THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

This exam paper must not be removed from the venue

Venue _____

Seat Number _____

Student Number |__|__|__|__|__|__|__|__|

Family Name _____

First Name _____

School of Mathematics & Physics

EXAMINATION

Semester Two Final Examinations, 2018

STAT1201 Analysis of Scientific Data

This paper is for St Lucia Campus, Gatton Campus (External) and Gatton Campus students.

Examination Duration: 120 minutes

Reading Time: 10 minutes

Exam Conditions:

This is a Central Examination

This is a Closed Book Examination - specified materials permitted

During reading time - Write only on rough paper provided

This examination paper will be released to the Library

Materials Permitted In The Exam Venue:

(No electronic aids are permitted e.g. laptops, phones)

Calculators - Casio FX82 series or UQ approved (labelled)

An annotated copy of *A Portable Introduction to Data Analysis* (any edition) is also permitted.

Materials To Be Supplied To Students:

None

Instructions To Students:

There are **50** marks available on this exam from **4** questions.

Write your answers in the spaces provided in pages 2–14 of this examination paper. Show your working and state conclusions where appropriate.

Pages 15–19 give formulas and statistical tables. Those pages will not be marked.

The textbook can have any amount of annotation on its pages. Loose sheets of paper or post-it notes are not permitted. Page tabs are allowed.

For Examiner Use Only

Question	Mark
1	
2	
3	
4	

Total _____

Question 1

13 marks

A recent study aimed to evaluate the clinical effects of the drug rivastigmine on the symptoms of schizophrenia when used alongside standard antipsychotic medication. A total of 36 patients with a diagnosis of schizophrenia entered into a 12-week, double-blind, clinical trial with random assignment to rivastigmine or placebo. Scores on the Mini Mental State Examination (MMSE) before and after the trial were used as the primary outcome measure.

(a) The baseline MMSE score for the 18 patients in the rivastigmine group had a mean of 24.1 and a standard deviation of 2.9 while the 18 patients in the placebo group had a mean MMSE score of 23.0 and a standard deviation of 3.2. Does this give any evidence of a difference in the baseline MMSE scores between the two groups? State the null and alternative hypotheses, and use an appropriate test statistic to determine the P-value. What do you conclude? [5 marks]

(b) Interpret your conclusion from (a) in the context of the random assignment to groups. [1 mark]

(c) After the 12-week trial, the 18 subjects in the rivastigmine group had an improved mean MMSE score of 27.1. The mean of the 18 changes in MMSE score was 3.0 with standard deviation 3.7. Does this give any evidence that mean MMSE score increases for patients with rivastigmine? State the null and alternative hypotheses, and use an appropriate test statistic to determine the P-value. What do you conclude? [4 marks]

(d) Out of the 18 patients in the rivastigmine group, 16 of them experienced an increase in MMSE score while 2 of them decreased. Based on this, find the *P*-value for a sign test of whether MMSE score tends to increase for patients with rivastigmine. What do you conclude? [2 marks]

