8029A Semester 2 2011

The University of Sydney

Faculties of Arts, Economics, Education, Engineering and Science

MATH1905

Statistics

November 2011

Lecturer: J.T. Ormerod

Time Allowed: 90 minutes

Family Name: .....

Other Names: .....

SID: ..... Seat Number: .....

This examination has two sections: Multiple Choice and Extended Answer.

The Multiple Choice Section is worth 35% of the total examination; there are 20 questions of equal value; Answers to the Multiple Choice questions must be entered on the Multiple Choice Answer Sheet.

The Extended Answer Section is worth 65% of the total examination; there are 4 questions; the questions are not of equal value; all questions may be attempted; working must be shown.

Approved non-programmable, non-graphics calculators may be used. Statistical tables and notes for use in this examination are printed after the last question in the extended answer section in this booklet.

THE QUESTION PAPER MUST NOT BE REMOVED FROM THE EXAMINATION ROOM.

Marker's use only	

Extended Answer Section

Answer these questions in the answer book(s) provided.
Ask for extra books if you need them.

Extended Answer Question Paper

1. (14 marks in all)

(a) (8 Marks) The following R-output gives daily temperature, x_i in degrees Fahrenheit, and Ozone level, y_i in parts per billion in New York over 16 successive days.

```
> x = c(61,61,67,81,79,76,82,90,87,82,77,72,65,73,76,84)
> y = c(4,32,23,45,115,37,29,71,39,23,21,37,20,12,13,135)
```

Additionally you might find the following R output of use:

```
> sum(x)
[1] 1213
> sum(y)
[1] 656
> sum(x^2)
[1] 93125
> sum(y^2)
[1] 46868
> sum(x*y)
[1] 52111
> sort(round(lm(y~x)$resid,0))
 10  15   7   9  14  11   1   4   6   3   8  13  12   2   5  16
-31 -28 -25 -25 -23 -22  -7  -7  -4   0   1   1   4  21  67  77
```

(i) Calculate the correlation coefficient. How would the correlation coefficient change if the daily temperatures were measured in degrees Celsius instead of degrees Fahrenheit?

(ii) Calculate the least squares regression fit and the proportion variability of y's is explained by the regression line.

(iii) Use the R output above to calculate a 5 number summary. Use the 5 number summary to comment on the distribution of the residuals.

(b) (6 Marks) Consider the paired data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

(i) For a least squares regression fit show that the residuals $\widehat{e}_i = y_i - a - bx_i$ satisfy

$$\sum_{i=1}^{n} \widehat{e}_i = 0 \qquad \text{and} \qquad \sum_{i=1}^{n} \widehat{e}_i x_i = 0.$$

(ii) Hence, show that $\{x_i\}$ and $\{\widehat{e}_i\}$ are uncorrelated.

2. (18 marks in all)

(a) (9 Marks) The clinically accepted value for mean blood pressure in healthy males aged 18 to 22 years is 120 mm Hg. It is widely claimed that examination stress causes blood pressure to rise above 120 mm Hg. To test this theory, 10 healthy male students have their blood pressure taken just prior to a Statistics quiz. The sample mean and sample standard deviation of these measurements are 135.1 and 17.42 respectively. Assume that the measurements for each student are independent and normally distributed.

(*i*) What are appropriate null and alternative hypotheses to test this claim?

(*ii*) State an appropriate test statistic to test this hypothesis and the null distribution of this test statistic.

(*iii*) Calculate the test statistic chosen in (ii), the corresponding P-value and form an appropriate conclusion.

(b) (9 Marks) Consider the geometric distribution $P(X = x) = pq^x$, $x = 0, 1, 2, \ldots$, $0 \leq p \leq 1$ and $q = 1 - p$.

(*i*) Show that the probability generating function is given by $\pi(s) = p/(1 - qs)$ for $|s| < 1/q$.

(*ii*) Use Part (i) to show that $E(X) = q/p$ and $\text{Var}(X) = q/(p^2)$.

(*iii*) Use Part (ii) and Chebyshev's inequality to bound $P(|X - q/p| > 1)$.

