Imperial College London

MSc in Statistics
Global MSc in Statistics
Department of Mathematics
Imperial College London

Preparing for
the 2022-23 Postgraduate Degrees
MSc in Statistics & Global MSc in Statistics
at Imperial College London

This document is intended to help you prepare for the MSc in Statistics and Global MSc in Statistics degrees at Imperial College London, and in particular for the core courses in the Autumn term.

This document includes study questions and references to

1. help you determine areas of statistics in which your working knowledge is at a sufficiently advanced level;
2. provide suggestions for self-study in areas of statistics that you find challenging;
3. allow you to arrive reassured that you are well equipped to embark on an exciting year of postgraduate study in Statistics at Imperial College London.

We thus recommend that you attempt the questions first, and then follow up with the recommended reading on any difficulties, in order to give you the best possible start in October.

1 Probability for Statistics

The *Probability for Statistics* module in the Autumn term assumes familiarity with standard concepts in probability theory, and fluency in applying and manipulating these concepts. The document "Probability for Statistics Refresher" covers the basic probability concepts that we expect students to be familiar with before starting the course. Overall, the *Probability for Statistics* module will cover material similar to the textbooks "Essentials of probability theory for statisticians" by Proschan and Shaw (2016) and "Probability and random processes" by Grimmett and Stirzaker (2001).

If you had significant exposure to classical probability theory in your previous studies, we strongly recommend that you take a good look at your undergraduate sources – textbooks, lecture notes etc – and refresh your working knowledge and calculus skills.

For those who have some but not much exposure to classical Probability Theory, we recommend reading the first eight chapters of "Probability with Martingales" by David Williams to familiarize themselves with rigorous probability theory.

For those who have little or no exposure to probability theory, we recommend reading the first four chapters of "All of Statistics" by Wasserman (Wasserman, 2003).

If you want advanced texts for personal interest, we recommend "Probability: Theory and Examples" by Rick Durett or "Probability and Measure Theory" by Billingsly. These texts are very good references but their difficulty level is higher than that of the module, so they are entirely optional readings.

The following exercises should help you review assumed working knowledge and identify areas in which self-study may be required:

1. (a) Formally define the following concepts: A probability space. A random variable. Independence of two events A and B.
   (b) Let Ω be a nonempty set:
      i. Show that the collection $\mathcal{B} = \{\emptyset, \Omega\}$ is a sigma algebra.
      ii. Let $\mathcal{B} = \{$ all subsets of Ω, including Ω itself$\}$. Show $\mathcal{B}$ is a sigma algebra.
      iii. Show the intersection of two sigma algebras is a sigma algebra.
   (c) Consider the probability space $(\Omega, \mathcal{B}, \mathrm{P})$ with $A, B \in \mathcal{B}$. Using only the Kolmogorov axioms prove
      i. $P(A) \leq 1$
      ii. If $A \subset B$, then $P(A) \leq P(B)$, and
      iii. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

   Purpose: Review of basic properties of a probability space and of sigma algebras. The first four chapters of Williams (1991) may be helpful.

2. The Library Cafe has two tills at which you can pay. One has a slow server and the other a fast server. Upon arrival at the tills, you are routed to server i with probability p_i for $i = 1, 2$, where $p_1 + p_2 = 1$. The service time at server i is exponentially distributed with parameter μ_i for $i = 1, 2$. What is the probability density of your service time at the tills?

   Purpose: Tests knowledge of mixture distributions. For fun think, about what would happen with a countably infinite number of tills. What about an uncountably infinite number of tills? What would happen if you were allowed to choose the till?

3. The *Poisson distribution* $P(\lambda)$ with *parameter* $\lambda > 0$ is defined by

   $$P(X = \lambda) = e^{-\lambda}\lambda^k/k! \qquad (k = 0, 1, 2, \ldots)$$

   Show that if X and Y are independent and distributed as $P(\lambda)$ and $P(\mu)$ respectively, then $X + Y$ is $P(\lambda + \mu)$.

