MSc in Statistics
Preparing for the Course

Department of Mathematics
Imperial College London

2016/17

This document is intended to help you prepare for the MSc in Statistics (in particular for the core courses in the autumn term) - it should help you to check and improve your background knowledge and skills.

You will find some questions below - you should attempt the questions, and follow up with the recommended reading on any difficulties. It might be best to start with the starred questions ($*$).

Importantly: try to improve the areas where you think you have the greatest gaps.

Engaging with the material will help you get off to a good start!

1 Programming Background (Applied + Computational Statistics)

Most of the practical applications in the MSc will be done using the R-language. For this it obviously would be advantageous if you were to start the course already with a firm grasp of this language (most importantly its use as a programming and data manipulation language).

One excellent guide on R is "An Introduction to R", written byt he R Core Team. It is available at http://cran.r-project.org/manuals.html, direct links are at
`http://cran.r-project.org/doc/manuals/r-release/R-intro.html`
`http://cran.r-project.org/doc/manuals/r-release/R-intro.pdf`
Even if you are a very experienced programmer, but have not done extensive work in R, it would be good to work through this.

As an easy start you might find the interactive tutorials offered by the swirl package (`http://swirlstats.com`) useful to work through. Go to `http://swirlstats.com/students.html` for instructions on how to do this. 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`).

Regarding which editor you could be using:

- the default editor under Windows is decent.
- RStudio is a more full fledged editor (`http://www.rstudio.com/`). This is probably the best choice if you are not very experienced One useful feature is its LaTeX integration.
- You can use the general purpose editor Emacs or XEmacs together with the extension ESS (`http://ess.r-project.org/`). Getting to use Emacs efficiently requires a bit of effort but once you have mastered it it can be very powerful (and as it is general purpose you will be able to use it for (most) other programming languages).

2 Probability Background

Most students will have met Probability somewhere before. For these, we strongly recommend having a good look at your undergraduate (or even school) sources – textbooks, lecture notes etc. A good working knowledge of what you have met before (but, having been already examined on it, may have forgotten!) will get you off to a flying start. For those of you with no previous Probability background: we strongly recommend getting hold of a decent book (some examples follow below), and investing a serious amount of your time in reading it (at least in part), to get you to where the others will be.

The 'default option' for an undergraduate probability book in the UK is Grimmett and Stirzaker (2001). Further good elementary books include: Haigh (2002), Chung and Ait-Sahlia (2007).

1. A crime has been committed and a suspect is being held by police. He is either guilty, G, or not, G^C, and the probability of his being guilty on the basis of current evidence is $P(G) = p$, say. Forensic evidence is now produced which shows that the criminal must have a property, A, which occurs in a proportion, π, of the general population. Suppose that if the suspect is innocent he can be treated as a member of the general population, so that $\mathrm{P}\left(A \mid G^C\right) = \pi$.

   The suspect is now interrogated and found to have property A. Show that the odds on his guilt have now risen from $\lambda_0 = p/(1-p)$ to $\lambda_1 = \lambda_0/\pi$. [*The odds on an event E are defined to be the ratio* $P(E)/P(E^C)$.]

2. $(*)$ A diagnostic test has a probability 0.95 of giving a positive result when applied to a person suffering from a certain disease, and a probability 0.10 of giving a (false) positive when applied to a non-sufferer. It is estimated that 0.5 % of the population are sufferers. Suppose that the test is now administered to a person about whom we have no relevant information relating to the disease (apart from the fact that he/she comes from this population). Calculate the following probabilities:

   (a) that the test result will be positive;
   (b) that, given a positive result, the person is a sufferer;
   (c) that, given a negative result, the person is a non-sufferer;
   (d) that the person will be misclassified.

3. The normal distribution is found very widely in nature, and in statistical data. Why is this?

4. $(*)$

   (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. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

5. 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 $P(\lambda + \mu)$.

6. If f is a probability density function (briefly, density – non-negative function integrating to 1), its *characteristic function* (CF) is its Fourier transform $\hat{f}$:

$$\hat{f}(t) := \int_{-\infty}^{\infty} e^{itx} f(x)dx.$$

   Show that if f is standard normal – $f(x) = \phi(x) := e^{-\frac{1}{2}x^2}/\sqrt{2\pi}$ – then its CF is $e^{-\frac{1}{2}t^2}$. Comment on the similarity between density and CF.

7. (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)

8. Writing the expectation (or mean) of a random variables X as $E[X]$, the *variance var X* is defined by

$$var\ X := E[(X - E[X])^2].$$

For random variables X, Y, the *correlation coefficient* ρ is

$$\rho := E[(X - E[X])(Y - E[Y])]/\sqrt{varX\ varY}.$$

(i) Show that

$$-1 \leq \rho \leq +1.$$

(ii) Show that $r = \pm 1$ if and only if (iff) there is a linear relationship between X and Y.

3 Statistical Methodology

Preliminary reading for the course *Fundamentals of Statistical Inference* should be geared towards basic knowledge of the main ideas of statistical inference, and revision of the necessary underlying probability/distribution theory. Very suitable are Chapters 1-12 of Wasserman (2003). Chapters 1-5 of that book give a concise account of probability/distribution theory background, and Chapters 6-12 give an introduction to key elements of the course. A very suitable, more detailed, reference is Casella and Berger (2002).

Distribution theory required by the course 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.

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, method of moments); hypothesis testing (including notions of size, power, critical region, optimal construction for simple hypotheses); confidence sets (properties, construction via pivotal quantities and inversion of hypothesis tests).

Appropriate levels of knowledge would be such as enables answering the following questions.

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 $\hat{\mu}$?

4 Applied Statistics

This material will require the use of a statistics package. Ideally, we would like you to use **R** to attempt these questions and, as practice for coursework assignments, write a report, using LaTeX, that contains your answers.

Suitable background reading is Venables and Ripley (2004); Faraway (2004, 2005).

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				

   (a) 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.

   (b) Do you have any reservations about applying routine analysis to this data, as is?

   Comments: Basic familiarity with manipulating and examining simple data.

2. ($*$) 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.

Comments: Simple ML computations, and selection of appropriate hypothesis testing procedure.

3. ($*$) 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).

Comments: Regression, transformations of variables, data analysis.

5 Computational Statistics

Some key computational methodology in statistics is based on discrete time Markov chains. Many of you will have come across discrete time, discrete state Markov chains, and it would be useful to refresh your knowledge about this (e.g. by studying (Norris, 1997, Chapter 1)).

The questions below are intended to give you some practice with simple simulations.

1. ($*$) 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)?

2. 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, $(X_1, \dots, X_{50})$ 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.

References

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

Chung, K.-L. and F. Ait-Sahlia (2007). *Elementary probability theory with stochastic processes* (4th ed.). Springer.

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.

Haigh, J. (2002). *Probability models*. Springer.

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

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.

Dr Ed Cohen
Course Director MSc in Statistics

July 2016