The University of Sydney
School of Mathematics and Statistics

Tutorial Week 13

MATH1905: Statistics (Advanced) Semester 2, 2017

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

This week's exercise set is long due to the quiz in week 12. Please seek extra assistance from either the lecturer or a tutor if you need it.

1. Verify the following computing formula for Pearson's statistic. If $O_1, \ldots, O_g$ and $E_1, \ldots, E_g$ satisfy $\sum_{i=1}^{g} O_i = \sum_{i=1}^{g} E_i = n$, show that

$$\sum_{i=1}^{g} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{g} \frac{O_i^2}{E_i} - n\,.$$

2. One of Mendel's breeding trials had the following results:

Type of pea	Frequency
Smooth yellow	315
Wrinkled yellow	101
Smooth green	108
Wrinkled green	32
Total	556

What frequencies are expected under a model 9 : 3 : 3 : 1? Examine the model for goodness of fit.

3. A pre-election Gallup poll on voting patterns in the next election resulted in the following table. Test if the political party preference is the same in each state.

Political Party	NSW	VIC	QLD	WA	Total
Labor	105	120	105	70	400
Liberal	120	100	130	150	500
Greens	25	30	15	30	100
Total	250	250	250	250	1000

4. Are the events "Income Level" and "Church Attendance" independent?

	Income level			
Church attendance	Low	Middle	High	Total
Never	27	48	15	90
Occasional	29	58	13	100
Regular	24	74	12	110
Total	80	180	40	300

5. In a backcross experiment to investigate the genetic linkage between two factors A and B in a species of flower, some researchers classified 400 offspring by phenotype as follows

AB	Ab	aB	ab
128	86	74	112

(a) Under the "no linkage" model, the four phenotypes are equally likely. Show that this model is a poor fit.

(b) If linkage is in the "coupling phase", the probabilities of the four phenotypes are

AB	Ab	aB	ab
$\frac{1}{2}(1-p)$	$\frac{1}{2}p$	$\frac{1}{2}p$	$\frac{1}{2}(1-p)$

where p is the "recombination fraction" and is estimated by the overall proportion of Ab and aB. Show that this "coupling phase" model fits the data well.

6. A study of grand juries in a county of California compared the demographic characteristics of jurors with the general population, to see if the jury panels were representative. The breakdown by age for the county is known from Public Health Department data, and the breakdown by age for a random sample of 66 jurors (over 21) is also given. Are the data consistent with the theory that these 66 jurors were selected at random from the population (over 21) of the county?

Age	County percentage	Number of jurors
21 to 40	42	5
41 to 50	23	9
51 to 60	16	19
61 and over	19	33

7. 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

Is there any evidence that cancer at the edge of surgery affects the chance of the cancer being controlled?

8. Suppose that the probability of success at each repetition of an experiment is q. Perform a one-sided test of $H_0 : q = 0.6$ against $H_1 : q > 0.6$ given the following data.

(a) There were 8 successes observed in 10 trials.

(b) There were 40 successes observed in 50 trials.

(c) Comment on your answers.

9. The proportion of deaths due to lung cancer in working males aged 15–64 in Australia during the period 1970–1972 was 10%. There is reason to believe that working for an extended period in a chemical plant increases the risk of lung cancer. Accordingly, several chemical plants were investigated, and it was found that of 90 deaths occurring among 15–64 year old male workers who had worked for at least 1 year in the plants, 19 were due to lung cancer. Report a p-value from this study for testing that the death rate is really 10% against the alternative that the rate is greater than 10%. Interpret your result.

10. Suppose the probability of success at each repetition of an experiment is q.

(a) Perform a (two-sided) test of $H_0 : q = 0.5$ vs $H_1 : q \neq 0.5$ if 60 successes are observed in 100 trials

(b) Compare the "exact" p-value to one obtained using a normal approximation with continuity correction.