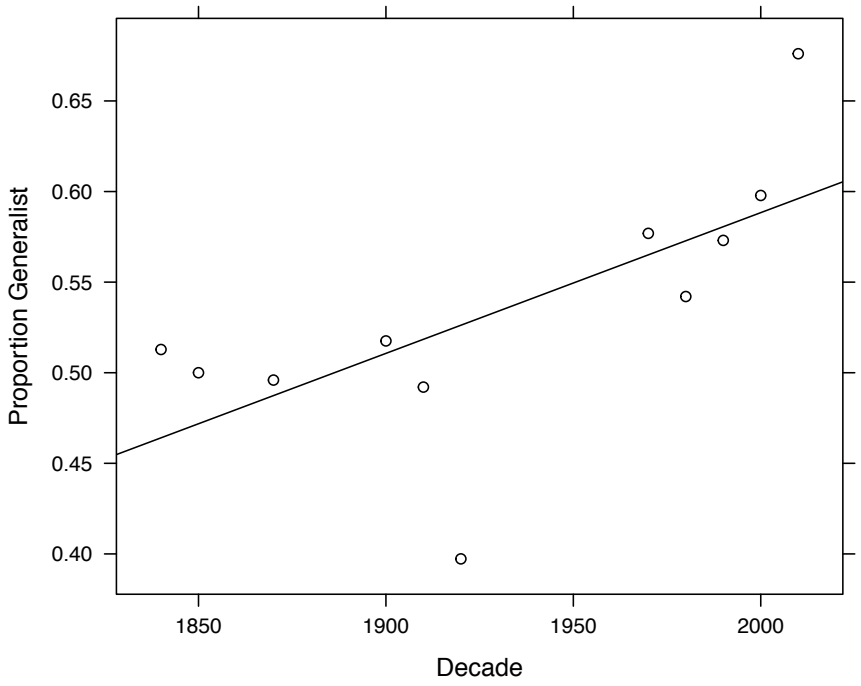
(e) Do your results from (c) and (d) show that rivastigmine is effective at improving the mean MMSE score for patients with schizophrenia? Briefly justify your answer. [1 mark]

Question 2

11 marks

Environmental changes strongly impact the distribution of species but little is known about how communities of species in different habitats transform over long time frames. To address this, a recent study analyzed changes in the species composition of a southeastern German butterfly community over nearly two centuries (1840–2013). They classified all species observed over this period, according to various ecological and behavioural traits, as either a habitat specialist (preferring a particular type of habitat) or a habitat generalist.

For each of 11 decades of available data, the researchers calculated the proportion of the species that were generalists. The figure below displays the proportion of generalists observed for each decade along with the least-squares line:



The output on the next page shows the results of a linear regression in R for the relationship between the proportion of generalist species and decade:

```
lm(formula = PropGeneral ~ Decade)

Residuals:
      Min        1Q    Median        3Q       Max
-0.128950 -0.016963  0.008562  0.020019  0.079940

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.9651151  0.5515491  -1.750   0.1141
Decade       0.0007767  0.0002855   2.721   0.0236 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05601 on 9 degrees of freedom
Multiple R-squared:  0.4513,  Adjusted R-squared:  0.3903
F-statistic: 7.401 on 1 and 9 DF,  p-value: 0.02359
```

(a) Briefly interpret the value -0.9651151 in the regression output. [1 mark]

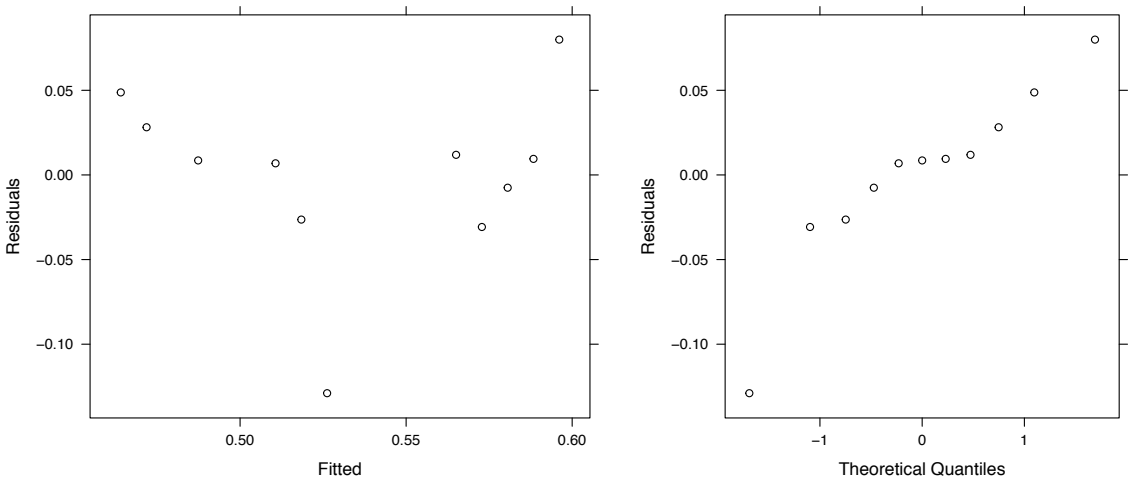
(b) Briefly interpret the value 0.0007767 in the regression output. [1 mark]

