國立清華大學103學年度碩士班考試入學試題

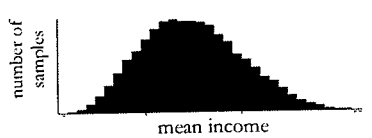
系所班組別：服務科學研究所碩士班 甲組（服務管理組）

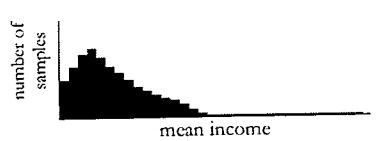
考試科目（代碼）：統計學（4701）

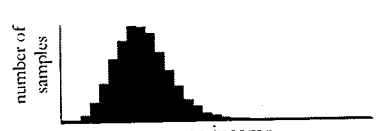
For each question on this exam, **please only circle one answer** that seems most appropriate to you. Wherever you have to calculate values, **please show your calculations**. For questions that do not require much calculation, **please give an explanation** for why you picked your answer (use brief explanations in very simple English). If an answer is wrong, any extra work or explanation you have given will be considered for partial credit. If an answer is correct but you have not shown your calculations or given adequate explanation, you may not get full points.

1. [10 %] The following data distribution shows the yearly incomes of all 50,000 alumni who have graduated from a university in Taiwan and are currently employed.

Let us say we were to take one thousand sub-samples from this population data of alumni (each sub-sample has 50 alumni, who are randomly picked with replacement). Consider how the statistical mean of the sub-samples would be distributed. Pick which distribution below would most likely reflect the distribution of the sampling mean across sub-samples.

(a)


(b)


(c)


(d)


(e) none of the above could possibly be the distribution of sampling means.

Please explain why you picked your answer.

You are attending a marketing presentation where your colleague is describing how she computed statistics describing sales on your company's online platform:
" *After customers have selected products they wish to purchase, they must go through our website's electronic shopping-cart checkout. I wanted to know the central tendency of checkout totals, so I calculated the total of each checkout, summed up these checkout totals, and divided by the number of checkouts. There were 13,500 checkouts last week, with an average checkout total of NTD 833.57. I also wanted to get an idea of the dispersion of checkout totals. To do this, I first calculated the total of each of the 13,500 electronic shopping-cart checkouts. Then, I computed the average of all checkout totals. After that, I estimated the difference of each checkout total from the average checkout total and then I took the average of these differences. This gives us a dispersion of NTD 7.97* "

She then displays the R command she used to calculate the "central tendency of checkout totals" and the "dispersion of checkout totals":

```
central = sum(checkouts)/length(checkouts)
dispersion = sum(abs(checkouts - mean(checkouts)))/length(checkouts)
```

Your boss turns to you and quietly admits that he is very confused by how "central tendency of checkout totals" and "dispersion of checkout totals" were calculated. He asks you if there is another description of these two statistics that were just computed, so that he can read more about it later. How would you answer him?

2. [10 %] The "central tendency of checkout totals" was computed as a:
(a) mean of checkout totals
(b) median of checkout totals
(c) mean absolute deviation of checkout totals
(d) standard deviation of checkout totals
(e) variance of checkout totals
(f) variability of checkout totals
(g) no idea – none of the above!

3. [10 %] The "dispersion of checkout totals" was computed as a:
(a) mean of checkout totals
(b) median of checkout totals
(c) mean absolute deviation of checkout totals
(d) standard deviation of checkout totals
(e) variance of checkout totals
(f) variability of checkout totals
(g) no idea – none of the above!

4. [10 %] Your boss needs your help again. This time, he is trying to understand what a regression line is, but does not know where to start. Which of the following most accurately describes how the line, or surface, produced by ordinary least squares regression fits the data?

(a) it minimizes distances between itself and data points in the direction of a dependent variable.
(b) it shows the direction of maximum variance in a multi-dimensional set of variables.
(c) it is the best fit for all variables in a data set (although the fit is not usually "perfect").
(d) it best predicts future values of a dependent variable, given independent variables.
(e) it proves whether changes in independent factors cause a dependent variable to change.

