1. [$10 \times 1 = 10$ marks] Multiple choice questions. Circle the letter correponding to the correct answer.

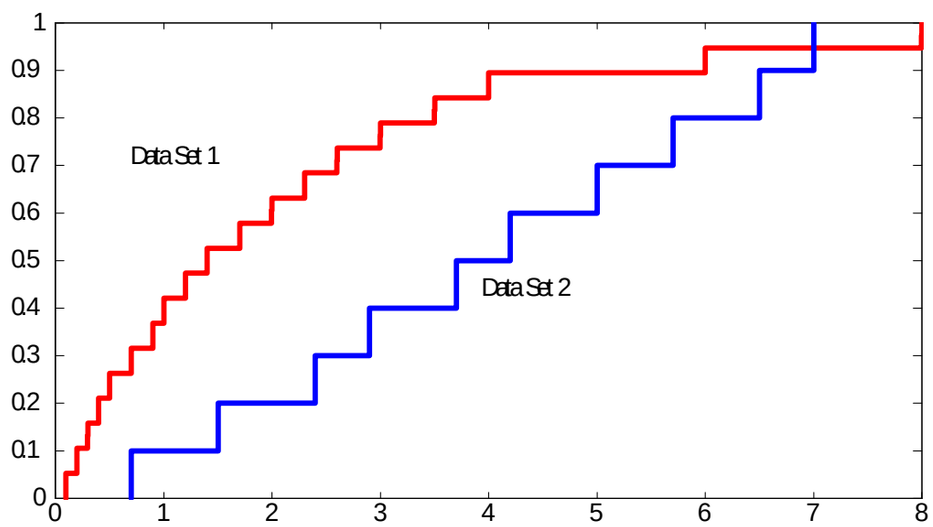
(a) Which one of the following situations can be appropriately modeled using a Binomial distribution?

A: time between arrivals of buses at a bus stop
B: number of tosses needed until we get 10 Heads in total when tossing a coin
C: closing price in dollars of a stock
D: number of calls received by a call center in one hour
E: number of people in a sample drawn at random from a large population that have a certain disease

(b) Which one of the following statements is **FALSE**?

A: Pie charts and bar charts are suitable for representing categorical data.
B: In a relative frequency histogram the height of each rectangle is equal to k times the number of observations in the interval for some positive constant k.
C: The sample variance of a data set cannot be determined from a boxplot.
D: A run chart is a good way to summarize data collected over time.

(c) In the graph below the empirical cumulative distribution function is graphed for two different data sets. The observations in each data set are unique.



Which one of the following statements is **TRUE**?

A: There are more observations in Data Set 2 than in Data Set 1.
B: All the values in both data sets are positive.
C: For Data Set 2, $\hat{F}(5) = 0.6$.
D: The skewness of Data Set 1 is negative.

(d) Which one of the following statements is **FALSE**?

A: $L(\theta)$ and $l(\theta) = \log L(\theta)$ are maximized for the same value of θ.
B: $L(\theta)$ and $l(\theta) = \log L(\theta)$ have the same concavity near their maximum value.
C: $L(\theta)$ and $l(\theta) = \log L(\theta)$ have the same shape.
D: $l(\theta) = \log L(\theta)$ is a one-to-one function of $L(\theta)$.

(e) Which one of the following statements is **FALSE**?

A: If y successes are observed in n Bernoulli trial with $P(\text{Success}) = \theta$ then the maximum likelihood estimate of θ is $\hat{\theta} = y/n$.

B: For an observed random sample $y_1, y_2, \ldots, y_n$ from a Poisson(θ) distribution the maximum likelihood estimate of θ is $\hat{\theta} = \bar{y}$.

