May 2012 Question 2 Correlation and Regression

- The sample size $n = 10$.
- The ***independent*** variable, usually denoted x, is the "cause variable" or "predictor variable".
- The ***dependent*** variable, usually denoted y, is the "effect variable".
- Here the Maths achievement test score is the independent variable and the final grade in statistics is the dependent variable.
- A big hint is given in the notation of the question.

Sums of Squares Identities

Before we do anything, We need to compute the following sums of squares identities

- s_{xx}
- s_{yy}
- s_{xy}

Calculation 1

$$s_{xx} = \sum(x^2) - \frac{\sum(x)^2}{n}$$

$$s_{xx} = 23.634 - \frac{\sum(x)^2}{10}$$

Calculation 2

$$s_{yy} = \sum(y^2) - \frac{\sum(y)^2}{n}$$

Calculation 3

$$s_{xy} = \sum(xy) - \frac{\sum(x) \times \sum y}{n}$$

Part iv - Prediction

- Suppose the regression equation is as follows

$$\hat{y} = 40.78424 + 0.76556x$$

- If a student scored 5 marks on the achievement test (i.e. $x = 5$), predict the students statistics grade.

$$\hat{y}_{(x=5)} = 40.78424 + (0.76556 \times 5)$$

- Solving using a calculator we get

$$\hat{y}_{(x=5)} = 44.61204$$

- The score should be approximately 44.61.

May 2013 Question 6b Correlation and Regression

Calculate the correlation coefficient and interpret the value.

Residence	X	Y

May 2012 Question 4 Normal Distribution / Theory

Revision of Normal Distribution

0.1 Important rules for normal distribution

- **Complement Rule**
  For some value A, and for any continuous distribution X (including any normal distribution and the Z distribution) we can say.

$$P(X \leq a) = 1 - P(X \geq A)$$

- **Symmetry Rule**
  For the standard normal (Z) distribution only, we can say

$$P(Z \leq -A) = P(Z \geq A)$$

  or conversely

$$P(Z \geq -A) = P(Z \leq A)$$

- **Interval Rule**

  Suppose we have an interval for the random variable X defined by the

  - the lower bound L
  - the upper bound U

$$L \leq X \leq U$$

  The probability of being inside this interval is the **complement** of being outside the interval. The event of being outside the interval is the union of two disjoint events.

  - The probability of X being too low for the interval (i.e. less than the interval minimum L)

$$P(X \leq L)$$

  - The probability of X being too high for the interval (i.e. less than the interval maximum U)

$$P(X \geq U)$$

$$P(U \leq X \leq L) = 1 - (P(X \leq L) + P(X \geq U))$$

Parameter Values

Given the parameters of the normal distribution X in the question.

- Normal Mean $\mu = 73$ points
- Normal Standard Deviation $\sigma = 8$ points
- $P(X \leq 91)$
- $P(65 \leq X \leq 89)$

Find the Z score for X = 91.

$$Z = \frac{x - \mu}{\sigma} = \frac{91 - 73}{8} = \frac{18}{8} = 2.25$$

Therefore we can say :
$P(X \leq 91) = P(Z \leq 2.25)$

From the tables $P(Z \leq 2.25) = 0.9877$ Therefore the probability of getting a grade lower than 91 is 0.9877 (i.e 98.77%)

What is the probability of getting a score between 65 and 89. Writing this mathematically:

$$P(65 \leq X \leq 89)$$

- How many people get a score greater than 89? $(P(X \geq 89))$
- How many people get a score less than 65? $(P(X \leq 65))$

To compute $P(X \geq 89)$ first compute the Z-score.

$$Z = \frac{x - \mu}{\sigma} = \frac{89 - 73}{8} = \frac{16}{8} = 2$$

$P(X \geq 89) = P(Z \geq 2) = 0.0225$.
To compute $P(X \leq 65)$ first compute the Z-score.

$$Z = \frac{x - \mu}{\sigma} = \frac{65 - 73}{8} = \frac{-8}{8} = -1$$

$P(X \leq 65) = P(Z \leq -1)$

- We use the **symmetry rule**

$$P(Z \leq -1) = P(Z \geq +1)$$

- so we can say $P(X \leq 65) = P(Z \geq +1)$
- From the statistical tables $P(Z \geq +1) = 0.1583$.

Theory Components

- Distinguish between a bimodal distribution and a unimodal distribution
- Compare and contrast interval and ordinal data.

