



MAST20005 Final Exam

Statistics (University of Melbourne)

Semester 2 Assessment, 2020

School of Mathematics and Statistics

MAST20005 Statistics

Reading time: 30 minutes — Writing time: 3 hours — Upload time: 30 minutes

This exam consists of 19 pages (including this page)

Permitted Materials

- This exam and/or an offline electronic PDF reader, one or more copies of the masked exam template made available earlier, blank loose-leaf paper and a Casio FX-82 calculator.
- One double sided A4 page of notes (handwritten or printed).

Instructions to Students

- There are 7 questions with marks as shown. The total number of marks available is 80.
- You should attempt all questions.
- Full reasoning must be shown and penalties may be imposed for poorly presented, unclear, untidy and badly worded solutions.
- During writing time you may only interact with the device running the Zoom session with supervisor permission. The screen of any other device must be visible in Zoom from the start of the session.
- If you have a printer, print the exam one-sided. If you cannot print, download the exam to a second device, which must then be disconnected from the internet.
- Write your answers in the boxes provided on the exam that you have printed or the masked exam template that has been previously made available. If you are unable to answer the whole question in the answer space provided then you can append additional handwritten solutions to the end after the 19 numbered pages. If you do this you **MUST** make a note in the correct answer space or page for the question, warning the marker that you have appended additional remarks at the end.
- If you have been unable to print the exam and do not have the masked template write your answers on A4 paper. The first page should contain only your student number, the subject code and the subject name. Write on one side of each sheet only. Start each question on a new page and include the question number at the top of each page.
- Assemble all exam pages (or masked template pages) in correct page number order and the correct way up. Add any extra pages with additional working at the end. Use a mobile phone scanning application to scan all pages to a single PDF file. Scan from directly above to reduce keystone effects. Check that all pages are clearly readable and cropped to the A4 borders of the original page. Poorly scanned submissions may be impossible to mark.
- Submit your PDF file to the Canvas Assignment corresponding to this exam using the Gradescope window. Before leaving Zoom supervision, confirm with your Zoom supervisor that you have Gradescope confirmation of submission.

Question 1 (11 marks)

We will fit the following regression model: $Y_i = \alpha + \beta x_i + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, \dots, 10$. Consider the following R output:

```
> summary(x)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  6.18  29.21   60.10   55.15   83.90   94.47
> sd(x)
[1] 31.56569
> sort(x)
[1]  6.18 20.17 26.55 37.21 57.29 62.91 66.08 89.84 90.82 94.47
> f = lm(y ~ x)
> summary(f)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.45528     3.05901   0.476   0.647
x             4.01616     0.04874  82.394 5.25e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.616 on 8 degrees of freedom
Multiple R-squared:  0.9988,    Adjusted R-squared:  0.9987
F-statistic: 6789 on 1 and 8 DF,  p-value: 5.251e-13
```

- (a) For each of the following quantities, state or calculate its value if possible, or otherwise explain why it is not possible.
- (i) $x_{(4.5)}$
 - (ii) \bar{x}
 - (iii) α
 - (iv) $\hat{\sigma}^2$
 - (v) Sample correlation coefficient, r

- (b) For each of the following pairs of hypotheses, carry out the test if it is possible, using a 5% significance level, or otherwise explain what further information you need in order to do it.
- (i) $H_0: \alpha = 0$ versus $H_1: \alpha \neq 0$
 - (ii) $H_0: \alpha = 0$ versus $H_1: \alpha > 0$
 - (iii) $H_0: \alpha = 4$ versus $H_1: \alpha \neq 4$

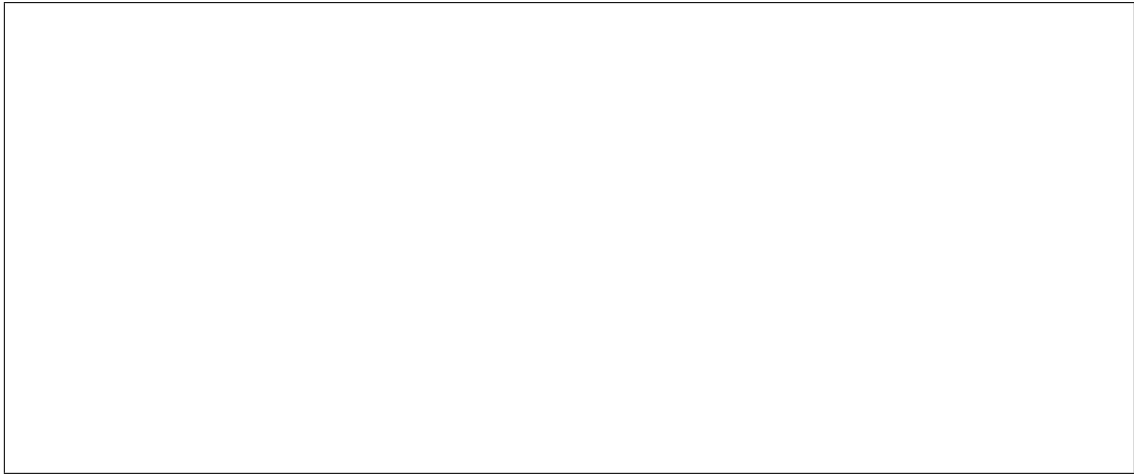
Question 2 (8 marks)