   Purpose: Review of basic distributional properties of the Poisson distribution. There are various ways of showing this, but there is a rather quick argument available.

4. (i) Obtain the *convolution formula*: if X, Y are independent random variables, with densities f and g, then $X + Y$ has density h, where

   $$h(x) = \int_{-\infty}^{\infty} f(x - y)g(y)dy.$$

   (ii) With

   $$\Gamma(\lambda) := \int_0^{\infty} e^{-x}x^{\lambda-1}dx,$$

the *gamma distribution* with *parameter* $\lambda > 0$ has density

$$f(x) = e^{-x}x^{\lambda-1}/\Gamma(\lambda) \qquad (x > 0).$$

By finding the density of $X + Y$ where X, Y are independent and Gamma distributed with parameters $\alpha, \beta > 0$, or otherwise, obtain *Euler's integral for the Beta function*:

$$B(\alpha, \beta) := \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta).$$

Purpose: Tests convolution of random variables and algebraic manipulations.

5. (i) For the *Cauchy* distribution, with density

$$f(x) := \frac{1}{\pi(1+x^2)},$$

show that the CF is $e^{-|t|}$. (You may find Complex Analysis useful here. One background source is the undergraduate course M2P3 Complex Analysis, see `http://www.ma.ic.ac.uk/~bin06/M2PM3-Complex-Analysis`.)
(ii) For f the *symmetric exponential* density,

$$f(x) := \frac{1}{2}e^{-|x|},$$

show that the CF is $1/(1+t^2)$.
(iii) Comment on the relationship between (i) and (ii)

Purpose: Review of complex integration, you will not be expected to perform very difficult complex integration in this module.

6. The *Student* t*-distribution* with n *degrees of freedom*, $t(n)$, is defined as that of the ratio $t := X\sqrt{n-1}/U$, where X and U are independent, $X \sim N(0,1)$ and U has *chi-square* distribution with n *degrees of freedom* (df), $\chi^2(n)$. The density of $t(n)$ is

$$f(x) = \frac{\Gamma(\frac{1}{2}+\frac{1}{2}n)}{\sqrt{n}\Gamma(\frac{1}{2})\Gamma(\frac{1}{2}n)}\left(1+\frac{x^2}{n}\right)^{-\frac{1}{2}(n+1)}.$$

Show that $t(n)$ converges to standard normal as $n \to \infty$.

Purpose: Revision of commonly used asymptotic arguments. You may find Stirling's approximation useful, as well as the limit definition of the exponential function.

7. Suppose that $X_1, \ldots, X_N$ are a random sample from the uniform distribution $U(0, \theta)$, where $\theta > 0$ is unknown.

   (a) What's the maximum likelihood estimator of θ given $X_1, \ldots, X_N$?

   (b) Suppose we adopt a Bayesian approach to estimating θ. Assume that we put a Pareto prior distribution for theta, i.e.

$$p(\theta) = P(a, b, K) = \frac{b^K}{(a^{K+1})} \qquad (1)$$

   if $a \geq b$ and 0 otherwise.

(c) Given $D = \{X_1, \ldots, X_N\}$ calculate the joint distribution $p(D, \theta)$, and deduce the posterior distribution $p(\theta|D)$.

(d) Write down the predictive density of observing a new sample x, i.e. $p(x|X_1, \ldots, X_N)$.

Purpose: Review of Bayesian statistics. If Bayesian statistics is new to you, then Chapter 11 of Wasserman (2003) can serve as an introduction.

8. The gas in a closed container is composed of a very large number of atoms. The atoms are in perpetual motion (why?), and collide with each other. The atoms are subject to Newton's Laws of Motion, and these are *time-reversible* (invariant under reversal of the direction of the arrow of time). But the behaviour of a gas is in general *not* time-reversible (think of observing smoke spreading). How can you account for this discrepancy?

   Purpose: Tests for concepts related to the reversibility and stationarity of stochastic processes. This may be unfamiliar to some, but we will define these concepts in class.