May 2012 Question 5 Normal Distribution

Given

- X is the variable of interest.
- Normal Mean $\mu = 25.5$ mpg
- Normal Standard Deviation $\sigma = 4.5$ mpg
- Find x such that $P(X \geq x) = 0.30$

Solution

From the Standard Normal Tables, find the value of z that would give us

$$P(Z \geq z) = 0.30$$

Or if you are using the other type of tables

$$P(Z \leq z) = 0.70$$

May 2013 Question 3 Normal Distributions

Important Information from the Question

- Normal Mean $\mu = 1000$ units
- Normal Standard Deviation $\sigma = 200$ units

Objectives

Compute the following :

- $P(X \geq 1400)$ More than 1400
- $P(X \leq 500)$ Less than 500

Part 1 - More than 1400

Firstly compute the z score for 1400.

$$Z_{1400} = \frac{X - \mu}{\sigma} = \frac{1400 - 1000}{200} = \frac{400}{200} = 2$$

So the **Z-score** in this case is 2.
This much we can say

$$P(X \geq 1400) = P(Z \geq 2)$$

$P(Z \geq 2)$ can be determined using statistical tables. Depending on which statistical tables you are using, you will get one of the following answers. (Note the second and third statements are examples of complementary probabilities.)

- $P(0 \leq Z \leq 2) = 0.4775$
- $P(Z \leq 2) = 0.9775$
- $P(Z \geq 2) = 0.0225$

The last expression is useful here. Recall that $P(X \geq 1400) = P(Z \geq 2)$. Therefore

$$P(X \geq 1400) = 0.0225$$

Return on Investment Question

- The company needs to recover its investment in one year (i.e. make 50000).
- As each product sells for 2 dollars profit, the company needs to sell 25,000 units to recover its investment.
- we need to compute the probability of selling more than 25,000 units.

$$P(X \geq 25000)$$

- We are told the normal mean for demand $\mu = 20000$ and the normal standard deviation $\sigma = 4000$.
- The first step is to compute the ***z-score***

$$z = \frac{x - \mu}{\sigma} = \frac{25000 - 20000}{4000} = \frac{5000}{4000} = 1.25$$

May 2013 Question 4 Probability

	within W	outside O	Totals
Correct time C	83	51	134
Delayed D	24	12	36
Totals	107	63	170

- Probability of departing at correct time

$$P(C) = 107/170$$

- Probability of being delayed and flying outside europe

$$P(D \text{ and } O) = 12/170$$

- Probability of

$$P()$$

- Probability of

$$P()$$

- Probability of

$$P()$$

May 2013 Question 5 Regression and Correlation

	Experience	No. of Rejects
A	4	22
B	5	20
C	7	18
D	9	15
E	9	16
F	10	11
G	14	10

- $\sum x = 58$
- $\sum y = 118$
- $\sum x^2 = 548$
- $\sum y^2 = 1910$

- $\sum xy = 843$
- $s_{xx} = 67.428$
- $s_{yy} = 118$
- s_{xy} = -85

The correlation coefficient

The correlation coefficient is $r = -0.9529$.

$$r = \frac{s_{xy}}{\sqrt{(s_x x \times s_y y)}} = \frac{-85}{\sqrt{(67.428 \times 118)}}$$

$$r = \frac{-85}{sqrt(7956.504)} = \frac{-85}{89.199}$$

$$r = -0.9529$$

The coefficient of determination

The coefficient of determination r^2 is computed as the square of the correlation coefficient.

$$r^2 = (-0.9529)^2 = 0.90801$$

May 2013 Question 6a Sampling

- (purely) random samplings
- quota sampling
- stratified sampling

Stratified Sampling

A stratified sample is a mini-reproduction of the population. Before sampling, the population is divided into characteristics of importance for the research. For example, by gender, social class, education level, religion, etc. Then the population is randomly sampled within each category or stratum. If 38% of the population is college-educated, then 38% of the sample is randomly selected from the college-educated population.

May 2012 Question 3 Time Series Analysis

	Q_1	Q_2	Q_3	Q_4
2007	20	35	26	18
2008	18	36	24	15
2009	14	34	25	14
2010	15	32	23	12

Calculate the trend using the moving averages method

Year	Time Period	Y	Moving Total	Moving Average	Centered MA
(column 1)	(column 2)	(column 3)	(column 4)	(column 5)	(column 6)
2007	1	20			
2007	2	35	99	24.75	
2007	3	26	97	24.25	24.5
2007	4	18	98	24.5	24.375
2008	5	18	96	24	24.25
2008	6	36	93	23.25	23.625
2008	7	24	89	22.25	22.75
2008	8	15	87	21.75	22
2009	9	14	88	22	21.875
2009	10	34	87	21.75	21.875
2009	11	25	88	22	21.875
2009	12	14	86	21.5	21.75
2010	13	15	84	21	21.25
2010	14	32	82	20.5	20.75
2010	15	23			
2010	16	12			

Calculate the seasonal variation for each Quarter

Where trend values are present, subtract the trend values for the observed values $y - t$. Format them in a table like what is given belwo.

	period	y-t	period	y-t	period	y-t	period	y-t
2007	1	...	2	...	3		4	
2008	5		6		7		8	
2009	9		10		11		12	
2010	9		10		11	...	12	...

Outline three reasons why seasonal variations should be measures

1.

2.

3.

