Statistics and data analysis 2019

Final Exam (Bet)

Guidelines

- There are **4** (**FOUR**) questions in the exam. You need to answer **all** of them (no choice).
- You can respond in English and/or Hebrew.
- Write the answers to the questions in exam notebooks. Don't use the exam printout.
- Justify all your answers. Even though many of the questions are not purely mathematical, you should mathematically explain your answers. You may assume results proven (or stated as a fact) in class or in the homework (unless the question instructs otherwise).
- Make sure you write in a clear and legible way. Grading will also depend on the clarity and not only on correctness.
- You can use the reference and formulae sheet as provided, including the standard normal table.
- Use normal approximation when appropriate and needed.
- You can use hand held calculators.
- No other auxiliary material can be used during the exam.
- The total time of the exam is 3 (three) hours.
- Good luck!

Question 1 (25 pts)

A. (6 pts)

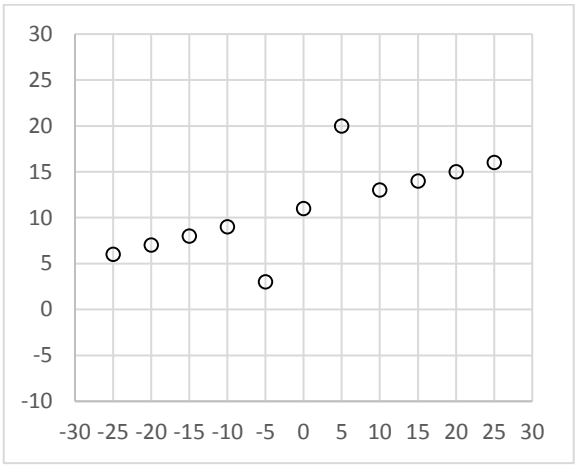
Consider the pairs of observed measurements below. There are three of them. Determine a matching between Pearson and Spearman correlation values in the rows of Table 1 below and the letter enumeration (A to C in Fig 1) of the depicted cases. Indicate the matching clearly in your notebook.

Table 1:

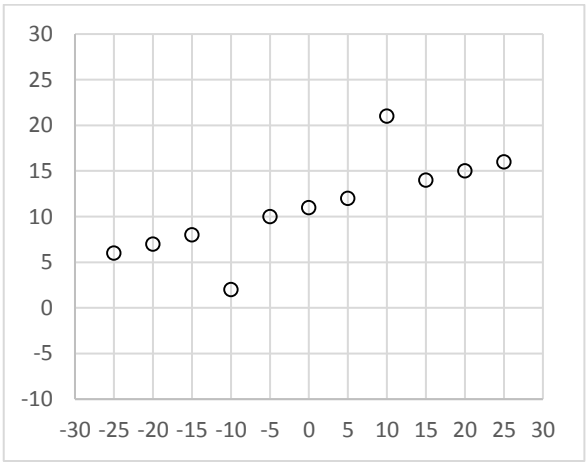
Number (to be matched to the figures)	Pearson correlation	Spearman correlation
1	0.87	0.98
2	0.75	0.82
3	0.79	0.89

Fig 1 (A-C):

