THE UNIVERSITY OF SYDNEY
SCHOOL OF MATHEMATICS AND STATISTICS

Practice Quiz Week 7

MATH1905: Statistics (Advanced) Semester 2, 2017

Web Page: http://sydney.edu.au/science/maths/MATH1905
Lecturer: Michael Stewart

Full Name .... SID ....

Day .... Time .... Room...........................

Tutor .... Signature...

Time allowed: 40 minutes

1. **This quiz is closed book. You may not use a computer.**
2. Full marks will only be given if you obtain the correct answer **and** your working is sufficient to justify your answer.
3. Partial marks may be awarded for working.
4. Please write carefully and legibly.
5. All of your answers should be written using ink and **not** pencil, with your final answer placed in the answer box.
6. All working must be done on the quiz paper in the indicated space.
7. Each question is worth **2 marks**.
8. Only University of Sydney approved calculators may be used (must have a sticker).
9. All pages (including working) of the quiz paper must be handed in at the end of the quiz.

This quiz paper has 12 pages (this cover sheet + 10 pages of questions + 1 page of statistical formulae) and 10 questions.

1. A vector `x` in R yields the following output:

```
length(x)
```

```
[1] 50
```

```
sum(x)
```

```
[1] 249
```

```
sum(x^2)
```

```
[1] 1453
```

Determine (to 2 decimal places) the mean and sample standard deviation of `x`.

Mean of `x` is	Sample SD of `x` is
4.98	2.08

Please show your working below this line

Writing $\bar{x}$ and s^2 for the mean and (sample) variance respectively, we have

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{249}{50} = 4.98$$

and

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right] = \frac{1}{49}\left[1453 - \frac{249^2}{50}\right] = 10649/2450 \approx 4.347$$

Thus the sample SD is $\sqrt{4.347} \approx 2.08$.

2. The chief accountant of a large company collected the following information on advertising expenditure (in thousands of dollars) and revenue (in ten thousands of dollars) for 7 of its popular products as shown below:

Revenue y	6	5.4	7.4	4.7	4.9	4.6	7
Expenditure x	2.8	2.5	2.5	2.6	2.6	2.7	2.5

Determine the correlation coefficient (to 3 decimal places). You may find the R output below useful:

```
y=c(6,5.4,7.4,4.7,4.9,4.6,7)
x=c(2.8,2.5,2.5,2.6,2.6,2.7,2.5)
```

```
var(x)
```

```
[1] 0.01333333
```

```
var(y)
```

```
[1] 1.268095
```

```
sum((x-mean(x))*(y-mean(y)))
```

```
[1] -0.32
```

The correlation coefficient is

-0.410

Please show your working below this line

The final number in the output is precisely the quantity S_{xy}. The quantity S_{xx} is related to s_x^2, the sample variance of the x_i's, via

$$s_x^2 = \frac{S_{xx}}{n-1} = \frac{S_{xx}}{6},$$

so

$$S_{xx} = 6 \times \cancel{0.13} = 0.08\,;$$

0.013

similarly

$$S_{yy} = 6 \times 1.268 \approx 7.608\,.$$

Therefore the correlation coefficient

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \approx \frac{-0.32}{\sqrt{0.08 \times 7.608}} \approx -0.410\,.$$

3. Two events A and B are such that $P(A) = 10/31$, $P(B) = 12/31$ and $P(A \cup B) = 16/31$. Determine $P(A|B)$, that is the conditional probability of A given B.

$P(A|B) = 1/2$

Please show your working below this line

Note firstly that since, by the "general addition rule", we have

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

we can deduce that $P(A \cap B) = P(A) + P(B) - P(A \cup B) = 10/31 + 12/31 - 16/31 = 6/31$. Thus $P(A|B) = \frac{P(A \cap B)}{P(B)} = (6/31)/(12/31) = 1/2$.

4. An urn contains 13 balls: 4 are red, 6 are blue and 3 are white. A random sample of size 3 is taken **with** replacement. Determine the probability (to 4 decimal places, or as an exact ratio) that all balls in the sample are the same colour.

