UNIVERSITY OF LONDON | INTERNATIONAL PROGRAMMES

Course information 2015–16
ST104b Statistics 2 (half course)

This half course requires the student to develop the concepts introduced in ST104a Statistics 1 of measurement and hypothesis testing.

Rule
ST104a Statistics 1 must be taken before or at the same time as ST104b Statistics 2.

Aims and objectives
The aim of this half course is to develop students' knowledge of elementary statistical theory. The emphasis is on topics that are of importance in applications to econometrics, finance and the social sciences. Concepts and methods that provide the foundation for more specialised courses in statistics are introduced.

Essential reading
For full details please refer to the reading list.

Newbold, P., W. Carlson and B. Thorne *Statistics for Business and Economics.* (London: Pearson).

Assessment
This half course is assessed by a two hour unseen written examination.

Learning outcomes
At the end of this half course, and having completed the Essential reading and activities, students should be able to:

- ☑ apply and be competent users of standard statistical operators and be able to recall a variety of well-known distributions and their respective moments
- ☑ explain the fundamentals of statistical inference and apply these principles to justify the use of an appropriate model and perform tests in a number of different settings
- ☑ demonstrate understanding that statistical techniques are based on assumptions and the plausibility of such assumptions must be investigated when analysing real problems.

Students should consult the *Programme Regulations for degrees and diplomas in Economics, Management, Finance and the Social Sciences* that are reviewed annually. Notice is also given in the *Regulations* of any courses which are being phased out and students are advised to check course availability.

Syllabus

This is a description of the material to be examined, as published in the *Programme handbook*. On registration, students will receive a detailed subject guide which provides a framework for covering the topics in the syllabus and directions to the essential reading.

Probability: *Set theory: the basics; Axiomatic definition of probability; Classical probability and counting rules; Conditional probability and Bayes' theorem.*

Random variables: *Discrete random variables; Continuous random variables.*

Common distributions of random variables: *Common discrete distributions; Common continuous distributions.*

Multivariate random variables: *Joint probability functions; Conditional distributions; Covariance and correlation; Independent random variables; Sums and products of random variables.*

Sampling distributions of statistics: *Random samples; Statistics and their sampling distributions; Sampling distribution of a statistic; Sample mean from a normal population; The central limit theorem; Some common sampling distributions; Prelude to statistical inference.*

Point estimation: *Estimation criteria: bias, variance and mean squared error; Method of moments estimation; Least squares estimation; Maximum likelihood estimation.*

Interval estimation: *Interval estimation for means of normal distributions; Use of the chi-squared distribution; Confidence intervals for normal variances.*

Hypothesis testing: *Setting p-value, significance level, test statistic; t tests; General approach to statistical tests; Two type of error; Tests for normal variances; Comparing two normal means with paired observations; Comparing two normal means; Tests for correlation coefficients; Tests for the ratio of two normal variances.*

Analysis of variance: *One-way analysis of variance; Two-way analysis of variance.*

Linear regression: *Simple linear regression; Inference for parameters in normal regression models; Regression ANOVA; Confidence intervals for E(y); Prediction intervals for y; Multiple linear regression models.*

Examiners' commentaries 2015

ST104b Statistics 2

Important note

This commentary reflects the examination and assessment arrangements for this course in the academic year 2014–15. The format and structure of the examination may change in future years, and any such changes will be publicised on the virtual learning environment (VLE).

Information about the subject guide and the Essential reading references

Unless otherwise stated, all cross-references will be to the latest version of the subject guide (2014). You should always attempt to use the most recent edition of any Essential reading textbook, even if the commentary and/or online reading list and/or subject guide refer to an earlier edition. If different editions of Essential reading are listed, please check the VLE for reading supplements – if none are available, please use the contents list and index of the new edition to find the relevant section.

Comments on specific questions – Zone A

Candidates should answer all **FOUR** questions: **QUESTION 1** of Section A (40 marks) and all **THREE** questions from Section B (60 marks in total). **Candidates are strongly advised to divide their time accordingly.**

Section A

Answer all parts of question 1 (40 marks in total).

Question 1

(a) Consider a sequence of random variables $X_1, X_2, X_3, \ldots$ that are independent and normally distributed with mean 0 and variance 1. Using as many of these random variables as you like construct a random variable that is a function of $X_1, X_2, X_3, \ldots$ and has a t distribution with 11 degrees of freedom. *Any correct answer will be given full marks. You do not have to think of the smallest number of random variables that make this possible.*

(5 marks)

Reading for this question

Section 6.11.2 of the subject guide defines the (Student's) t distribution as a standard normal variable divided by the square root of an independent chi-squared variable divided by its degrees of freedom.