The School of Mathematics and Statistics conducted a survey about the time spent commuting by students each day. A total of 40 students were surveyed and the responses are summarised in the table below.

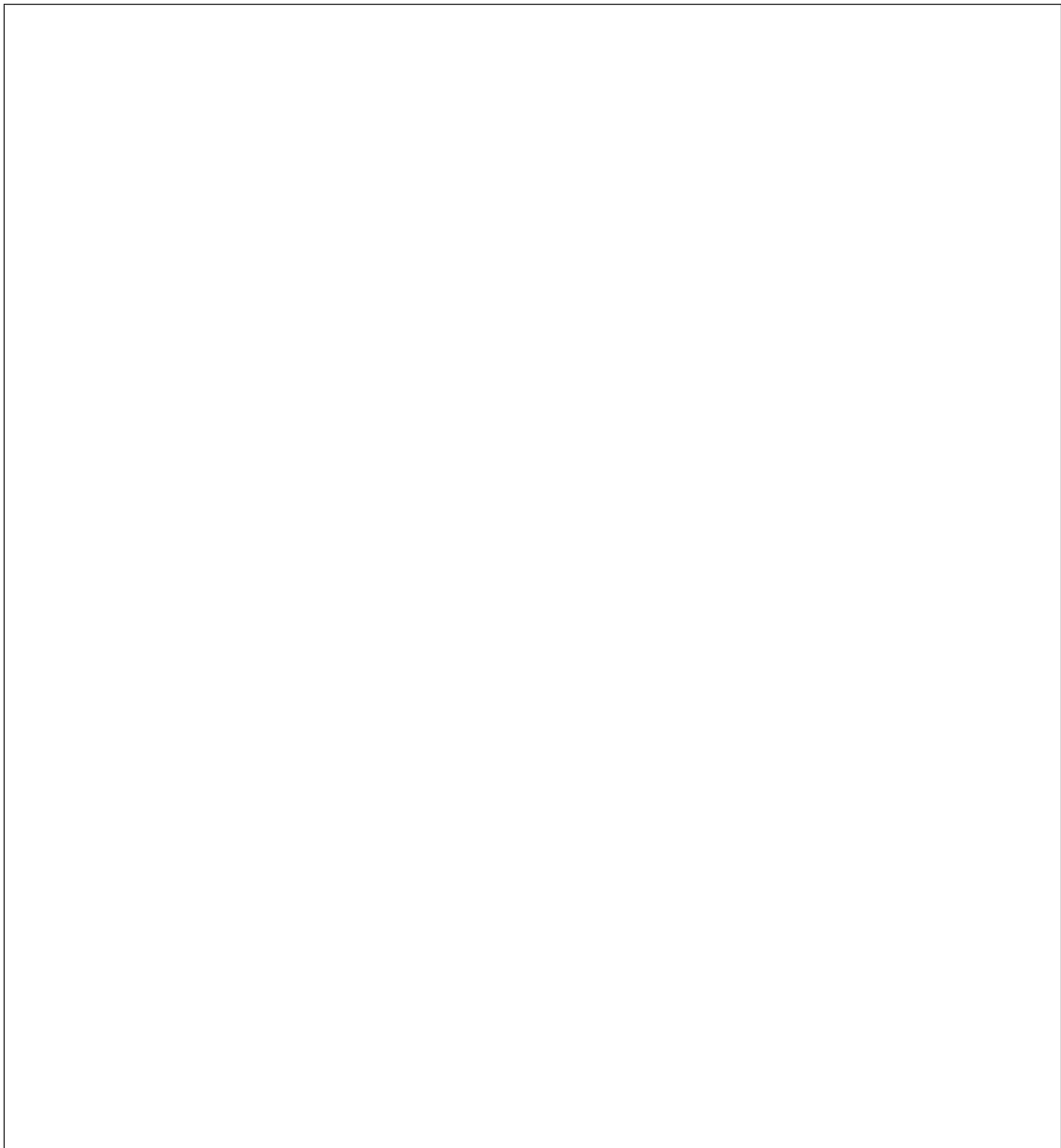
	0–1 hours	1–2 hours	2–3 hours	3+ hours
Number of students	10	16	10	4

