The University of Sydney
School of Mathematics and Statistics

Solutions to Practice Quiz

STAT5002: Introduction to Statistics Semester 1, 2025

Lecturer: Tiangang Cui

Information about the sample quiz

- The actual quiz will be 50 minutes long and will consist of 15 questions. You can expect to spend roughly 10 minutes for every three questions.
- This practice sheet contains 12 questions in total, and you are expected to complete it in 40 minutes.
- The purpose of the practice sheet is to give you an idea of the format, types of questions, and the involvement of R coding in the mid-term test.
- For a comprehensive review of the mid-term quiz, you should also go over the weekly online quizzes and do the independent practice sheets.
- In the mid-term test (quiz), all questions will be **multiple-choice questions** with four answer options and exactly one correct answer.
- Answers to the questions below will be posted at the end of week 7.

Questions

1. In a dataset of size 8, the mean is 7 and standard deviation is 4. We add 4 to each observation in the dataset. The new mean and SD are respectively

(a) 7 and 4
(b) 11 and 8
(c) 7 and 8
(d) 11 and 4

Solution: Adding 4 to each observation leads to a new data set with $y_i = x_i + 4$. Then, the mean of $(y_1, \ldots, y_8)$ is

$$\bar{y} = \frac{1}{8}\sum_{i=1}^{8}(x_i + 4) = 4 + \frac{1}{8}\sum_{i=1}^{8} x_i = 4 + \bar{x},$$

and the standard deviation of $(y_1, \ldots, y_8)$ is

$$SD_y = \sqrt{\frac{1}{8-1}\sum_{i=1}^{8}(y_i - \bar{y})^2} = \sqrt{\frac{1}{8-1}\sum_{i=1}^{8}(x_i - \bar{x})^2} = SD_x.$$

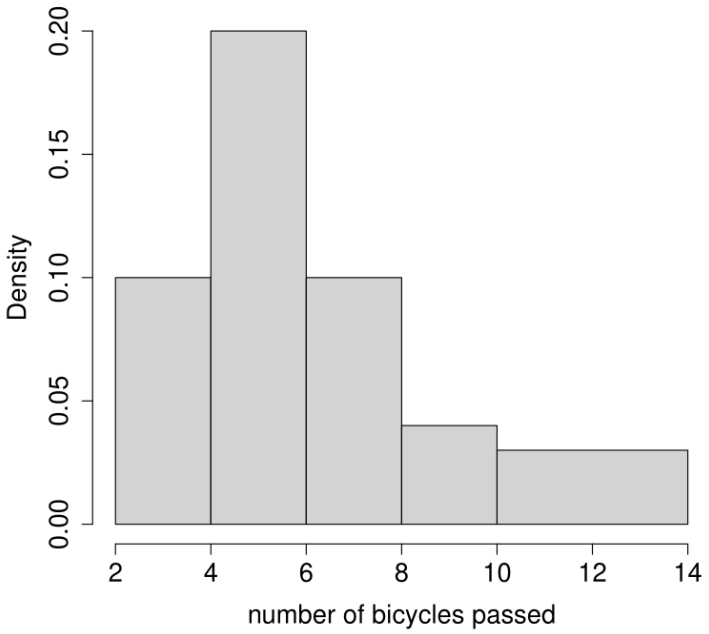
So the correct anwser is (d) — (11 and 4).

2. Which of the following statements is definitely correct?

(a) If A and B are mutually exclusive events, then A and B are independent.
(b) If we take a sample of points of size 10 from a scatterplot with correlation coefficient 1, then the sampled points lie on a straight line.
(c) The slope of a regression line relating height and age among boys is 1.3 times higher than the slope of a regression line relating the height and age among girls. Therefore, the correlation coefficient between height and age among boys is 1.3 times higher than correlation coefficient between height and age among girls.
(d) All the other listed options are incorrect.

Solution: (b) is correct.

3. Each day for a 25-day period, Jimmy counted the number of bicycles he passed on his drive to work. A histogram of his counts is plotted below. How many days did he pass less than 6 bicycles? (Assume each bin is of the left-closed and right-open form $[a, b)$.)



(a) 8 (b) 10
(c) 12 (d) 15

Solution: The area of the bins less than 6 is $0.1 \times 2 + 0.2 \times 2 = 0.6$. Then the proportion 0.6 times the total number of days 25 gives 15. So (d) – (15) is the correct answer.

4. A study conducted at the University of Sydney shows that the average height of the female staff members is 165.52cm with a SD of 5.59cm. Consider this the full population of the female staff members of the University of Sydney and that the height can be described by a normal model. Which one is the correct R code for calculating the percentage of female staff members with height between 152.40cm and 167.64cm?

(a) qnorm$(167.64, 165.52, 5.59)$-qnorm$(152.4, 165.52, 5.59)$
(b) pnorm$(167.64, 165.52, 5.59^2)$-pnorm$(152.4, 165.52, 5.59^2)$
(c) qnorm$(167.64, 165.52, 5.59^2)$-qnorm$(152.4, 165.52, 5.59^2)$
(d) pnorm$(167.64, 165.52, 5.59)$-pnorm$(152.4, 165.52, 5.59)$

Solution: (d) is the correct answer.

