Imperial College London

Problem Sheet 1 Solutions

MATH50011
Statistical Modelling 1

Week 1

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?

Solution. If the less-knowledgeable participants and the oxen-raising participants both guess the correct weight on average, then the model will be reasonable.

However, suppose that we assume $(Y|X = 1) \sim N(\mu, \sigma_1^2)$ and $(Y|X = 0) \sim N(\mu, \sigma_0^2)$ for $X \sim Bernoulli(\pi)$. The marginal cdf of Y, $P(Y \leq y)$, can be written as

$$P(Y \leq y|X = 1)P(X = 1) + P(Y \leq y|X = 0)P(X = 0) = \pi\Phi\left(\frac{y-\mu}{\sigma_1}\right) + (1-\pi)\Phi\left(\frac{y-\mu}{\sigma_0}\right)$$

which is not the cdf of a normal distribution (unless $\sigma_0 = \sigma_1$). Hence, we need to be careful about how we describe the model used.

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?

Solution. If x_i is fixed, then each time we repeat the same study the sequence $x_1, x_2, \ldots$ will be identical. This determinism only makes sense if we have designed an experiment where we carefully control the values of x_i that get sampled.

In observational studies, the x_i are usually treated as the realization of a random variable X_i so that we are sampling iid random vectors (Y_i, X_i).

However, if we are interested in the association between Y_i and X_i we usually only need to model the distribution of $(Y_i|X_i = x_i)$. In such cases where we condition on the values of $X_i = x_i$, we can usually treat the covariates as fixed for the purpose of estimation/inference.

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

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

Solution. Let $Z = T - \theta$. We have $E(Z) = bias(T)$, $Var(Z) = Var(T)$ and $E(Z^2) = MSE(T)$. This means that

$$Var(T) = Var(Z) = E(Z)^2 - \{E(Z)\}^2 = MSE(T) - bias(T)^2.$$

The result follows by solving for $MSE(T)$.

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?

Solution. First, consider $\bar{Y}$. By properties of $E(\cdot)$ and $Var(\cdot)$ for independent random variables, we have

$$\begin{aligned}
E(\bar{Y}) &= E(n^{-1}\sum Y_i) = n^{-1}\sum E(Y_i) = n^{-1}n\lambda^{-1} = \lambda^{-1} \\
bias(\bar{Y}) &= E(\bar{Y}) - \lambda^{-1} = 0 \\
Var(\bar{Y}) &= Var(n^{-1}\sum Y_i) = n^{-2}\sum Var(Y_i) = n^{-1}\lambda^{-2} \\
MSE(\bar{Y}) &= Var(\bar{Y}) + \{bias(\bar{Y})\}^2 = n^{-1}\lambda^{-2}.
\end{aligned}$$

For T, we have

$$\begin{aligned}
E(T) &= E(n\bar{Y}/(n+1)) = nE(\bar{Y})/(n+1) = \frac{n}{n+1}\lambda^{-1} \\
bias(\bar{Y}) &= E(T) - \lambda^{-1} = \frac{-1}{n+1}\lambda^{-1} \\
Var(T) &= Var(\frac{n}{n+1}\bar{Y}) = \frac{n^2}{(n+1)^2}Var(\bar{Y}) = \frac{n}{(n+1)^2}\lambda^{-2} \\
MSE(T) &= Var(T) + \{bias(T)\}^2 = \frac{n}{(n+1)^2}\lambda^{-2} + \frac{1}{(n+1)^2}\lambda^{-2} = \frac{1}{n+1}\lambda^{-2}.
\end{aligned}$$

While $\bar{Y}$ is unbiased and T is biased, T has lower MSE for all values of λ.

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 $\text{Var}(S) = 0$.

Solution.

(a) $E(\bar{Y}^2) = Var(\bar{Y}) + [E(\bar{Y})]^2 = n^{-1}\sigma^2 + \mu^2 \neq \mu^2$ unless $\sigma^2 = 0$.

(b) $Var(S) = E(S^2) - (E(S))^2 = \sigma^2 - (E(S))^2$ so

$$E(S) = \sqrt{\sigma^2 - Var(S)} = \sigma \Leftrightarrow Var(S) = 0.$$

6. 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 $\text{Var}_\theta(T_j) = \sigma_j^2$ for $j = 1, 2$ and $\text{corr}(T_1, T_2) = \rho$.