Probability all are the same colour is:

$$\frac{307}{2197} \approx 0.1397\,.$$

Please show your working below this line

In general, for any non-negative integers x, y, z with $x + y + z = 3$,

$$p(x, y, z) = P(x \text{ red, } y \text{ blue, } z \text{ white}) = \frac{3!}{x!y!z!}\left(\frac{4}{13}\right)^x\left(\frac{6}{13}\right)^y\left(\frac{3}{13}\right)^z,$$

i.e. a multinomial probability. The desired probability is in fact

$$\begin{aligned}
p(3,0,0) + p(0,3,0) + p(0,0,3) &= \left(\frac{4}{13}\right)^3 + \left(\frac{6}{13}\right)^3 + \left(\frac{3}{13}\right)^3 \\
&= \frac{4^3 + 6^3 + 3^3}{13^3} = \frac{64 + 216 + 27}{2197} = \frac{307}{2197} \approx 0.1397
\end{aligned}$$

5. It is known that 6% of the children in a particular community suffer from a particular blood disorder. A test performed in a clinic correctly diagnoses 97% of children with this disorder as "positive" for the disorder, but also misdiagnoses 9% of children who do not have the disorder as "positive" for the disorder.

A child (randomly chosen from the community) is diagnosed "positive" by the clinic. Write down (to 3 decimal places) $P(D|+)$, that is the (conditional) probability that they actually have the disorder, given they have a positive test result.

$$\boxed{P(D|+) = 0.408}$$

Please show your working below this line

Consider the events

$$\begin{aligned} D &= \text{ the randomly chosen child has the disorder} \\ + &= \text{ the randomly chosen child has a positive test result}\,; \end{aligned}$$

then the information in the question can be translated as follows:

$$\begin{aligned} P(D) &= 0.06\,, \\ P(+|D) &= 0.97\,, \\ P(+|D^c) &= 0.09\,. \end{aligned}$$

Also,

$$P(D^c) = 1 - P(D) = 0.94\,.$$

Using Bayes' rule,

$$\begin{aligned} P(D|+) &= \frac{P(D \cap +)}{P(+)} \\ &= \frac{P(D \cap +)}{P(D \cap +) + P(D^c \cap +)} \\ &= \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|D^c)P(D^c)} \\ &= \frac{0.97 \times 0.06}{(0.97 \times 0.06) + (0.09 \times 0.94)} \\ &= \frac{0.0582}{0.0582 + 0.0846} \\ &\approx 0.408\,. \end{aligned}$$

6. A random variable X only taking values 0,1,. . . ,5 has the following probability distribution ($P(X = 5)$ is obscured by the $*$):

x	0	1	2	3	4	5
$P(X = x)$	0.2	0.05	0.25	0.35	0.1	*

Given that the mean or the expected value of X, $E(X) = 2.25$, determine $P(X = 5)$ and $Var(X)$.

$P(X = 5) =$	$Var(X) =$
0.05	1.9875

Please show your working below this line

The visible probabilities add to

$$0.2 + 0.05 + 0.25 + 0.35 + 0.1 = 0.95\,,$$

so $P(X = 5) = 0.05$.

The mean square is

$$\begin{aligned} E(X^2) &= (0 \times 0.2) + (1 \times 0.05) + (4 \times 0.25) + (9 \times 0.35) + (16 \times 0.1) + (25 \times 0.05) \\ &= 0 + 0.05 + 1 + 3.15 + 1.6 + 1.25 = 7.05\,. \end{aligned}$$

Therefore

$$Var(X) = E(X^2) - [E(X)]^2 = 7.05 - (2.25^2) = 7.05 - 5.0625 = 1.9875\,.$$

7. A fair six-sided die is thrown twice independently. Let A be the even that the sum of the two numbers showing face-up is strictly less than 6. Determine $P(A)$.

$$\boxed{P(A) = \frac{5}{18}.}$$

Please show your working below this line

Let S denote the sum of the two numbers showing face-up. There are 36 equally likely possible outcomes. One of these gives $\{S = 2\}$: (1,1). Two give $\{S = 3\}$: (1,2) and (2,1). Three give $\{S = 4\}$: (1,3), (2,2) and (3,1). Four give $S = 5$: (1,4), (2,3), (3,2) and (4,1). Since 10 of the 36 equally likely outcomes give $\{S \leq 5\}$, the answer is

$$\frac{10}{36} = \frac{5}{18}\,.$$

8. 10 tickets of equal size, feel, each have a number written on them. The numbers are stored in the R vector `tickets`, whose summary statistics are given below:

```
tickets
```

```
[1]  1  2  4  4  5  7  7  9 10 15
```

```
sum(tickets)
```

```
[1] 64
```

```
var(tickets)
```

```
[1] 17.37778
```

A ticket is drawn at random (so that each is equally likely). Let X denote the (random) number showing on the selected ticket. Determine $Var(X)$ (to 2 decimal places).