(c) Does the regression analysis provide evidence of a change in the proportion of butterfly species that are generalist over time? State the null and alternative hypotheses, and report the appropriate test statistic and P-value from the output. What do you conclude? [3 marks]

(d) Give a 95% confidence interval for the underlying slope of the linear relationship between the proportion of generalist species and time. [2 marks]

(e) In 1920 the observed proportion of habitat generalists was 0.397. What is the residual associated with 1920 in the regression model? [1 mark]

(f) The following figures were generated by R to help check the assumptions underlying the linear regression:

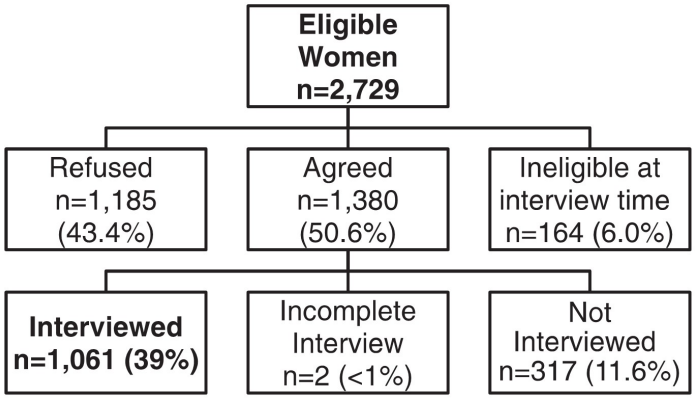


Comment on the validity of the assumptions underlying linear regression for this data with reference to these figures. [3 marks]

Question 3

12 marks

A 2014 study aimed to assess the relationship between volume and type of alcohol consumed during pregnancy in relation to miscarriage. A total of 2,729 women who had positive pregnancy tests at clinics were identified for participation but only 1,061 were ultimately interviewed, as shown in the following diagram from the paper:



In addition to recording whether each woman had a miscarriage, the researchers recorded variables such as their age, smoking status and the average number of alcoholic drinks per week.

(a) Was this an observational or experimental study? Briefly justify your answer. [1 mark]

Of the 208 women who were aged 36 years or above (36+), 52 had a miscarriage, while for the remaining 853 women who were under 36 (<36), 120 had a miscarriage.

(b) Overall, what proportion of women in the study had a miscarriage? [1 mark]

(c) What is the estimated difference in the rates of miscarriage between women 36+ and women <36? [1 mark]

(d) Give a 95% confidence interval for the true difference in the rates of miscarriage between women 36+ and women <36. What does the interval say about the relationship between age and the rate of miscarriage? [4 marks]

(e) The following table gives the summary of results related to the main research question of whether the volume of alcohol consumed is associated with the rate of miscarriage:

	Miscarriage	
Average number of alcoholic drinks per week	Yes	No
4+ drinks per week	11	21
1-3 drinks per week	66	337
No alcohol intake	95	531