Approaching the question

The simplest answer is:

$$\frac{\sqrt{11}X_{12}}{\sqrt{\sum_{i=1}^{11} X_i^2}}$$

but there are other answers such as the following based on the t test. Define:

$$\bar{X} = \frac{1}{12}\sum_{i=1}^{12} X_i$$

and:

$$S^2 = \frac{\sum_{i=1}^{12}(X_i - \bar{X})^2}{11}.$$

The random variable:

$$\frac{\sqrt{11}\bar{X}}{S}$$

has a t_{11} distribution. Also variations of the above would be acceptable.

(b) Let X have a binomial distribution with parameters 6 and p. We want to test the null hypothesis $p = 1/2$ against the alternative $p = 3/4$. We reject the null hypothesis if and only if $X > 4$. Calculate the size and power of the test.

(6 marks)

Reading for this question

Section 9.10 of the subject guide covers the size of the test (the probability of a Type I error) and introduces the power function.

Approaching the question

The size is the probability we reject the null hypothesis when it is true:

$$P(X > 4 \,|\, p = 1/2) = 6\left(\frac{1}{2}\right)^6 + \left(\frac{1}{2}\right)^6 = \frac{7}{64} = 0.1094.$$

The power is the probability we reject the null hypothesis when the alternative is true:

$$P(X > 4 \,|\, p = 3/4) = 6\left(\frac{3}{4}\right)^5 \frac{1}{4} + \left(\frac{3}{4}\right)^6 = \frac{729}{4096} = 0.5339.$$

(c) State and prove Bayes' theorem for two events A and B. *Make sure you state any assumptions made.*

(5 marks)

Reading for this question

Section 2.9.2 covers Bayes' theorem.

Approaching the question

We have:

$$P(A \,|\, B) = \frac{P(A)\,P(B \,|\, A)}{P(B)}.$$

This is true because:

$$P(A \,|\, B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)\,P(B \,|\, A)}{P(B)}.$$

The assumptions are $P(A) > 0$ and $P(B) > 0$.

(d) **A biologist conducts an experiment to find out if a tumour is benign or malignant. 20% of tumours are malignant. If the tumour is malignant, the experiment will report it is benign with probability 0.1, malignant with probability 0.8 and produce an inconclusive result with probability 0.1. If it is benign, the experiment will report it is benign with probability 0.6, malignant with probability 0.2 and produce an inconclusive result with probability 0.2. The experiment is performed twice; the first time it produces an inconclusive result and the second time it reports that the tumour is malignant. Assuming the results of the two experiments are independent, what is the probability the tumour is malignant given the results of the experiments?**

(8 marks)

Reading for this question

Conditional probability is discussed in Section 2.9 of the subject guide.

Approaching the question

The probability the tumour is malignant and we observe the results we have seen is:

$$0.2 \times 0.1 \times 0.8 = 0.016.$$

The probability the tumour is benign and we observe the results we have seen is:

$$0.8 \times 0.2 \times 0.2 = 0.032.$$

The probability we observe the results we have seen is:

$$0.016 + 0.032 = 0.048.$$

Therefore, the conditional probability the tumour is malignant given the results is:

$$\frac{0.016}{0.048} = \frac{1}{3}.$$

(e) **The random variables ε_i for $i = 1, 2, 3$ and 4, are independent with mean 0 and variance 1 and α, β and γ are unknown parameters. Suppose you are given observations $x_1, x_2, x_3$and x_4 such that**

$$\begin{aligned} x_1 &= \alpha + \beta + \varepsilon_1 \\ x_2 &= 2\alpha + \gamma + \varepsilon_2 \\ x_3 &= 3\alpha + \varepsilon_3 \\ x_4 &= 4\alpha + \varepsilon_4. \end{aligned}$$

The least squares estimators of α, β and γ are $\widehat{\alpha}$, $\widehat{\beta}$ and $\widehat{\gamma}$. Prove that

$$\widehat{\alpha} = \frac{3x_3 + 4x_4}{25}.$$

(8 marks)

Reading for this question

Section 7.8 of the subject guide introduces least squares estimation, and this question is similar to Sample examination question 5 at the end of Chapter 7.