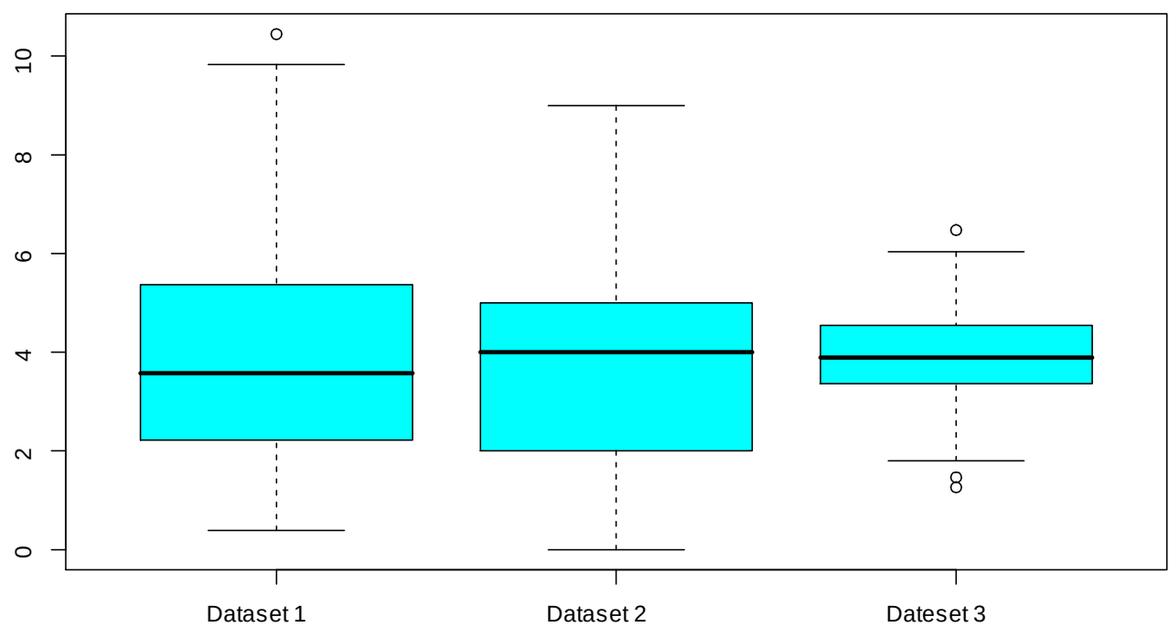
C: For an observed random sample $y_1, y_2, \ldots, y_n$ from a Exponential(θ) distribution the maximum likelihood estimate of θ is $\hat{\theta} = \bar{y}$.

D: For an observed random sample $y_1, y_2, \ldots, y_n$ from a $G(\mu, \sigma)$ distribution the maximum likelihood estimate of $\theta = (\mu, \sigma^2)$ is $\hat{\theta} = \left(\hat{\mu}, \hat{\sigma}^2\right) = \left(\bar{y},\ \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2\right)$.

(f) Suppose a and b are positive constants. The function $G(\theta) = \theta^a (1-\theta)^b$, $\theta > 0$ is maximized for:

A: $\frac{a}{a+b}$
B: $\frac{a}{b}$
C: $\frac{b}{a}$
D: b
E: None of the above.

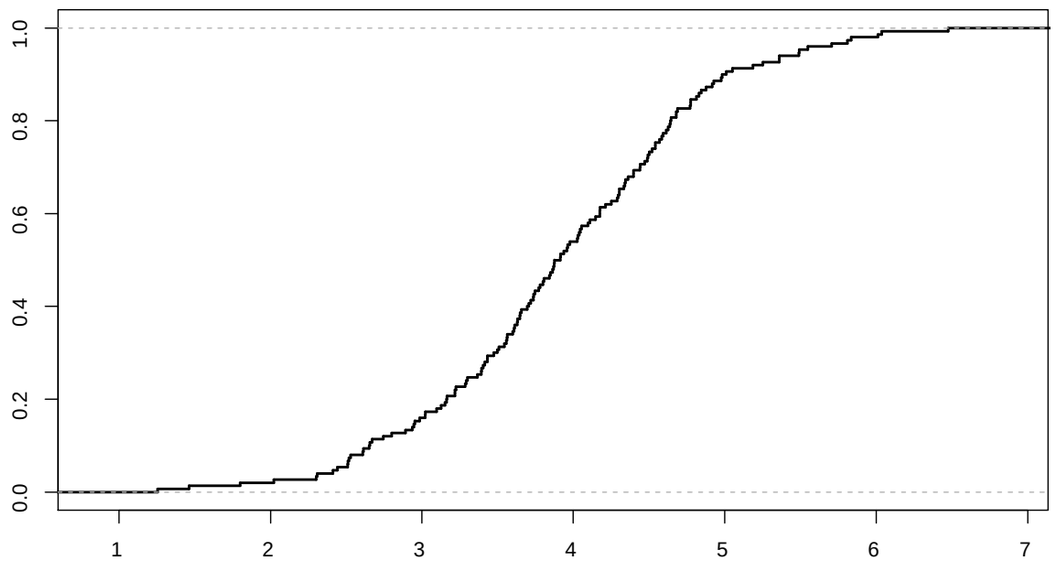
(g) In the figure below are boxplots for 3 different datasets. Assume all 3 datasets are unimodal.