- (a) Let p be the proportion of students taking 1–2 hours to commute. Calculate a 95% confidence interval for p .

- (b) Let p_1, p_2, p_3, p_4 be the proportion of all students (not just those surveyed) that would be in each of the groups defined in the table above. Using the survey data and a 10% significance level, test whether the commuting time groups follow the distribution $p_1 = p_2 = p_3 = p_4 = 1/4$.



- (c) Using a 5% significance level, test whether the commuting time follows the exponential distribution with the mean equal to 2.



Question 3 (11 marks)

Let X be the length of a tomato seed (in millimeters). Ten randomly selected seeds are measured, giving the following measurements:

6.2	8.4	11.2	5.7	7.8
8.3	9.4	7.5	12.0	7.7

For these data, we have $\bar{x} = 8.42$ and $s = 1.97$. You may assume that $X \sim N(\mu, \sigma^2)$.

- (a) Calculate a 95% confidence interval for μ .

- (b) Calculate a 95% confidence interval for σ^2 .

- (c) Let X^* be the length of a new seed. Assuming $\sigma = 2$ is known, calculate a 95% prediction interval for X^* .

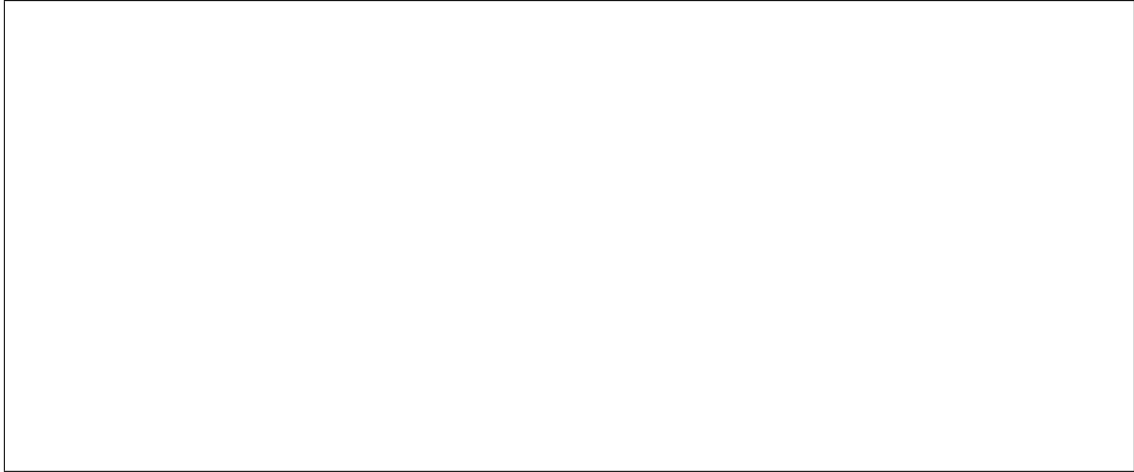
- (d) Let X^* be the length of a new seed. If σ is unknown, calculate a 95% prediction interval for X^* .

Question 4 (16 marks)

Let X_1, \dots, X_n be a random sample from the following distribution with parameter λ and pdf:

$$f(x | \lambda) = \lambda e^{-\lambda(x-2)}, \quad x \geq 2.$$

- (a) Determine a sufficient statistic for λ .



- (b) Find the maximum likelihood estimator (MLE) of λ .

