Instructors: M. Pivovarova/M.Tanaka

Duration: 80 minutes.

Allowed aids: A non-programmable calculator and aid sheets provided.

Format: This test consists of two parts and a SCANTRON form. For both parts combined there are a total of 80 possible points.

ENTER YOUR NAME AND STUDENT # ON BOTH THE PINK SCANTRON FORM AND PART 1 BEFORE THE END OF THE EXAM IS ANNOUNCED.

Part 2: 14 multiple choice questions worth 2.5 points each for a total of 35 points

- A correct answer is worth 2.5 points and an incorrect answer is worth 0 points
- Answers must be properly recorded on the pink SCANTRON form to earn marks
- Print your **LAST NAME** and **INITIALS** in the boxes AND darken each letter in the corresponding bracket below each box; Sign your name in the **SIGNATURE** box
- Print your 9 digit **STUDENT NUMBER** in the boxes AND darken each number in the corresponding bracket below each box
- Your **FORM NUMBER** is **01**
- Use only a pencil or blue or black ball point pen
    - **Pencil strongly recommended** because it can be erased
- Make dark solid marks that fill the bubble completely
- Erase completely any marks you want to change
- Crossing out a marked box is not acceptable and is incorrect
- If more than one answer is marked then that question earns 0 points

►For Question (1)

(1) Roughly how many observations in a sample described below lie within 3 standard deviations from the mean?

Percentiles		Smallest		
1%	2	1		
5%	3	1		
10%	4	1	Obs	1160
25%	7	1	Sum of Wgt.	1160
50%	14		Mean	13.40086
		Largest	Std. Dev.	7.3311
75%	18	35		
90%	28	35		
95%	28	35		
99%	35	36		

(A) 1021
(B) 1102
(C) 1157
(D) 1159
(E) 1160

►For Question (2) Consider the following tabulation of variable X:

x	Freq.	Percent	Cum.
-4	7	5.47	5.47
-3	8	6.25	11.72
-2	18	14.06	25.78
-1	14	10.94	36.72
0	16	12.50	49.22
1	6	4.69	53.91
2	17	13.28	67.19
3	7	5.47	72.66
4	19	14.84	87.50
6	16	12.50	100.00
Total	128	100.00	

(2) What is the height of the bar for the density histogram of variable X which starts at -2.545 and ends at -0.545?

(A) 0.0781
(B) 0.1250
(C) 0.2083
(D) 0.2578
(E) 0.3672

(3) Since the mode is the most frequently occurring value of observations in the data, it:

(A) is always equal to the mean
(B) is always larger than the median
(C) must have value of at least two
(D) is always larger than the mean
(E) none of the above answers is correct

►**For Questions(4)-(5):** The study has been undertaken to learn about the short-term health effects of smoking. Each person in a random sample was asked to report the average number of cigarettes smoked per day (smoke) and the number of days absent from work due to colds last year (days).

(4) This is an example of _____study with _____data and _____variables.

(A) Experimental, cross-sectional, interval
(B) Observational, cross-sectional, interval
(C) Observational, panel, interval
(D) Experimental, panel, nominal
(E) Observational, cross-sectional, nominal

(5) Based on the data from the sample, the following least squares line (OLS) is estimated:

$$Days\_hut = 0.2 * Smoke + 7.3$$

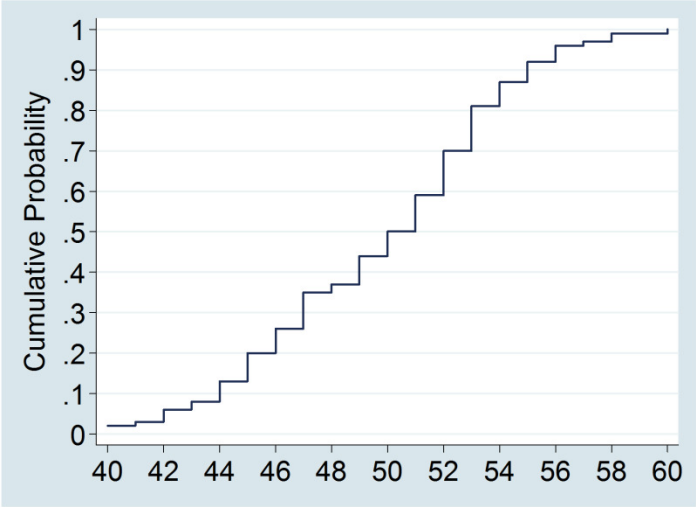
What is the correct interpretation of the OLS line?