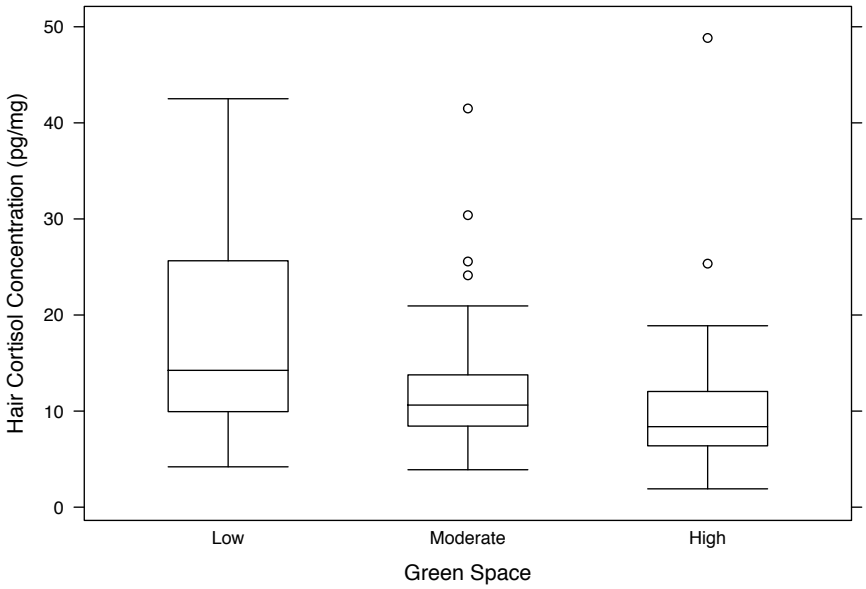
Based on this table, is there evidence of an association between the number of alcoholic drinks and rate of miscarriage? [5 marks]

Question 4

14 marks

Neighbourhood green space has been positively associated with health. This may be due to stress reducing effects of nature but researchers have usually had to rely on self-reported measures in studying this relationship. Hair cortisol concentration (HCC) provides a novel method of measuring chronic stress since high levels of cortisol secretion have been associated with stress and this builds up over time in hair.

Researchers in the UK collected data from 135 healthy employed adults. Postcodes were used to determine the proportion of green space in each participants' home neighbourhood and this was split into three equally sized groups labelled as 'Low', 'Moderate' and 'High'. Hair samples (3 cm) were taken from the scalp and HCC was determined to reflect the past three months of cortisol secretion. The following plot shows a summary of the results:



The following table shows the corresponding summary statistics:

Green Space	n	$\bar{x}$	s
Low	45	18.1	10.69
Moderate	45	12.5	6.91
High	45	10.4	7.71

(a) Assuming a population standard deviation of 8 pg/mg, the researchers calculated that a sample size of 45 in each group would allow them to detect a 3 pg/mg difference between the three groups with 90% power. Briefly explain what '90% power' means. [1 mark]

(b) Briefly describe the distribution of hair cortisol concentration (pg/mg) for the 45 participants from neighbourhoods with low amounts of green space. [1 mark]

(c) The researchers were interested in whether hair cortisol concentration was related to the level of green space in the neighbourhood. State the null hypothesis in words for this study. [1 mark]

(d) An analysis of variance was conducted to compare the mean hair cortisol concentration (pg/mg) between the three levels of green space. This gave a sum of squares for the group variable of 1412 with a residual sum of squares of 9744. Based on these values and the study design, complete the following ANOVA table. [3 marks]

Source	DF	SS	MS	F
Residuals				
Total				

(e) What is the R^2 value for this model? Briefly interpret the value. [2 marks]

(f) Is there significant evidence of a difference in mean hair cortisol concentration across the three levels of green space at the 5% level? Justify your conclusion. [2 marks]

(g) It can be calculated that the least-significant difference (LSD) for a comparison between two groups at the 5% level is 3.58 pg/mg. Based on this value, which pairwise comparisons between levels of natural environment are significantly different? [2 marks]

(h) State the assumptions underlying one-way analysis of variance. Do these assumptions seem reasonable for this study? [2 marks]

END OF EXAMINATION

Formulas and Statistical Tables

BASICS

$$\overline{x} = \frac{\sum x_j}{n} \qquad s = \sqrt{\frac{\sum (x_j - \overline{x})^2}{n-1}}$$

STANDARDISING

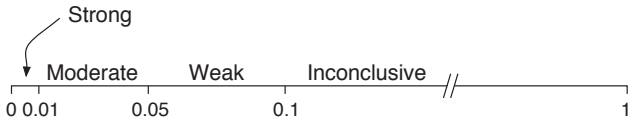
If $X \sim \text{Normal}(\mu, \sigma)$ then $Z = \frac{X-\mu}{\sigma} \sim \text{Normal}(0,1)$