9. You've been told that the average length of a fox is 15 inches. You think you've seen a fox on the Queen's Lawn, but are not sure. You estimate it is at least 20 inches long.

   (a) Give a nontrivial upper bound on the probability that a fox could be at least 20 inches long.

   (b) The standard deviation this height distribution is 2 inches. Find a lower bound on the probability that a given fox is between 10 and 20 inches long.

   (c) Now assume this distribution is normal. Repeat the calculation from the second question. How close was your bound to the true probability?

   Purpose: Applications of concentration inequalities and evaluating the tightness of these inequalities. Chapter 4 of Wasserman (2003) may be helpful.

10. Each person has two genes which are relate to a disorder. Each gene is either N or C. Each child receives one gene from each parent. If your genes are NN or NC or CN then you are normal; if they are CC then you have the disorder.

    (a) Neither of Sally's parents has the disorder. Nor does she. However, Sally's sister Hannah does have the disorder. Find the probability that Sally has at least one C gene (given that she does not have the disorder herself).

    (b) In the general population the ratio of N genes to C genes is about 49 to 1. You can assume that the two genes in a person are independent. Harry does not have the disorder. Find the probability that he has at least one C gene (given that he does not have the disorder).

    (c) Harry and Sally plan to have a child. Find the probability that the child will have the disorder (given that neither Harry nor Sally has it).

    Purpose: An application of probability theory to genetics and manipulations of conditional probabilities.

2 Fundamentals of Statistical Inference

The *Fundamentals of Statistical Inference* module in the Autumn term will assume good working knowledge of the main ideas of statistical inference, and fluency in the necessary underlying probability/distribution theory. Strongly recommended are Chapters 1-12 of Wasserman (2003). Chapters 1-5 of that book give a concise account of probability/distribution theory background which will be assumed, and Chapters 6-12 give an introduction to key elements of the

Fundamentals of Statistical Inference module. A very suitable, more detailed, reference is Casella and Berger (2002).

Distribution theory required by the module will include: standard probability distributions and relationships between them; identification and manipulation of distributions by the moment generating function; sampling from normal distributions and distributions related to the normal. These ideas will be used regularly, and should therefore be studied in preparation.

The course will assume familiarity with basic ideas of statistical inference, in particular: key notions of frequentist and Bayesian inference; point estimation (bias, mean squared error, construction of estimators by maximum likelihood); hypothesis testing (including notions of size, power, critical region, optimal construction for simple hypotheses using the Neyman-Pearson Lemma); confidence sets (properties, construction via pivotal quantities and inversion of hypothesis tests). These concepts will be reviewed early in the course.

Consider the following questions to help you review assumed working knowledge and identify areas in which we recommend further self-study:

1. What does it mean to say that an estimator $\hat{\theta}$ of a parameter θ is *unbiased*? What is meant by the *mean squared error* of $\hat{\theta}$? What is meant by *consistency* of an estimator?

   Let $X_1, \ldots, X_n$ be independently, identically distributed (IID) with the uniform distribution on $(0, \theta)$. Let $\hat{\theta}_1 = \max\{X_1, \ldots, X_n\}$ and $\hat{\theta}_2 = 2\bar{X}$, with $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$. Calculate the biases and mean squared errors of $\hat{\theta}_1$ and $\hat{\theta}_2$. Which estimator do you prefer?

2. What are the main characteristics of the Bayesian approach to statistical inference?

   Let $X_1, \ldots, X_n$ be IID $N(\theta, \sigma^2)$, and suppose that the prior distribution for θ is $N(\mu, \tau^2)$, where σ^2, μ, τ^2 are known. Determine the posterior distribution for θ, given $X_1, \ldots, X_n$. How would a (i) frequentist, (ii) Bayesian statistician estimate θ?

3. In the context of hypothesis testing, define the following terms: (i) simple hypothesis; (ii) critical region; (iii) size; (iv) power and (v) type II error probability.

