### THE UNIVERSITY OF SYDNEY SCHOOL OF MATHEMATICS AND STATISTICS

## MATH1015 BIOSTATISTICS

LECTURERS:

Time Allowed: Reading - 10 minutes; Writing - 1.5 hours	
EXAM CONDITIONS: This is a closed-book examination — no material permitted is not permitted at all during reading time.	. Writing
Family Name:	
Other Names: Seat Number:	
Please check that your examination paper is complete (20 pages) and indicate by signing I have checked Strangillation of the paper and fifting Complet and Help	g below.
https://powcoder.com	
Add WeChat powcoder	Marker's usi
This examination has two sections: Multiple Choice and Extended Answer.	

The Multiple Choice Section is worth 50% of the total examination.

There are 20 questions. The questions are of equal value.

All questions may be attempted.

Answers to the Multiple Choice questions must be entered on the Multiple Choice Answer Sheet before the end of the examination.

The Extended Answer Section is worth 50% of the total examination.

There are 3 questions. The questions are of equal value.

All questions may be attempted. Working must be shown.

THE QUESTION PAPER MUST NOT BE REMOVED FROM THE

**EXAMINATION ROOM.** 

ONLY						

#### Multiple Choice Section

In each question, choose at most one option.

Your answers must be entered on the Multiple Choice Answer Sheet.

- 1. In a medical trial, the placebo effect occurs when
  - (a) the subject is scared of the trial
  - (b) the subject responds to the idea of the treatment
  - (c) the subject is already sick
  - (d) the subject knows the investigator
  - there is a historical treatment group
- 2. In a study on smoking and lung cancer, a possible confounder is NOT
  - a gene for smoking
- (b) brand of cigarettes
- (c) diet

- (d)alcohol
- (e) exercise

### t Project Exam Help

- 3. Simpson's Paradox occurs when
  - relationships between percentages in subgroups are reversed when the subgroups are ttps://powcoder.com
  - relationships between percentages in subgroups are the same when the subgroups (b) are combined.
  - (c) relationships Active percentages in a to prove the character when the subgroups are separated
  - relationships between percentages in subgroups disappear when the subgroups are separated.
  - a clear trend in individual groups of data is revealed when the groups are pooled together.
- 4. Suppose a set of bivariate data has a correlation coefficient of 0.90. Which statement is true?
  - (a) 90% of the points are highly correlated.
  - 90% of the points fall on a line.
  - The linear regression line has a slope of 0.9.
  - 90\% of the points can be predicted by a linear regression line.
  - The data may have a strong linear trend.
- 5. In a dataset of size 6, the mean is 7 and standard deviation is 4. We add 3 to each observation in the data. The new mean and standard deviation are respectively
  - (a) 7 and 4
- (b) 10 and 4 (c) 10 and 7 (d) 7 and 7
- (e) 10 and 13

- 6. A box has a mean 7 and standard deviation 2. Which distribution can best approximate the distribution for the mean of 100 draws?
  - (a) Normal with mean 7 and standard deviation 2.
  - Normal with mean 7 and standard deviation 0.2.
  - Normal with mean 700 and standard deviation 20.
  - T with degrees of freedom 99, mean 7 and standard deviation 2.
  - T with degrees of freedom 100, mean 700 and standard deviation 20.
- 7. For a quantitative data set, the mean and median are the same. Which statement is true?
  - (a) The histogram is skewed
- (b) The scatterplot shows a linear trend
- The barplot is balanced
- The boxplot is symmetric
- not enough information
- 8. For a quantitative data set, the interquartile range (iqr) is 1. Which statement is true?

The data is skewed.

- (a)
- The standard deviation is 1.
- (d) The boxplot has length 1.
- The box in the trops the powcoder.com
- **9.** If  $Z \sim N(0,1)$  then  $P(-3 \le Z \le 3)$  is closest to
  - 0.25

Add WeChat pow.goder 0.997

**10.** Which is the output for the following R command?

pnorm(0)

- (a) 0
- (b) 0.5
- (c) 0.7
- (d) 0.9
- (e) 1
- 11. Using the linear regression line, what expression would predict y when x=2.

```
##
## Call:
## lm(formula = y ~ x)
## Coefficients:
##
   (Intercept)
                            X
        1.8403
                       0.8655
##
```

