Imperial College
London

Problem Sheet 1

MATH50011
Statistical Modelling 1

Weeks 1 & 2

Lecture 1 (Statistical models)

1. Suppose that in Example 1 it is known that most participants have little knowledge about oxen but some participants raise oxen for a living. Under what assumptions will the proposed $N(543.4, \sigma^2)$ distribution still be a reasonable model?

2. In Example 2 of the lecture notes, we consider models where the distribution of Y_i depends on a fixed covariate x_i. Does treating Y_i as random and x_i as fixed make more sense for an observational study or a designed experiment?

Lecture 2 (Estimators)

3. Let T be an estimator of a parameter $g(\theta)$. Show that

$$\text{MSE}_\theta(T) = \text{Var}_\theta(T) + \text{bias}_\theta(T)^2.$$

4. Let $Y_1, \dots, Y_n$ be a random sample of size n from the Exponential(λ) distribution, for some $\lambda > 0$. The pdf of Y_i is then

$$f(y;\lambda) = \lambda e^{-\lambda y}, \quad y > 0$$

   and zero for $y \leq 0$.

   Two possible estimators for the mean $1/\lambda$ of an Exponential(λ) distribution from the random sample are $\bar{Y} = n^{-1}\sum_{i=1}^n Y_i$ and $T = n\bar{Y}/(n+1)$.

   Find the bias, variance, and mean square error of these estimators.

   What do you notice?

5. Let $Y_1, \dots, Y_n$ be a random sample with $E(Y_i) = \mu$ and $\text{Var}(Y_i) = \sigma^2$. Show that

   (a) $\bar{Y}^2$ is not unbiased for μ^2 unless $\sigma^2 = 0$;

(b) The sample standard deviation

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

is not an unbiased estimator for σ unless $\mathrm{Var}(S) = 0$.

6. **(Challenging)** Let T_1 and T_2 be two statistics. Suppose that T_1 is an unbiased estimator for θ and that $E_\theta(T_2) = 0$ for all θ. Also let $\mathrm{Var}_\theta(T_j) = \sigma_j^2$ for $j = 1, 2$ and $\mathrm{corr}(T_1, T_2) = \rho$.

   (a) Compare the bias, variance, and MSE of T_1 and $T_1 + T_2$ for θ;

   (b) Calculate the bias and variance of $T_1 + \alpha T_2$ where α is a constant;

   (c) Find the value $\tilde{\alpha}$ of α that minimises $\mathrm{MSE}_\theta(T_1 + \alpha T_2)$;

   (d) Compare the MSE of $T_1 + \tilde{\alpha} T_2$ and T_1 as ρ varies between -1 and 1.

Lecture 3 (CRLB)

7. In the lecture notes, we sketched the proof of the Cramér-Rao lower bound (CRLB) for continuous random variables. Prove the CRLB for discrete random variables with finite support. (Recall that the *support* of X is the set of values where the pdf/pmf is greater than zero.)

8. Find the CRLB for estimating θ based on a random sample of size n from the following distributions

   (a) Exponential(θ);

   (b) Normal(θ, σ^2) with known $\sigma^2 > 0$;

   (c) Bernoulli(θ); (see Example 8)

   (d) Poisson(θ).

9. For which of the distributions in 8(a-d) can the sample mean be used to construct an unbiased estimator T with variance equal to the CRLB for estimating θ?

10. **(Challenging)** Suppose that we wish to estimate θ based on a random sample $X_1, \dots, X_n$ of Bernoulli(θ) random variables. However, we are only able to obtain a random sample $(Y_i, R_i), \dots, (Y_n, R_n)$ where the R_i's are iid Bernoulli(p_0) for known p_0, independent of the X_i and $Y_i = R_i X_i$ for $i = 1, \dots, n$. Compare the CRLBs for estimating θ based on

    (a) The full data distribution of the X_i's;

    (b) The marginal distribution of the Y_i's;

    (c) The joint distribution of the (Y_i, R_i)'s.

Lecture 4 (Consistency)

11. Show that an asymptotically unbiased estimator sequence need not be consistent. (Hint: consider estimating μ based on a sequence of independent rv's $X_i \sim N(\mu, 2i)$ for $i = 1, 2, 3, \dots$)

12. Show that a consistent estimator sequence T_n need not be asymptotically unbiased. (Hint: consider a sequence (T_n, Y_n) with $Y_n \sim$ Bernoulli(1/n) and $T_n | Y_n = 0 \sim N(\theta, \sigma^2/n)$ and $T_n | Y_n = 1 \sim N(n^2, 1)$.)

13. **(Challenging)** Let $X_1, X_2, \dots$ be iid Uniform$(0, \theta)$ random variables and define $\hat{\theta}_n = \max\{X_1, \dots, X_n\}$.

    (a) Show that $\hat{\theta}_n$ is asymptotically unbiased and consistent.

    (b) Find a sequence of constants a_n such that $a_n \hat{\theta}_n$ is unbiased and consistent.

    (c) Compare the MSE of $\hat{\theta}_n$ and $a_n \hat{\theta}_n$.

14. **(Challenging)** Let $X_1, X_2, \dots$ be iid Bernoulli(θ) random variables and consider estimating $g(\theta) = \text{Var}(X_1) = \theta(1 - \theta)$. Define the sample mean $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$.

    (a) Show that $T_n = \bar{X}_n(1 - \bar{X}_n)$ is asymptotically unbiased and consistent.

    (b) Find a sequence of constants a_n such that $a_n T_n$ is unbiased and consistent.

    (c) Compare the MSE of T_n and $a_n T_n$.

    Hint: you may use the fact that

    $$Var(S_n^2) = \frac{\mu_4}{n} - \frac{\sigma^4(n-3)}{n(n-1)}$$

    where $\sigma^2 = Var(X_i)$ and $\mu_4 = E\{(X_i - \mu)^4\}$.

R lab: Descriptive statistics

This exercise is intended to reinforce concepts through use of the R software package.

15. The podcast *Planet Money* hosted a competition similar to Example 1. Here, $n = 17,183$ contestants guessed the weight (in lbs) of Penelope the cow.

    The data from the competition is in the file `Planet Money Cow Data.csv` on Blackboard. The file consists of a single column with 17,184 rows (Note: the first row is the column name "guess").

    (a) Set your working directory to *the same folder containing the data* downloaded from Blackboard. Then read the data into R and store it in an object called `cow` using the command

    ```
    cow <- read.csv("Planet Money Cow Data.csv")
    ```

(b) Run the commands `class(cow)` and `dim(cow)` to verify that the object `cow` is stored as a `data.frame` with dimensions $17,183 \times 1$.

(c) Use the command `table(is.na(cow$guess))` to tabulate ('table') the number of missing values ('is.na') in the column containing the variable `guess` (`cow$guess`). There should be no missing values in the data.

(d) Experiment with the functions `summary()`, `boxplot()`, `hist()` to generate summary statistics and plots for the guesses. To learn more about the functions, type e.g. `?summary` into the `R` console.

(e) Write a brief description of the data based on your statistics and plots from part (d), including the sample mean and standard deviation. Comment on the suitability of the normal distribution as a model for the guesses.

(f) It is known that Penelope weighs $\mu = 1,355$ lbs. How many standard errors from μ is the sample mean? The functions `sqrt()`, `mean()`, and `sd()` may be useful.
(Hint: recall that the *standard error* of an estimator T is $\sqrt{\mathrm{Var}(T)}$.)