(A) On average, individuals who smoke are more likely to be absent from work due to colds.
(B) On average, smoking one more cigarette per day increases the number of days absent from work due to colds per year by 0.2.
(C) Individuals who do not smoke are absent from work due to colds 7.3 days per year.
(D) Individuals who smoke 10 cigarettes per day are absent from work due to colds 9.3 days per year.
(E) All of the above.

(6) The population is bell-shaped, and roughly 95% of the population lies within 26 and 50. The mean and variance of the population are:

(A) 36 and 4
(B) 38 and 6
(C) 38 and 16
(D) 36 and 36
(E) 38 and 36

►**For Question (7):** Consider the following ogive of the sample:



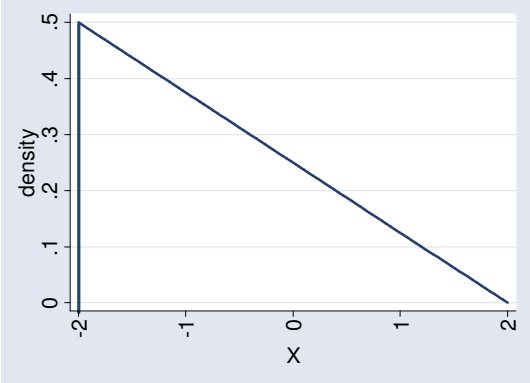
(7) Which of the following is NOT a true statement about these data:

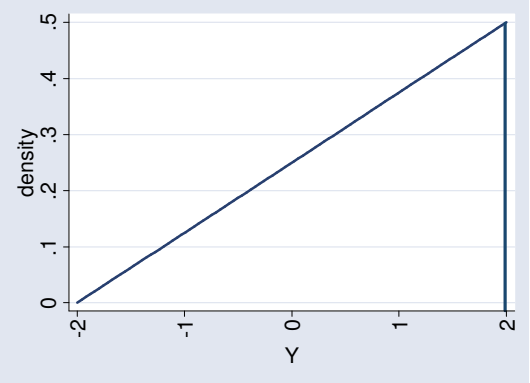
(A) The interquartile range is 7
(B) The 10^{th} percentile is 44
(C) The median is above 50
(D) The data are interval
(E) All of the above

(8) Variable Z has unimodal and positively skewed distribution. Which one of the statements about the distribution is true?

(A) There are fewer observations below mean than above.
(B) There are fewer observations above mean than below.
(C) Median is greater than range
(D) Median is equal to mean.
(E) None of above.

►**For Questions(9):** Consider following density histograms for variable X and Y.





I. X and Y have identical interquartile range.
II. X and Y have identical variance
III. X and Y have identical mean.
IV. X and Y have identical median.

(9) Which of above four statements is true?

(A) Only I
(B) Only II
(C) Only III
(D) I and II
(E) III and IV

(10) In a survey, 1500 individuals in three cities in Canada are asked if they have more than one credit card. The result shows that 37 percent of respondents own more than one credit card. Of all the respondents, 47 percent live in Toronto, 21 percent live in Montreal, and 32 percent live in Vancouver. If location of the respondent and the number of credit cards he/she owns are independent, what is the probability that a randomly chosen respondent lives in Vancouver and has more than one credit card?

(A) 0.86
(B) 0.37
(C) 0.32
(D) 0.20
(E) 0.12

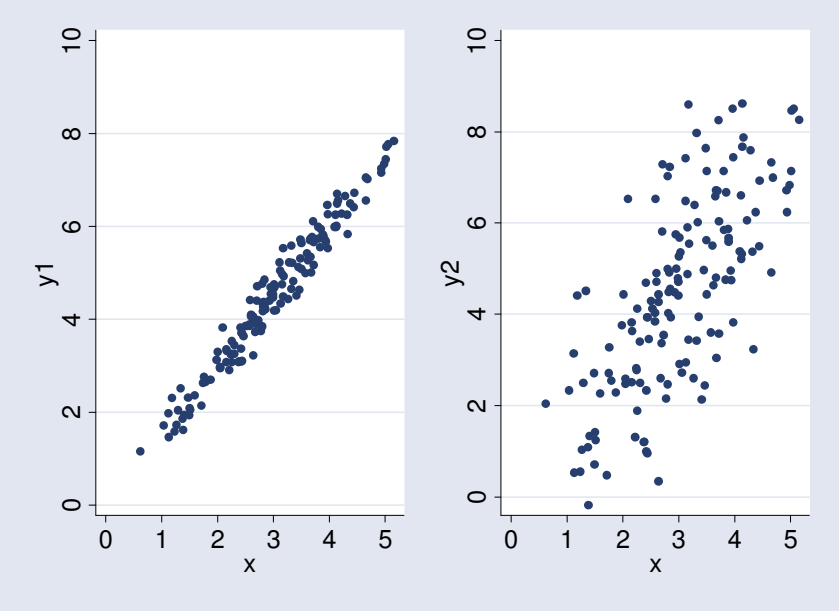
►For Questions(11): A retail chain is analyzing purchase patterns of customers in different branches in September 2010. The sample means and sample standard deviations of spending per customer per visit at each branch are summarized in the table below.