   Let X be a *single* random variable, with a distribution F. Consider testing the null hypothesis $H_0 : F$ is standard normal, $N(0, 1)$, against the alternative hypothesis $H_1 : F$ is double exponential, with density $\frac{1}{4} e^{-|x|/2}$, $x \in \mathbf{R}$. Find the test of size $\alpha, \alpha < 1/4$, which maximizes power, and show that the power is $e^{-t/2}$, where $\Phi(t) = 1 - \alpha/2$ and Φ is the distribution function of $N(0, 1)$.

4. Let $X_1, \ldots, X_n$ be IID $N(\mu, \sigma^2)$, with σ^2 known and μ unknown. How would you test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$? Find, in terms of σ^2, how large the size n of the sample must be in order for there to exist a 95% confidence interval for μ of length no more than some given $\epsilon > 0$.

   How would you test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ in the situation where σ^2 is *unknown*?

5. What is meant by a *confidence set* for an unknown parameter? What is meant by the *maximum likelihood estimator* of an unknown parameter?

   Let $X_1, \ldots, X_n$ be IID from the exponential distribution with density $f(x; \theta) = \theta e^{-\theta x}$, $x > 0$. Find the maximum likelihood estimator of θ. Is it biased? What is the distribution of $n/\hat{\theta}$?

What is the distribution of $\theta/\hat{\theta}$? Show how to construct a $100(1-\alpha)\%$ confidence interval for θ of the form $(0, c\hat{\theta})$, for a constant $c > 0$ depending on α.

A Gamma distribution, $\Gamma(k, \lambda)$, has probability density function of the form $f(x; k, \lambda) = \lambda^k x^{k-1} e^{-\lambda x}/\Gamma(k),\ x > 0$. Taking a Bayesian point of view, suppose your prior distribution for θ is $\Gamma(k, \lambda)$. What is the posterior distribution for θ?

6. Let the random variable X have probability density function

$$f(x; \mu, \lambda) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left\{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right\},\ x > 0, \text{ where } \ \mu > 0,\ \lambda > 0.$$

(i) Verify that the moment generating function of X is of the form

$$M_X(t) = \exp\left\{\frac{\lambda}{\mu}\left(1 - \sqrt{1 - 2t\mu^2/\lambda}\right)\right\},$$

and hence, or otherwise, find the mean of X.

(ii) Identify the distribution of Y, where

$$Y = \frac{\lambda(X-\mu)^2}{\mu^2 X}.$$

(iii) Let $X_1, \ldots, X_n$ be IID, with common density function $f(x; \mu, \lambda)$. Find the form of the maximum likelihood estimators $\hat{\mu}, \hat{\lambda}$ of μ, λ. What is the distribution of $\hat{\mu}$?

3 Applied Statistics

The *Applied Statistics* and *Computational Statistics* modules in the Autumn term will heavily use the **R** programming language, as well as many elective modules in the Spring term. Several Spring term modules will also use **python**, **TensorFlow**, **Stan** etc., but for the start of term 1 a firm grasp of **R** only is advantageous. Suitable background reading for the *Applied Statistics* and *Computational Statistics* modules are Venables and Ripley (2004); Faraway (2004, 2005).

If you are not well familiar with **R**, consider the “Introduction to R” document. Another excellent guide to **R** is ”An Introduction to R”, written by the **R** Core Team, and available at `http://cran.r-project.org/doc/manuals/r-release/R-intro.html`. Another good starting point are likely the interactive tutorials offered by the **R** swirl package, available at `http://swirlstats.com/students.html`. If you have a lot of time available, you may want to do the Coursera course on R programming (`https://www.coursera.org/course/rprog`). Throughout the degree, the **R** Cheat Sheets will prove useful, and these are available at `https://www.rstudio.com/resources/cheatsheets`.

If you are a more experienced programmer, but have not done extensive work in **R**, it would also be a good idea to consider the “Introduction to R” document.

To use **R**, there are many options including the default editor that comes with the **R** installation, though **RStudio** is probably the best choice, and is available at `http://www.rstudio.com/`. Useful features of **RStudio** are LaTeX, github, and Rmarkdown integration.