3. (15 marks in all)

(a) (10 marks) It has been claimed that at least 60% of all purchasers of a certain computer program will call the manufacturer's hotline within one month of purchase. A random sample of 12 purchasers of this software is drawn and 3 of those in the sample had contacted the hotline within one month of purchase. Does this provide evidence that the claim of a 60% contact rate is an overestimate? Let p be the true proportion of all purchasers who contact the hotline.

(*i*) Calculate an approximate 95% confidence interval for p.

(*ii*) Form an appropriate hypotheses and perform a statistical test for the above situation stating any assumptions you may require.

(*iii*) Do the results from part (i) and part (ii) agree? Justify your answer.

(b) (5 marks) Suppose that the probability that a randomly chosen individual having a particular disease is 0.02.

(*i*) Write an expression for the exact probability that at most 2 cases in a random sample of 100 has the disease.

(*ii*) Use the Poisson approximation to the Binomial distribution to approximate this probability. Why is this approximation appropriate?

4. (18 marks in all)

(a) (5 marks) A new casino game involves rolling 3 dice. The winnings are directly proportional to the total number of sixes rolled. Suppose a gambler plays the game 100 times, with the following observed counts:

Sixes Rolled	0	1	2	3
Number of Rolls	48	35	15	3

You may use the R output:

```
> dbinom(0:3,3,1/6)
[1] 0.57870370 0.34722222 0.06944444 0.00462963
```

The casino becomes suspicious of the gambler and wishes to determine whether the dice are fair. Use a statistical test to determine where whether the dice are fair or not.

(b) (9 marks) Two pathology labs, lab A and lab B, are compared to see which of the labs report their results for a specific test faster. Samples from 20 patients are collected and then 10 of these samples are sent to lab A and 10 of these samples are sent to lab B. Summary values for the times, in days, for each lab to report its results are summarised in the table below:

	Size	Mean	Median	Variance
Lab A	10	20.23	19.45	2.74
Lab B	10	18.68	17.98	1.64

You may assume that the measurements for lab A and lab B are normally distributed. Suppose we wish to test whether there is a difference in the times each lab reports its results.

(*i*) State an appropriate null and alternative hypothesis, defining any parameters used.

(*ii*) State an appropriate test statistic to test this hypothesis and the null distribution of this test statistic, stating any additional assumptions required.

(*iii*) Calculate the test statistic chosen in (ii), the corresponding P-value and form an appropriate conclusion.

(c) (4 marks) Let Ω be a sample space, $A \subset \Omega$, $B \subset \Omega$ and $P(\cdot)$ be a probability function satisfying the axioms of probability

- For any event $A \subset \Omega$, $P(A) \geq 0$,
- $P(\Omega) = 1$
- If A and B are mutually exclusive events ($A \cap B = \emptyset$), then

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B).$$

Use the above axioms to show that $P(A) \leq 1$ for all events A.

FORMULA SHEET FOR MATH1905 STATISTICS

- **Calculation formulae:**
  - *For a sample* $x_1, x_2, \ldots, x_n$

Sample mean $\overline{x}$	$\dfrac{1}{n}\sum_{i=1}^{n} x_i$
Sample variance s^2	$\dfrac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \dfrac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right] = \dfrac{1}{n-1}S_{xx}.$

  - *For paired observations* $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

S_{xy}	$\sum_{i=1}^{n} x_i y_i - \dfrac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$
S_{xx}	$\sum_{i=1}^{n} x_i^2 - \dfrac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$
S_{yy}	$\sum_{i=1}^{n} y_i^2 - \dfrac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2$
r	$\dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$

For the regression line $y = a + bx$:	
b	$\dfrac{S_{xy}}{S_{xx}}$
a	$\overline{y} - b\overline{x}$

- **Some probability results:**

For any two events A and B	$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$ and $\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B\|A)$
If A and B are mutually exclusive (m.e.)	$\mathrm{P}(A \cap B) = 0$ and $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$
If A and B are independent	$\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B)$

- If $X \sim B(n, p)$, then :
  $\mathrm{P}(X = i) = \binom{n}{i} p^i (1-p)^{n-i}$, $i = 0, \ldots, n$, $\mathrm{E}(X) = np$ and $\mathrm{Var}(X) = np(1-p)$