- $1.8403 0.8655 \times 2$
- $1.8403 + 0.8655 \times 2$ (b)
- (c)  $0.8655 + 1.8403 \times 2$

- (d)  $0.8655 - 1.8403 \times 2$
- (e) Not enough information

12. A box contains the numbers 0,2,3,4,6 and 25 draws are made with replacement. The expected sum of draws and the standard error are respectively

```
library(multicon)
box=c(0,2,3,4,6)
mean(box)
## [1] 3
popsd(box)
## [1] 2
```

- (a) 3/25 and  $2/\sqrt{25}$  (b)  $3 \times \sqrt{25}$  and  $2/\sqrt{25}$  (c)  $3 \times \sqrt{25}$  and  $2 \times \sqrt{25}$
- (d)  $3 \times 25$  and  $2 \times \sqrt{25}$  (e)  $3 \times 25$  and  $2 \times 25$
- 13. A box contains the numbers 0,2,3,4,6 and 100 draws are made with replacement. Which R codes calculate the probability that the mean of draws lies within 2.9 to 3.1?

  - (e) pt((3.1-3)/(2/100),100)-pt((2.9-3)/(2/100),100)

- Add WeChat powcoder

  14. A box contains the numbers 0,2,3,4,6, each repeated 200 times and 20 draws are made without replacement, which of the following statements about standard error for the sum of draws is TRUE?
  - (a) It becomes one fourth if the sample size is half.
  - (b) It drops to 0 if the sample size is 200.
  - It remains unchanged compared to the with replacement case.
  - It enlarges by a factor of  $\frac{1000-20}{1000-1}$  compared to the with replacement case.
  - (e) It shrinks by a factor of  $\sqrt{\frac{1000-20}{1000-1}}$  compared to the with replacement case.
- 15. A box contains nine "0" and one "1" and 16 draws are made with replacement. Which distribution will best approximate the distribution for the average of draws?

```
box=c(0,0,0,0,0,0,0,0,0,1)
library(multicon)
mean(box)
## [1] 0.1
```

```
popsd(box)
## [1] 0.3
```

- (a) Normal with mean 0.025 and standard deviation 0.3/4.
- (b) Normal with mean 0.1 and standard deviation 0.3/16.
- (c) T with degrees of freedom 15, mean 1.6 and standard deviation 0.3\*4.
- Unknown skew distribution with mean 0.1 and standard deviation 0.3/4. (d)
- Unknown skew distribution with mean 0.4 and standard 0.3\*4
- 16. A box contains nine "0" and one "1" and 16 draws are made with replacement. Using the simulation result below, what is the probability that the *sum* of draws is at least 6?

```
box=c(0,0,0,0,0,0,0,0,0,1)
sim=replicate(10000,sum(sample(box,16,rep=T)))
        signment Project Exam Help
## sim
## 1840 3382 hetps://spowcoder.com
```

- Add WeChat powcoder (e) 0.99980.0002
- 17. A marketing company is surveying consumers' preference for Coke over Pepsi. If p =P(Customer prefers Coke to Pepsi), the null and alternative hypotheses are respectively

  - (a)  $H_0: p = 0.5$  and  $H_1: p \neq 0.5$  (b)  $H_0: p = 0.5$  and  $H_1: p > 0.5$

  - (c)  $H_0: p \neq 0.5$  and  $H_1: p = 0.5$  (d)  $H_0: p = 0.5$  and  $H_1: p < 0.5$
  - (e)  $H_0: p < 0.5$  and  $H_1: p > 0.5$
- 18. A marketing company is surveying consumers' preference for Coke over Pepsi. Which box could NOT model the null hypothesis?
  - (a) 0.1
- (b) 0,1,1
- (c) 0,0,1,1
- (d) 0.0001.111 (e) 0.0000.111.111
- 19. A marketing company is surveying consumers' preference for Coke over Pepsi. Out of 100 consumers surveyed, 60 prefer Coke to Pepsi. Which formula gives the test statistic?

```
box = c(0,1)
mean(box)
## [1] 0.5
```

- (a)  $\frac{60-50}{10}$  (b)  $\frac{50-0.6}{10}$  (c)  $\frac{0.6-0.5}{0.05}$  (d)  $\frac{0.6-0.5}{0.5}$  (e)  $\frac{0.5-0.6}{5}$
- **20.** A test was conducted to test the hypotheses  $H_0: \mu_A = \mu_B$  vs  $H_1: \mu_A \neq \mu_B$  where  $\mu_A$  and  $\mu_B$  represent the population mean for group A and B respectively.

