Mathematics 1

Sheet 27: Statistical Inference

Touch here for an explanatory video on this topic.

1. At Trinity College, 58% of the students travel by tram. A group of 100 Trinity College students was selected in a random sample, and 65 traveled by tram. In this situation:

   (a) What is the value of the population proportion, p, of students who travel by tram?

   (b) What is the value of the sample proportion, $\hat{p}$, of students who travel by tram?

2. State whether the following information will give a population proportion or a sample proportion.

   (a) The birth-dates of all the current Trinity College Foundation students were checked, and it was found that 6.8% of the students are 16 years old.

   (b) A survey of 100 primary school teachers showed that 60% were female and 40% were male.

3. A confectionery store has 120 chocolate bars for sale. Of these, 7 are peppermint flavoured. Sally takes a random sample of 8 chocolate bars, and writes down that the possible values for the sample proportion variable $\hat{P}$ of peppermint flavoured chocolate bars in the sample are $0, \frac{1}{8}, \frac{1}{4}, \frac{3}{8}, \frac{1}{2}, \frac{5}{8}, \frac{3}{4}, \frac{7}{8}$ and 1.

   Is Sally correct?

4. A survey about work locations is being conducted in a large town in regional Victoria. Suppose that there are n completed surveys, and suppose we define two random variables X and $\hat{P}$, as follows:

   - X denotes the number of completed surveys in which the person has indicated that he/she travels to Melbourne for work, and
   - $\hat{P} = \dfrac{X}{n}$.

   (a) State the condition under which X is hypergeometric.

   (b) State the conditions under which X is binomial, or can be approximated well by a binomial distribution.

   (c) Under what conditions is the distribution of $\hat{P}$ approximately normal?

5. In a neighbourhood of 120 households, 47 households own dogs. Suppose that a random sample of 4 households is chosen, and that X is defined to be the number of households in the sample that have dogs.

   (a) Calculate the probability distribution of the sample proportion variable, $\hat{P}$, to 4 decimal places.

   (b) Is it sensible to approximate this distribution with a binomial distribution? If so, then calculate a binomial distibution which approximates $\hat{P}$, to 4 decimal places.

6. Consider a town which has a population containing 10 000 adults. Suppose that 30% of the adults in the town have high blood-pressure. That is, 3000 of the adults have high blood-pressure, and the other 7000 adults do not. Now suppose that 25 blood-pressure tests were completed, for adults selected randomly from the town.

   (a) Calculate the probability that exactly 12 of the tests indicated high blood-pressure, if the sample of adults to test was chosen *without* replacement. Write your answer accurate to 4 decimal places.

   (b) Calculate the probability that exactly 12 of the tests indicated high blood-pressure, if the sample of adults to test was chosen *with* replacement. Write your answer accurate to 4 decimal places.

   Notice that the answer to Question 6(a) is similar to the answer to 6(b). This is because, when the population is much larger than the size of the sample (as was the case in Exercise 6), then the hypergeometric and binomial distributions are similar to each other.

7. Suppose 75% of TCFS students study Maths 1. A random sample of 25 TCFS students is chosen. Consider the sample proportion variable $\hat{P}$ of students who study maths.

   (a) What is the expected value of $\hat{P}$?

   (b) What is the standard deviation of $\hat{P}$? Give your answer to 4 decimal places.

8. Five random samples of 10 people were selected from the population of students at a university. The number of students who studied science was examined and the following sample proportions were found:

$$0.3,\ 0.2,\ 0.5,\ 0.3,\ 0.1\,.$$

The process was repeated with five random samples of 100 students being selected, and this time the following sample proportions were found:

$$0.25,\ 0.32,\ 0.27,\ 0.35,\ 0.31\,.$$

(a) Calculate the average and the standard deviation for each set of sample proportions. Give your answers to 2 decimal places.

(b) Suppose that the population proportion, p, of students who study science is 0.31.

Assume that the number of students in a sample who study science has a binomial distribution, and consider the sample proportion variable $\hat{P}$ of students who study science.

i. For samples of 10 people, find:
   - the long term average of $\hat{P}$, and
   - the standard deviation of $\hat{P}$.

   Give your answers to 2 decimal places.

ii. Similarly, for samples of 100 people, find:
   - the long term average of $\hat{P}$, and
   - the standard deviation of $\hat{P}$.

   Give your answers to 2 decimal places.

iii. Compare the values from (b) to the averages and standard deviations calculated in (a).