- **Some test statistics** and sampling distributions under appropriate assumptions and hypotheses:

$\overline{X} \sim \mathcal{N}\left(\mu, \dfrac{\sigma^2}{n}\right)$
$\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$
$\dfrac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$

$\dfrac{\overline{X} - \overline{Y}}{S_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \sim t_{n_x+n_y-2}$, where
$S_p^2 = [(n_x - 1)S_x^2 + (n_y - 1)S_y^2]/(n_x + n_y - 2)$
$\sum_i \dfrac{(O_i - E_i)^2}{E_i} \sim \chi^2_\nu$, for appropriate ν

TABLE 1. **Some values of the standard normal distribution:** $\Phi(z) = F(z) = \mathrm{P}(Z \leq z)$, where $Z \sim \mathcal{N}(0,1)$. The point tabulated is $1-p$, where $\mathrm{P}(Z \leq z) = 1-p$.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990

TABLE 2. **Quantiles of the $\mathcal{N}(0,1)$ distribution:** Some percentage points of the standard normal. The point tabulated is z, where $\mathrm{P}(Z > z) = p$, where $Z \sim \mathcal{N}(0,1)$.

p								
0.25	0.15	0.10	0.05	0.025	0.01	0.005	0.0025	0.001
0.674	1.036	1.282	1.645	1.960	2.326	2.576	2.807	3.090

TABLE 3. **Critical values of the t test:** Some percentage points of the t-distribution with ν degrees of freedom. The point tabulated is t, where $\mathrm{P}(t_\nu > t) = p$.

	p								
ν	0.25	0.15	0.10	0.05	0.025	0.01	0.005	0.0025	0.001
1	1.000	1.963	3.078	6.314	12.706	31.821	63.656	127.321	318.309
2	0.817	1.386	1.886	2.920	4.303	6.965	9.925	14.089	22.328
3	0.765	1.250	1.638	2.353	3.182	4.541	5.841	7.453	10.214
4	0.741	1.190	1.533	2.132	2.776	3.747	4.604	5.598	7.173
5	0.727	1.156	1.476	2.015	2.571	3.365	4.032	4.773	5.894
6	0.718	1.134	1.440	1.943	2.447	3.143	3.707	4.317	5.208
7	0.711	1.119	1.415	1.895	2.365	2.998	3.499	4.029	4.785
8	0.706	1.108	1.397	1.860	2.306	2.896	3.355	3.833	4.501
9	0.703	1.100	1.383	1.833	2.262	2.821	3.250	3.690	4.297
10	0.700	1.093	1.372	1.812	2.228	2.764	3.169	3.581	4.144
20	0.687	1.064	1.325	1.725	2.086	2.528	2.845	3.153	3.552
30	0.683	1.055	1.310	1.697	2.042	2.457	2.750	3.030	3.385
50	0.679	1.047	1.299	1.676	2.009	2.403	2.678	2.937	3.261
∞	0.674	1.036	1.282	1.645	1.960	2.326	2.576	2.807	3.090

TABLE 4. **Quantiles of the χ^2_ν distribution:** Some percentage points of the χ^2-distribution with ν degrees of freedom. The point tabulated is x, where $\mathrm{P}(\chi^2_\nu > x) = p$.

	p					
ν	0.25	0.15	0.10	0.05	0.025	0.01
1	1.323	2.072	2.706	3.841	5.024	6.635
2	2.773	3.794	4.605	5.991	7.378	9.210
3	4.108	5.317	6.251	7.815	9.348	11.345
4	5.385	6.745	7.779	9.488	11.143	13.277
5	6.626	8.115	9.236	11.070	12.833	15.086
6	7.841	9.446	10.645	12.592	14.449	16.812
7	9.037	10.748	12.017	14.067	16.013	18.475
8	10.219	12.027	13.362	15.507	17.535	20.090
9	11.389	13.288	14.684	16.919	19.023	21.666
10	12.549	14.534	15.987	18.307	20.483	23.209

End of Extended Answer Section