### Based on the relations; whoo we code it a comms FALSE?

- (a) The test statistic is -4.9994.
- (b) The p-value Act of We Chat powcoder
- (c) Equality of variance assumption is made.
- (d) The data are against  $H_0$ .
- (e) The 95% confidence interval of  $\mu_A \mu_B$  excludes 0.

#### **Extended Answer Section**

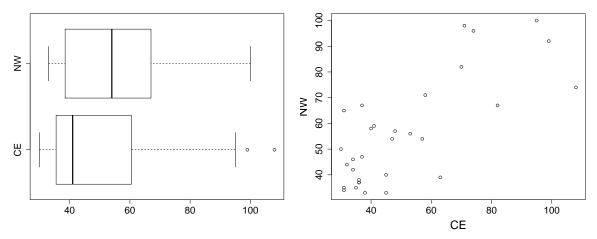
Answer these questions in the spaces provided.

1. We are interested in monitoring the air quality index (AQI) in the month of July 2015 between two regions: Sydney's central-east (CE) and Sydney's north-west (NW). Due to the fact that data readings from different pollutants have different underlying units of measure, the AQI is a derived value based on multiple data readings that enables easier comparison across regions and time. In general an AQI score above 100 and below 150 indicates a 'poor' air quality level and that people in the sensitive group (e.g. people with asthma, older adults and children) should consider either cutting back or rescheduling strenuous outdoor activities. The general public are usually not affected by the air quality within this AQI range.

Source: http://www.environment.nsw.gov.au/AQMS/search.htm

```
head(data)
    CEAQI NWAQI
        ssignment Project Exam Help
[1,] 9/8
[2,] 32
[3,] 70
       82
[4,] 74
           https://powcoder.com
[5,] 95 100
[29,] 41 59
           Add WeChat powcoder
[30,] 48 57
CE=data$CEAQI
NW=data$NWAQI
summary(CE)
    Min. 1st Qu. Median
##
                         Mean 3rd Qu.
                                       Max.
##
   30.00
          35.50
                 41.00
                        50.77
                               60.50
                                      108.00
summary(NW)
##
    Min. 1st Qu.
                Median
                         Mean 3rd Qu.
                                       Max.
          38.50
                 54.00
                        56.52
##
   33.00
                               67.00
                                     100.00
n = nrow(data)
table(CE >= 100)
        TRUE
## FALSE
##
    30
          1
```

table(NW >= 100)## FALSE TRUE 30 1



Look over the R output and then answer the following questions.

Assignment Project Exam Help
Begin by examining the air quality index (AQI) of Sydney's central-east (CE) region.

- - What is the mean AQI score for Sydrey's CE region during July? https://powcoder.com

## Add WeChat powcoder

(ii) Comment on the shape of the boxplot for Sydney's CE region during July.

- (iii)Would it be better to report the mean or median as the measure of centre for Sydney's CE region? Explain.
- How many days in July was the air quality considered 'poor' in Sydney's CE region?

- (b) The scatter plot graphically shows the air quality index (AQI) relationship between the two Sydney regions in the month of July.
  - (i) From the scatter plot, suggest a value for the correlation coefficient between the two Sydney regions. Explain what it represents.

(ii) Suppose on the 28th of July, the air quality monitoring instrument was not working in Sydney's NW region, but the AQI value in Sydney's CE region was recorded as 40. Using the R output below, give an expression without evaluation to estimate the AQI value in Sydney's NW region.

A round (L\$coeff.3) Project Exam Help

## (Intercept)
## 19.887 0.714

https://powcoder.com

### Add WeChat powcoder

(iii) Assuming a normal distribution, if the percentile of AQI in CE is 50% (median) for a certain day, what would you expect the percentile of AQI in NW?

