2024 Statistics Midterm Exam

Name: ________________ Student ID:____________________

1. Over a seven-week period, Professor Cheng assigned the students in her Statistics class the following pages of reading.

Week number	X (Number Pages of Reading Assigned)
1	40
2	50
3	70
4	50
5	40
6	20
7	20

(1) What is the statistical unit in this problem? (1 pt)

A week, or 7 weeks

(2) What is the variable in this problem? (1pt)

Number Pages of Reading Assigned

(3) What is the measurement type of the variable in this problem? (1 pt)

Interval scale

(4) Calculate the mean number of pages of reading assigned per week. (2 pt)

$$\bar{X} = \frac{\sum X_i}{7} = \frac{290}{7} = 41.4$$

(5) List the mode(s) of this problem. (2 pt)

Mode = 20, 40, 50

(6) What is the median number of pages of reading in a week. (2 pt)

X (Array)
20
20
40
40

50
50
70

Median position = (7+1)/2 = 4

Median = 40

(7) Calculate the variance of the above distribution. (2 pt)

$$variance = S^2 = \frac{\sum_i (X_i - \bar{X})^2}{N} = \frac{13900 - \frac{290^2}{7}}{7} = 269.4$$

(8) Calculate the sum of the deviations of X (number of pages of reading assigned each week) about the mean of X. (2 pt)

0

2. Suppose that a random sample of 75 of Professor Cheng's Statistics students is drawn and each of these students is asked how many of total number of pages of assigned reading he/she actually read. The following data are compiled.

Number of Pages Actually Read	Number of Students
201-250	0
151-200	10
101-150	20
51-100	30
1-50	10
0	5

(1) What is the statistical unit in this problem? (1 pt)

A student, or 75 students

(2) What is the variable in this problem? (1 pt)

Number of pages actually read

(3) What is the range of the variable in this problem? (1 pt)

200

(4) Calculate the mean for the above distribution. (2 pt)

$$\bar{X} = \frac{Midpoint\ X * frequency}{n}$$

$$= \frac{225.5 * 0 + 175.5 * 20 + 125.5 * 20 + 75.5 * 30 + 25.5 * 10 + 0 * 5}{75}$$

$$= \frac{6785}{75} = 90.5$$

(5) What is the median for the above distribution. (2 pt)

Number of Pages Actually Read	Number of Students (f)	Midpoints (X)	X * f	Cumulative frequency
201-250	0	225.5	0	
151-200	10	175.5	1755	75

101-150	20	125.5	2510	65
51-100	30	75.5	2265	45
1-50	10	25.5	255	15
0	5	0	0	5

Median position = (75+1)/2 = 38

Median =$Md = L_{Md} + \left(\frac{\frac{N}{2} - cf_{Md}}{f_{Md}}\right) i_{Md} = 50.5 + \left(\frac{\frac{75}{2} - 15}{30}\right)(105.5 - 50.5) = 88$

(6) List the mode(s) of this problem. (2 pt)

Mode = 75.5

3. Data from the above sample of students in Prof. Cheng's Statistics class were compiled to explore the relationship between the total number of pages of reading a student had actually read and his/her performance on a quiz given at the end of the unit of study.

Number of Pages Actually Read	Performance on Quiz (Number of Students)	
	Failed	Passed
150 pages or less	13	52
More than 150 pages	2	8

Indicate whether each of the following statement is true (T) or false (F) in terms of these data. If any part of a statement is untrue, it should be marked false (F).

__T__ a. The statistical unit in these data is a student. (1 pt)

__F__ b. Those students who read 150 pages or less were more likely to fail the quiz than were those who read more than 150 pages. (1 pt)

__T__ c. The ratio of those who passed to those who failed the test is 4.0. (1 pt)

__T__ d. In this sample, there is no relationship between number of pages a student read and his/her performance on the quiz. (1 pt)

__F__ e. The mode for the above data is 52. (1 pt)

__T__ f. "Performance on Quiz" is the dependent variable in the above table; "Number of pages actually read" is the independent variable. (1 pt)

__T__ g. Performance on Quiz is measured by a two-category nominal scale on the above table. (1 pt)

__T__ h. The mean number of pages actually read by students who failed the course cannot be determined from the above table. (1 pt)