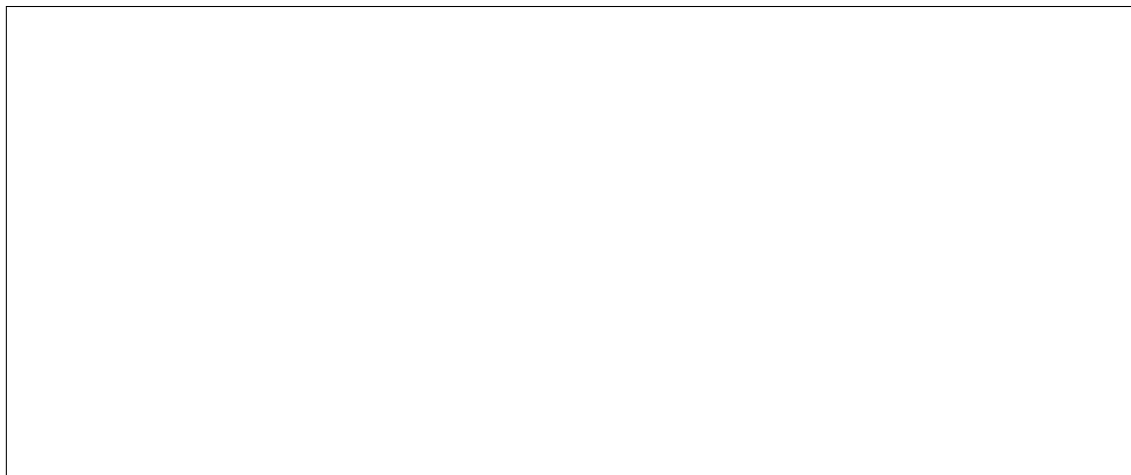


(c) Is this MLE unbiased?

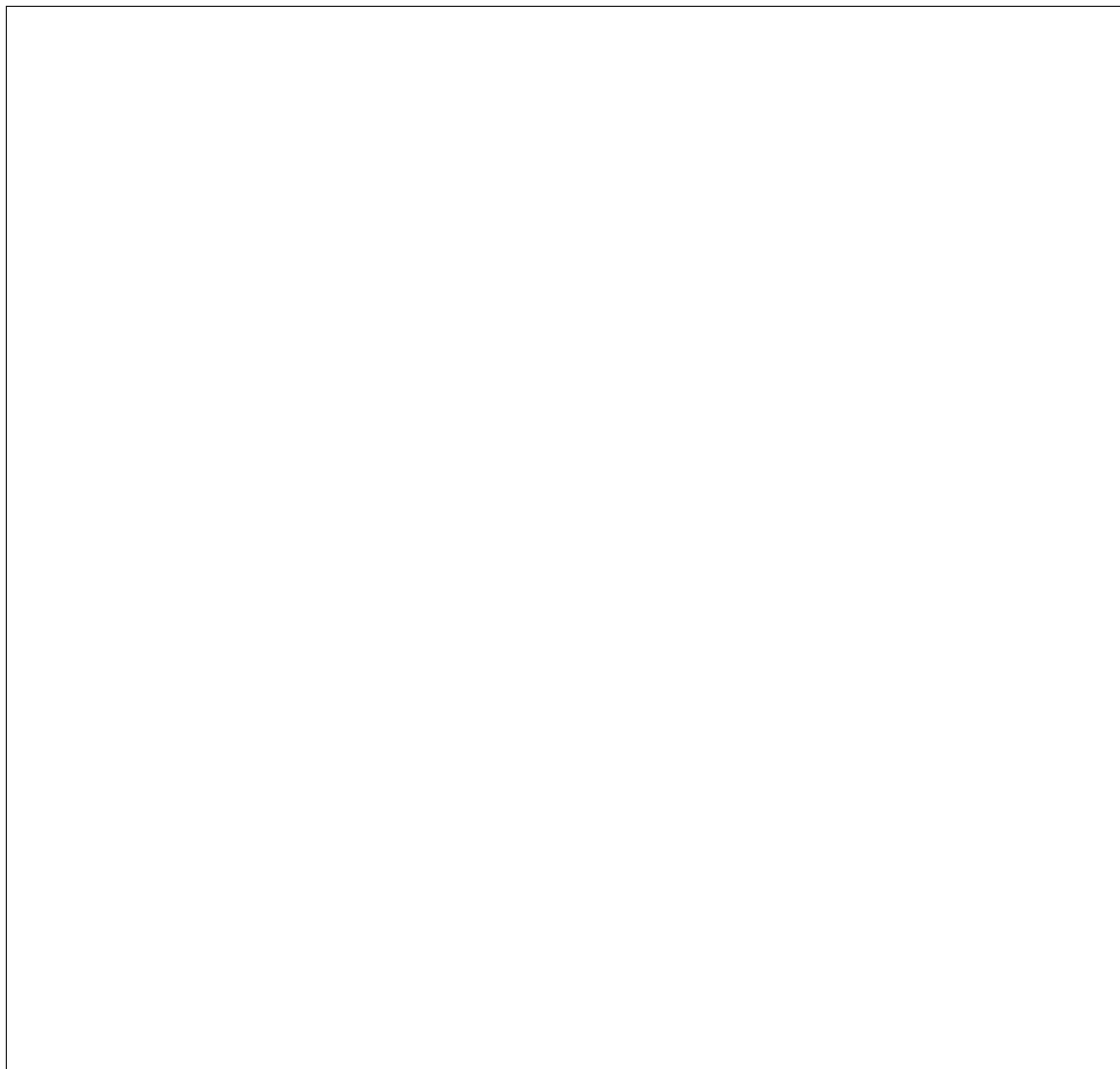
(d) Find the Cramér–Rao lower bound for unbiased estimators of λ .

