8029 Semester 2 2009

The University of Sydney

Faculties of Arts, Economics, Education,
Engineering and Science

MATH1905

Statistics

November 2009

Lecturer: S. Müller

Time Allowed: **90 minutes**

This examination consists of 7 pages, numbered from 1 to 7.

There are 4 questions, numbered from 1 to 4.

Calculators will be supplied; no other calculators are permitted.
Statistical tables and notes for use in this examination are printed after the last question in the extended answer section in this booklet.

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

Extended Answer Question Paper

1. (20 marks in all) In an experiment, rats were placed in a shuttlebox (maze) and were timed in successive attempts. An unsuccessful attempt was followed by an electric shock for the duration of the next attempt if an attempt was longer than 5 seconds. The data give the average times y and the number of shocks x for all attempts.

```
> x  # Shocks
 [1]  0  1  2  3  4  5  6  7  8  9 10 11 12
> y  # Time (average time to complete the maze)
 [1] 11.4 11.9  7.1 14.2  5.9  6.1  5.4  3.1  5.7  4.4
[11]  4.0  2.8  2.6
```

(a) (5 marks) The minimum, Q_1, median, Q_3 and maximum are calculated with R:

```
> quantile(y,type=2)
  0%  25%  50%  75% 100%
 2.6  4.0  5.7  7.1 14.2
```

Using this information sketch a boxplot of time (y) and comment on the plot with regards to symmetry and outliers present.

(b) (5 marks) Sketch the scatter plot of the data with Schocks, x, on the horizontal axis. Comment on the relationship of Schocks and Time.

(c) (4 marks) Given that $\sum x = 78$, $\sum y = 84.6$, $\sum x^2 = 650$, $\sum y^2 = 716.86$ and $\sum xy = 364.1$ find the regression line to predict Time from Schocks and add the line to your scatter plot. Use the regression to predict the average time to complete the maze corresponding to the number of Shocks of 14.

(d) For paired observations $(x_1, y_1), \ldots, (x_n, y_n)$ we denote $\hat{y}_i = a + bx_i$ as the ith predicted value of the regression line.

(i) (3 marks) Show that $\displaystyle\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \frac{S_{xy}^2}{S_{xx}}$.

(ii) (3 marks) Explain how the result in (i) relates to r^2, the squared correlation coefficient.

2. (16 marks in all)

(a) For 100 used-car salesman data is available on the number of attempts needed in order to sell their first car. The number of attempts, x, before the first success were

No. of attempts (x):	0	1	2	3	4	5	6	7 or more
Frequency:	15	16	17	14	14	8	8	8

Suppose that we can model the number of unsuccessful attempts before the first success as a Geometric random variable X with success probability p, where

$$P(X = j) = (1 - p)^j p, \quad j = 0, 1, \ldots.$$

(*i*) (2 marks) Because $\mathrm{P}(X = 1) = (1-p)p$ we can estimate the success probability p by solving $\frac{16}{100} = \mathrm{P}(X = 1)$ for p. Determine $\hat{p}$, the estimated p.

(*ii*) (6 marks) However, in the salesman-community it is in general believed that the success probability is 0.15. Therefore X should be modeled with a Geometric distribution with $p = 0.15$. Complete the table of expected frequencies below and test the goodness of fit of the Geometric distribution with $p = 0.15$ as a model for the number of failures before the first successful sale.

No. of attempts (x):	0	1	2	3	4	5	6	7 or more
Expected Frequency:	15.00	12.75	?	9.21	7.83	6.66	5.66	?

For specifying if P-value $\geq 5\%$ or P-value $\leq 5\%$ use the following R-output:

```
> qchisq(0.95,7)
[1] 14.06714
> qchisq(0.95,8)
[1] 15.50731
> qchisq(0.95,9)
[1] 16.91898
```

(b) The random variable X has probability generating function

$$\pi(s) = \frac{p}{1 - s(1 - p)}.$$

(*i*) (3 marks) Use this function to find $\mathrm{E}(X)$.

(*ii*) (5 marks) Given that $\pi''(s) = 2p(1 - p)^2[1 - s(1 - p)]^{-3}$ find $\mathrm{Var}(X)$.