BINOMIAL RANDOM VARIABLES

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad \text{(but is usually available in tables)}$$

$$\text{E}(X) = np \quad \text{sd}(X) = \sqrt{np(1-p)} \quad \hat{P} = \frac{X}{n} \quad \text{E}(\hat{P}) = p \quad \text{sd}(\hat{P}) = \sqrt{\frac{p(1-p)}{n}}$$

P-VALUES AND ERRORS



	Decision	
	Retain	Reject
H_0 is true	Correct $(1-\alpha)$	Type I Error (α)
H_0 is false	Type II Error (β)	Correct $(1-\beta)$

TESTS AND CONFIDENCE INTERVALS BASED ON STANDARD ERRORS

$$t = \frac{\text{estimate} - \text{hypothesised}}{\text{se(estimate)}} \qquad \text{estimate} \pm t^* \text{se(estimate)}$$

$$\text{se}(\overline{x}) = \frac{s}{\sqrt{n}} \qquad \text{se}(\overline{x}_1 - \overline{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad \text{se}(r) = \sqrt{\frac{1-r^2}{n-2}}$$

$$\text{se}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \qquad \text{se}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Use t for means, correlation and regression. Use z for proportions.

REGRESSION

$$y = b_0 + b_1 x \qquad y = b_0 + b_1 x + b_2 x_1 \qquad x_1 = \begin{cases} 1, & \text{if Group B} \\ 0, & \text{if Group A} \end{cases}$$

POOLED VARIANCE

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

ANOVA TABLES

$$\text{DFT} = n-1 \qquad \text{DFG} = k-1$$

$$\text{MS} = \frac{\text{SS}}{\text{DF}} \qquad R^2 = \frac{\text{SSG}}{\text{SST}} \qquad s_p = \sqrt{MSR} \qquad F = \frac{\text{MSG}}{\text{MSR}}$$

BONFERRONI CORRECTION FOR k COMPARISONS

$$\alpha = \frac{0.05}{k}$$

ODDS AND ODDS RATIOS

$$\text{Odds} = \frac{p}{1-p} \qquad \text{OR} = \frac{\text{Odds for group B}}{\text{Odds for group A}} \qquad \text{se}(\ln(\text{OR})) = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$$

CHI-SQUARED TESTS

$$\text{expected} = \frac{(\text{row total}) \times (\text{column total})}{\text{overall total}} \qquad \chi^2 = \sum \frac{(\text{observed}-\text{expected})^2}{\text{expected}}$$

$$\text{df} = (\#\text{ rows}-1) \times (\#\text{ columns}-1)$$

SIGN TEST

Count of positive values is $X \sim \text{Binomial}(n, 0.5)$

SIGNED-RANK TEST

$S =$ sum of ranks corresponding to positive differences

$$\text{E}(S) = \frac{n(n+1)}{4} \qquad \text{sd}(S) = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

Binomial Distribution

This table gives $P(X \geq x)$, where $X \sim \text{Binomial}(n, p)$.

		p							
n	x	0.01	0.05	0.10	0.20	0.25	0.30	0.40	0.50
18	1	0.165	0.603	0.850	0.982	0.994	0.998	1.000	1.000
	2	0.014	0.226	0.550	0.901	0.961	0.986	0.999	1.000
	3	0.001	0.058	0.266	0.729	0.865	0.940	0.992	0.999
	4		0.011	0.098	0.499	0.694	0.835	0.967	0.996
	5		0.002	0.028	0.284	0.481	0.667	0.906	0.985
	6			0.006	0.133	0.283	0.466	0.791	0.952
	7			0.001	0.051	0.139	0.278	0.626	0.881
	8				0.016	0.057	0.141	0.437	0.760
	9				0.004	0.019	0.060	0.263	0.593
	10				0.001	0.005	0.021	0.135	0.407
	11					0.001	0.006	0.058	0.240
	12						0.001	0.020	0.119
	13							0.006	0.048
	14							0.001	0.015
	15								0.004
	16								0.001

Critical values of the F distribution

This table gives f^* such that $P(F_{n,d} \geq f^*) = p$.