Which one of the following statements is **FALSE**?

A: Dataset 3 has the smallest varibility.
B: Dataset 1 is negatively skewed.
C: Dataset 3 is the most bell-shaped.
D: Dataset 2 has the largest sample median
E: Dataset 1 has the largest range.

(h) Which of the following commands in R would produce the following plot for the variable y?



A: ecdf(y)
B: hist(y)
C: boxplot(y)
D: plot(ecdf(y))
E: None of the above.

(i) Consider the following R console output:

fivenum(y)
[1] 0.0013 0.2805 0.6747 1.3374 8.6917

The IQR of the variable y is given by:

A: 0.6747
B: 1.0568
C: 8.6904
D: None of the above.

(j) The correlation between two variables x and y can be computed in R using which of the following commands:

A: cov(x, y)
B: cor(x, y)
C: cov(x, y)(var(x)*var(y))
D: cor(x, y)/(var(x)*var(y))
E: None of the above.

2. [10 marks] In modelling the number of transactions of a certain type received by a central computer for a company with many on-line terminals the Poisson distribution can be used. If the transactions arrive at random at the rate of θ per minute then the probability of y transactions in a time interval of length t minutes is

$$P(Y = y; \theta) = f(y; \theta) = \frac{(\theta t)^y}{y!} e^{-\theta t} \quad \text{for } y = 0, 1, \ldots \text{ and } \theta > 0. \tag{1}$$

(a) [5] Suppose $y_1, y_2, \ldots, y_n$ were the number of transaction recorded in n independent $t = 1$ minute intervals. Find the the maximum likelihood estimate of θ based on the model (1) and these data. Clearly show all your steps.

Since $t = 1$, the log likelihood function is

$$L(\theta) = \prod_{i=1}^{n} \frac{\theta^{y_i}}{y_i!} e^{-\theta} = \left(\prod_{i=1}^{n} \frac{1}{y_i!} \right) \theta^{\sum_{i=1}^{n} y_i} e^{-n\theta} \quad \text{for } \theta > 0$$

or more simply (ignoring constants with respect to θ)

$$L(\theta) = \theta^{\sum_{i=1}^{n} y_i} e^{-n\theta} = \theta^{n\bar{y}} e^{-n\theta} \quad \text{for } \theta > 0 \text{ since } n\bar{y} = \sum_{i=1}^{n} y_i$$

The log likelihood function is

$$l(\theta) = n\bar{y} \log \theta - n\theta$$

Solving

$$l'(\theta) = \frac{n\bar{y}}{\theta} - n = \frac{n\bar{y} - n\theta}{\theta} = \frac{n(\bar{y} - \theta)}{\theta} = 0$$

gives the maximum likelihood estimate

$$\hat{\theta} = \bar{y}$$

(b) [4] Suppose that for $n = 200$ independent $t = 1$ minute intervals the observed frequencies were those given in the table below. Assuming $\theta = \hat{\theta} = 2.1$, complete the following table of expected frequencies. Comment on how well the model fits the data.

	0	1	2	3	4	≥ 5	Total
Observed Frequency	28	45	56	40	21	10	200
Expected Frequency	24.491	51.432	54.003	37.802	19.846	12.425	200

$$e_j = 200 \cdot \frac{(2.1)^j}{j!} e^{-2.1} \quad \text{for } j = 0, 1, 2, 3, 4, 5$$

The agreement between the observed and expected frequencies seems quite good. The model appears to fit the data well.