May 2013 Question 6 Time Series Analysis

	Q_1	Q_2	Q_3	Q_4
2010	14	16	9	14
2011	16	17	12	17
2012	18	20	13	18

Calculate the trend using the moving averages method

Year	Time Period	Y	Moving Total	Moving Average	Centered MA
(column 1)	(column 2)	(column 3)	(column 4)	(column 5)	(column 6)
2008	1	14			
2008	2	16	53	13.25	
2008	3	9	55	13.75	13.5
2008	4	14	56	14	13.875
2009	5	16	59	14.75	14.375
2009	6	17	62	15.5	15.125
2009	7	12	64	16	15.75
2009	8	17	67	16.75	16.375
2010	9	18	68	17	16.875
2010	10	20	69	17.25	17.125
2010	11	13			
2010	12	18			

- Observed values are re-arranged in column 3. An identifier for each observation is listed in column 2.
- Compute the sum of each set of four observations.
  - The first calculation is the sum of observations 1 to 4.(column 4) We then divide this by four to find the moving average (column 5)

$$\frac{14+16+9+14}{4} = \frac{53}{4} = 13.25$$

  - The second calculation is the sum of observations 2 to 5.(column 4) We then divide this by four to find the moving average (column 5)

$$\frac{16+9+14+16}{4} = \frac{55}{4} = 13.75$$

- The second calculation is the sum of observations 3 to 6.(column 4) We then divide this by four to find the moving average (column 5)

$$\frac{9+14+16+17}{4} = \frac{56}{4} = 14$$

- We continue on in this fashion until we get to the last step.
- The last calculation is the sum of observations 9 to 12.(column 4) We then divide this by four to find the moving average (column 5)

$$\frac{18+20+13+18}{4} = \frac{69}{4} = 17.25$$

- Columns 4 and 5 are now completed.
- The ***trend*** is the average of each consecutive pairs of values in column 5.
  - The first value in the trend is the average of 13.25 and 13.75. i.e. 13.5
  - The second value in the trend is the average of 13.75 and 14. i.e. 13.875
- Once we have complete this column we have the trend.

Calculate the seasonal variation for each Quarter

Forecast the number of batteries made in each quarter of 2013

May 2012 Question 7 Price Indices

Some Theory

The two most basic formulae used to calculate price indices are the ***Paasche index*** (after the economist Hermann Paasche) and the ***Laspeyres index*** (after the economist Etienne Laspeyres).

The Paasche index is computed as

$$P_P = \frac{\sum(p_n \cdot q_n)}{\sum(p_0 \cdot q_n)}$$

while the Laspeyres index is computed as

$$P_L = \frac{\sum(p_n \cdot q_0)}{\sum(p_0 \cdot q_0)}$$

- p_0 and q_0 price and quantity at base period.

- p_n and q_0n price and quantity at period n.

A Laspeyres Index is known as a base-weighted or xed-weighted index because the price increases are weighted by the quantities in the base period.

Question

	P_0 (2009)	Q_0 (2009)	P_n (2010)	Q_n (2010)
A	36	100	40	95
B	80	12	90	10
C	45	16	41	18
D	5	1100	6	1200

Price Index Formulae

- **Laspeyres Price Index**

$$L_{PI} = \frac{\sum p_n \times q_0}{\sum p_0 \times q_0}$$

- **Paasche Price Index**

$$P_{PI} = \frac{\sum p_n \times q_n}{\sum p_0 \times q_n}$$

Laspeyres Price Index - Calculation

	p_0	q_0	p_n	q_n	$p_0 \times q_0$	$p_n \times q_0$
A	36	100	40	94	3600	4000
B	80	12	90	10	960	1080
C	45	16	41	18	720	656
D	5	1100	6	1200	5500	6600
				sum	10780	12336

$$L_{PI} = \frac{\sum p_n \times q_0}{\sum p_0 \times q_0} \times 100 = \frac{12336}{10780} \times 100 = 114.43$$

Paasche Price Index - Calculation

	p_0	q_0	p_n	q_n	$p_n \times q_n$	$p_0 \times q_n$
A	36	100	40	94	3760	3384
B	80	12	90	10	900	800
C	45	16	41	18	738	810
D	5	1100	6	1200	7200	6000
				sum	12598	10994

$$P_{PI} = \frac{\sum p_n \times q_n}{\sum p_0 \times q_n} \times 100 = \frac{12598}{10994} \times 100 = 114.59$$

Interpret both indices

Explain why the Laspeyres Index is considered a "pure" index, but not the Paasche index

- Laspeyres is "pure" in that it measures like with like from period to period whereas with Paasche's index, the weighting will change.

May 2013 Question 2 Price Indices

Comparison of Paasche and Laspeyres Indices

which is likely to give a bigger answer, and why?

- Laspeyres is "pure" in that it measures like with like from period to period whereas with Paasche's index, the weighting will change.

Calculation of Indices

First - we are given the following information

	P_0 (2009)	Q_0 (2009)	P_n (2010)	Q_n (2010)
A	50	200	60	300
B	80	100	100	200
C	100	300	120	400