3. (13 marks in all)

(a) (3 marks) Of 12 patients, 7 have a positive test result and 5 have a negative result. If 3 patients are selected at random from the 12 patients, what is the probability that none of them will have positive results.

(b) (3 marks) A drug is expected to have positive results in 50% of the cases. If 12 patients are treated with the drug, what is the chance that at most 2 will have positive results. Obtain an answer using the normal approximation with correction for continuity and the normal table.

(c) (2 marks) For the experiment in (b) there were 2 patients of the 12 with positive results. Explain how this provides evidence that the drug has positive results in less than 50% of cases.

(d) (5 marks) Let A and B be two independent events. Show that $\mathrm{P}(A^c \cap B^c) = \mathrm{P}(A^c)\,\mathrm{P}(B^c)$, where A^c denotes the complement of A. (Hints: you can use the three axioms of probability, the multiplication rule, $1 - \mathrm{P}(A) = A^C$, de Morgan etc).

4. (16 marks in all) For each of ten patients, the average number of extra hours' sleep was measured using each of two different drugs, A and B (negative results mean that the drug actually *reduced* sleeping time overall). For each patient, the average gain for each drug as well as the difference between these is given below for each patient. Sums and sums of squares are given for each column.

Patient	A	B	$B - A$
1	0.7	1.9	1.2
2	−1.6	0.8	2.4
3	−0.2	1.1	1.3
4	−1.2	0.1	1.3
5	−0.1	−0.1	0.0
6	3.4	4.4	1.0
7	3.7	5.5	1.8
8	0.8	1.6	0.8
9	0.0	4.6	4.6
10	2.0	3.4	1.4
Sum	7.5	23.3	15.8
Sum sq.	34.43	90.37	38.58

(a) (3 marks) Obtain a 95% one-sided confidence interval of the form $[c, \infty)$ for the true but unknown average gain (population mean) of drug B.

(b) Perform an appropriate two-sided t-test of a hypothesis that there is no difference between the drugs. In particular, address the following three points:

(*i*) (1 mark) State the assumptions under which the test is valid;

(*ii*) (2 marks) introduce an appropriate statistical model and state appropriate null and alternative hypotheses in terms of a parameter in the model;

(*iii*) (3 marks) compute and interpret the resulting P-value;

(c) The t-test requires a distributional assumption to be valid. An alternative test which does not have such an assumption is the sign test.

(*i*) (2 marks) Set up appropriate null and alternative hypotheses for this sign test.

(*ii*) (2 marks) Compute and interpret the resulting P-value.

(d) (3 marks) Based on the following R-output discuss briefly how you might decide which test is better to use.

```
> quantile(A,type=2)
   0%   25%   50%   75%  100% 
-1.60 -0.20  0.35  2.00  3.70 
> quantile(B,type=2)
   0%   25%   50%   75%  100% 
-0.10  0.80  1.75  4.40  5.50 
> quantile(B-A,type=2)
  0%  25%  50%  75% 100% 
 0.0  1.0  1.3  1.8  4.6 
```

FORMULA SHEET FOR MATH1905 STATISTICS

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

Sample mean $\bar{x}$	$\frac{1}{n}\sum_{i=1}^{n} x_i$
Sample variance s^2	$\frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\Big(\sum_{i=1}^{n} x_i\Big)^2\right].$

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

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

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

- **Some probability results:**

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

- If $X \sim \mathcal{B}(n, p)$, then :

$$P(X = r) = \binom{n}{r} p^r (1-p)^{n-r},\ r = 0, \ldots, n, \quad E(X) = np \quad \text{and} \quad var(X) = np(1-p)$$

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

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

$$\frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \sim t_{n_x+n_y-2}, \text{ where}$$

$$S_p^2 = [(n_x - 1)S_x^2 + (n_y - 1)S_y^2]/(n_x + n_y - 2)$$

$$\sum_i \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_\nu, \ \text{for appropriate } \nu$$

Table 1. **Some values of the standard normal**: $\Phi(x) = F(z) = P(Z \leq z)$, where $Z \sim \mathcal{N}(0,1)$.

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. **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

This is the last page of the question paper.