5. The weight of the box is normally distributed with a mean of 5 kilograms and a SD of 2 kilograms. The box is going to be shipped to a country where weights are commonly reported in pounds. To convert from kilograms to pounds, one must use the approximate formula lb $= 2 \times$ kg, where kg is the weight in kilograms and lb is the weight in pounds. What is the probability that the weight of the box in pounds is greater than 8 pounds, which is the weight limit for packages being shipped to that country? You may find the following R output useful:

```
> round(pnorm(0.5), 2)
[1] 0.69
> round(pnorm(1), 2)
[1] 0.84
> round(pnorm(1.5), 2)
[1] 0.93
> round(pnorm(2), 2)
[1] 0.98
```

(a)	0.69	(b)	0.31
(c)	0.07	(d)	0.16

Solution: The weight 8 lb is $8/2 \approx 4$ kg, which is about $\frac{4-5}{2} \approx -0.5$ standard unit. We want to work out the area under the standard normal curve $Z \sim N(0, 1)$ for $Z > -0.5$. Using the symmetry, this is the same as the area for $Z < 0.5$, which is given by pnorm(1) ≈ 0.69. So (a) is the correct answer.

6. Which of the following statements is definitely correct?

(a) A study conducted at the University of Sydney collected 352 height measurements of staff members. If the histogram of reported heights is skewed to the left, then the median is smaller than the mean.

(b) If two lists of numbers have the same mean and same standard deviation, then they must have the same first quartile.

(c) Let A and B be the 90th percentiles of two different normal curves. Both curves have mean 0, and the standard deviations are 1 and 5, respectively. Then A is smaller than B.

(d) The median is more sensitive to extreme values than the mean.

Solution: (c) is the correct answer.

(a): incorrect as the median should be larger than the mean for left skewed data.
(b): the first quartile is not related to the mean (nor the median).
(c): this is can be shown by draw two normal curves, the 90th percentile of the one with larger spread will be above that of smaller spread.
(d): mean is more sensitive to outliers.

7. Which of the following statements is definitely correct?

(a) In a data set, we observe that its interquartile range is greater than its sample standard deviation, so it is reasonable to assume that the data should be normally distributed.

(b) All the other options are incorrect.

(c) If X and Y are mutually exclusive events, then we must have $P(X|Y) = P(X)$.

(d) In a data set, we observe that its sample median is greater than its sample mean, so the data is skewed to the left.

Solution: (d) is the correct answer.

(a): it is inconclusive.
(c): If X and Y are mutually exclusive, then $P(X|Y) = 0$.

8. A random sample with replacement of size $n = 16$ is to be taken from a box containing $N = 49$ tickets, each bearing a number. The list of numbers has SD $\sigma = 3$. The standard error of the **average** of the numbers drawn is

(a) 12

(b) 21

(c) $\frac{3}{7}$

(d) $\frac{3}{4}$

Solution: (d) is the correct answer.

9. We are given a fair die, each of whose six faces (numbered $1, 2, 3, 4, 5, 6$) has an equal chance of landing facing up. Roll the die three times. Which of the following statements is definitely correct?

(a) The probability that any of these three rolls results in an outcome greater than 2 is more than 0.9.

(b) The probability that any of these three rolls results in an outcome greater than 4 is more than 0.9.

(c) The probability that any of these three rolls results in an outcome greater than 2 is less than 0.5.

(d) All the other listed options are correct.

Solution: (a) is the correct answer. With 1/3 chance, we have a single outcome that is less than or equal to 2. so $1 - (1/3)^3 = 0.963$ is the probability that any of these three rolls results in an outcome greater than 2.

10. Select the correct statement from below.

(a) In a data set, we observe that its interquartile range is greater than its sample standard deviation, so it is reasonable to assume that the data should be normally distributed.

(b) All other listed options are incorrect.

(c) If X and Y are mutually exclusive events, then we must have $P(X|Y) = P(X)$.

(d) Given a data set, we observe that its sample median is greater than its sample mean, so the data is skewed to the left.

Solution: (d) is the correct answer.

11. A random sample of size 900 is taken from a large population with mean $\mu = 5$ and standard deviation $\sigma = 3$. Consider the R output below.

```
 > pnorm(1.111)
[1] 0.8667158
```

The probability that the sample sum is between 4400 and 4500 is closest to

(a) 0.87

(b) 0.13

(c) 0.37

(d) 0.26

Solution: (c) is the correct answer. The mean and the standard error of the sample sum is $n\mu = 4500$ and $\sigma\sqrt{n} = 90$, respectively. The sample sum between 4400 and 4500 is equivalently bounded by -1.111 and 0 under the standard normal curve. So the area between -1.111 and 0 is $pnorm(0) - pnorm(-1.111) = 0.5 - pnorm(-1.111)$.

We have pnorm(-1.111) = 1 - pnorm(1.111) = 0.13, so the area between -1.111 and 0 is 0.37.

12. A sample of size 900 from a population with known variance $\sigma^2 = 9$ produces a sample mean of 8. Construct a 90% confidence interval for the population mean μ. You may use the outputs of the following R code for this question.

```
> round(qnorm(0.025), 2)
[1] -1.96
> round(qnorm(0.05), 2)
[1] -1.64
> round(qnorm(0.95), 2)
[1] 1.64
> round(qnorm(0.975), 2)
[1] 1.96
```

(a) $[7.804, 8.196]$

(b) We don't have sufficient information to find the interval, as the population mean is unknown.

(c) $[6.36, 9.64]$

(d) $[7.836, 8.164]$

Solution: (d) is the correct answer. The standard error of the sample mean is $\frac{\sigma}{\sqrt{n}} = 0.1$. For 90% confidence interval, we choose a multiplier 1.64, corresponding to 5% area below and above the multiplier, which leaves 90% in the middle. This gives the confidence interval $\bar{x} \pm 1.64 \times 0.1$.