Consider the following questions to help you review assumed working knowledge and identify areas in which we recommend further self-study. We recommend to write short reports in LaTeX that contain your answers.

1. The following data were obtained in an experiment

0.27695500	1.19025212	1.15439013	0.68360395	1.29513634	0.84684675
0.76268877	0.38309755	0.22700716	0.27854125	0.38530675	0.48182418
0.20216833	0.89146250	0.77185243	0.00134230	0.00132544	0.00132002
0.00132965	1.74544500				

By means of graphical and numerical summary, describe the main features of these data. Pay careful attention to the precision of the numbers, and consider any appropriate transformations.

2. Compare and contrast a two-sample t-test with the Kolmogorov-Smirnov two-sample test. Conduct a one-sample KS test to determine whether the data in Q1 are likely to have arisen from a Gamma distribution:

$$f(x) = \frac{1}{s^a \Gamma(a)} x^{a-1} e^{-x/s}$$

where $a = 2$ and $s = 1/2$.

Background Reading: Kanji (2006), Casella and Berger (2002, Chapter 8)

3. Consider a series of independent trials, each of which results in success or failure, with common probability of success θ. The distribution of the number of failures X before the first success has probability mass function

$$P(X = x) = (1 - \theta)^x \theta$$

for $x = 0, 1, 2, \ldots$, with $\theta > 0$.

Suppose a random sample of n observations from this distribution is obtained. Show (and verify) that the maximum likelihood estimator is

$$\hat{\theta} = \frac{1}{\bar{x} + 1}$$

where $\bar{x}$ is the sample mean.

A locksmith keeps records about the number of test keys he tries, that fail, in each lock before the lock opens. A random selection of these records yields the following counts

x	0	1	2	3	≥ 4
count	12	11	5	5	7

Suppose it is claimed that the number of test keys follows the distribution above. Perform an appropriate test, at the 5% significance level, to test this claim. Clearly describe the test and your conclusions.

4. Suppose we have a sample of observations, $\{(y_i, x_i)\}_{i=1}^n$, where $y_i \in \mathbb{R}$ is the *response* of the ith experimental unit, and $x_i \in \mathbb{R}$ is the corresponding *covariate*. Consider the model

$$Y = \alpha + \beta X + \varepsilon$$

   (a) Show that the least squares estimators are

$$\hat{\alpha} = \bar{x} - \hat{\beta}\bar{y}$$

   and

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})}$$

(b) Write down the elements of this model, and the solution, in the standard matrix formulation.

(c) Suppose $\varepsilon \sim N(0, \sigma^2)$. Comment on the relationship between the maximum likelihood estimators, and the least squares estimators. Are other assumptions required?

Background Reading: Casella and Berger (2002, Chapter 11), Weisberg (2005)

5. The following data are response (y) and covariate (x)pairs:

x	y
2.6550866	3.644926
3.7212390	4.659500
5.7285336	4.654434
9.0820779	4.316676
2.0168193	4.484011
8.9838968	5.255947
9.4467527	4.119184
6.6079779	3.231716
6.2911404	1.658974
0.6178627	8.966016

(a) Construct a plot displaying response against covariate.

(b) Suppose the model

$$Y = \alpha + \frac{\beta}{X} + \varepsilon,$$

where $\varepsilon \sim N(0, \sigma^2)$, is known to be adequate for these data. Compute least squares estimate of the model parameters.

(c) Is this a linear model?

(d) Add the fitted regression line to the plot in part (a).

4 Computational Statistics

The *Computational Statistics* module in the Autumn term will focus on introducing algorithmic statistical methods. Prior knowledge of statistical inference techniques such as the ones mentioned in the sections above will be assumed. Most algorithms will be presented in pseudo-code, and students need to be familiar with the **R** language, be able to write for, while loops, and if statements, and they should be able to write simple **R** functions involving these elements. The *Computational Statistics* module will build on key concepts of discrete-time discrete-state Markov chains. If you are not well familiar with these, consider (Norris, 1997, Chapter 1)).

