

MSc/Diploma in Applied Statistics Week 0 Review Problems 2009

Students undertaking the MSc in Applied Statistics often have very different academic and professional backgrounds. However, lecturers must assume that everybody has a certain level of knowledge and familiarity with statistics and probability on which the course can build. The main purpose of these exercises is to help you and your supervisor identify if there are any areas of assumed knowledge in which you may benefit from further revision before the coursework begins in earnest.

If you have difficulty in answering the questions, the necessary background material can generally be found in most introductory texts on probability and statistics. References given below for each question refer to the following books, which are available in the library: (R) *Mathematical Statistics and Data Analysis* by John A. Rice, (DeG) *Probability and Statistics* by Morris DeGroot (2nd edition).

Please attempt all of the questions and hand in your answers, neatly written, to the box inside the entrance of 1 South Parks Road before 10 am on Monday of week 1.

1. Random Variables (R:ch2, DeG:ch3)

- (a) Describe the features of the density function $f_X(x)$ in figure 1, and sketch the corresponding distribution function $F_X(x)$.

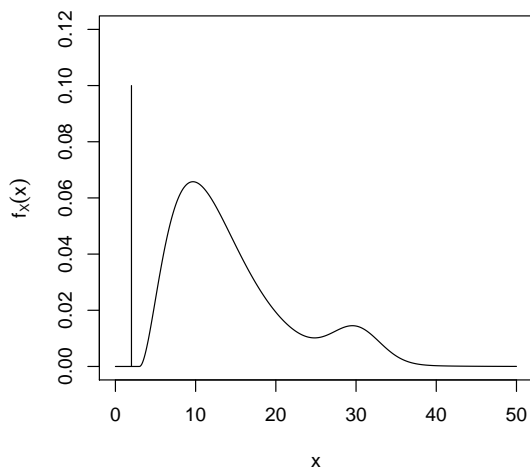


Figure 1: The density function of a random variable X .

2. Expectation (R:ch4, DeG:ch4)

- (a) Define the expected value of a function $g(X)$ of a random variable X .
(b) Suppose X and Y are continuous random variables, whose first and second moments exist. Show that

$$E_Y[E_{X|Y}[X]] = E_X[X],$$

where $E[\cdot]$ denotes expected value, and explain why this result might be useful.

- (c) Derive the moment generating function (mgf) of an exponential random variable.
- (d) Show that if X_1, X_2, \dots, X_n are independent, identically distributed exponential random variables, then

$$Y = \sum_{i=1}^n X_i$$

has a gamma distribution.

3. Bayes rule (R:ch1, DeG:ch2)

- (a) Suppose A and $B_i, i = 1, 2, \dots, n$ are events with non-zero probability, and that $\{B_1, B_2, \dots, B_n\}$ is a partition of the sample space Ω . State and prove Bayes rule, expressing $P(B_j|A)$ in terms of $p(B_i)$ and $p(A|B_i), i = 1, 2, \dots, n$.
- (b) A group of n players is divided into two teams (red and blue) according to the following procedure: First, a number X is chosen randomly from the set $\{1, 2, \dots, n-1\}$, with all values equally likely. Then X of the n players are chosen to form the red team, with all possible sets of size X equally likely. The remaining $n - X$ players form the blue team.
- Find the mean size of the red team.
 - Consider one particular player, called player A. Find the probability that the team containing player A has size k , for $k = 1, 2, \dots, n-1$.
 - Find the mean size of the team containing player A.
 - After the teams have been chosen, each team selects a captain uniformly at random from the members of the team. Find the conditional distribution of the size of the team containing player A, given that player A is the captain.

Hint: You may use (without proof) the following identities:

$$\sum_{i=1}^n i = \frac{n(n-1)}{2}, \quad \sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{6}.$$

4. Likelihood (R:ch8, DeG:ch6) Estimation (DeG:ch7)

- (a) Suppose X has a pareto distribution with parameters $\alpha > 0$ and β , so

$$f(x) = \begin{cases} \frac{\alpha\beta^\alpha}{x^{\alpha+1}} & x > \beta, \\ 0 & x < \beta \end{cases}$$

Find the mean and variance of X .

- (b) If X_1, X_2, \dots, X_n is a random sample from a Pareto distribution, find the maximum likelihood estimator of α , assuming that β is known.
- (c) Explain what a likelihood function is, and why the maximum likelihood estimator might be a reasonable choice of estimator.
- (d) Given i.i.d. random variables X_1, X_2, \dots, X_n assumed to be from a gaussian distribution, $S^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / (n-1)$ is often used as an estimator of the variance σ^2 . Show that this estimator is unbiased, and explain why it might be preferred to the maximum likelihood estimator $(n-1)/nS^2$.

5. **Linear models, matrix algebra** (R:ch14, DeG:ch10)

- (a) Consider the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_{p-1} x_{i(p-1)} + \epsilon_i, \epsilon_i \sim N(0, \sigma^2), i = 1, \dots, n.$$

Write this model in matrix form, carefully stating the dimension and elements of the matrices you define.

- (b) Using matrix algebra, find the value of the parameters $\beta_0, \dots, \beta_{p-1}$ which minimise the sum of squared errors ϵ_i .
- (c) When fitting a linear model, it is common to use the “least squares” estimate of the parameters $\beta_0, \dots, \beta_{p-1}$. Why do you think this is the case, and do you think this choice is justified?

6. **Hypothesis testing** (R:ch9, DeG:ch8) **confidence intervals** (DeG:ch7) **statistical tests** (R: ch11,ch13, DeG:ch8,ch9)

- (a) Define the type-1 error of a hypothesis test, the power of a test, and describe what a p-value and a confidence interval are.
- (b) The data in the table below give the silver content of coins from the 1st and 4th coinages of King Manuel 1, Comnemus (1143-1180).

1st coinage	5.9	6.8	6.4	7.0	6.6	7.7	7.2	6.9	6.2
4th coinage	5.3	5.6	5.5	5.1	6.2	5.8	5.8		

- i. State how you would check the validity of the assumption that the data from each coinage is normally distributed.
- ii. Assuming the data is normally distributed, perform appropriate statistical tests of the hypotheses that the means and variances of the silver content of each coinage are equal.
- iii. Interpret your results in the context of the question.