), the probability of observing x_1, x_2 and x_3 is

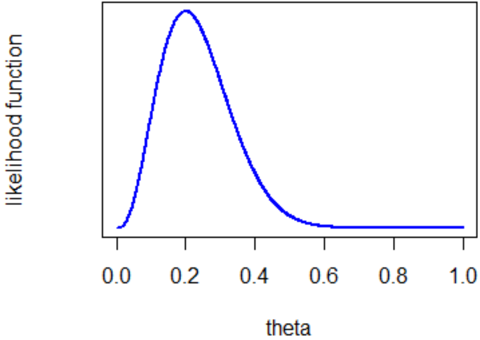
$P(X_1 = x_1, X_2 = x_2, X_3 = x_3; \theta) =$

$$\begin{aligned} &= P(X_1 = 3; \theta)P(X_2 = 4; \theta)P(X_3 = 8; \theta) \\ &= (1-\theta)^{3-1}\theta(1-\theta)^{4-1}\theta(1-\theta)^{8-1}\theta \\ &= (1-\theta)^{12}\theta^3 \end{aligned}$$

This is the **likelihood function**

the likelihood is a function of the unknown parameter θ.

Plotting the likelihood function



- the likelihood function is maximized at the value $\theta = 0.2$
- Hence, 0.2 is the maximum likelihood estimate of θ

Point estimation: another example

- Suppose an experiment to study the incidence of a certain tumour in mice
- a binary random variable (tumour/not tumour) can be used to represent each mouse in a random sample with size N
- the **total** number of mice with tumour can be modelled with a binomial distribution
- given that 6 out of 54 mice were observed with tumour, the probability of this event is

$$P(X = k) = \binom{54}{6} \theta^6 (1 - \theta)^{54-6}$$

where θ represents the unknown proportion of mice with tumour in the **population.** This is the likelihood function of θ for the data observed, and it is maximized at $\hat{\theta} = 0.11$
Hence, $\hat{\theta} = 0.11$ is the maximum likelihood estimate of the proportion of the mice with tumour in the population.

notes on point estimation/MLEs

- maximum likelihood estimators have important statistical properties (e.g., unbiasedness and eficiency)
- calculus or numeric procedures are used to obtain MLEs
- shape of the likelihood function plays a vital role in statistical inference (more on this later)
- it can be difficult to derive a proper likelihood to represent the data (example: epidemic modelling)
- **different samples provide different likelihoods hence different MLEs** - How to consider this feature into our inferences?

Two main approaches can be used: Frequentist (Classical) Inference and Bayesian Inference

Interval estimation: the frequentist approach

interval estimation idea:
given the available data and assuming an underlying model for the variation observed, what **range of values** is plausible for the indexing parameter?

This range of values involve a degree of uncertainty.

goal: for an **unknown** and **fixed** parameter θ, we want a range (a,b) such that

$$P[a < \theta < b] = \delta$$

- *a* and *b* are functions of the data: ***a* and *b* are random quantities**
- δ is a probability: usually 0.95
- example: when $\delta = 0.95$, the range of values $[a, b]$ is called a **95% confidence interval for the parameter θ**

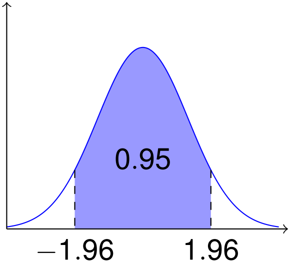
Confidence intervals example: the frequentist approach

- Suppose that observations from a data set follow a normal distribution with unknown parameter μ and unknown variance σ^2
- $X_1, X_2, ..., X_N$ are random variables which represents the measurements of a random sample of size N
- the sample mean $\overline{X} = (X_1 + X_2 + \ldots + X_N)/N$ is the **maximum likelihood estimator** of the population mean μ
- Since $\overline{X}$ is a function of random variables, this **estimator** is also a **random variable** which follows a **probability distribution**

Confidence intervals example: the frequentist approach

- When the sample size is large, it can be shown $\overline{X}$ has approximately a normal distribution unknown mean μ and variance $\hat{\sigma}^2/n$.
- $\hat{\sigma}^2$ is an estimate of the variance (the sample variance)
- this is a key result in Statistics (Central-limit theorem) and **it holds for any data distribution**

density of a N(0,1)



a 95% confidence interval for the mean μ is

$$[\overline{X} - 1.96\hat{\sigma}/\sqrt{n}, \overline{X} + 1.96\hat{\sigma}/\sqrt{n}]$$

Interpretation of a confidence interval
the frequentist approach

- Confidence intervals are functions of the data available: they are random quantities
- In the frequentist approach, the distribution parameters are assumed fixed

frequentist interpretation of a 95% CI for the mean μ:

If a large number of confidence intervals were calculated using independent random samples of the population, 95% of them would contain the true mean μ

notes on frequentist interval estimation

for normal distributed variables

- exact confidence intervals can be obtained even when the sample size is small: in this case a Student's t-distribution is used
- confidence intervals for the variance can be calculated using a chi-square distribution

Other distributions

- large-sample confidence intervals can be used to calculate CI for proportions (practical)
- definition of large-sample can vary. Usually n=30 is enough for means of normal distributions and also proportions

point and interval estimation can be also applied to regression analysis

Summary of key ideas: Frequentist inference

- Statistics allow the incorporation of uncertainty about the quantities of interest into models
- Any statistical conclusion involves uncertainty
- Frequentist inference allows parameter estimation by using vailable data only (using likelihood)
- Maximum likelihood estimates are the ones that best explain the observed data
- confidence intervals represent the uncertainty regarding Frequentist inference: plausible **range of values** for a model parameter

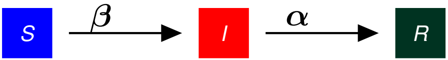
Tutorial 9a

Analysing fish infection data

- Looking at likelihood plots
- Comparing frequentist estimates based on different sample sizes

Representing infectious disease dynamics: stochastic compartmental models

“diversity” represented through disease states - ex. SIR model



possible individual states

- susceptible (S)
- infected (I)
- recovered (R)

- Infections occur with rate $\beta S(t)I(t)$ (Poisson process)
- Recoveries occur with rate $\gamma \rightarrow$ infectious period follows an exponential distribution (general stochastic model)

Assumptions

- closed population, homogeneous mixing and no latent period
- all individuals are **equally** susceptible and infectious

Inference problem: estimation of β, γ and R_0

Estimating R_0 (in general)

There are several approaches for estimating R_0, depending on **assumptions** and **data limitations** (see Keeling and Rohani's book)

For example, R_0 can be estimated using reported cases, seroprevalence data, Average age at infection and final size data

most of simple methods for estimating R_0 are based on deterministic models - cannot accommodate uncertainty regarding parameter estimation

R_0 can be also estimated from transmission experiments (see Diekmann *et al*, 2013) - **practical**

Likelihood function for a SIR model

The likelihood function depends on the **structure** and **availability** of the epidemic data

Some possible scenarios:

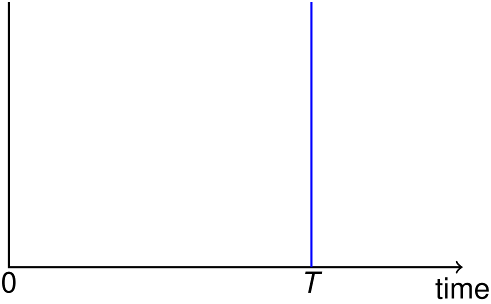
- **best scenario:** exact infection and removed times (complete observation)
- Infection times observed at fixed time periods (e.g an individual infected between week 1 and 2)
- only removed times observed
- only final numbers infected

Not suitable for genetic analysis

Infectious disease data in genetic analysis

usual phenotype: binary disease status at a **single time** T

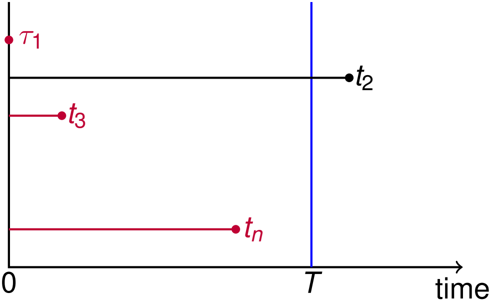
animal	**phenotype**
1	diseased
2	healthy
3	diseased
⋮	⋮
n	diseased



Infectious disease data in genetic analysis

usual phenotype: binary disease status at a **single time** T

animal	**phenotype**
1	diseased
2	healthy
3	diseased
⋮	⋮
n	diseased

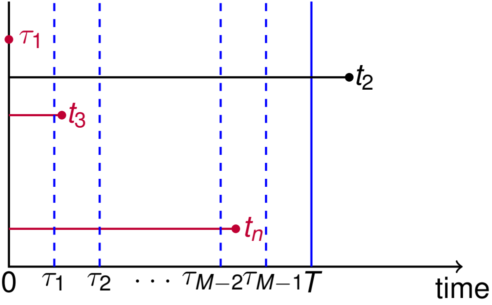


- time to infection is more informative about disease traits (rarely observed but can be inferred - to be seen in the MCMC lecture)

Infectious disease data in genetic analysis

usual phenotype: binary disease status at a **single time** T

animal	**phenotype**
1	diseased
2	healthy
3	diseased
⋮	⋮
n	diseased

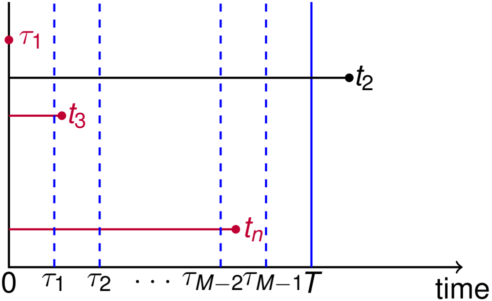


- time to infection is more informative about disease traits (rarely observed but can be inferred - to be seen in the MCMC lecture)
- disease status at sampling times $(0, \tau_1, \tau_2, \ldots T)$ are required