(c) [1] What is the maximum likelihood estimate of the probability that during a 2 minute interval there are no transactions?

By the Invariance Property of maximum likelihood estimates, the maximum likelihood estimate of the probability that during a $t = 2$ minute interval there are no transactions is

$$\frac{\left(2\hat{\theta}\right)^0}{0!} e^{-2\hat{\theta}} = e^{-2(2.1)} = e^{-4.2} = 0.015$$

3. [10 marks] Suppose $y_1, y_2, \ldots, y_n$ is an observed random sample from the $G(0, \sigma)$ distribution with probability density function

$$f(y;\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-y^2/(2\sigma^2)} \text{ for } y \in \Re \text{ and } \sigma > 0.$$

(a) [5] Find the likelihood function $L(\sigma)$ and the maximum likelihood estimate $\hat{\sigma}$ based on the observed data $y_1, y_2, \ldots, y_n$. Clearly show all your steps.

The likelihood function is

$$\begin{aligned} L(\sigma) &= \prod_{i=1}^{n} f(y_i;\sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-y_i^2/(2\sigma^2)} \quad \text{for } \sigma > 0 \\ &= (2\pi)^{-n/2} \frac{1}{\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2\right). \end{aligned}$$

or more simply

$$L(\sigma) = \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2\right) \quad \sigma > 0.$$

The log likelihood is

$$l(\sigma) = -n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2 \quad \sigma > 0.$$

Solving

$$l'(\sigma) = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n} y_i^2 = \frac{1}{\sigma^3}\left(-n\sigma^2 + \sum_{i=1}^{n} y_i^2\right) = 0$$

gives the maximum likelihood estimate

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2}$$

(b) [3] Show that the relative likelihood function $R(\sigma)$ is given by

$$R(\sigma) = \left(\frac{\hat{\sigma}}{\sigma}\right)^n \exp\left\{\frac{n}{2}\left[1 - \frac{(\hat{\sigma})^2}{\sigma^2}\right]\right\} \quad \text{for } \sigma > 0.$$

Note: $\exp(x) = e^x$

$$\begin{aligned} R(\sigma) &= \frac{L(\sigma)}{L(\hat{\sigma})} = \frac{\sigma^{-n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2\right)}{\hat{\sigma}^{-n}\exp\left(-\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n} y_i^2\right)} \quad \sigma > 0 \\ &= \left(\frac{\hat{\sigma}}{\sigma}\right)^n \frac{\exp\left(-\frac{1}{2\sigma^2}n\hat{\sigma}^2\right)}{\exp\left(-\frac{n}{2}\right)} \quad \text{since } n\hat{\sigma}^2 = \sum_{i=1}^{n} y_i^2 \\ &= \left(\frac{\hat{\sigma}}{\sigma}\right)^n \exp\left\{\frac{n}{2}\left[1 - \frac{(\hat{\sigma})^2}{\sigma^2}\right]\right\} \end{aligned}$$

as required.

(c) [2] Suppose $\hat{\sigma} = 1.2$ for a given data set. If $Y \backsim G(0, \sigma)$ then determine the maximum likelihood estimate of $P(Y > 0.3; \sigma)$.

By the Invariance Property of maximum likelihood estimates, the maximum likelihood estimate of $P(Y > 0.1; \sigma)$ is

$$\begin{aligned} P(Y > 0.1; \hat{\sigma}) &= P\left(Z > \frac{0.3 - 0}{1.2}\right) \quad \text{where } Z \backsim G(0,1) \\ &= 1 - P(Z < 0.25) = 1 - 0.5987 = 0.4013 \approx 0.401 \end{aligned}$$