The questions below are intended to give you some practice with simple computational routines, and identify areas in which we recommend further self-study.

1. Consider a sample $X_1, \ldots X_n$ of i.i.d. random variables following a geometric distribution with probability mass function

$$P(X = x) = \theta \, (1 - \theta)^x \quad \forall x \in \{0, 1, 2, 3, \ldots\}$$

where the success probability $0 < \theta < 1$ and is unknown.

(a) Write down the log-likelihood function $L(\theta | \{x_1, \ldots, x_n\})$ of a sample $\{x_1, \ldots, x_n\}$.

(b) Determine the maximum likelihood estimator (MLE) of the success probability θ. Prove that it is indeed a maximum.

(c) Write a R code in order to

i. Generate a sample of size $n = 1000$ of i.i.d. random variables distributed according to a geometric distribution with a success probability $\theta = 0.2$ by using the function rgeom.

ii. Plot the log-likelihood function $L(\theta|\{x_1, \ldots, x_n\})$ for this sample as a function of θ.

iii. Estimate by MLE the parameter θ. Add the point $(\hat{\theta}, L(\hat{\theta}|\{x_1, \ldots, x_n\})$ where $\hat{\theta}$ is the obtained MLE) on the log-likeliood plot from the previous question.

2. Consider the following random walk: $X_0 \sim \mathrm{N}(0, 1)$ and $X_t|(X_0, \ldots X_{t-1}) \sim \mathrm{N}(X_{t-1}, 1)$. (Here the notation $\mathrm{N}(\mu, \sigma^2)$ indicates the Normal distribution with mean μ and variance σ^2.)

(a) Derive the marginal distribution of X_t for each t.

(b) Derive the correlation between X_t and X_{t+1}.

(c) Use a software package such as **R** to simulate 1000 realizations of the random walk, $(X_1, \ldots, X_{50})$. Make normal quantile plots of the 1000 draws of X_1. Make similar plots for X_2, X_{10} and X_{50}.

(d) Based on your simulation, plot the correlation between X_0 and X_t for $t = 1, \ldots, 10$. How does this compare to your answer to part (b)?

3. Consider the following random walk: $X_0 \sim \mathrm{N}(100, 1)$ and $X_t|(X_0, \ldots X_{t-1}) \sim \mathrm{N}(\rho X_{t-1}, 1 - \rho^2)$.

(a) Derive the marginal distribution of X_t for each $t = 1, 2, 3$, as a function of ρ.

(b) Derive the correlation between X_1 and X_2, as a function of ρ.

(c) Use a software package such as **R** to simulate 1000 realizations of the random walk, $(X_1, \ldots, X_{50})$ with $\rho = 0.5$. Make a normal quantile plot of the 1000 draws of X_{50} Repeat with $\rho = 0.8, 0.95$ and 0.99. What pattern do you see?

(d) Using simulation, plot the correlation between X_0 and X_t for $t = 1, \ldots, 50$ with $\rho = 0.5$. Repeat with $\rho = 0.8$, 0.95 and 0.99. What pattern do you see?

(e) Now simulate a single chain of length 1000, $(X_1, \ldots, X_{50})$ with $\rho = 0.5$. Use computer software (e.g., the `acf` command in **R**) to plot the autocorrelation function of this chain. Repeat with $\rho = 0.8$, 0.95 and 0.99. How do these plots compare with your plots from part (d)?

(f) What is the stationary distribution of the random walk?

(g) Comment on how long the chain must be, as a function of ρ, before it returns a sample from its stationary distribution.

4. Consider the *log-linear model*

$$Y \sim \mathrm{Poisson}(\lambda_i) \text{ with } \log(\lambda_i) = \alpha + \beta x_i \text{ for } i = 1, \ldots, 5,$$

where $x = (1, 5, 1, 2, -1)$ and $(\alpha, \beta) = (1, -0.5)$.

(a) Compute the expected Fisher information for (α, β) under this model.

