Name:_____

STAT614.02 - Summer 2022
Final Exam

1. *One important theme in statistics is to study a possible relationship between two variables, say both are numerical. Sometimes, we get lucky and the simplest type of the models, namely a linear one, can approximate the behaviour well enough.*
   Consider the following data set, where eleven cereal brands ($\blacksquare_i$) are surveyed and x denotes the cost per box (in USD) and y denotes the amount of fiber per cup (in grams).

	$\blacksquare_1$	$\blacksquare_2$	$\blacksquare_3$	$\blacksquare_4$	$\blacksquare_5$	$\blacksquare_6$	$\blacksquare_7$	$\blacksquare_8$	$\blacksquare_9$	$\blacksquare_{10}$	$\blacksquare_{11}$
x	10.0	8.0	13.0	9.0	11.0	14.0	6.0	4.0	12.0	7.0	5.0
y	8.04	6.95	7.58	8.81	8.33	9.96	7.24	3.45	10.84	4.82	5.68

   a) *Let's start with summarizing and standardizing data.* Find the sample mean $\bar{x}$ and the sample standard deviaton s_x. Compute z-scores for the second and fourth individuals of the sample and fill the table below:

	$\blacksquare_1$	$\blacksquare_2$	$\blacksquare_3$	$\blacksquare_4$	$\blacksquare_5$	$\blacksquare_6$	$\blacksquare_7$	$\blacksquare_8$	$\blacksquare_9$	$\blacksquare_{10}$	$\blacksquare_{11}$
x	10.0	8.0	13.0	9.0	11.0	14.0	6.0	4.0	12.0	7.0	5.0
$\bar{x}$	?										
s_x	?										
z_x	0.302	?	1.206	?	0.603	1.508	-0.905	-1.508	0.905	-0.603	-1.206

b) *Since you're at it, find out if there are any outliers:* Determine the five-number summary for x:

10 8 13 9 11 14 6 4 12 7 5

?	?	?	?	?
minimum	Q1 (first quartile)	Q2 (median)	Q3 (third quartile)	maximum

(i) What is the range?

(ii) What is the interquartile range (IQR)?

(iii) Are there any outliers (mild or extreme)?

c) *We are ready to move on.* Complete and use the following table to compute the Pearson's correlation coefficient r.

What is the level of linear correllation between x and y? $\left(|r| \in \begin{cases} [0, 0.5) : \text{weak} \\ [0, 0.8) : \text{moderate} \\ [0.8, 1] : \text{strong} \end{cases} \right)$

	☕$_1$	☕$_2$	☕$_3$	☕$_4$	☕$_5$	☕$_6$	☕$_7$	☕$_8$	☕$_9$	☕$_{10}$	☕$_{11}$
z_x	0.302	?	1.206	?	0.603	1.508	-0.905	-1.508	0.905	-0.603	-1.206
z_y	0.265	-0.271	0.039	0.644	0.408	1.210	-0.128	-1.595	1.643	-1.320	-0.896
$z_x \cdot z_y$	0.080	?	0.047	?	0.246	1.825	0.116	2.405	1.487	0.796	1.081

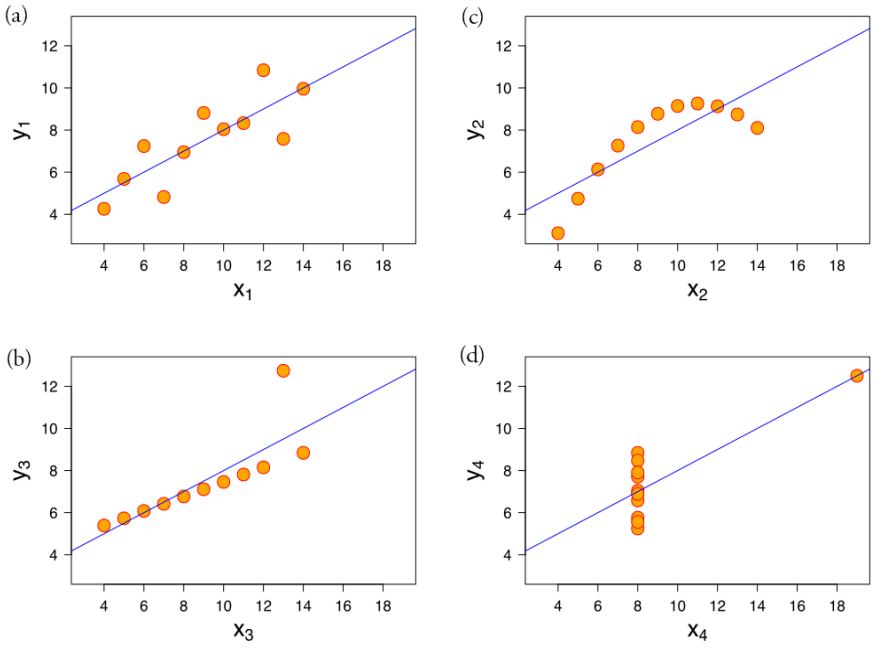
d) *Well, not perfect, but still workable.* Find the equation of the least-square line $\hat{y} = a + bx$, where $\bar{y} = 7.50$ and $s_y = 2.032$.