$Var(X) \approx 15.64$.

Please show your working below this line

$Var(X)$ is simply the *population variance* of the numbers on the tickets. This is given by $\frac{9}{10}$`var(tickets)` $\approx \frac{9}{10} \times 17.37778 \approx 15.64$.

9. Emails arrive in an inbox at a rate of 1.5 per minute and the number over any time period is well modelled as a Poisson random variable. If X is the number of emails arriving in the next three minutes, determine $P(X = 3)$ (to 3 decimal places).

$P(X = 3) \approx 0.169.$

Please show your working below this line

The expected number of emails in the next 3 minutes is 4.5, so we take the distribution of X as being Poisson(4.5). Thus for any non-negative integer x,

$$P(X = x) = \frac{e^{-4.5}4.5^x}{x!}$$

and so the desired probability is $\frac{e^{-4.5}(4.5)^3}{6} \approx 0.169.$

10. A random variable X only taking non-negative integer values has probability generating function given by

$$\pi_X(s) = E\left(s^X\right) = (4-3s)^{-1}\,.$$

Deduce the k-th derivative $\pi_X^{(k)}(s) = \frac{d^k}{ds^k}\pi_X(s)$ and hence determine the probability distribution of X i.e. write $P(X=x)$ as a function of x.

$\pi_X^{(k)}(s) =$	$P(X=x) =$
$(3^k)(k!)(4-3s)^{-(k+1)}$	$\left(\frac{3}{4}\right)^x \frac{1}{4}$

Please show your working below this line

Taking the first few derivatives we see

$$\begin{aligned}
\pi_X'(s) &= -(4-3s)^{-2}(-3) = 3(4-3s)^{-2} \\
\pi_X''(s) &= 3(-2)(4-3s)^{-3}(-3) = (3^2)2(4-3s)^{-3} \\
\pi_X'''(s) &= 3^2(-3)2(4-3s)^{-4}(-3) = (3^3)3!(4-3s)^{-4} \\
\pi_X^{(4)}(s) &= 3^3(-4)(3!)(4-3s)^{-5}(-3) = (3^4)4!(4-3s)^{-5}
\end{aligned}$$

So it seems we might have

$$\pi_X^{(k)}(s) = (3^k)(k!)(4-3s)^{-(k+1)}\,, \qquad (*)$$

and indeed we can check that this pattern is preserved after one further differentiation:

$$\begin{aligned}
\pi_X^{(k+1)}(s) &= (3^k)[-(k+1)](k!)(4-3s)^{-(k+2)}(-3) \\
&= (3^{k+1})[(k+1)!](4-3s)^{-(k+2)}\,;
\end{aligned}$$

this is equivalent to a proof via induction. Thus (**??**) is indeed the desired derivative.

Finally note that since $\pi_X^{(k)}(0) = k!P(X=k)$, the probability distribution is given by

$$P(X=x) = \frac{\pi_X^{(x)}(0)}{x!} = 3^x 4^{-(x+1)} = \left(\frac{3}{4}\right)^x \frac{1}{4}\,.$$

Formula sheet for MATH1905 Statistics

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

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

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

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

For the least-squares line $y = a + bx$:

b	$\frac{S_{xy}}{S_{xx}}$
a	$\overline{y} - b\overline{x}$

- **Some probability results:**

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

- If $Y \sim \text{Pois}(\lambda)$, $P(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!}$ for $y = 0, 1, 2, \ldots$, $E(Y) = \lambda$ and $Var(Y) = \lambda$.
- If $X \sim B(n, p)$, $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$, for $x = 0, 1, \ldots, n$, $E(X) = np$ and $Var(X) = np(1-p)$.

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

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

$\frac{\overline{X} - \overline{Y}}{S_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \sim t_{n_x+n_y-2}$, where $S_p^2 = \frac{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}{n_x + n_y - 2}$
$\hat{\alpha} \sim N\left(\alpha, \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]\right)$; $\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right)$; $\hat{\sigma}^2 = \frac{\sum_i \hat{\varepsilon}_i^2}{n-2} \sim \frac{\sigma^2\chi^2_{n-2}}{n-2}$
$\sum_i \frac{(O_i - E_i)^2}{E_i} = \sum_i \frac{O_i^2}{E_i} - n \sim \chi^2_\nu$, for appropriate ν