9. A fitness club conducted a survey of 200 randomly chosen clients who are enrolled for a particular fitness program. It found that 40 of the surveyed clients have not achieved the goal set before the start of the program. Let p be the probability of success of the program. Find, to 4 decimal places, an approximate 95% confidence interval for p.

10. A survey of 1000 university students finds that 38% are excited by the opportunity to take a statistics class. From this, construct an approximate 95% confidence interval on the true proportion of university students who are excited to take a statistics class.
Write your answer to 2 decimal places.

11. A drug company is conducting a test on a drug that improves a person's ability to do mathematics. Out of a sample of 100 participants, 84 found the drug effective.

    (a) Calculate the sample proportion of people that have found the drug effective.

    (b) Calculate an approximate 95% confidence interval for the population proportion of people that find the drug effective.
    Write your answer to 5 decimal places.

(c) What interpretation can be made about the population proportion that find the drug effective?

(d) Discussion (not examinable): Suppose that the drug company will only continue with the trial if it is very certain that the population proportion is greater than 60%. How would you advise the drug company in this situation?

12. Consider a sample containing 200 TCFS students. Suppose that in this sample there are 37 students who are staying in home-stay accommodation. For this sample, find the following confidence intervals for the proportion of TCFS students who are staying in home-stay accommodation:

(a) an approximate 90% confidence interval,

(b) an approximate 95% confidence interval,

(c) an approximate 99% confidence interval,

using 3 decimal places in your answers.

(d) Now suppose that it is known that the *true* proportion of TCFS students staying in home-stay accommodation is 24%. Compare the width of the three confidence intervals found in (a)-(c), and check whether those intervals actually contain this known population proportion p.

13. Suppose that $(0.2835\,,\,0.3165)$ is an approximate 90% confidence interval for an unknown population proportion parameter p.

Using this confidence interval, determine

(a) the sample proportion $\hat{p}$, and

(b) the sample size n.

14. Suppose that $(0.743\,,\,0.857)$ is an approximate 99.56% confidence interval for an unknown population proportion parameter p.

Using this confidence interval, determine

(a) the sample proportion $\hat{p}$, and

(b) the relevant value of z_k

(Hint: See Example 9(b) of Chapter 25),

and

(c) the sample size n.

15. Two opinion polls, seeking information about the proportion of voters supporting a certain political party, were carried out.

    (a) For the first poll, one thousand randomly chosen people were interviewed and 368 of them said they intended to support the party.

        Calculate an approximate 95% confidence interval for the population proportion supporting the party. Give your answer to 3 decimal places.

    (b) The result of the second poll was reported in a newspaper. The report said that support for the party was

        "between 33% and 37%,

        based on a sample of 3786 people".

        If we express this information as an approximate k% confidence interval, calculate the value of k (to the nearest integer).

16. An electricity company is interested in finding an approximate 90% confidence interval for the proportion of households that have solar panels installed. A random sample of 500 households was examined, and 120 households were found to have solar panels installed.

    (a) Find, to 3 decimal places, an approximate 90% confidence interval for the proportion of households that have solar panels installed.

    (b) The electricity company wants a margin of error of at most 0.01 in the approximate 90% confidence interval. Find the smallest sample size, n, required to achieve this.

17. An apple farmer inspects a large random sample of his apples and finds that 20% contain excellent worms. He wants to charge his customers extra, and claims that the 90% confidence interval for the population proportion of apples that contain worms has a margin of error of at most 0.01. At least how many apples did the farmer inspect?

18. A recent survey of n meat-eaters showed that an approximate 90% confidence interval for the proportion that prefer eating beef rather than chicken is $(0.35\,,\,0.45)$. Find the value of n.

    (Round up to the next integer to get the answer.)

19. A survey of Trinity College students found that an approximate 95% confidence interval for the proportion of students who like eating KFC was $(0.29277\,,\ 0.56723)$.

    Unfortunately the incompetent KFC employees who conducted the survey lost the original results.

    (a) Find the sample proportion $\hat{p}$ for this survey.

    (b) Find the sample size n.

20. **Maths 1 Extension (Not Examinable):**

    A point with coordinates (x, y) is randomly chosen, with the restriction that $0 < x < 1$ and $0 < y < 1$.