5. [10 %] A car tire company in Hsinchu getting a bad reputation because people believe they are creating low quality tires using poor materials. In a recent interview on TV, the company's CEO claimed that their tires can safely travel at least 60,000 km on the highway.

But a skeptical journalist wants to test this claim. He has purchased 36 tires from Chekzar and tested them in a special facility. On average, the tires in his sample lasted 58,341.69 km before failure, with a standard deviation of 3632.53 km.

Is the claim of the tire company's CEO to be believed, or can it be challenged?

(a) Can reject the claim at 0.1% significance
(b) Can only reject the claim at 1% significance, not less
(c) Can only reject the claim at 5% significance, not less
(d) Cannot reject the CEO's claim
(e) Not enough information to answer this question

Please explain how you arrived at your answer.

In a separate test, the memory usage of our computer scientist's image processing algorithm is tested against the memory usage of other algorithms, and against memory usage when different types of image content are considered. The results of an ANOVA test are shown below. The factor `Algorithm` refers to the choice of algorithm used, and the factor `Content` refers to different types of image content.

	DF	Sum-Sq	Mean-Sq	F-value	Pr(>F)
Algorithm	2	45300	22650	10.270	0.00476
Content	2	6100	3050	1.383	0.29944
Algorithm*Content	4	11200	2800	1.270	0.35033
Residuals	9	19850	2206		

6. [10 %] From the ANOVA results, which factors would you say were significantly related to memory usage at 5% significance or less?
(a) The choice of algorithm
(b) The choice of image content
(c) The interaction of algorithm and content
(d) More than one of the above
(e) None of the above

Please explain how you arrived at your answer.

7. [10 %] From the ANOVA results, how much of the variance of memory usage is explained by the content of the images? (note: units are omitted)
(a) 2
(b) 6100
(c) 3050
(d) 1.383
(e) 0.29944
(f) Not enough information to answer this question

Please explain how you arrived at your answer.

國立清華大學103學年度碩士班考試入學試題

系所班組別：服務科學研究所碩士班 甲組（服務管理組）

考試科目（代碼）：統計學（4701）

A sports coach is monitoring the performance (recorded as points) of an athlete every day at practice, and has also kept record of how much sleep (recorded as hours) she got each night before practice, and how much caffeine (recorded in milliliters) she consumed each morning before practice. The coach has conducted a regression where the dependent variable was performance, and the independent variables were sleep, caffeine and the interaction between sleep and caffeine. Fully standardized results are shown:

```
                Estimate Std. Error t-value Pr(>|t|)
(Intercept)      0.01109    0.09832   0.113   0.9105
sleep            0.43004    0.09986   4.307 4.89e-05
caffeine        -0.01669    0.10433  -0.160   0.8734
sleep*caffeine  -0.18082    0.08386  -2.156   0.0342

Residual standard error: 0.8782 on 76 degrees of freedom
R-squared: 0.258,    Adj. R-squared: 0.2287
```

8. [10 points] How many days worth of data would you say the coach has recorded?
(a) 50 days
(b) 60 days
(c) 70 days
(d) 80 days
(e) Not enough information to answer this question

9. [10 points] We often describe the similarity of two variables are using a *correlation coefficient.* However, we can also describe how similar a variable is to a set of many other variables using a *multiple correlation coefficient.* From the regression results above, what would you estimate is the multiple correlation coefficient of memory usage when compared to sleep and caffeine?
(a) 0.878
(b) 0.508
(c) 0.413
(d) 0.258
(e) 0.229
(f) Not enough information to answer this question

10. [10 points] If the athlete in question kept his caffeine intake to his average, but increased his amount of sleep by one standard deviation, by how much would you predict his performance would change?
(a) Performance would increase by 0.23 standard deviations
(b) Performance would increase by 0.25 standard deviations
(c) Performance would increase by 0.43 standard deviations
(d) Performance would increase by 1 standard deviation
(e) Not enough information to answer this question

Please explain how you arrived at all your answers on this page.