Approaching the question

We want to minimise:

$$S = (x_1 - \alpha - \beta)^2 + (x_2 - 2\alpha - \gamma)^2 + (x_3 - 3\alpha)^2 + (x_4 - 4\alpha)^2.$$

Differentiating with respect to α and equating to 0, we have:

$$-2(x_1 - \widehat{\alpha} - \widehat{\beta}) - 4(x_2 - 2\widehat{\alpha} - \widehat{\gamma}) - 6(x_3 - 3\widehat{\alpha}) - 8(x_4 - 4\widehat{\alpha}) = 0.$$

Differentiating with respect to β and equating to 0, we have:

$$-2(x_1 - \widehat{\alpha} - \widehat{\beta}) = 0.$$

Differentiating with respect to γ and equating to 0, we have:

$$-2(x_2 - 2\widehat{\alpha} - \widehat{\gamma}) = 0.$$

Substituting the last two equations into the previous one, we have:

$$50\widehat{\alpha} = 6x_3 + 8x_4$$

and, therefore:

$$\widehat{\alpha} = \frac{3x_3 + 4x_4}{25}.$$

(f) Continuing from (e) show that $\widehat{\alpha}$ is an unbiased estimator of α.

(3 marks)

Reading for this question

Unbiased estimators are defined in Section 7.6 of the subject guide.

Approaching the question

We have:

$$\mathrm{E}(\widehat{\alpha}) = \frac{3 \times 3\alpha + 4 \times 4\alpha}{25} = \alpha$$

so it is unbiased.

(g) Continuing from (e) and (f), find the covariance of $\widehat{\beta}$ and $\widehat{\gamma}$.

(5 marks)

Reading for this question

Section 5.8.1 of the subject guide defines the covariance of two random variables.

Approaching the question

From (e) we can see that:

$$\widehat{\beta} = x_1 - \widehat{\alpha}$$

and:

$$\widehat{\gamma} = x_2 - 2\widehat{\alpha}.$$

We note that x_1, x_2 and $\widehat{\alpha} = (3x_3 + 4x_4)/25$ are three independent random variables, so we conclude that:

$$\mathrm{Cov}(\widehat{\beta}, \widehat{\gamma}) = \mathrm{Cov}(x_1 - \widehat{\alpha}, x_2 - 2\widehat{\alpha}) = 2\mathrm{Var}(\widehat{\alpha}) = 2 \times \frac{3^2 + 4^2}{(25)^2} = \frac{2}{25} = 0.08.$$

Section B

Answer all **three** questions in this section (60 marks in total).

Question 2

(a) Let $X_1, X_2, \ldots X_n$ be a random sample from the following distribution

$$\mathbf{Pr}(X = x) = \frac{1}{\alpha + 1}\left(\frac{\alpha}{\alpha + 1}\right)^x$$

for $x = 0, 1, 2, \ldots$ with $\alpha > 0$. Show the maximum likelihood estimator of α is the sample average and prove that it is unbiased. *You might need the identity*

$$\sum_{i=1}^{\infty} i\theta^{i-1} = \frac{1}{(1-\theta)^2}.$$

(12 marks)

Reading for this question

Section 7.9 of the subject guide covers maximum likelihood estimation.

Approaching the question

The likelihood function is:

$$\prod_{i=1}^{n}\left(\left(\frac{1}{\alpha+1}\right)\left(\frac{\alpha}{\alpha+1}\right)^{X_i}\right) = \left(\frac{1}{\alpha+1}\right)^{n+\sum X_i}\alpha^{\sum X_i}$$

and the log-likelihood is:

$$l(\alpha) = -\left(n + \sum X_i\right)\ln(\alpha+1) + \sum X_i \ln(\alpha).$$

Differentiating and equating to 0, we have:

$$-\frac{n+\sum X_i}{\widehat{\alpha}+1} + \frac{\sum X_i}{\widehat{\alpha}} = 0$$

and therefore:

$$\widehat{\alpha} = \frac{\sum X_i}{n}.$$

We also have:

$$\begin{aligned}\mathrm{E}(X_i) &= \sum_{x=1}^{\infty} x\frac{1}{\alpha+1}\left(\frac{\alpha}{\alpha+1}\right)^x = \frac{\alpha}{(\alpha+1)^2}\sum_{x=1}^{\infty} x\left(\frac{\alpha}{\alpha+1}\right)^{x-1} \\ &= \frac{\alpha}{(\alpha+1)^2}\left(1-\frac{\alpha}{\alpha+1}\right)^{-2} = \frac{\alpha}{(\alpha+1)^2}\frac{1}{(\alpha+1)^{-2}} = \alpha\end{aligned}$$

and therefore:

$$\mathrm{E}(\widehat{\alpha}) = \frac{n\alpha}{n} = \alpha$$

hence it is unbiased.