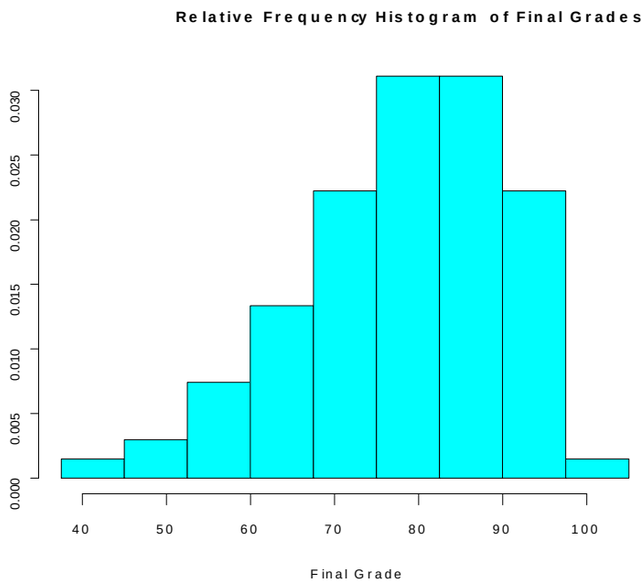
4. [15 marks] The data below are the final grades of 90 students in a second year statistics course:

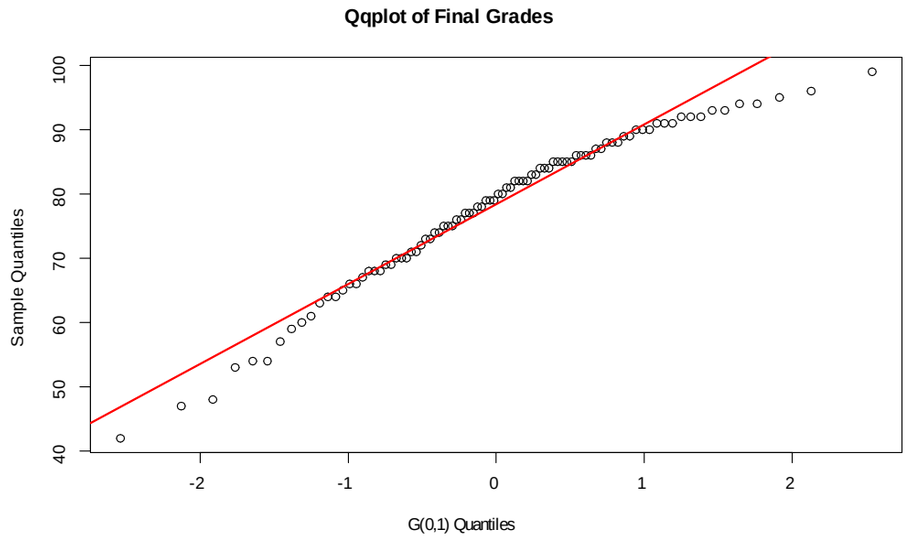
99	96	95	94	94	93	93	92	92	92	91	91	91	90	90	90	89	89	88	88
88	87	87	86	86	86	86	85	85	85	85	85	84	84	84	83	83	82	82	82
82	81	81	80	80	79	79	79	78	78	77	77	77	76	76	75	75	75	74	74
73	73	72	71	71	70	70	70	69	69	68	68	68	67	66	66	65	64	64	63
61	60	59	57	54	54	53	48	47	42										

For these data

$$\sum_{i=1}^{90} y_i = 6987, \quad \sum_{i=1}^{90} y_i^2 = 555863, \quad \text{and sample kurtosis} = 3.0211$$

A relative frequency histogram and qqplot for these data are given below:





Answer questions $(a)-(e)$ based on the information given on the previous page.

(a) [3] The five-number summary for these data is:

____42____, ____69.5____, ____79.5____, ____87____, ____99____.

$$\begin{aligned} q(0.25) &= \frac{1}{2}\left(y_{(22)}+y_{(23)}\right)=\frac{1}{2}(69+70)=69.5,\ \ q(0.5)=\frac{1}{2}\left(y_{(45)}+y_{(46)}\right)=\frac{1}{2}(79+80)=79.9 \\ q(0.75) &= \frac{1}{2}\left(y_{(68)}+y_{(69)}\right)=\frac{1}{2}(87+87)=87 \end{aligned}$$