    (a) What is the probability that the point lies inside the circle whose equation is
    $$x^2 + y^2 = 1\,?$$

    (b) In a computer simulation 1000 such points were generated and 784 of them lay inside the circle. Use this information to obtain an estimate for π, and to calculate an approximate 90% confidence interval for π.
    Give your answers to 3 decimal places.

    (c) Show that 295 065 points need to be selected so that an approximate 90% confidence interval for π will have a margin of error less than 0.005.

Answers:

1. (a) $p = 0.58$ (b) $\hat{p} = 0.65$

2. (a) p, population proportion (b) $\hat{p}$, sample proportion

3. No. In this example, we cannot have $\hat{P} = 1$ since there were only 7 peppermint flavoured chocolate bars in total.

4. (a) X is hypergeometric if the selection of people to do the survey was done *without* replacement.

   (b) X is binomial if the selection of people to do the survey was done *with* replacement. However, if the selection of people to do the survey was done *without* replacement (and so X is hypergeometric), then X is approximated well by a binomial distribution if we have
   $$\text{the sample size } n < \frac{\text{population size}}{10}\,.$$

   (c) Let $\hat{p}$ denote the sample proportion of people who travel to Melbourne to work. Then the sample proportion variable $\hat{P}$ is approximately normal if $n\hat{p} > 15$ and $n(1-\hat{p}) > 15$.

5. (a)

x	$\hat{p}$	$\Pr(X = x) = \Pr(\hat{P} = \hat{p})$
0	0	$\dfrac{{}^{47}\mathrm{C}_0 \times {}^{73}\mathrm{C}_4}{{}^{120}\mathrm{C}_4} = 0.1325$ (4 d.p.)
1	$\frac{1}{4}$	$\dfrac{{}^{47}\mathrm{C}_1 \times {}^{73}\mathrm{C}_3}{{}^{120}\mathrm{C}_4} = 0.3559$ (4 d.p.)
2	$\frac{2}{4}$	$\dfrac{{}^{47}\mathrm{C}_2 \times {}^{73}\mathrm{C}_2}{{}^{120}\mathrm{C}_4} = 0.3458$ (4 d.p.)
3	$\frac{3}{4}$	$\dfrac{{}^{47}\mathrm{C}_3 \times {}^{73}\mathrm{C}_1}{{}^{120}\mathrm{C}_4} = 0.1441$ (4 d.p.)
4	1	$\dfrac{{}^{47}\mathrm{C}_4 \times {}^{73}\mathrm{C}_0}{{}^{120}\mathrm{C}_4} = 0.0217$ (4 d.p.)

(b) We have $n = 4$ and the population-size is 120.

Thus we see that $n < \dfrac{\text{population size}}{10}$, and so we can approximate the distribution of $\hat{P}$ well with a binomial distribution. Let X^* be a binomial random variable with $n = 4$ and $p = \dfrac{47}{120}$, and let $\hat{P}^* = \dfrac{X^*}{4}$.

Then the probability distribution of $\hat{P}^*$ is as follows:

x	$\hat{p}$	$\Pr(X^* = x) = \Pr(\hat{P}^* = \hat{p})$
0	0	${}^4C_0 \left(\frac{47}{120}\right)^0 \left(\frac{73}{120}\right)^4 = 0.1370$ (4 d.p.)
1	$\frac{1}{4}$	${}^4C_1 \left(\frac{47}{120}\right)^1 \left(\frac{73}{120}\right)^3 = 0.3527$ (4 d.p.)
2	$\frac{2}{4}$	${}^4C_2 \left(\frac{47}{120}\right)^2 \left(\frac{73}{120}\right)^2 = 0.3406$ (4 d.p.)
3	$\frac{3}{4}$	${}^4C_3 \left(\frac{47}{120}\right)^3 \left(\frac{73}{120}\right)^1 = 0.1462$ (4 d.p.)
4	1	${}^4C_4 \left(\frac{47}{120}\right)^4 \left(\frac{73}{120}\right)^0 = 0.0235$ (4 d.p.)

6. (a) 0.0267 (4 d.p.)

   (b) 0.0268 (4 d.p.)

7. (a) $E(\hat{P}) = p = 0.75$

   (b) $\sigma = \sqrt{\dfrac{p(1-p)}{n}} = \sqrt{\dfrac{0.75 \times 0.25}{25}} = 0.0866$ (4 d.p.)

8. (a) Using the statistical functions on our calculators (as we learnt in Chapter 16) we obtain the following results:

   5 samples of 10: $\bar{x} = 0.28, \; s = 0.15$ (2 d.p.)