(e) Find the p quantile, π_p .

(f) Find the pdf of $X_{(2)}$.



(g) Let $n = 10$ and $\lambda = 1/8$, find $\Pr(X_{(2)} > 4)$.



Question 5 (13 marks)

Tingjin wants to compare three types of fertiliser. Each type of fertiliser is applied to five plants, and the growth of each plant (in centimetres) is measured after one month.

Fertiliser	Plant					Statistics (per-fertiliser)	
	1	2	3	4	5	Mean	Standard deviation
Type A	4.5	2.7	5.7	7.5	4.0	4.88	1.82
Type B	3.7	0.4	4.5	3.1	0.9	2.52	1.79
Type C	8.5	2.6	6.3	4.3	3.8	5.10	2.32

- (a) Assuming a normal distribution and equal variances for the the plant growth, test whether the growth with Type A is larger than that with Type B ($\alpha = 0.05$).

- (b) Use the sign test with $\alpha = 0.05$ to test if the median plant growth with Type A is larger than 3 centimetres.

- (c) Consider a one-way ANOVA model, please complete the following ANOVA table:

Source	df	SS	MS	<i>F</i>
Treatment (fertiliser type)				
Error				
Total				

- (d) Is there evidence of any differences in the average plant growth between the three types of fertiliser?

Question 6 (7 marks)

Let $X_1, \dots, X_n \sim N(0, 1/\theta)$ be a random sample.

- (a) Let $f(\theta) = e^{-\theta}$ with $\theta \geq 0$ be the prior distribution for θ .
Derive the posterior distribution for θ .

- (b) Find the posterior mean for θ .

- (c) Find a conjugate prior distribution for θ .

- (d) Let $f(\theta) = 1/\theta$ with $\theta \geq 0$ be the prior distribution for θ .
Explain why $f(\theta)$ is called an improper prior.

Question 7 (14 marks)

Consider the following regression model without an intercept term:

$$Y_i = \beta x_i + \epsilon_i, \quad i = 1, \dots, n,$$

where $\epsilon_1, \dots, \epsilon_n \sim N(0, \sigma^2)$ are independent random variables.

- (a) Find the distribution of Y_i .

- (b) Find the estimator of β using least squares estimation.

- (c) Show that this estimator (from part (b)) is unbiased.

- (d) Find the joint density function of Y_1, \dots, Y_n .

- (e) Assuming β is known, find the maximum likelihood estimator of σ^2 .

- (f) Is this estimator (from part (e)) unbiased? Justify your answer.

End of Exam — Total Available Marks = 80

Turn the page for appended material

Appendix

Distributions

- The pdf of $X \sim N(\mu, \sigma^2)$ is,

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$

- The pdf of an exponential distribution with mean equal to $1/\lambda$ is,

$$f(x) = \lambda \exp\{-\lambda x\}, \quad (x \geq 0).$$

- The pdf of $X \sim \text{Gamma}(\alpha, \beta)$ is,

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-x\beta}, \quad (x \geq 0)$$

and it has mean $\mathbb{E}(X) = \alpha/\beta$ and $\text{var}(X) = \alpha/\beta^2$.

- The pdf of $X \sim \text{Beta}(\alpha, \beta)$ is,

$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}, \quad (0 \leq x \leq 1)$$

and it has mean $\mathbb{E}(X) = \alpha/(\alpha + \beta)$ and $\text{var}(X) = \alpha\beta/\{(\alpha + \beta)^2(\alpha + \beta + 1)\}$.