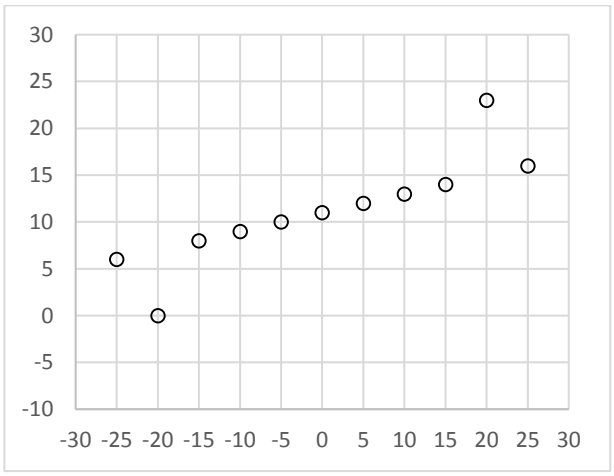
A



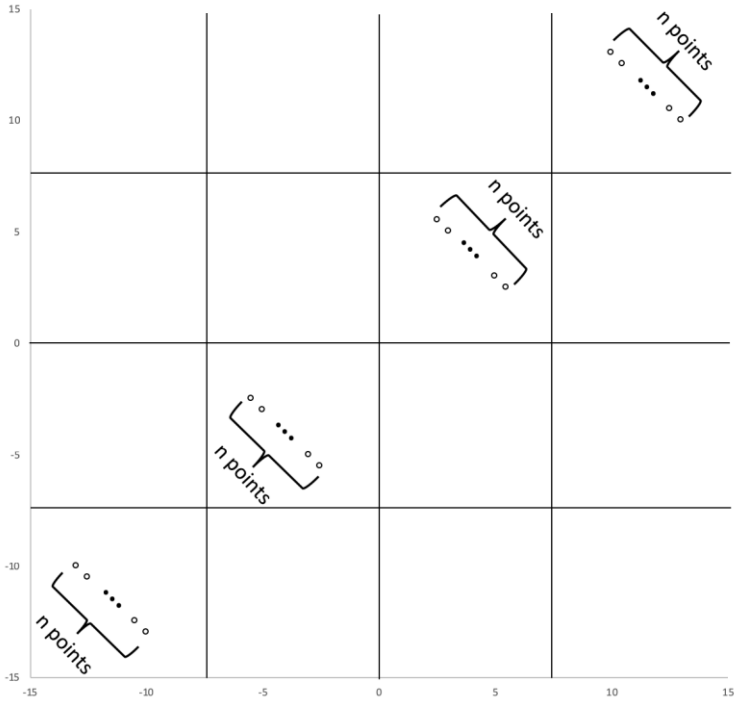
B



C



B. (7 pts) Consider the following dataset D(n), defined by the following picture:



Let $\tau(n)$ = Kendall correlation of the dataset D(n).
Find:

$$\lim_{n\to\infty} \tau(n)$$

Prove your answer.

C. (12 pts)

Safety tests were conducted for cars made in the Randomistan Opel factory and in the Germany Opel factory. Higher safety scores are better. The house statistician in the company had decided to declare the Randomistan cars safer if the WRS p-value of the observed data is better than 0.15.

For each one of the situations described below state whether Randomistan cars are declared safer.

In #3 also state what would happen if one were to use Student t-test rather than a WRS test.

Explain your calculations and answers.

1. (3 pts)

Randomistan	9.3	8.8	8.5		
Germany	9.1	8.2	8.1	8	7.9

2. (3 pts)

Randomistan	8.8	8.6	8.1		
Germany	9.2	9.1	9	8.9	5

3. (6 pts)

Randomistan	10	0.06	0.05					
Germany	9.95	9.9	9.85	0.04	0.03	0.02	0.01	0

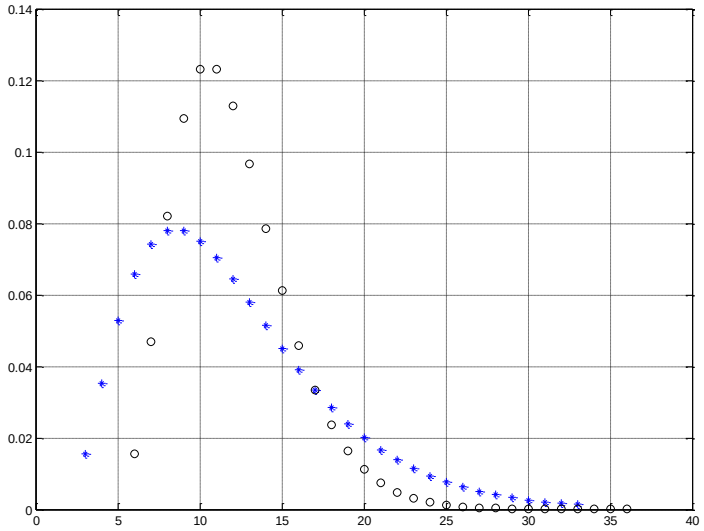
Question 2 (25 pts)

A.