For another day, the percentile AQI in CE is 90%, which expression, a, b or c, will give the expected AQI in NW? Explain briefly.

```
a=qnorm(0.9)*cor(CE,NW)
b=pnorm(qnorm(0.9)*cor(CE,NW))
c=mean(NW)+qnorm(0.9)*cor(CE,NW)*sd(NW)
```

- 2. In a presidential election, there are three candidates and 100,000 eligible voters. It is believed 80% of them will vote and among those who vote, 40% will favor candidate A and 30% will favor candidate B. Before the election, an opinion survey is conducted to evaluate the percentage of support for candidate A. 1000 voters (who will vote) are selected and 42% of them support candidate A.
  - (a) In the following R codes which define the box model, how many 1 and 0 should be included to represent the selection of voters (who will vote) in this opinion survey?

```
box=c(rep(1,??),rep(0,??))
mean(box)

## [1] 0.4

library(multicon)
popsd(box)

## [1] 0.4898979
```

## Assignment Project Exam Help

### https://powcoder.com

(b) From the following R Wpt, Central the Part Central the proportion of support of candidate A and explain the calculation.

```
N=length(box)
n=1000
popsd(box)/sqrt(n)

## [1] 0.01549193

popsd(box)/sqrt(n)*sqrt((N-n)/(N-1))

## [1] 0.0153949
```

(c) A simulation experiment is conducted with 10000 replicates to check the standard error calculation for the estimate of the proportion of support for candidate A in the opinion poll. State the first and second missing values.

```
set.seed(12345)
means = replicate(??, mean(sample(box, ??, rep = F)))
sd(means)
## [1] 0.01537127
```

(d) A betting company offers the following game:
every \$1 bet on candidate A will return \$1.5 (including the \$1 bet) if A wins,
every \$2 to 1 including the \$1 bet) if A wins and
every \$1 bet on candidate C will return \$1.2 (including the \$1 bet) if C wins

What is the expected gain or loss if a person bets \$10 on candidate A? https://powcoder.com

### Add WeChat powcoder

(e) To estimate the difference between the proportions of support for candidate A and B, what tickets should be put in the box?

3. To determine if a new type of fertilizer is different from the type currently in use, researcher Adam took 12 two-acre plots of land scattered throughout a farm. Each plot was divided into two equal-size subplots, one of which was treated with the current fertilizer and the other of which was treated with the new fertilizer. Wheat was planted, and the crop yields were measured. He entered the data in R as below.

```
current=c(53,32,68,72,61,69,42,39,60,72,75,66)
new=c(60,49,66,73,59,67,61,60,58,75,72,68)
diff=new-current
diff
## [1] 7 17 -2 1 -2 -2 19 21 -2 3 -3 2
```

(a) Is this study a control experiment or observational study? Explain briefly.

## Assignment Project Exam Help

(b) Adam wants to test if the two fertilizers have the same effects on the yield. Explain why he choose the predsame was considered and the choose the constant of the choose the choose the constant of the choose the choo

### Add WeChat powcoder

(c) State the null and alternative hypotheses of the test.

(d) He uses R to perform the following preliminary studies.

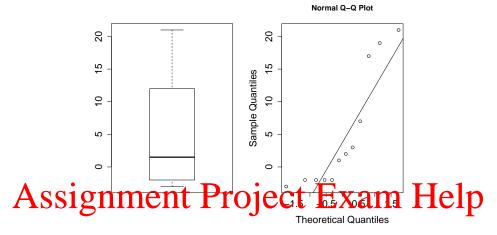
```
shapiro.test(diff)

##

## Shapiro-Wilk normality test
##
```

```
## data: diff
## W = 0.79007, p-value = 0.00724

par(mfrow=c(1,2))
boxplot(diff)
qqnorm(diff)
qqline(diff)
```



Interpret these plots and test result.

https://powcoder.com

# Add WeChat powcoder

(e) Adam further performs the following test. Report test statistic and P-value. Then draw conclusion of the test.

```
t.test(diff)

##

## One Sample t-test

##

## data: diff

## t = 1.8946, df = 11, p-value = 0.08471

## alternative hypothesis: true mean is not equal to 0

## 95 percent confidence interval:

## -0.7949713 10.6283047

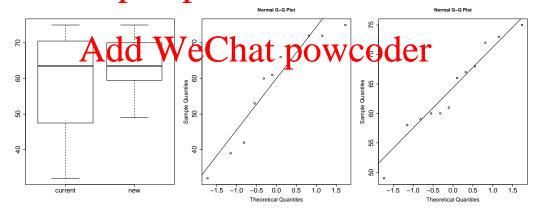
## sample estimates:

## mean of x

## 4.916667
```