   (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 $\text{MSE}_\theta(T_1 + \alpha T_2)$;

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

Solution.

(a) Since T_1 is unbiased, $MSE(T_1) = \sigma_1^2$. For $T_1 + T_2$, we find

$$\begin{aligned} E(T_1 + T_2) &= E(T_1) + E(T_2) = \theta + 0 = \theta \\ bias(T_1 + T_2) &= E(T_1 + T_2) - \theta = 0 \\ Var(T_1 + T_2) &= \sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2 \\ MSE(T_1 + T_2) &= \sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2 \end{aligned}$$

since $T_1 + T_2$ is again unbiased. Comparing the MSE of T_1 and $T_1 + T_2$ is equivalent to comparing their variances. We have

$$Var(T_1 + T_2) - Var(T_1) = \sigma_2^2 + 2\rho\sigma_1\sigma_2$$

which is less than zero if $-1 < \rho < -\frac{1}{2}\frac{\sigma_2}{\sigma_1}$ and greater than zero if $-\frac{1}{2}\frac{\sigma_2}{\sigma_1} < \rho < 1$.

(b) By similar calculations we have

$$\begin{aligned} E(T_1 + \alpha T_2) &= E(T_1) + \alpha E(T_2) = \theta + 0 = \theta \\ bias(T_1 + \alpha T_2) &= E(T_1 + \alpha T_2) - \theta = 0 \\ Var(T_1 + \alpha T_2) &= \sigma_1^2 + \alpha^2\sigma_2^2 + 2\alpha\rho\sigma_1\sigma_2 \\ MSE(T_1 + \alpha T_2) &= \sigma_1^2 + \alpha^2\sigma_2^2 + 2\alpha\rho\sigma_1\sigma_2\,. \end{aligned}$$

(c) To find a minimum we set the first derivative equal to zero

$$\frac{d}{d\alpha}MSE(T_1 + \alpha T_2) = 2\alpha\sigma_2^2 + 2\rho\sigma_1\sigma_2 \equiv 0$$

and find that $\tilde{\alpha} = -\rho\sigma_1/\sigma_2$ is the minimizer since $\frac{d^2}{d\alpha^2}MSE(T_1 + \alpha T_2) = 2\sigma_2^2 > 0$ for all α.

(d) We have that $MSE(T_1 + \alpha T_2) = \sigma_1^2 + \tilde{\alpha}^2\sigma_2^2 + 2\tilde{\alpha}\rho\sigma_1\sigma_2 = \sigma_1^2(1 - \rho^2) \leq \sigma_1^2 = MSE(T_1)$ with equality if and only if $\rho \in \{-1, 1\}$.

R lab: Descriptive statistics

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

7. 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").

Solution.

(a-c) See the code used in `Rlab-Week-1.R` file.

(d) There are many possible descriptive statistics that could be reported. Some combination of measures of center (e.g. mean, median) and spread (e.g. standard deviation, interquartile range, min/max) would be fairly typical.

Sample Size	Mean	Median	Std. Dev.	IQR	Min	Max
17183	1287	1245	622	635	1	14555

The following four types of plots are all potentially useful ways to visualize the sample. We display them in Figure 1 below.

A. The default histogram has far too few bins to be of use, so we increased the number of breaks to 75.

B. The boxplot shows us the minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum of the sample.

C. The density plot is a smooth alternative to the histogram. Both options estimate the pdf without assuming a parametric form.

D. The quantile-quantile (Q-Q) plot plots the sample quantiles against the quantiles of a N(0,1) distribution. Major deviation from a linear relationship may indicate that the sample was not drawn from a normal distribution.

(e) Planet Money's contest had 17,183 participants guess the weight of Penelope the cow. The guesses ranged all the way from 1 lb to 14,555 lbs. The average guess was 1,287 lbs with a standard deviation of 622 lbs. It is clear from any one of the histogram, boxplot or density estimate that most of the data is concentrated near the mean but with a long upper tail. This extreme tail would be surprising if the data were drawn from a normal distribution. Alternatively, the Q-Q plot in Figure 1D. shows that, based on deviation from the straight line, the upper sample quantiles do not agree well with the normal distribution.

