The University of Sydney

Faculties of Arts, Economics, Education,
Engineering and Science

MATH1905

Statistics

November 2010 Lecturer: S. Müller

Time Allowed: 90 minutes

Family Name: ....

Other Names: ....

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

Marker's use only	

This examination consists of 8 pages, numbered from 1 to 8.

There are 4 questions, numbered from 1 to 4.

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

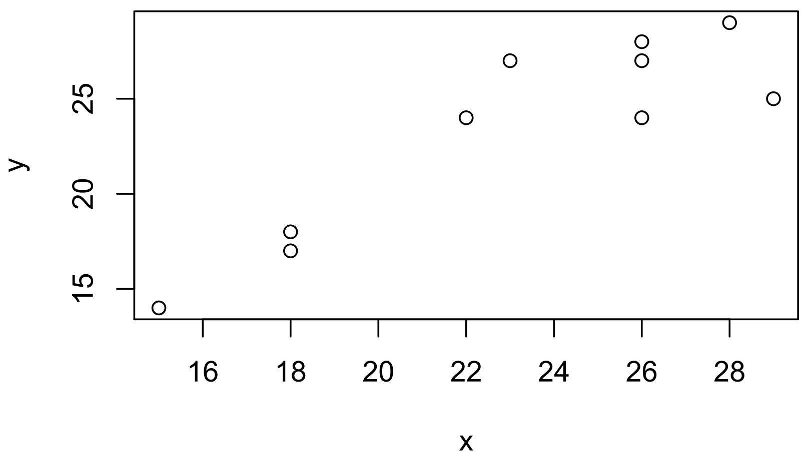
Extended Answer Question Paper

1. (16 marks in all) The following R-output gives the maximum temperatures x_i and y_i, in °C, on two successive days (day 1 and day 2) in 10 Australian weather observation stations, $i = 1, \ldots, 10$:

```
> x = c(22,18,26,26,29,15,23,18,28,26)
> y = c(24,17,28,24,25,14,27,18,29,27)
```

Additionally you might find the following R output of use:

```
> (10-1)*var(x)
[1] 202.9
> 9*var(y)
[1] 240.1
> sum(x)
[1] 231
> sum(y)
[1] 233
> sum(x^2)
[1] 5539
> sum(y^2)
[1] 5669
> sum(x*y)
[1] 5580
```



(a) (3 marks) Calculate the coefficient of correlation.

(b) (4 marks) Find the equation of the least squares line of temperature on day 2 (y) on temperature of day 1 (x).

(c) (2 marks) The simplest weather forecast is 'tomorrow will be like today'. Use the regression line to give an improved forecast in one of those weather stations, if today's temperature there is 15°C

(d) (6 marks) The residuals $e_1, \ldots, e_{10}$ from the least squares fit are as follows (to 1dp):

```
> round(lm(y~x)$res,1)
   1    2    3    4    5    6    7    8    9   10
 1.8 -1.3  1.9 -2.1 -4.0 -1.4  3.8 -0.3  0.9  0.9
```

Draw a residual plot and comment on the appropriateness of the linear regression.

(e) (1 mark) About what percentage of the variability of y's is explained by the regression line?

2. (15 marks in all) The following data are measurement of weight gain (in gm) after 10 male rats and 10 female rats were given the same diet over the same period of time. The 10 male rats and 10 female rats were chosen independently.

Male (x)	2.6	4.8	12.5	8.7	9.7	8.2	9.4	8.7	9.2	10.0
Female (y)	8.1	7.6	10.5	8.9	11.2	6.9	11.7	12.6	10.3	7.1

(a) (8 marks) Assume that both samples can be modelled by a normal distribution with the same population variance. Given that $\bar{x} = 8.38$, $s_x^2 = 7.724$, $\Sigma_{i=1}^{10} y_i = 94.9$, $\Sigma_{i=1}^{10} y_i^2 = 938.03$, is there evidence of a difference in weight gains between male rats and female rats?

(b) (7 marks) Let $X \sim \mathcal{N}(\mu, \sigma^2)$ with $\mu = 1$ and $\sigma^2 = 1/2$.

(i) Use Chebyshev's inequality to bound $\mathrm{P}(X > 2)$.

(ii) Is Chebyshev's bound obtained in *(i)* sharp in this case, i.e. are the two probabilities the same? (Justify your answer)

3. (17 marks in all) The number of radioactive counts in 100 one minute intervals for a particular machine were

No. of counts (x):	0	1	2	3	4	5
Observed frequency:	15	28	23	16	12	6

Suppose that we can model the number of counts in one minute by a Poisson random variable X, where

$$\mathrm{P}(X = i) = e^{-\lambda}\lambda^i / i!, \quad i = 0, 1, 2, \ldots.$$

(a) (4 marks) Prove that the probability generating function of the random variable X is $\pi(s) = e^{\lambda(s-1)}$.

(b) (2 marks) Use the probability generating function in part (a) to show that $\mathrm{E}(X) = \lambda$.

(c) (2 marks) It is known that $\mathrm{Var}(X) = \mathrm{E}(X) = \lambda$, determine $\mathrm{E}(X^2)$.

(d) (2 marks) Calculate the expected number of 0's in a sample of size 100 from a Poisson random variable with mean 2.

(e) (7 marks) You are given that the above sample average is $\bar{x} = 2$. Complete the table of expected frequencies below and test the goodness of fit of the Poisson distribution as a model for the number of radioactive counts.

No. of counts (x):	0	1	2	3	4	≥ 5
Expected frequency:	?	27.07	27.07	18.04	9.02	?

4. (17 marks in all) When cancerous tumours are removed from the colon it is not always possible to remove all cancerous cells without removing too much of the patient's vital organs. Consider the following data:

		Was the cancer controlled?	
		Yes	No
Was cancer present	Yes	8	182
at the edge of surgery?	No	11	58

(a) (8 marks) Is there any evidence that cancer at the edge of surgery affects the chance of the cancer being controlled?

(b) (5 marks) Consider the 69 patients who had 'no cancer present at the edge of surgery' only:

(i) Provide a conservative 95% confidence interval for the proportion of having controlled cancer.

(ii) Determine how much smaller the length of the conservative confidence interval in *(i)* is when the sample size was 100 instead of 69 patients.

(c) (4 marks) Let A and B be two independent events. Show that $\mathrm{P}(A \cap B^c) = \mathrm{P}(A) \times \mathrm{P}(B^c)$, where B^c denotes the complement of B.

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	$\frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right] = \frac{1}{n-1}S_{xx}.$

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

S_{xy}	$\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$
S_{xx}	$\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$
S_{yy}	$\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2$
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 :
  $P(X = i) = \binom{n}{i} p^i (1-p)^{n-i},\ i = 0, \ldots, n,\quad E(X) = np \quad \text{and} \quad \mathrm{Var}(X) = np(1-p)$

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

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

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

Table 1. **Some values of the standard normal distribution**: $\Phi(x) = F(z) = \mathrm{P}(Z \leq z)$, 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 $P(t_\nu > t) = p$.

	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 $P(\chi^2_\nu > x) = p$.

	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