   5 samples of 100: $\bar{x} = 0.30, \; s = 0.04$

   (b) i. $E(\hat{P}) = 0.31, \; \sigma = 0.15$ (2 d.p.)

   ii. $E(\hat{P}) = 0.31, \; \sigma = 0.05$ (2 d.p.)

   iii. The mean found in (a) of the data with the larger sample size was closer to the long-term average (found in (b)) than the mean of the data with the smaller sample-size.

   Also, the standard deviation found in (a) of the data with the larger sample-size was smaller than the standard deviation of the data with the smaller sample-size.

   This was also the case for the standard deviation found in (b)(ii) (larger sample-size) compared to the standard deviation found in (b)(i) (smaller sample-size), and illustrates that *a larger sample-size tends to lead to sample proportions which are clustered closer to the mean.*

9. First note that since $n = 200$ and $\hat{p} = \dfrac{160}{200}$, then we have

$$n\hat{p} > 15 \text{ and } n(1-\hat{p}) > 15.$$

Thus we can write down an approximate 95% confidence interval of the form

$$\left(\hat{p} - z_k\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \, , \, \hat{p} + z_k\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

using $z_k = 1.96$ (given on the table on the Formula Sheet, but originally obtained from the normal tables).

This gives an approximate 95% confidence interval of

$$(0.7446\, , \, 0.8554) \text{ (to 4 d.p.).}$$

10. $(0.35\,,\,0.41)$ (2 d.p.)

11. (a) The sample proportion is $\hat{p} = \dfrac{84}{100} = 0.84$.

    (b) An approximate 95% confidence interval for the population proportion p is $(0.76\,815\,,\,0.91\,185)$ (5 d.p.).

    (c) If many samples were considered and used to calculate *many* 95% confidence intervals in this same way, then approximately 95% of those confidence intervals will include the population proportion.

    (d) The company could consider 99% confidence intervals (or higher percentage confidence intervals, if they prefer, such as 99.9% confidence intervals). However, those intervals might be unhelpfully wide. The company could decrease the width of the intervals by increasing the sample-size. Then, by looking at many such confidence intervals, the company can observe whether (or not) the vast majority of the intervals only contain values which are > 0.6 . Just considering a single confidence interval would be very risky, as that particular sample might not be 'typical' (and so could give a misleading impression).

12. (a) $(0.140\,,\,0.230)$ (to 3 d.p.)

    (b) $(0.131\,,\,0.239)$ (to 3 d.p.)

    (c) $(0.114\,,\,0.256)$ (to 3 d.p.)

    (d) The higher the percentage is, the wider the confidence interval is. The population proportion, $p = 0.24$, is only in the 99% confidence interval.

13. (a) The sample proportion is the midpoint of the confidence interval, and so can be computed as the average of the endpoints of the interval. Thus

$$\hat{p} \;=\; \frac{0.2835 + 0.3165}{2} \;=\; 0.3\,.$$

    (b) The sample size is $n = 2100$.

14. (a) The sample proportion is the midpoint of the confidence interval, and so can be computed as the average of the endpoints of the interval. Thus

$$\hat{p} \;=\; \frac{0.743 + 0.857}{2} \;=\; 0.8\,.$$

    (b) Using the normal tables in reverse, we find that $z_k = 2.85$.

    (c) The sample size is $n = 400$.

15. (a) $(0.338, 0.398)$ (to 3 d.p.) .

    (b) $k = 99$ (since $z_k = 2.58$ which corresponds to a 99% confidence interval).

16. (a) $(0.208\,,\,0.272)$ (to 3 d.p.)

    (b) The smallest value of n is 4966.

17. 4356

18. $n = 262$

19. (a) $\hat{p} = 0.43$

    (b) $n = 50$

20. (a) The probability that the point lies within the circle is $\frac{\pi}{4}$.

    (b) The estimate for π obtained is 3.136, and an approximate 90% confidence interval for π is

$$(3.050, 3.222) \;\; \text{(to 3 d.p.)}.$$

    (c) If the margin of error for the interval for π is < 0.005 then the margin of error for the interval for $\frac{\pi}{4}$ needs to be $< \frac{0.005}{4}$. Therefore we solve the following inequality for n :

$$1.65 \times \sqrt{\frac{0.784 \times 0.216}{n}} \;<\; \frac{0.005}{4}\,.$$