(b) Three restaurants A,B and C have recorded some customers' bills. Restaurant A recorded bills from 5 customers with amounts $x_{A1}, x_{A2}, \ldots, x_{A5}$; restaurant B recorded bills from 8 customers with amounts $x_{B1}, x_{B2}, \ldots, x_{B8}$ and restaurant C recorded bills from 9 customers with amounts $x_{C1}, x_{C2}, \ldots, x_{C9}$. Explain how you would use this information to construct a one way ANOVA table and use it to test whether the three restaurants are equally expensive against the alternative that they are not. The size of the test should be 0.05 and you should provide the critical value.

(10 marks)

Reading for this question

Section 10.7 of the subject guide covers one-way analysis of variance.

Approaching the question

We need to calculate the following:

$$\begin{aligned}\bar{X}_A &= \frac{1}{5}\sum_{i=1}^{5} X_{Ai}\\ \bar{X}_B &= \frac{1}{8}\sum_{i=1}^{8} X_{Bi}\\ \bar{X}_C &= \frac{1}{9}\sum_{i=1}^{9} X_{Ci}\\ \bar{X} &= \frac{\sum_{i=1}^{5} X_{Ai} + \sum_{i=1}^{8} X_{Bi} + \sum_{i=1}^{9} X_{Ci}}{22}\end{aligned}$$

Alternatively:

$$\bar{X} = \frac{5\bar{X}_A + 8\bar{X}_B + 9\bar{X}_C}{22}.$$

We then need the between-groups sum of squares, B:

$$B = 5(\bar{X}_A - \bar{X})^2 + 8(\bar{X}_B - \bar{X})^2 + 9(\bar{X}_C - \bar{X})^2$$

and the within-groups sum of squares, W:

$$W = \sum_{i=1}^{5}(X_{Ai} - \bar{X}_A)^2 + \sum_{i=1}^{8}(X_{Bi} - \bar{X}_B)^2 + \sum_{i=1}^{9}(X_{Ci} - \bar{X}_C)^2.$$

Alternatively, one could calculate only one of the two, and calculate the total sum of squares (TSS):

$$\text{TSS} = \sum_{i=1}^{5}(X_{Ai} - \bar{X})^2 + \sum_{i=1}^{8}(X_{Bi} - \bar{X})^2 + \sum_{i=1}^{9}(X_{Ci} - \bar{X})^2$$

and use the relationship $\text{TSS} = B + W$ to calculate the other. We then construct the ANOVA table:

Source	Degrees of Freedom	SS	Mean Square	F value
Between	2	b	$b/2$	$f = 19b/2w$
Error	19	w	$w/19$	
Total	21	$b + w$		

We then compare f to the value from tables of the $F_{2,\,19}$ distribution, i.e. $F_{0.05,\,2,\,19} = 3.52$. We will reject the null hypothesis that there is no difference if $f > 3.52$.

Question 3

The random variable X has density

$$\beta x + \gamma x^2$$

over the region $0 < x < 1$. The density is 0 elsewhere.

You are also given that $\text{E}(X) = 25/36$.

(a) Show that $\beta = 4/3$ and $\gamma = 1$.

(6 marks)

(b) Find the value of $\mathbf{Cov}\,(X, 1/X)$.

(5 marks)

(c) Does $\mathbf{Var}\,(1/X)$ exist? Explain your answer.

(3 marks)

(d) Calculate $\mathbf{Pr}\left((2X-1)^2 < 1/4 \,|\, X < 1/2\right)$.

(5 marks)

Reading for this question

Section 3.7 of the subject guide introduces continuous random variables.

Approaching the question

(a) For the expression to be a valid density function, we need it to integrate to 1, so:

$$\int_0^1 (\beta x + \gamma x^2)\,\mathrm{d}x = 1$$

and, therefore:

$$\frac{\beta}{2} + \frac{\gamma}{3} = 1.$$

Also:

$$\int_0^1 x(\beta x + \gamma x^2)\,\mathrm{d}x = \frac{25}{36}$$

and therefore:

$$\frac{\beta}{3} + \frac{\gamma}{4} = \frac{25}{36}.$$

Solving we get $\beta = 4/3$ and $\gamma = 1$.

(b) We have:

$$\mathrm{E}\,(1/X) = \int_0^1 \left(\frac{4}{3} + x\right)\,\mathrm{d}x = \frac{4}{3} + \frac{1}{2} = \frac{11}{6}.$$

Hence:

$$\mathrm{Cov}\,(X, 1/X) = 1 - \mathrm{E}(X)\,\mathrm{E}\,(1/X) = 1 - \frac{11}{6} \times \frac{25}{36} = -\frac{59}{216}.$$

(c) It does not exist because $\mathrm{E}\left(1/X^2\right)$ does not exist as the integral:

$$\int_0^1 \frac{1}{x}\,\mathrm{d}x$$

is divergent.