		n								
d	p	1	2	3	4	5	6	7	8	9
132	0.100	2.74	2.34	2.13	1.99	1.89	1.82	1.76	1.72	1.68
	0.050	3.91	3.06	2.67	2.44	2.28	2.17	2.08	2.01	1.95
	0.010	6.83	4.77	3.93	3.46	3.16	2.94	2.78	2.65	2.54
	0.001	11.3	7.28	5.75	4.92	4.39	4.02	3.74	3.52	3.35
133	0.100	2.74	2.34	2.13	1.99	1.89	1.82	1.76	1.72	1.68
	0.050	3.91	3.06	2.67	2.44	2.28	2.17	2.08	2.01	1.95
	0.010	6.83	4.77	3.93	3.46	3.16	2.94	2.78	2.65	2.54
	0.001	11.3	7.28	5.74	4.91	4.38	4.01	3.74	3.52	3.35
134	0.100	2.74	2.34	2.13	1.99	1.89	1.82	1.76	1.72	1.68
	0.050	3.91	3.06	2.67	2.44	2.28	2.17	2.08	2.01	1.95
	0.010	6.83	4.77	3.93	3.46	3.16	2.94	2.78	2.65	2.54
	0.001	11.3	7.28	5.74	4.91	4.38	4.01	3.74	3.52	3.35

Probabilities for the Standard Normal distribution

This table gives $P(Z \geq z)$ for $Z \sim \text{Normal}(0, 1)$.

	Second decimal place of z									
z	0	1	2	3	4	5	6	7	8	9
0.0	0.500	0.496	0.492	0.488	0.484	0.480	0.476	0.472	0.468	0.464
0.1	0.460	0.456	0.452	0.448	0.444	0.440	0.436	0.433	0.429	0.425
0.2	0.421	0.417	0.413	0.409	0.405	0.401	0.397	0.394	0.390	0.386
0.3	0.382	0.378	0.374	0.371	0.367	0.363	0.359	0.356	0.352	0.348
0.4	0.345	0.341	0.337	0.334	0.330	0.326	0.323	0.319	0.316	0.312
0.5	0.309	0.305	0.302	0.298	0.295	0.291	0.288	0.284	0.281	0.278
0.6	0.274	0.271	0.268	0.264	0.261	0.258	0.255	0.251	0.248	0.245
0.7	0.242	0.239	0.236	0.233	0.230	0.227	0.224	0.221	0.218	0.215
0.8	0.212	0.209	0.206	0.203	0.200	0.198	0.195	0.192	0.189	0.187
0.9	0.184	0.181	0.179	0.176	0.174	0.171	0.169	0.166	0.164	0.161
1.0	0.159	0.156	0.154	0.152	0.149	0.147	0.145	0.142	0.140	0.138
1.1	0.136	0.133	0.131	0.129	0.127	0.125	0.123	0.121	0.119	0.117
1.2	0.115	0.113	0.111	0.109	0.107	0.106	0.104	0.102	0.100	0.099
1.3	0.097	0.095	0.093	0.092	0.090	0.089	0.087	0.085	0.084	0.082
1.4	0.081	0.079	0.078	0.076	0.075	0.074	0.072	0.071	0.069	0.068
1.5	0.067	0.066	0.064	0.063	0.062	0.061	0.059	0.058	0.057	0.056
1.6	0.055	0.054	0.053	0.052	0.051	0.049	0.048	0.047	0.046	0.046
1.7	0.045	0.044	0.043	0.042	0.041	0.040	0.039	0.038	0.038	0.037
1.8	0.036	0.035	0.034	0.034	0.033	0.032	0.031	0.031	0.030	0.029
1.9	0.029	0.028	0.027	0.027	0.026	0.026	0.025	0.024	0.024	0.023
2.0	0.023	0.022	0.022	0.021	0.021	0.020	0.020	0.019	0.019	0.018
2.1	0.018	0.017	0.017	0.017	0.016	0.016	0.015	0.015	0.015	0.014
2.2	0.014	0.014	0.013	0.013	0.013	0.012	0.012	0.012	0.011	0.011
2.3	0.011	0.010	0.010	0.010	0.010	0.009	0.009	0.009	0.009	0.008
2.4	0.008	0.008	0.008	0.008	0.007	0.007	0.007	0.007	0.007	0.006
2.5	0.006	0.006	0.006	0.006	0.006	0.005	0.005	0.005	0.005	0.005
2.6	0.005	0.005	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004
2.7	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003
2.8	0.003	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
2.9	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.001	0.001	0.001
3.0	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
3.1	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
3.2	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
3.3										