e) Use the equation of the regression line to answer the following questions:

(i) How much fiber (per cup) would a box of cereal would contain, if it costs \$7.5 per box?

(ii) On average, how much more (or less) fiber correspond to one dollar increase in the price of cereals?

f) *Statiscian F. Anscombe constructed in 1973 four data sets with the same* $\bar{x}$, $\bar{y}$, s_x, s_y, r, *(hence* R^2*) and* $\hat{y} = a + bx$. *Those four data sets had the following scatterplots:*



Fill in the blanks: _____ is the scatterplot for our data set, ___ has an outlier in y direction *(which needs to be removed from the sample for a more accurate model)*, ___ has an influential observation *(which needs to be removed from the sample for a more accurate model)*, and _____ can be described much better by a nonlinear model.

2. Circle either T(rue) or F(alse).

T / F If the data for a variable is numerical and finite, then the corresponding variable is a discrete variable.

T / F The difference between a bar chart and a histogram is that one can do mathematics on the horizontal axis of a histogram, but not on the horizontal axis of a bar chart.

T / F If you're given a relative frequency distribution and the sample size, you can construct the frequency distribution.

T / F The population standard deviation is always greater than or equal to the sample standard deviation, as the population contains all individuals of the sample and more.

T / F The population range is always greater than or equal to the sample range, as the population contains all individuals of the sample and more.

T / F z-scores cannot be computed if the distribution is not approximately normal.

3. As part of a study described in the report "I Can't Get My Work Done!", each person in a sample of 258 cell phone users age 20 to 39 was asked if they use their cell phones to stay connected while they are in bed. The same question was also asked of each person in a sample of 129 cell phone users age 40 to 49. 168 cell phone users in the first group said that they sleep with their cell phones, whereas 61 cell phone users from the second group said that they sleep with their cell phones.

   a) Find $\hat{p}_1$ and $\hat{p}_2$ (*where p denotes the proportion of cell phone users that stay connected while they are in bed*).

   b) Construct a 90% confidence interval for $p_1 - p_2$. (*Assume that the samples are representative of the corresponding populations, they are independent and one can easily check that the sample sizes are large enough by verifying that $n_1\hat{p}_1, n_1(1-\hat{p}_1), n_2\hat{p}_2$ and $n_2(1-\hat{p}_2)$ being all greater than 10).*

4. The Economist collects data each year on the price of a Big Mac in various countries around the world. The price of a Big Mac for a sample of McDonald's restaurants in Europe in January 2014 resulted in the following Big Mac prices (after conversion to U.S. dollars):

   5.18, 4.95, 4.07, 4.68, 5.22, 4.67, 4.14, 4.98, 5.15, 5.56, 5.36, 4.60

   The mean price of a Big Mac in the U.S. in January 2014 was \$4.62. Assume it is reasonable to regard the sample as representative of European McDonald's restaurants. Does the sample provide convincing evidence that the mean January 2014 price of a Big Mac in Europe is greater than the reported U.S. price? Test the relevant hypotheses using $\alpha = .05$.
   To make your life easier, $s = \$0.462$ and $\bar{x} = \$4.88$.