

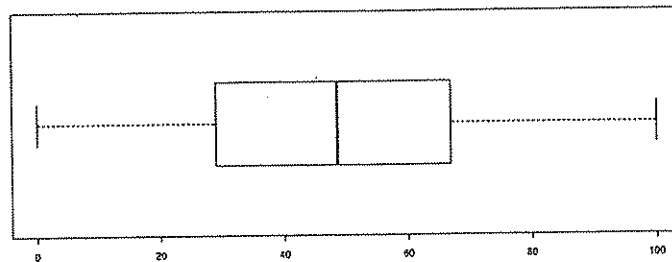
考試科目	統計學	所(組)別	統計學系	考試時間	101 年 11 月 4 日 星期日 10:00 - 11:40
------	-----	-------	------	------	-------------------------------------

Part 1

1. (20%) The following 100 observations are the mid-term scores (in ascending order) of Statistics for 2012 Fall semester. The sample average is 49.72 and the sample standard deviation is 25.05.

0	23	44	57	75
3	25	45	58	76
11	26	45	58	76
12	28	45	58	77
14	29	46	58	78
15	29	46	59	79
15	31	47	59	80
15	32	48	61	80
16	33	48	63	81
17	33	48	63	83
19	33	49	64	88
19	34	50	65	90
19	35	50	65	92
19	36	50	67	92
19	39	50	67	93
20	40	51	67	95
21	41	52	68	96
21	41	54	70	98
22	42	54	73	98
23	44	55	74	100

- (a) The distribution of mid-term scores is believed to be uniform, normal, or exponential distribution. Using the descriptive statistics (without plotting the data) to judge which distribution is the most likely.
- (b) We can also use the box plot (or box-and-whisker plot) to describe the mid-term scores (the graph below). Define the numbers shown in the box plot (i.e., the definition of 5 vertical lines) and use the data in the table above to fill in the required statistics.



- (c) What is the empirical rule? Do the scores satisfy the empirical rule?
- (d) Judging from the box plot in (b), do you think that the mid-term scores are a random sample from a normal distribution? State your reasons.

考試科目	統計學	所(組)別	統計學系	考試時間	101 年 11 月 4 日 星期日 10:00 - 11:40
------	-----	-------	------	------	-------------------------------------

Part 2

1. (5%) Let X follow a Binomial(80,0.7) distribution. Approximate the probability $P(X > 60)$ by the normal approximation to the binomial.

2. (10%) The distribution of the number of traffic accidents per month at a particular crossroad over the past 586 months was recorded as follows.

#accidents	0	1	2	3	4	5
Months	229	221	93	35	7	1

(a) Would you model this data by a binomial distribution, a Poisson distribution, a normal distribution, an exponential distribution, or other distributions? (Need to explain.)

(b) Obtain an appropriate parameter associated with the distribution you suggested in part (a).

3. (5%) A study was conducted to estimate the average size of households in Taipei. A total of 500 people were randomly selected from the population of Taipei and the average size was found to be 3.5. Would an estimate computed in this manner be appropriate for the true average size? Or would it tend to be larger or smaller than the true average size? Briefly explain why.

考試科目	統計學	所(組)別	統計學系	考試時間	101年11月4日 星期日 10:00-11:40
------	-----	-------	------	------	------------------------------

Part 3

1. To study the characteristics of hospitalizations by different family income levels, the data of 13,860 low-income patients hospitalized and 600,903 non-low-income patients hospitalized in 2006-2007 were randomly collected in the National Health Insurance Research database (全民健保資料庫). Assume the research interest lies on the **comorbidity** (共病現象), which indicates the presence of one or more diseases in addition to a primary disease. **The research hypothesis is that the inpatients with low-income had more comorbid conditions.** Denote the number of diseases that coexist with the primary disease of an individual by X . The following are the frequency table of X by the income level in the data.

X	Low income	Non-low income
	Frequency (Proportion)	Frequency (Proportion)
0	2,132 (15%)	136,073 (23%)
1	2,529 (18%)	133,452 (22%)
2	2,469 (18%)	107,335 (18%)
3	2,335 (17%)	83,419 (14%)
4	4,395 (32%)	140,624 (23%)
Total number	13,860 (100%)	600,903 (100%)

- (a) (10 pts.) Compare the means of X of the two income groups to test the research hypothesis at the significance level 5%.
- (b) (5 pts.) Define the comorbidity probability as $P(X>0)$. Compare the comorbidity probabilities of the two income groups to test the research hypothesis at the significance level 5%.
- (c) (5 pts.) Compare the method and conclusion of (a) and (b). Which one do you prefer?

考 試 科 目	統計學	所(組)別	統計學系	考 試 時 間	101 年 11 月 4 日 星期日 10:00 - 11:40
---------	-----	-------	------	---------	-------------------------------------

Part 4

1. Write down the equation of multiple linear regression model and its required assumptions. (6%)
2. Give a brief description of the diagnose procedure to determine whether the above required assumptions are satisfied. (4%)
3. Consider the following situation :
When the regression output (t tests) reveals that none of the independent variables is significantly related to the dependent variable. However, the F test indicates that the model is valid. Moreover, the coefficient of determination (R^2) is high, which tells us that the model is a good fit.
How can the model be valid and fit well, when none of the independent variables that made up the model is linearly related to dependent variable? Is there any statistics that can be used to detect this phenomenon? (5%)
4. When the data is gathered sequentially over a series of time period, and multiple linear regression model is applied to the data. Which of the required assumptions may be violated? Is there any statistics that can be used to detect this phenomenon? (5%)

考試科目	統計學	所(組)別	統計學系	考試時間	101年11月4日 星期日 10:00-11:40
------	-----	-------	------	------	------------------------------

Part 5

1. After getting trounced by your little brother in a children's game, you suspect the die may be unfair. To check, you roll it 60 times, recording the number of times each face appears. Do these results cast doubt on the die's fairness?

Face	Count
1	11
2	7
3	9
4	15
5	12
6	6

- a) (2%) If the die is fair, how many times would you expect each face to show?
- b) (3%) To see if these results are unusual, will you test goodness-of-fit, homogeneity, or independence? Explain.
- c) (2%) State your hypotheses.
- d) (2%) How many degrees of freedom are there?
- e) (3%) Find the P-value.
- f) (2%) State your conclusion.
- g) (3%) If you had rolled the die 180 times, and counted exactly three times as many in each category (i.e. 33, 21, 27, 45, 36, 18), what conclusion would you have reached?
- h) (3%) Is there a discrepancy between the two conclusions? Does it surprise you? Explain.