(f) His colleague Alex argues that the 12 plots of land are essentially the same. Interpret the plots and test result below and explain which t-test he should choose.

```
var.test(current,new)
##
   F test to compare two variances
##
## data: current and new
## F = 3.6703, num df = 11, denom df = 11, p-value = 0.0412
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
    1.056592 12.749462
## sample estimates:
## ratio of variances
            3.670284
par(mfrow=c(1,3))
boxplot(current, new)
Assignment Project Exam Help
qqline(current)
qqnorm(new)
qqline(new
           ttps://powcoder.com
```



(g) Report the test result from Alex. Is the result consistent with the one from Adam?

```
t.test(current,new,var.equal=F)

##

## ??? Two Sample t-test

##

## data: current and new

## t = -1.0481, df = 16.58, p-value = 0.3096

## alternative hypothesis: true difference in means is not equal to 0

## 95 percent confidence interval:

## -14.833280    4.999947

## sample estimates:

## mean of x mean of y

## 59.08333   64.00000
```

### Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

#### Note

This paper is not a past paper, rather some examples of the style of questions you can expect.

#### (1) Multiple choice question

Assess specific learning outcomes across the whole course. You will have 20 questions.

#### (2) Extended answer questions

Require you to explain your statistical thinking and interpret R Output in the context of a real data story. This assesses your ability to read simple R Code and apply your statistical thinking in a specific data story context. You will have 3 questions.

#### Answer with explanation:

### Assignment Project Exam Help

- 1. (a) (i) 50.77
  - (ii) Right-skewed with two outliers.
  - (iii) It wall be strength from the CE region.
  - (iv) One day.
  - (b) (i) Aroun 0 d. dt meuts dat then the AQI of the other region will tend to increase.
    - (ii)  $19.887 + 0.714 \times 40$
    - (iii) First answer: Median of AQI in CE will give median of AQI in NW.

Explanation: since a regression line has center  $(\bar{x}, \bar{y})$ . Referring to symmetric distribution, these means are also medians. Hence the median in CE will gives median in NW. You may check with pnorm(qnorm(0.5)\*cor(CE,NW)) which also gives 0.5 since qnorm(0.5) is 0 and pnorm(0) is 0.5. Try.

Second answer:

a gives standardised value,

b gives percentile and

c gives original value of AQI in NW (reverse of standardisation). So c is the answer.

Explanation: From L10, P.31, the steps are

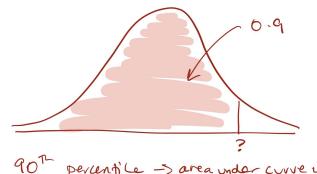
Step 1: Find the z score in the x direction:  $z_x$ .

Step 2: Find the predicted z score in the y direction:  $z_y = r * z_x$ .

Step 3: Translate  $z_y$  back to the percentile in the y direction.

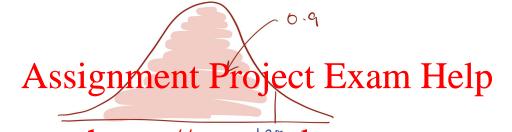
That means

In step 1, we need to find the z score for the 90th percentile rank in the CE direction. The **qnorm** function can help us with this.



90th percentile -> area under curve up to particular point is 0.9.

Using qnorm(0.9), we find this value is 1.28.



In step 2, now a is describing this which is to multiply the z-score by the correlation coefficient, qnorm(0.9)\*cor(CE,NW).

In step 3 A is to not the list as to perwise in the QI scale. We use the pnorm function for this as usual.

Then c takes the lecture slides one step further, by giving us the data point for AQI, as opposed to the percentile (note the question does not ask for AQI percentile, but rather the predicted AQI value). If

$$z \ score = \frac{data \ point \ - \ mean}{sd}$$

Then after some rearranging, we have

$$data point = (z score \times sd) + mean$$

which is what the R code in c is giving you.

- 2. (a) 32000 and 48000 respectively (total  $100000 \times 0.8 = 80000$ ; Vote A:  $80000 \times 0.4 = 32000$ ; Not vote A:  $80000 \times 0.6 = 48000$ )
  - (b) 0.0153949 because finite population adjustment factor should be applied when sampling is done without replacement.
  - (c) 10000 and 1000 respectively.
  - (d) The expected gain is  $10[0.5 \times 0.4 + 0.6 \times (-1)] = -4$ , a loss of \$4.
  - (e) The box should contain 32000 "1" (A "1", B "0"; difference "1 0"),

24000 "-1" (A "0", B "1"; difference "0-1") and 24000 "0" (A "0", B "0"; difference "0-0").