Critical values of Student's T distribution

This table gives t^* such that $P(T \geq t^*) = p$, where $T \sim$ Student(df).

	Probability p								
df	0.25	0.10	0.05	0.025	0.01	0.005	0.001	0.0005	0.0001
1	1.000	3.078	6.314	12.71	31.82	63.66	318.3	636.6	3183
2	0.816	1.886	2.920	4.303	6.965	9.925	22.33	31.60	70.70
3	0.765	1.638	2.353	3.182	4.541	5.841	10.21	12.92	22.20
4	0.741	1.533	2.132	2.776	3.747	4.604	7.173	8.610	13.03
5	0.727	1.476	2.015	2.571	3.365	4.032	5.893	6.869	9.678
6	0.718	1.440	1.943	2.447	3.143	3.707	5.208	5.959	8.025
7	0.711	1.415	1.895	2.365	2.998	3.499	4.785	5.408	7.063
8	0.706	1.397	1.860	2.306	2.896	3.355	4.501	5.041	6.442
9	0.703	1.383	1.833	2.262	2.821	3.250	4.297	4.781	6.010
10	0.700	1.372	1.812	2.228	2.764	3.169	4.144	4.587	5.694
11	0.697	1.363	1.796	2.201	2.718	3.106	4.025	4.437	5.453
12	0.695	1.356	1.782	2.179	2.681	3.055	3.930	4.318	5.263
13	0.694	1.350	1.771	2.160	2.650	3.012	3.852	4.221	5.111
14	0.692	1.345	1.761	2.145	2.624	2.977	3.787	4.140	4.985
15	0.691	1.341	1.753	2.131	2.602	2.947	3.733	4.073	4.880
16	0.690	1.337	1.746	2.120	2.583	2.921	3.686	4.015	4.791
17	0.689	1.333	1.740	2.110	2.567	2.898	3.646	3.965	4.714
18	0.688	1.330	1.734	2.101	2.552	2.878	3.610	3.922	4.648
19	0.688	1.328	1.729	2.093	2.539	2.861	3.579	3.883	4.590
20	0.687	1.325	1.725	2.086	2.528	2.845	3.552	3.850	4.539
30	0.683	1.310	1.697	2.042	2.457	2.750	3.385	3.646	4.234
40	0.681	1.303	1.684	2.021	2.423	2.704	3.307	3.551	4.094
50	0.679	1.299	1.676	2.009	2.403	2.678	3.261	3.496	4.014
60	0.679	1.296	1.671	2.000	2.390	2.660	3.232	3.460	3.962
∞	0.674	1.282	1.645	1.960	2.326	2.576	3.090	3.291	3.719

Critical values of the χ^2 distribution

This table gives x^* such that $P(X^2 \geq x^*) = p$, where $X^2 \sim \chi^2$(df).

	Probability p								
df	0.975	0.95	0.25	0.10	0.05	0.025	0.01	0.005	0.001
1	0.001	0.004	1.323	2.706	3.841	5.024	6.635	7.879	10.83
2	0.051	0.103	2.773	4.605	5.991	7.378	9.210	10.60	13.82
3	0.216	0.352	4.108	6.251	7.815	9.348	11.34	12.84	16.27
4	0.484	0.711	5.385	7.779	9.488	11.14	13.28	14.86	18.47
5	0.831	1.145	6.626	9.236	11.07	12.83	15.09	16.75	20.52