(b) A data set simulated under this model is $Y = (3, 2, 1, 0, 4)$. Use **R** to fit the model to this data. Report the observed Fisher information matrix and 95% confidence intervals for (α, β).

(c) Simulate 1000 draws from the sampling distribution of the maximum likelihood estimates of (α, β). (Hint: You will need to simulate 1000 data sets under the model and fit the model to each.). Make a scatter plots of your 1000 draws.

(d) Comment on the bias, variance and mean square error of the two estimates.

(e) How do the observed and expected Fisher information matrices compare with the variance-covariance matrix of the sampling distribution of the estimates?

5. In a certain large statistics class there is a single mid-term exam and a final exam. Let X_1 denote the mid-term score and X_2 the final exam score. Suppose that the joint distribution of X_1 and X_2 is Bivariate Normal:

$$(X_1, X_2)^T \sim \mathrm{N}\left\{\mu = (40, 50)^T, \Sigma = \begin{pmatrix} 100 & 75 \\ 75 & 225 \end{pmatrix}\right\}. \tag{2}$$

(a) What is the marginal distribution of X_2?

(b) What is the correlation between X_1 and X_2?

(c) What is the conditional distribution of X_2 given X_1.

(d) What is the probability that a student who gets a 50 on the mid-term will get more than 60 on the final?

(e) What is the probability that X_2 is greater than X_1?

6. Consider a random variable $Y|\theta \sim \mathrm{Binomial}(n, \theta)$ and suppose we put a beta prior distribution on the probability of success, $\theta \sim \mathrm{Beta}(\alpha, \beta)$. Recall the p.d.f. of the beta distribution is given by

$$f_\Theta(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad \text{for} \quad 0 < \theta < 1, \tag{3}$$

where $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$. The mean and variance of a beta random variable are given by $E(\Theta) = \frac{\alpha}{\alpha+\beta}$ and $Var(\Theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

(a) Show $E(\Theta) = \frac{\alpha}{\alpha+\beta}$.

(b) Compute $E(Y)$. (This is the *unconditional expectation*.)

(c) Derive the conditional distribution of Θ given Y. What named distribution is this? This is the posterior distribution of Θ.

(d) Compute $E(\Theta|Y = y)$ and $Var(\Theta|Y = y)$.

(e) Derive the marginal distribution of Y. This is the prior predictive distribution.

7. Suppose $X_i|(\mu_i, \theta) \overset{\text{iid}}{\sim} \mathrm{N}(\mu_i, \sigma^2)$ where $\mu_i|\theta \overset{\text{iid}}{\sim} \mathrm{N}(\alpha, \tau^2)$ for $i = 1, \ldots, n$ with $\theta = (\sigma^2, \tau^2, \alpha)$.

(a) What is the distribution of X_i given θ?

(b) What is the joint distribution of $(X_1, \ldots, X_n)$ given θ?

(c) What is the joint distribution of $(\mu_1, \ldots \mu_n)$ given $(X_1, \ldots, X_n)$ and θ?

References

Casella, G. and R. Berger (2002). *Statistical Inference*. Duxbury.

Faraway, J. J. (2004). *Linear models with R*. CRC Press.

Faraway, J. J. (2005). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. CRC press.

Grimmett, G. and D. Stirzaker (2001). *Probability and random processes* (3rd ed.). OUP.

Kanji, G. (2006). *100 Statistical tests* (third ed.). Sage.

Norris, J. R. (1997). *Markov Chains*. Cambridge University Press.

Proschan, M. and P. Shaw (2016). *Essentials of Probability Theory for Statisticians*. Chapman & Hall.

Venables, W. and B. Ripley (2004). *Modern Applied Statistics in Splus and R*. Springer.

Wasserman, L. (2003). *All of Statistics: A Concise Course in Statistical Inference*. Springer.

Weisberg, S. (2005). *Applied Linear Regression* (third ed.). Wiley.

Williams, D. (1991). *Probability with martingales*. Cambridge university press.