- 3. (a) Control experiment
  - (b) There are pairs of yield from using current and new fertiliser applied to the same plot of land with equal area.
  - (c)  $H_0: \mu_1 = \mu_2 \text{ vs } \mu_1 \neq \mu_2 \text{ where } \mu_1 \text{ and } \mu_2 \text{ are the mean yield from using current and new fertilisers.}$
  - (d) The boxplot is not symmetric and the some points in QQ plot does not lie close to the QQ line. Also the Shapiro Wilk test is significant with P-value less than 0.05. All these indicate that the data do not follow a normal distribution.
  - (e) The test statistic is 1.8946 and P value is 0.08471 > 0.05. There is not sufficient evidence in the data against  $H_0$  that the two fertilisers have the same effect on yield.

Note but not answer: Even though the test fails to find evidence against  $H_0$ , we can't claim that the two fertilisers are equal because we have already assumed  $H_0$  to be true in all calculation. Hence the failure of normality assumption is not an issue as the result is inconclusive (data consistent with  $H_0$ ).

In case of significant result (P value < 0.05), one should remark that the result of "significant difference between fertilisers" is valid only when the normality assumption for the data is satisfied. If not, the test statistic may not follow T distribution and the P value based on T distribution may be misleading. Then one may approximate the distribution of the test statistic using bootstrap simulation in the polyment the T calculates.

- (f) The boxplots have two different spread, the QQ plots show only the last point for current fertiliser deviates a lot from the QQ line and the F test has significant result against H<sub>1</sub> of goal variance. Alex should consider Welch T test which drops the equality of variance is subject to WCOCCI
- (g) The test statistic is -1.0481 and P value is 0.3096 > 0.05. Hence the data are consistent with  $H_0$  that the two fertilisers have the same effect on yield. Just further note: Although the degrees of freedom is higher than 11, there are more variability within each sample than within the sample of differences.

**End of Extended Answer Section** 

		Write yo SID here						I
THE UNIVERSITY OF SYNSCHOOL OF MATHEMATIC STATISTICS  MATH1015 Biostatis	S AND	Code your SID into the columns below each digit, by filling in the appropriate oval.	0 1 2 3 4 5 6 7 8					0 1 2 3 4 5 6 7 8
Family Name:								
Other Names:								
Seat Number: Assignm		_				lp		
https	s://p	owcode	er.	com		a h	a d	0
Answers	<b>W</b> <sub>2</sub> e		; > <b>&gt; W</b>	cod	Q11 <b>CT</b> 12	a b		e
Attempt every question. You will <b>not</b> be awarded negative marks for incorrect answers.	Q3 Q4 Q5				Q13 Q14 Q15			0 0 0
Fill in exactly one oval per question.	Q6 Q7 Q8		)		Q16 Q17 Q18			
If you make a mistake, draw a cross (X) through any mistakenly filled in oval(s) and then fill in your intended oval.	Q9 Q10		_		Q19 Q20			
An answer which contains two or more filled in (and uncrossed) ovals will be awarded no marks.								

## CORRECT RESPONSES TO MC COMPONENT OF MATH1015 Biostatistics

Sample Exam Questions: 2018

```
Q1 \longrightarrow b
 Q2 \longrightarrow b
 Q3 \longrightarrow a
 Q4 \longrightarrow e
 Q5 \longrightarrow b
 \mathrm{Q6} \longrightarrow \mathrm{b}
 Q7 \longrightarrow d
 Q8 \longrightarrow e
 Q9 \longrightarrow e
Q10 \longrightarrow b
Q11\longrightarrow b
Q12 Assignment Project Exam Help
\mathrm{Q}13\longrightarrow\mathrm{c}
\mathrm{Q}14 \longrightarrow \mathrm{e}
                     https://powcoder.com
Q15 \longrightarrow d
Q16 \longrightarrow c
\mathrm{Q}17\longrightarrow\mathrm{b}
                     Add WeChat powcoder
Q18 \longrightarrow b
Q19 \longrightarrow c
\mathrm{Q20}\longrightarrow\mathrm{c}
```