11. A biology laboratory is updating its weighing equipment and the new weights are calibrated against the old: thirteen molluscs are weighed on both scales to see how the scales compare, and the signs of the differences in weight are

$$-\quad +\quad +\quad +\quad +\quad +\quad -\quad +\quad +\quad 0\quad +\quad -\quad +$$

On this basis would you conclude that there is strong evidence that the scales measure differently on average (more precisely, that positive and negative discrepancies are *not* equally likely)?

12. *From the 1998 Examination* The proportion of defective items produced by a factory is 0.1. As part of the quality control procedure, a random sample of 12 items is inspected daily.

(a) State the distribution of X, the number of defective items in a random sample of size 12.

(b) Find the probability that there are fewer than two defective items on a particular day.

(c) A new manager introduces work practices which are expected to reduce the proportion of defective items produced. After a settling-in period, he asks for a random sample of 200 items to be inspected for defects. Test the effectiveness of the new work practices if it is found that there are only 11 defective items in the sample.

($Hint$: Let q be the proportion of defective items produced after the new work practices are introduced. Set up appropriate null and alternative hypotheses concerning q.)

13. 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 q be the true proportion of all purchasers who contact the hotline.

(a) Set up appropriate hypotheses to perform a statistical test.

(b) Why is a 1-sided test appropriate here?

(c) Calculate an exact p-value based on these data. Interpret your findings.

(d) Show that a normal approximation (with continuity correction) to the p-value is smaller than the exact p-value.

14. The clinically accepted value for mean blood pressure in healthy males aged 18 to 22 years is 120 mm Hg and the accepted standard deviation is 20 mm Hg. Assume that blood pressure for this age group is normally distributed.

(a) What proportion of healthy males of this age have a blood pressure above 145.6 mm Hg?

(b) It is widely believed that examination stress causes blood pressure to rise. To test this theory, 10 healthy male students have their blood pressure taken just prior to a statistics examination. The readings are simply recorded as *High* (above 145.6 mm Hg) or *Normal* (below 145.6 mm Hg). The actual measurements are not available.

(i) Set up the appropriate null and alternative hypotheses which might be tested using the data in this incomplete form. Define any symbols you use.

(ii) Test the above claim if 3 of the students are found to have {*High*} *blood pressure*.

15. *From the 1994 Examination* A random sample of 60 slow learners with a deficiency in reading was used to study a new teaching method, which is expected to be superior to the old method because of results of a pilot study. The subjects were matched by age, sex, background and current reading ability, and one of each pair was chosen at random to be taught by the old method. For the 30 pairs, the *differences*} in reading ability (on a standard reading test) after a period of two months were recorded. Of the 30 differences, 18 were positive, 2 were zero and 10 were negative. Use a statistical test to analyse the data, mentioning any assumptions.

16. In a biology laboratory that is updating its weighing equipment, the weights of twelve specimens using the new equipment are calibrated against the old.

Specimen	Old Scales (mg)	New Scales (mg)
1	6.14	6.13
2	5.90	5.88
3	7.15	7.14
4	8.86	8.87
5	4.99	4.99
6	6.74	6.72
7	7.81	7.78
8	8.15	8.12
9	6.37	6.38
10	8.80	8.78
11	6.26	6.22
12	6.97	6.93

The measurements in the table above can be read into R using:

```
old = c(6.14,5.9,7.15,8.86,4.99,6.74,7.81,8.15,6.37,8.8,6.26,6.97)
new = c(6.13,5.88,7.14,8.87,4.99,6.72,7.78,8.12,6.38,8.78,6.22,6.93)
```

On the basis of the twelve pairs of weights and the sign test would you conclude that there is strong evidence that the scales measure differently on average?

17. A consumer magazine performs a survey of its subscribers concerning two brands of product. Suppose that a random sample of 120 subscribers are surveyed and of these 71 prefer brand A and 49 prefer brand B.