(b) [2] For these data:

sample mean $=\bar{y}=$ ____77.633____ $\qquad \dfrac{6987}{90}=77.63333$

and

sample standard deviation $=s=$ ____12.288____ $\qquad s=\sqrt{\dfrac{1}{89}\left[555863-\dfrac{(6987)^2}{90}\right]}=12.28816$

(c) [1] For these data the sample skewness would be (Circle the letter correponding to your choice.):

A: negative
B: approximately zero
C: positive
D: not enough information to tell

(d) [2] For these data determine the proportion of observations in the interval $[\bar{y}-s,\bar{y}+s]$. Compare this with $P(Y\in[\mu-\sigma,\mu+\sigma])$ where $Y\backsim G(\mu,\sigma)$.

The proportion of observations in the interval $[\bar{y}-s,\bar{y}+s]=[65.345,\ 89.921]$ is $60/90=0.667$.
If $Y\backsim G(\mu,\sigma)$ then

$$\begin{aligned} P(Y\in[\mu-\sigma,\mu+\sigma]) &= P(|Y-\mu|\le\sigma)=P\left(\frac{|Y-\mu|}{\sigma}\le 1\right) \\ &= P(|Z|\le 1)=2P(Z\le 1)-1 \ \text{ where } \ Z\backsim N(0,1) \\ &= 2(0.84134)-1=0.68268 \\ &\approx 0.683 \end{aligned}$$

The proportion of observations in the interval (0.667) is slightly smaller than what would be expected for Gaussian data (0.683).

(e) [3] Find the interquartile range (IQR) for these data. Show that $IQR=1.349\sigma$ for data from a Gaussian distribution.

For these data the

$$IQR=87-69.5=17.5$$

To show that $IQR=1.349\sigma$ for Gaussian data we need to solve

$$0.5=P(|Y-\mu|\le c\sigma)=P\left(\frac{|Y-\mu|}{\sigma}\le c\right) \ \text{ for } c \text{ if } Y\backsim G(\mu,\sigma).$$

From $N(0,1)$ tables $P(|Z|\le c)=2P(Z\le c)-1=0.5$ holds if $c=0.6745$. Therefore

$$IQR=2(0.6745)\sigma=1.349\sigma$$

for Gaussian data.

(e) [4] Using both the numerical and graphical summaries for these data, assess whether it is reasonable to assume a Gaussian model for these data. You must support your conclusion with reasons. **Your reasons must be written in complete sentences and not in point form.**

For Gaussian data we expect the relative frequency histogram to be approximately symmetric. The relative frequency histogram for theses data is negatively skewed with a left tail.

For Gaussian data we expect the sample mean and median to be approximately equal. For these data the sample median $= 79.5 >$ mean $= 77.633$.

For Gaussian data we expect the sample kurtosis to be close to 3. The sample kurtosis for these data equals 3.0211 which is quite close to 3.

The points in the qqplot do not lie along a straight line. The shape of the qqplot is very U-shaped.

The proportion of observations in the interval $[\bar{y} - s, \bar{y} + s]$ (0.667) is slightly smaller than we would expect for Gaussian data (0.683).

For Gaussian Data we expect $1.349\sigma \approx 1.349s = 1.349\,(12.288) = 16.58$. For these data $IQR = 17.5$ which is larger than expected.

Based on these observations, the Gaussian model is not the best model for these data.

The relative frequency histogram suggests that a model which is negatively skewed would be more appropriate for these data.

Note: At least 4 points must be given which clearly indicate what is observed for these data and what is expected for Gaussian data.

5. [15 marks] Answer the question below based on the following article (condensed) which appeared in the Globe & Mail newspaper on January 22, 2014:

Early engagement key to getting girls into science careers, Canadian study says

Girls are almost three times more likely to consider careers in science, math and engineering if they participate in science fairs and summer camps – particularly in the early grades – according to a new Canadian report. The study by researchers at Mount Saint Vincent University in Halifax also suggests that good grades and teacher influence matters less than exposure to these outside-the-classroom activities.