(d) We have:

$$\begin{aligned} P\left((2X-1)^2 < 1/4 \,|\, X < 1/2\right) &= P\,(1/4 < X < 3/4 \,|\, X < 1/2) \\ &= \frac{P\,(1/4 < X < 3/4,\, X < 1/2)}{P\,(X < 1/2)} \\ &= \frac{P\,(1/4 < X < 1/2)}{P\,(X < 1/2)}. \end{aligned}$$

Now:

$$\begin{aligned} P\,(1/4 < X < 1/2) &= \int_{1/4}^{1/2} \left(\frac{4}{3}x + x^2\right)\,\mathrm{d}x \\ &= \frac{2}{3}\left(\left(\frac{1}{2}\right)^2 - \left(\frac{1}{4}\right)^2\right) + \frac{1}{3}\left(\left(\frac{1}{2}\right)^3 - \left(\frac{1}{4}\right)^3\right) \\ &= \frac{31}{192}. \end{aligned}$$

Also:

$$\begin{aligned} P\left(X < 1/2\right) &= \int_0^{1/2} \left(\frac{4}{3}x + x^2\right) \mathrm{d}x \\ &= \frac{2}{3}\left(\frac{1}{2}\right)^2 + \frac{1}{3}\left(\frac{1}{2}\right)^3 \\ &= \frac{5}{24} \end{aligned}$$

and, therefore:

$$P\left((2X-1)^2 < 1/4 \,|\, X < 1/2\right) = \frac{31/192}{5/24} = \frac{31}{40} = 0.7750.$$

Question 4

A box contains 4 red balls, 3 green balls and 3 blue balls. Two balls are selected without replacement. Let X represent the number of red balls in the sample and Y the number of green balls in the sample.

(a) Arrange the different pairs of values of (X, Y) as the cells in a table, each cell being filled with the probability of that pair of values occurring.

(6 marks)

(b) What does the random variable $Z = 2 - X - Y$ represent?

(3 marks)

(c) Calculate $\mathrm{Cov}(X, Y)$.

(5 marks)

(d) Calculate $\Pr(X = 1 \,|\, -2 < X - Y < 2)$.

(5 marks)

Reading for this question

Chapter 5 of the subject guide covers multivariate random variables.

Approaching the question

(a) We have:

$$\begin{aligned} P(X=0, Y=0) &= \frac{3}{10} \times \frac{2}{9} = \frac{6}{90} = \frac{1}{15} \\ P(X=0, Y=1) &= 2 \times \frac{3}{10} \times \frac{3}{9} = \frac{18}{90} = \frac{3}{15} \\ P(X=0, Y=2) &= \frac{3}{10} \times \frac{2}{9} = \frac{6}{90} = \frac{1}{15} \\ P(X=1, Y=0) &= 2 \times \frac{4}{10} \times \frac{3}{9} = \frac{24}{90} = \frac{4}{15} \\ P(X=1, Y=1) &= 2 \times \frac{4}{10} \times \frac{3}{9} = \frac{24}{90} = \frac{4}{15} \\ P(X=2, Y=0) &= \frac{4}{10} \times \frac{3}{9} = \frac{12}{90} = \frac{2}{15}. \end{aligned}$$

All other values have probability 0. We then construct the table of joint probabilities:

	$Y=0$	$Y=1$	$Y=2$
$X=0$	1/15	3/15	1/15
$X=1$	4/15	4/15	0
$X=2$	2/15	0	0

(b) The number of blue balls in the sample.

(c) We have:

$$\begin{aligned} \mathrm{E}(X) &= 1 \times \left(\frac{4}{15} + \frac{4}{15}\right) + 2 \times \frac{2}{15} = \frac{4}{5} \\ \mathrm{E}(Y) &= 1 \times \left(\frac{3}{15} + \frac{4}{15}\right) + 2 \times \frac{1}{15} = \frac{3}{5} \end{aligned}$$

and:

$$\mathrm{E}(XY) = 1 \times 1 \times \frac{4}{15} = \frac{4}{15}.$$

So:

$$\mathrm{Cov}(X,Y) = \frac{4}{15} - \frac{4}{5} \times \frac{3}{5} = -\frac{16}{75}.$$

(d) We have:

$$P(X = 1 \,|\, |X-Y| < 2) = \frac{4/15 + 4/15}{1/15 + 3/15 + 4/15 + 4/15} = \frac{2}{3}.$$