R output

```
> p1 <- c(0.01, 0.025, 0.05, 0.1, 0.9, 0.95, 0.975, 0.99)

> qnorm(p1)
[1] -2.326 -1.960 -1.645 -1.282  1.282  1.645  1.960  2.326

> qt(p1, df = 7)
[1] -2.998 -2.365 -1.895 -1.415  1.415  1.895  2.365  2.998
> qt(p1, df = 8)
[1] -2.896 -2.306 -1.860 -1.397  1.397  1.860  2.306  2.896
> qt(p1, df = 9)
[1] -2.821 -2.262 -1.833 -1.383  1.383  1.833  2.262  2.821
> qt(p1, df = 10)
[1] -2.764 -2.228 -1.812 -1.372  1.372  1.812  2.228  2.764
> qt(p1, df = 11)
[1] -2.718 -2.201 -1.796 -1.363  1.363  1.796  2.201  2.718
> qt(p1, df = 12)
[1] -2.681 -2.179 -1.782 -1.356  1.356  1.782  2.179  2.681
> qt(p1, df = 13)
[1] -2.650 -2.160 -1.771 -1.350  1.350  1.771  2.160  2.650
> qt(p1, df = 14)
[1] -2.624 -2.145 -1.761 -1.345  1.345  1.761  2.145  2.624
> qt(p1, df = 26)
[1] -2.479 -2.056 -1.706 -1.315  1.315  1.706  2.056  2.479
> qt(p1, df = 27)
[1] -2.473 -2.052 -1.703 -1.314  1.314  1.703  2.052  2.473
```

```

> qchisq(p1, df = 1)
[1] 0.0001571 0.0009821 0.0039321 0.0157908 2.7055435 3.8414588 5.0238862 6.6348966
> qchisq(p1, df = 2)
[1] 0.02010 0.05064 0.10259 0.21072 4.60517 5.99146 7.37776 9.21034
> qchisq(p1, df = 3)
[1] 0.1148 0.2158 0.3518 0.5844 6.2514 7.8147 9.3484 11.3449
> qchisq(p1, df = 7)
[1] 1.239 1.690 2.167 2.833 12.017 14.067 16.013 18.475
> qchisq(p1, df = 8)
[1] 1.646 2.180 2.733 3.490 13.362 15.507 17.535 20.090
> qchisq(p1, df = 9)
[1] 2.088 2.700 3.325 4.168 14.684 16.919 19.023 21.666
> qchisq(p1, df = 10)
[1] 2.558 3.247 3.940 4.865 15.987 18.307 20.483 23.209

> qf(p1, 1, 13)
[1] 0.0001632 0.0010206 0.0040868 0.0164196 3.1362051 4.6671927 6.4142543 9.0738057
> qf(p1, 1, 14)
[1] 0.0001628 0.0010178 0.0040756 0.0163739 3.1022134 4.6001099 6.2979386 8.8615927
> qf(p1, 2, 12)
[1] 0.01006 0.02537 0.05151 0.10629 2.80680 3.88529 5.09587 6.92661
> qf(p1, 2, 13)
[1] 0.01006 0.02537 0.05150 0.10622 2.76317 3.80557 4.96527 6.70096
> qf(p1, 3, 11)
[1] 0.03686 0.06957 0.11411 0.19148 2.66023 3.58743 4.63002 6.21673
> qf(p1, 3, 12)
[1] 0.03697 0.06975 0.11436 0.19173 2.60552 3.49029 4.47418 5.95254
> qf(p1, 3, 13)
[1] 0.03706 0.06991 0.11456 0.19195 2.56027 3.41053 4.34718 5.73938

> pbinom(0:5, 5, 0.5)
[1] 0.03125 0.18750 0.50000 0.81250 0.96875 1.00000
> pbinom(0:5, 10, 0.5)
[1] 0.0009766 0.0107422 0.0546875 0.1718750 0.3769531 0.6230469
> pbinom(0:5, 9, 0.5)
[1] 0.001953 0.019531 0.089844 0.253906 0.500000 0.746094
> pbinom(0:10, 10, 0.25)
[1] 0.056 0.244 0.526 0.776 0.922 0.980 0.996 1.000 1.000 1.000 1.000

> qgamma(p1, 8, 1)
[1] 2.906 3.454 3.981 4.656 11.771 13.148 14.423 16.000
> qgamma(p1, 9, 1)
[1] 3.507 4.115 4.695 5.432 12.995 14.435 15.763 17.403
> qgamma(p1, 10, 1)
[1] 4.130 4.795 5.425 6.221 14.206 15.705 17.085 18.783

```