The findings come at a time when governments are reaching out to young women in an effort to persuade them to consider the so-called STEM fields of learning – science, technology, engineering and mathematics – and organizations have stepped up their mentoring efforts. Learning experts say it is crucial to reach girls before their enthusiasm wanes and they drop science and math courses, which are optional in high school. "I think this is a wake-up call. We need to increase the engagement level, and we need to encourage it from a young age," said the study's lead investigator, Tamara Franz-Odendaal, an associate professor at the university.

Prof. Franz-Odendaal and her team surveyed about 600 students in Grades 7 through 9 last year from the provinces New Brunswick, Nova Scotia and Prince Edward Island. The data were collected using an online survey that students completed during school hours. They found girls who engaged in activities, such as science fairs, competitions and engineering summer camps, were 2.7 times more likely to consider a STEM career. For boys, the influence was statistically insignificant.

(a) [2] What type of study is this and why?

This is an obervational study because the researchers did not attempt to change or control any of the variates for the sampled units.

(b) [1] Define the Problem for this study.

The Problem is to examine the relationship between participation in activities such as science fairs, competitions and engineering camps and the likelihood of considering careers in science, math and engineering among students in the early grades.

(c) [1] Is the type of Problem descriptive, causative, or predictive? Explain why.

This is a causative type Problem since the researchers wanted to know whether participating in activities such as science fairs, competitions and engineering camps affected whether girls would be more likely to consider a STEM career.

(d) [2] What are the two most important variates in this study and what is their type?

One important variate is whether or not the student participated in activities such as science fairs, competitions and engineering camps. This is a categorical variate.

The other important variate was whether or not the student would conside a STEM career. This is a also a categorical variate.

(e) [1] Define a suitable target population for this study.

A suitable target population for this study is the set of all students in Grades 7 to 9 in the provinces of New Brunswick, Nova Scotia and Prince Edward Island.

OR

A suitable target population for this study is the set of all students in Grades 7 to 9 in Canada.

(f) [1] Define a suitable study population for this study.

A suitable study population for this study is the set of all students in Grades 7 to 9 in the schools chosen by the researchers in the provinces of New Brunswick, Nova Scotia and Prince Edward Island. The schools are not specified in the article but it would have been impossible for the researchers to go to every school in these 3 provinces.

OR

A suitable study population for this study is the set of all students in Grades 7 to 9 in the provinces of New Brunswick, Nova Scotia and Prince Edward Island.

(g) [2] Give a possible source of study error for this study in relation to your answers to to (e) and (f).

If the target population is the set all students in Grades 7 to 9 in the provinces of New Brunswick, Nova Scotia and Prince Edward Island and the study population is is the set of all students in Grades 7 to 9 in the schools chosen by the researchers in these provinces then a possible source of study error is that the students in the schools chosen by the researchers might be systematically different from the students in all schools. For example, it might be that the researchers only included schools in large cities and not schools in rural areas. Students in rural schools may have less access to science fairs, competitions and engineering camps.

OR

If the target population is the set all students in Grades 7 to 9 in Canada and the study population is is the set of all students in Grades 7 to 9 in the provinces of New Brunswick, Nova Scotia and Prince Edward Island then a possible source of study error is that the students in the the provinces of New Brunswick, Nova Scotia and Prince Edward Island might be systematically different from the students in Canada. For example, it might be that the schools in the provinces of New Brunswick, Nova Scotia and Prince Edward Island which have a much smaller population have less government funding for activities such as science fairs, competitions and engineering camps.

(h) [1] What information is given about the sampling protocol for this study?

The article indicates that the data were collected using an online survey and that the students completed the survey during school hours. No information is given about whether students were required to complete the survey or not.

(i) [2] Give a possible source of sample error for this study based on the information you have stated in (h).

A possible source of sample error is that the survey was a voluntary survey. It could be that students who completed the survey are students who are generally more engaged in all activities and therefore might also be more likely to engage in other activities such as science fairs, competitions and engineering camps as compared to the students in the study population who did not volunteer to complete the survey.

(j) [2] What type of numerical summary is the number 2.7 mentioned in the article?

The numerical summary 2.7 is a relative risk. It is the relative risk among girls of considering a STEM career in the group who participate in activities such as science fairs, competitions and engineering camps as compared to those who do not participate.