(f) The sample mean of 1,287 lbs is 14.3 standard errors below Penelope's true weight of 1,355 lbs. This is based on the calculation

$$\frac{\bar{y} - \mu}{sd(y)/\sqrt{n}} = \frac{1287 - 1355}{635/\sqrt{17183}} \approx -14.3$$

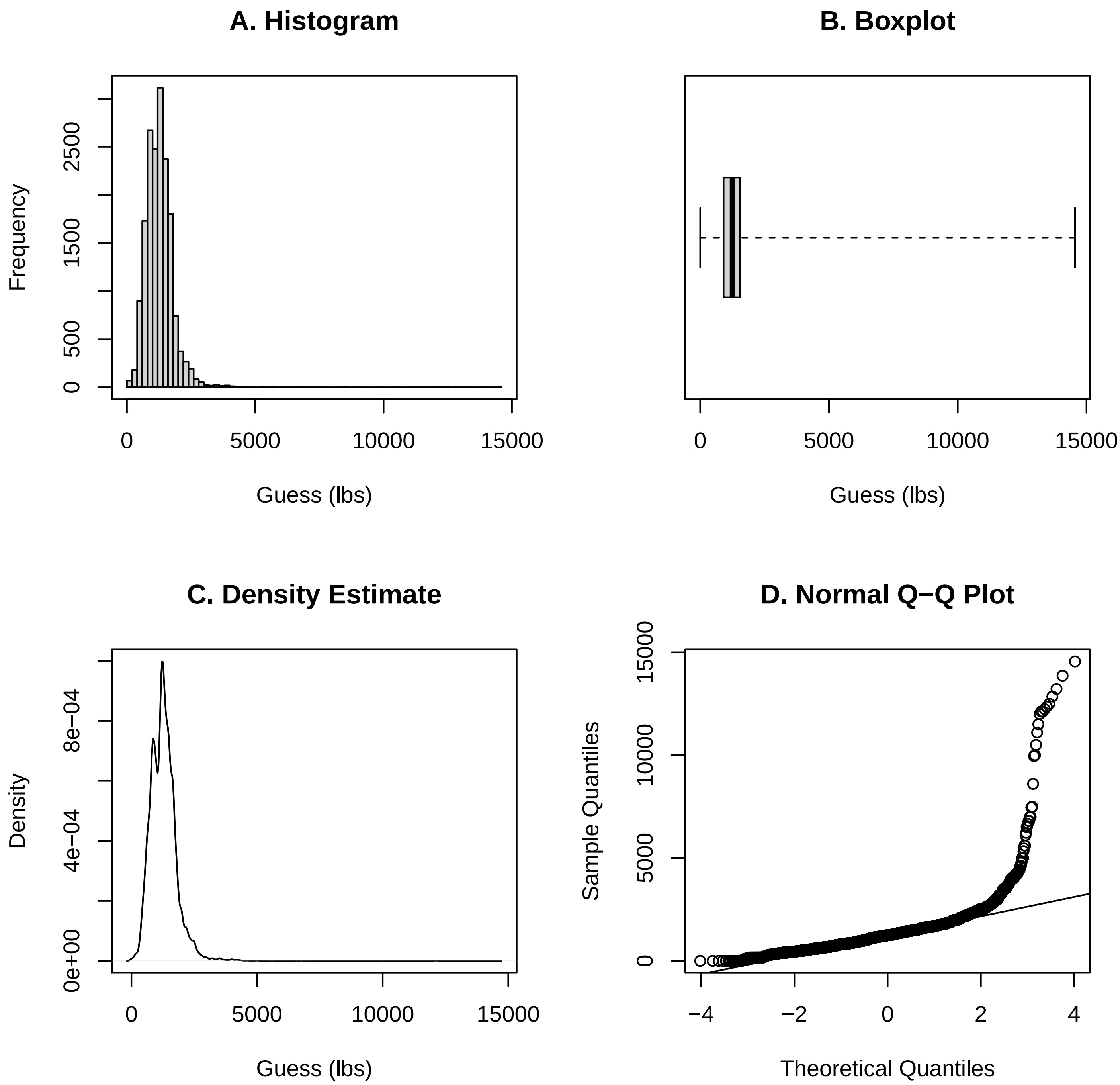

Figure 1: Four different plots using the Planet Money data. Panel A shows a histogram of the guesses (in lbs) with 75 bins. Panel B shows a boxplot of the data. Panel C shows a smooth density estimate, which exhibits similar features to the histogram. Panel D shows a normal Q-Q plot based on the data. In all panels, we see that there is a long upper tail that would be surprising if the data were drawn from a normal distribution.