(a) Write down a probability expression (without evaluating it) for the p-value of an (exact binomial) two-sided test of the hypothesis H_0: $q = 0.5$ where q denotes the proportion of the whole population of subscribers that prefer brand A.

(b) Evaluate the p-value expression obtained in part (a) using a normal approximation with continuity correction.

(c) Evaluate the p-value expression obtained in part (a) using a normal approximation *without* continuity correction.

(d) Verify (using R, via "manual" computing as well as using `chisq.test()`) that a (chi-squared) goodness-of-fit test using Pearson's statistic (and a chi-squared approximation for the p-value) based on these observed frequencies is equivalent to the binomial test in part (a) using a normal approximation *without* continuity correction %the third method of calculating the p-value (i.e. normal approximation *without* continuity correction). to approximate the p-value.

(e) Compare these two approximations to the *exact* p-value by computing the exact p-value and hence the *relative error* inherent in the two different normal approximations.

18. A certain trait has two forms, let's call them A and B. A certain theory proposes that the two forms should be present in the population in respective proportions 1:2. A random sample of 50 people from the population turns out to have 23 A's and 27 B's.

(a) Compute Pearson's statistic for comparing these observed frequencies to those expected under the theory.

(b) Using R obtain (a chi-squared approximation to) the p-value.

```
1-pchisq(stat,df=1)
```

```
[1] 0.05743312
```

(c) Write the *exact* p-value for the test based on Pearson's statistic as a probability involving a binomial random variable (**hint**: let X = the number of A's in the sample) and evaluate it using R.

(d) Verify that the *default* result of performing this test using `chisq.test()` gives the same result as that obtained in part (b).

(e) Verify that passing the optional parameter `simulate=TRUE` inside `chisq.test()` gives an approximate p-value much closer to the exact p-value obtained in part (c). You can increase the number of replicates (and hence increase the accuracy of the approximation) using e.g. `B=10000`.

19. (a) Obtain 95% and 99% Wilson's confidence intervals for q based on the scenario and data in question 17 above (make sure the `binom` package is available, either by selecting the "Packages" tab in the bottom right panel of RStudio and selecting the checkbox next to `binom` or just executing the command `library(binom)` at the console). Save the outputs of `binom.wilson()` as objects `w95` and `w99`. **Note:** the `binom` package is already installed on the Maths & Stats network, if you are doing this on your own machine you will need to install it first before performing the above steps. The easiest way is to use the RStudio menu item "Tools"–"Install Packages..." and then enter "`binom`" in the dialog box that pops up – you need to be online for this to work!

(b) The first block of R code below defines a vector `q0` of hypothesised values q_0 and obtains corresponding approximate chi-squared p-values for the test of $q = q_0$ based on Pearson's statistic for the scenario in question 17 and plots the p-values against `q0`.

```
q0=(1:999)/1000
Pearson.stats=((71-120*q0)^2)/(120*q0*(1-q0))
chisq.pvals=1-pchisq(Pearson.stats,df=1)
plot(q0,chisq.pvals,type="l")
```

(c) The second block below reproduces the same plot but zooms in and adds horizontal lines at 0.05 and 0.01.

```
plot(q0,chisq.pvals,type="l",ylim=c(0,0.1),xlim=c(0.4,0.8))
abline(h=0.05,col="red",lty=2)
abline(h=0.01,col="blue",lty=2)
legend("topright", leg=c("0.05","0.01"),col=c("red","blue"),lty=c(2,2))
```

Verify that the Wilson intervals obtained in the previous question correspond to the points of intersection of the horizontal lines and the p-value plot by appending the commands below:

```
abline(v=c(w95$lower,w95$upper),col="red",lty=2)
abline(v=c(w99$lower,w99$upper),col="blue",lty=2)
```

20. Repeat the previous two questions (i.e. compute 95% and 99% Wilson intervals, and produce p-value plots, etc) for the scenario in question 18. For definiteness, let q represent the proportion of form A.