__T__ i. 80% of those who read 150 pages or less passed the quiz. (1 pt)

4. Elevated serum cholesterol levels are often associated with cardiovascular disease. Cholesterol levels are often thought to be associated with type of diet, amount of exercise, and genetically related factors. The distribution of cholesterol levels in U.S. women aged 30–50 (i.e., middle-aged) is known to be approximately normally distributed with a mean of 190 mg/dL. A study was interested in determining if the mean cholesterol level among recent immigrants from China was different from cholesterol levels middle-aged women in the United States. Researchers did not have any prior information about these Chinese immigrants and they randomly selected and examined cholesterol levels among n = 100 female Chinese immigrants aged 30–50 who had immigrated to the United States in the past year. They were administered blood tests that yielded cholesterol levels having a mean of 178.2 mg/dL and a standard deviation of 45.3 mg/dL. Is there significant evidence in the data to demonstrate that the mean cholesterol level of the new immigrants differs from 190 mg/dL? Perform a statistic test. Use $\alpha = 0.05$ to determine the statistical significance. Precisely interpret the meaning of the statistical findings, and draw conclusions with estimates of the probability of having made Type I and Type II errors. (8 pt)

H_0: $\mu = \mu_0 = 190$
H_a: $\mu \neq 190$

With a sample size of n = 100, the Central Limit Theorem suggests that the sampling distribution of $\bar{y}$ is approximately normal.

T.S.: $z = \frac{\bar{y}-\mu_0}{\sigma/\sqrt{n}} = \frac{178.2-190}{45.3/\sqrt{100}} = -2.60$

R. R.: For two-tailed test, $\alpha = 0.05$, $\alpha/2 = 0.025$

$z_{\alpha/2} = z_{0.025} = \pm 1.96$

Because calculated z-value= $-2.60 <$ -1.96 ➔ Reject H_0
The data support the research hypothesis that there is difference between the mean cholesterol level among recent immigrants from China was different from cholesterol levels middle-aged women in the United States. Conclusion: The mean cholesterol level among recent immigrants from China was significantly smaller than the cholesterol levels middle-aged women in the United States. The probability of a Type I error, $\alpha \leq 0.047$, meaning that there is a $\leq$4.7% chance that we may be wrong in rejecting H_0. The probability of having made a Type II error, β=0

5. An analysis of income tax returns from the previous year indicates that for a given income classification, the amount of money owed to the government over and above the amount paid in the estimated tax vouchers for the first three payments is approximately normally distributed with a mean of $530 and a standard deviation of $200. The government wants to target that group of returns having the largest 25% of amounts owed. Find the measurements for those having the largest 25% of amounts owed. (2 pt)

Ans: We need to determine the measurement for the 75th percentile.

Check from the z-table, we find $z_{0.75} \cong 0.67$

➔ $y_{0.75} = \mu + z_{0.75} * \sigma = 530 + 0.67 * 200 = 664$

6. The Chunghua Telecom finds that 80% of their customers pay their monthly phone bill in full. Suppose two customers are chosen at random from the list of all customers of Chunghua Telecom.
   (1) What is the probability that both customers will pay their monthly phone bill in full? (2 pt)
   (2) What is the probability that at least one of them will pay in full? (2 pt)

Ans: (1) P(both customers pay in full) = 0.8*0.8=0.64

(2) P(at least one of two customers pay in full) = 1-P(neither customer pays in full)=1- (1-0.8)*(1-0.8) = 0.96

7. Educational researchers have long been interested in differences and similarities among people in the extent to which they can retain and recall information that they have previously learned. Retention has often been studied by having subjects learn "definitions" to 50 nonsense syllables and asking them to recall these definitions after a period of time. The results of numerous previous studies of people in the general population have shown that the number definitions recalled correctly by subjects at the end of 2 weeks is approximately normally distributed with a mean (μ) of 23 and standard deviation (σ) of 8. Answer the following in terms of this information.
   (1) What proportion of the population recall 13 definitions or more? (2 pt)
   (2) What is the probability of selecting a person at random from this population and obtaining one who recalled between 31 and 33 definitions correctly? (2 pt)
   (3) How many persons in a random sample of 100 persons would you expect to recall between 23 and 25 definitions correctly. Round your answer to the nearest integer. (2 pt)
   (4) What is the probability of selecting a random sample of 64 persons from the general population and obtaining a sample with a mean number of recalled definitions of between 21 and 24? (2 pt)
   (5) Let us suppose that you are interested in studying the degree to which NTU students' ability **exceeds** that of the general population in retaining and recalling information they have previously learned. To this end, you select a random sample of NTU students and have all of them learn the "definitions" of the 50 nonsense syllables standardly used to measure retention/recall. A test of their acquisition determines that all member of the sample have, by the end of the session learned all 50 definitions. After two weeks, your sample members are asked to recall these "definitions". Degree of retention of each student is indicated by the number of definitions he/she correctly recalls. Data on males and females were compiled as follows:

	Males	Females	Total
Number of Cases	40	30	70
Mean Retention Score	24.5	28.0	26.0
Sum of Retention Scores ($\sum y_i$)	980	840	1820
Sum of the Squares of the Retention Scores ($\sum y_i^2$)	25414	24564	49978

Using the information to test the statistical significance of the difference between the general population and NTU students in regard to retention scores. Use the 0.05 level to determine statistical significance. (6 pt)

Ans: (1) $z = \frac{13-23}{8} = -1.25$ ➔ Standard normal curve areas below $Z_{-1.25}$ = 0.1056

$P = 1 - 0.1056 = 0.8944$

(2) $z = \frac{31-23}{8} = 1.00$ ➔ Standard normal curve areas below $Z_{1.00}$ =0.8413

$z = \frac{33-23}{8} = 1.25$ ➔ Standard normal curve areas below $Z_{1.25}$ =0.8944

$P = 0.8944 - 0.8413 = 0.0531$

(3) $z = \frac{25-23}{8} = 0.25$ ➔ Standard normal curve areas below $Z_{0.25}$ =0.5987

(0.5987-0.5)*100 = 9.87 ≈ 10

(4) $z = \frac{21-23}{8/\sqrt{64}} = -2.00$ ➔ Standard normal curve areas below $Z_{-2.00}$ =0.0228

$z = \frac{24-23}{8/\sqrt{64}} = 1.00$ ➔ Standard normal curve areas below $Z_{1.00}$ =0.8413

$P = 0.8413 - 0.0228 = 0.8185$

(5) H_0: $\mu_{NTU} = \mu_0 = 23$

H_a: $\mu_{NTU} > \mu_0 = 23$

With a sample size of n = 70, the Central Limit Theorem suggests that the sampling distribution of $\bar{y}$ is approximately normal.

T.S.: $z = \frac{\bar{y}-\mu_0}{\sigma/\sqrt{n}} = \frac{26-23}{8/\sqrt{70}} = 3.1375$

R. R.: For right-tailed test, $\alpha = 0.05$, $z_\alpha = z_{0.05} = 1.645$

Because calculated z-value= 3.1375 > 1.645 ➔ Reject H_0

After checking the design of the experiments and the relevant assumptions are all ok, we conclude that the data support the research hypothesis. The population we sampled at NTU had a mean retention score (26) greater than the mean retention score (23) of the general population. However, we may have made a Type I error, α. The probability of a Type I error, α, is calculated as $\alpha = 1 - \frac{0.9991+0.9992}{2} =$ 0.0009, meaning that there is a $\leq$ 0.09% chance that we may be wrong in rejecting H_0. The probability of having made a Type II error, β=0

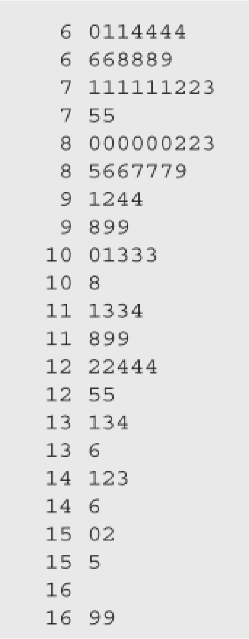
8. The data in the table are the maximum ozone readings (ppb) taken on 80 summer days in a large city. The readings are either two- or three-digit numbers. Use the first digit of the two-digit numbers and the first two digits of the three-digit numbers as the stem and the remaining digits as the leaf number to construct a stem-and-leaf plot. (5 pt)

Maximum ozone readings (ppb):

60	61	61	64	64	64	64	66	66	68
68	68	69	71	71	71	71	71	71	72
72	73	75	75	80	80	80	80	80	80
82	82	83	85	86	86	87	87	87	89
91	92	94	94	98	99	99	100	101	103
103	103	108	111	113	113	114	118	119	119
122	122	124	124	124	125	125	131	133	134
136	141	142	143	146	150	152	155	169	169

Ans:

Stem-and-leaf plot



```
 6  0114444
 6  668889
 7  111111223
 7  55
 8  000000223
 8  5667779
 9  1244
 9  899
10  01333
10  8
11  1334
11  899
12  22444
12  55
13  134
13  6
14  123
14  6
15  02
15  5
16
16  99
```

9. Draw a boxplot for the grades of the Statistics exam taken by college students and compare this with the boxplot for the grades of the Statistics exam by graduate students. **Mark the values at all relevant positions in the diagram**. (5 pt)
Grades of the Statistics exam taken by college students from a sample of 15 college students: 43, 80, 68, 92, 52, 88, 76, 73, 62, 81, 85, 32, 50, 63, 65
Grades of the Statistics exam taken by graduate students from a sample of 15 graduate students: 56, 68, 86, 72, 92, 98, 86, 79, 65, 91, 82, 70, 65, 44, 51

Ans:

(1) order the data.

college students: 32, 43, 50, 52, 62, 63, 65, 68, 73, 76, 80, 81, 85, 88, 92

graduate students: 30, 44, 51, 65, 65, 68, 70, 72, 79, 82, 86, 86, 91, 92, 98

(2) Calculate median and interquartile values.

college students: N=15, Q2=68, Q1=52, Q3=81, Interquartile range (IQR)=81-52=29, min=32, max=92

upper whisker= Q3+1.5IQR=85+1.5*23 =124.5 > Max grades=92, Use 92 as upper whisker

lower whisker= Q1-1.5IQR=52-1.5*23=8.5 < Min grades=32, Use 32 as lower whisker

graduate students: N=15, Q2=72, Q1=65, Q3=86, Interquartile range (IQR)=86-65=21, min=44, max=98

upper whisker= Q3+1.5IQR=85+1.5*23 =117.5 > Max grades=98, Use 98 as upper whisker

lower whisker= Q1-1.5IQR=65-1.5*21=33.5

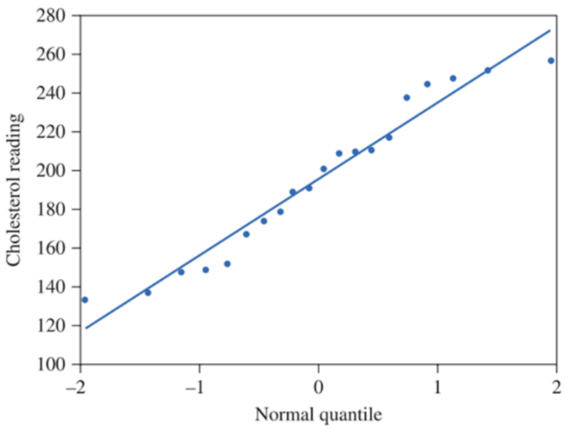
(3) Find outliers. college students: no outliers, graduate students: 30

(4) Plot these against a number line as a line and connect the median and quartiles together to make a box shape. Draw a point to represent the upper and lower values, with a straight line joining these up to the edges of the box. Plot the outliers as a point on the diagram.

10. It is generally assumed that cholesterol readings in large populations have a normal distribution. To evaluate this conjecture, the cholesterol readings of n=20 patients were obtained. The cholesterol readings are given in an ordered fashion from smallest to largest. Please calculate the normal quantile using the equation of $Q_{yi}=Q((i-0.5)/n)$ and **fill in the blank fields in the following table** to match the patients' cholesterol readings from the smallest to the largest. (6 pt)
In addition, use this information to **draw a normal quantile plot** for the cholesterol data, and **explain** how you assess whether the data selected from a population exhibits a normal distribution. (5 pt)

Patient	Cholesterol Reading	(i-0.5)/20	Normal Quantile
1	133	0.025	-1.960
2	137	0.075	-1.440
3	148	0.125	-1.150
4	149	0.175	-0.935
5	152	0.225	-0.755
6	167	0.275	-0.598
7	174	0.325	-0.454
8	179	0.375	-0.319
9	189	0.425	-0.189
10	192	0.475	-0.063
11	201	0.525	0.063
12	209	0.575	0.189
13	210	0.625	0.319
14	211	0.675	0.454
15	218	0.725	0.598
16	238	0.775	0.755
17	245	0.825	0.935
18	248	0.875	1.150
19	253	0.925	1.440
20	257	0.975	1.960

Ans:
In the scatter plot, we observe that the dots fall near the straight line. This indicates that the data selected from a population exhibits a normal distribution.


Cholesterol reading
Normal quantile