Infectious disease data in genetic analysis

usual phenotype: binary disease status at a **single time** T

animal	**phenotype**
1	diseased
2	healthy
3	diseased
⋮	⋮
n	diseased



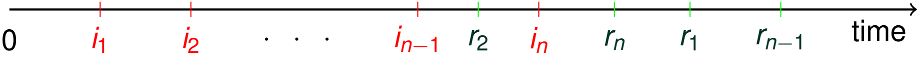
- time to infection is more informative about disease traits (rarely observed but can be inferred - to be seen in the MCMC lecture)
- disease status at sampling times $(0, \tau_1, \tau_2, \ldots T)$ are required
- better genetic analyses of disease traits when using binary longitudinal data (e.g, Anacleto *et al, 2015)*

Likelihood function for a SIR model (idea)

assumption: Complete observation of the data

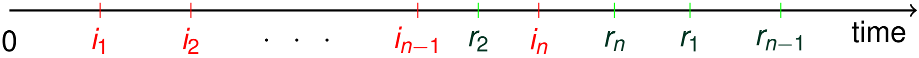
- infection times: $\boldsymbol{i} = (i_2, i_3, \dots i_n)$
- removal times: $\boldsymbol{r} = (r_1, r_2, \dots r_n)$

data observed until the end of the epidemic:



- **infection events are not independent!**
- likelihood can be decomposed into the contributions of the infectious and the removal processes
- order of the events and special properties of the stochastic process and associated distributions are important for derivation of the likelihood

Likelihood function for a SIR model (idea)



likelihood contribution of the infectious process

- Time between infections follows a exponential distribution (Poisson Process property)
- based on the conditional densities at small time steps

likelihood contribution of the removal processes

- based on the exponential distribution of the infectious periods

- infectious period of individual $j = r_j - i_j$

Different likelihood derivations depending on inference approach

- See Bailey and Thomas (1971), Becker (1989) and Britton and O'Neill (2002). See also Kypraios (thesis, 2007) for comparison.

Maximum likelihood estimates of SIR model parameters

Assume that

- epidemic started with an initially infected individual and was observed until its end at time T
- n_I and n_R are the total numbers of infecteds and recovereds
- S_t and I_t and R_t are the **numbers** of susceptibles, infecteds and recovereds at time t

$$\hat{\beta} = \frac{n_I}{\int_{I_1}^{T} S_t I_t dt}$$

$\hat{\beta}$'s denominator is the accumulated rate of contacts between susceptibles and infecteds

$$\hat{\gamma} = \frac{n_R}{\int_{I_1}^{T} R_t dt}$$

$\hat{\gamma}$'s denominator is the aggregated length of the infectious period

Maximum likelihood estimates of SIR model parameters

Rida (1991) showed that standard errors of $\hat{\beta}$ and $\hat{\gamma}$ are

$$\text{s.e}(\hat{\beta}) = \frac{\hat{\beta}}{\sqrt{n_I - 1}} \qquad \text{and} \qquad \text{s.e}(\hat{\gamma}) = \frac{\hat{\gamma}}{\sqrt{n_r}}$$

Hence, approximate confidence intervals for β and γ can be obtained by normal approximation

Approximate confidence intervals can also be derived for R_0 (see Diekmann *et al*, 2013)

Summary of key ideas: Frequentist inference for SIR model parameters

- challenges: high dependence between infection events and incomplete data
- several approaches for estimating R_0
- likelihood function of the SIR model is derived depending on inference approach
- likelihood function of the SIR can be decomposed into the contributions of the infectious and the removal processes - order of events are important!
- individual level data are required for (most of) genetic analysis of infectious diseases

References

Statistical Inference

- Wasserman, Larry. All of statistics: a concise course in statistical inference. Springer, 2013.

Frequentist Inference for stochastic epidemic models

- Becker, N. G., Britton, T. (1999). Statistical studies of infectious disease incidence. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61(2), 287-307.
- Diekmann, Odo, Hans Heesterbeek, and Tom Britton. Mathematical tools for understanding infectious disease dynamics. Princeton University Press, 2012.
- Britton, T. (1998). Estimation in multitype epidemics. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 60(4), 663-679.

Guidelines for Statistical Practice

- Kass, R.E., Caffo, B., Davidian, M., Meng, X.-L., Yu, B., and Reid, N. (2016) Ten simple rules for effective statistical practice, PLoS Computational Biology.

Tutorial 9b

- Frequentist inference for R_0
- Comparing frequentist estimates based on different sample sizes