	Mean	S.D.
Store A	$33	$6
Store B	$205	$50
Store C	$ 15	$4

(11) Which statement about the variability of customer's spending is true?

(A) The variation of the spending is the highest at store A.
(B) The variation of the spending is the highest at store B.
(C) The variation of the spending is the highest at store C.
(D) The variation of the spending is the same for all stores.
(E) The variation of the spending is the same for store A and B.

►**For Questions(12):** Consider following two graphs. Covariance of x and y1 is identical to covariance of x and y2.



(12) Which one of these statements best describes the difference between two graphs?

(A) variance of y1 is larger than variance of y2.
(B) mean of y1 is larger than mean of y2.
(C) variance of y1 is smaller than variance of y2.
(D) mean of y1 is smaller than mean of y2.
(E) median of y1 is larger than median of y2.

(13) Which one of the following statements about the nominal data is true?

(A) The median is often used to describe central tendency of nominal data.
(B) Calculating mean of nominal data is pointless.
(C) A box plot is a useful way of describing nominal data.
(D) The range is often used to describe variability of nominal data.
(E) All of above.

(14) Coefficient of correlation for variables X_1 and X_2 are -0.6. It implies that :

(A) there is negative linear relationship between X_1 and X_2.
(B) there is no nonlinear relationship between X_1 and X_2.
(C) Increase in X_2 is associated with increase in X_1 by 0.6 unit.
(D) there is no relationship between X_1 and X_2.
(E) Coefficient of determination is negative for X_1 and X_2.

Extra space for rough work:

Your form number is 01. Make sure to properly mark your name, student number, signature, and answers for Part 2 on the pink SCANTRON form before the end of the exam is announced. No extra time is permitted.

Population Mean: $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$

Sample Mean: $\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$

Population Variance: $\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$

Sample Variance: $s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{X})^2}{n-1} = \frac{1}{n-1}[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}]$

Population s.d.: $\sigma = \sqrt{\sigma^2}$

Sample s.d.: $s = \sqrt{s^2}$

Population coefficient of variation: $CV = \frac{\sigma}{\mu}$

Sample coefficient of variation: $cv = \frac{s}{\overline{X}}$

Population covariance: $\sigma_{XY} = \frac{\sum_{i=1}^{N} (x_i - \mu_X)(y_i - \mu_Y)}{N}$

Sample covariance: $s_{XY} = \frac{\sum_{i=1}^{n} (x_i - \overline{X})(y_i - \overline{Y})}{n-1}$

Population coefficient of correlation: $\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$

Sample covariance (shortcut): $s_{XY} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}\right]$

Sample coefficient of correlation: $r = \frac{s_{XY}}{s_X s_Y}$

Conditional Probability:

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$$

Complement rule:

$$P(A^C) = 1 - P(A) \qquad P(A^C \mid B) = 1 - P(A \mid B)$$

Multiplication Rule:

$$P(A \text{ and } B) = P(A \mid B) P(B)$$

Addition Rule:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

Expected Value: $E[X] = \mu = \sum_{all\ x} x p(x)$

Variance: $V[X] = E[(X-\mu)^2] = \sigma^2 = \sum_{all\ x} (x-\mu)^2 p(x)$

Covariance: $E[(X-\mu_X)(Y-\mu_Y)] = \sigma_{XY} = \sum_{all\ x} \sum_{all\ y} (x-\mu_X)(y-\mu_Y) p(x,y)$

Least squares line / linear regression line: $\hat{y} = a + bx \qquad b = \frac{\mathrm{cov}(x,y)}{s_x^2} \qquad a = \overline{y} - b\overline{x}$