1. (5 pts) Define two random variables X and Y that assume values on the non-negative integers so that:
   a. Both X and Y assume at least two values with non-zero probability (they are not constant)
   b. Let Z = X+Y. Then Z is uniformly distributed over the numbers {10, 11, 12, … , 101, 102}.
2. (5 pts) H(X) < 5? True or False? Explain.
3. (5 pts) H(Y) < 5? True or False? Explain.

B. Recall that the negative binomial distribution NegBinom(p,r) represents the number of times a coin with P(H) = p is tossed until the first time we see exactly r Hs.

1. (4 pts) Explain why if T ~ NegBinom(p,r) then
   $$P(T = t) = \binom{t-1}{r-1} p^r (1-p)^{t-r}$$
2. (3 pts) In the following figure we present the pdfs of two negative binomial distributions.
   One with r1 = 3 and p1 = ??
   The other with r2 = 6 and p2 = 0.5.
   Determine which is which (and explain your answer)



3. (3 pts) Given that the two distributions have the same mean, what is p1?

Question 3 (25 pts)

In this question $Poi(\lambda)$ stands for a Poisson distribution with mean λ.

Fred and Sid are repair technicians who work for Randobezeq – a phone company.
Fast Fred takes time which is $Poi(15)$ to repair a telephone line failure at a customer's home.
Slow Sid takes time which is $Poi(25)$ for the same task.

A. (5 pts) Fred is due to arrive to repair your phone at 10AM tomorrow. How confident can you be that he will be done by 10:10 if you know that he completes 1.8% of the cases in 7 minutes?
B. (5 pts) Given 2 Poisson distributions with the parameters λ_1, λ_2 and 2 mixture coefficients w_1, w_2, we define a Poisson mixture distribution, M, by writing its PDF:

$$P(i) = w_1 P_1(i) + w_2 P_2(i)$$

Prove that:

$$E(M) = w_1\lambda_1 + w_2\lambda_2$$

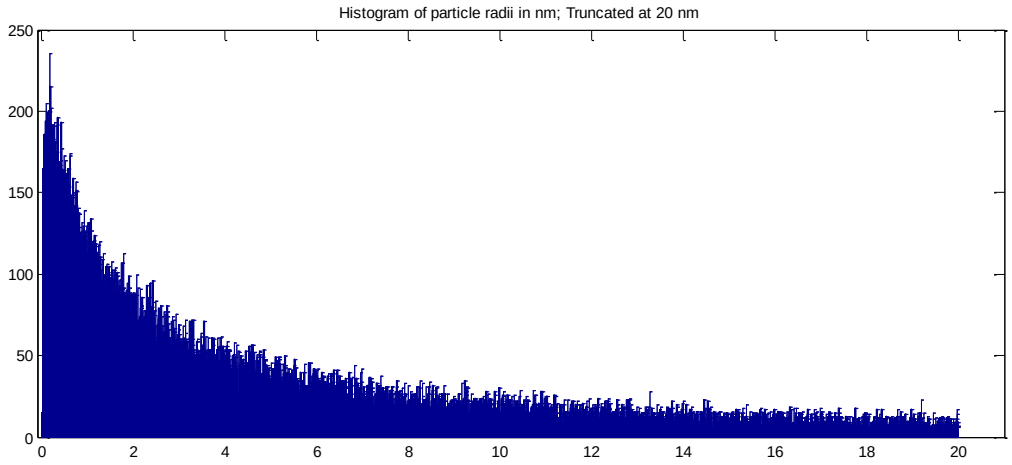
C. When a customer in North Randomistan orders a repair, then the distribution of the repair time is a Poisson mixture with mean = 21.
    1. (5 pts) What is the probability that following a call in North Randomistan Fred is providing the service?
    2. (10 pts) Under the condition of this question we can calculate that Fred completes 98% of the cases in 23 minutes or less and Sid completes 38.4% of the cases in 23 minutes or less.
    If a repair in North Randomistan starts at 10AM, which of the following times is the earliest time by which the customer can assume, with a 47% certainty, that the repair will already be done?
    State only one of the following options in your notebook and then justify and explain your answer.
    Options:

    10:10
    10:15
    10:17
    10:20
    10:23
    10:29

Question 4 (25 pts)

This question has 5 parts numbered A-E.

A scientist is generating nanoparticles for an experiment. She observes the following distribution of particle radii, in nms (nano-meters):

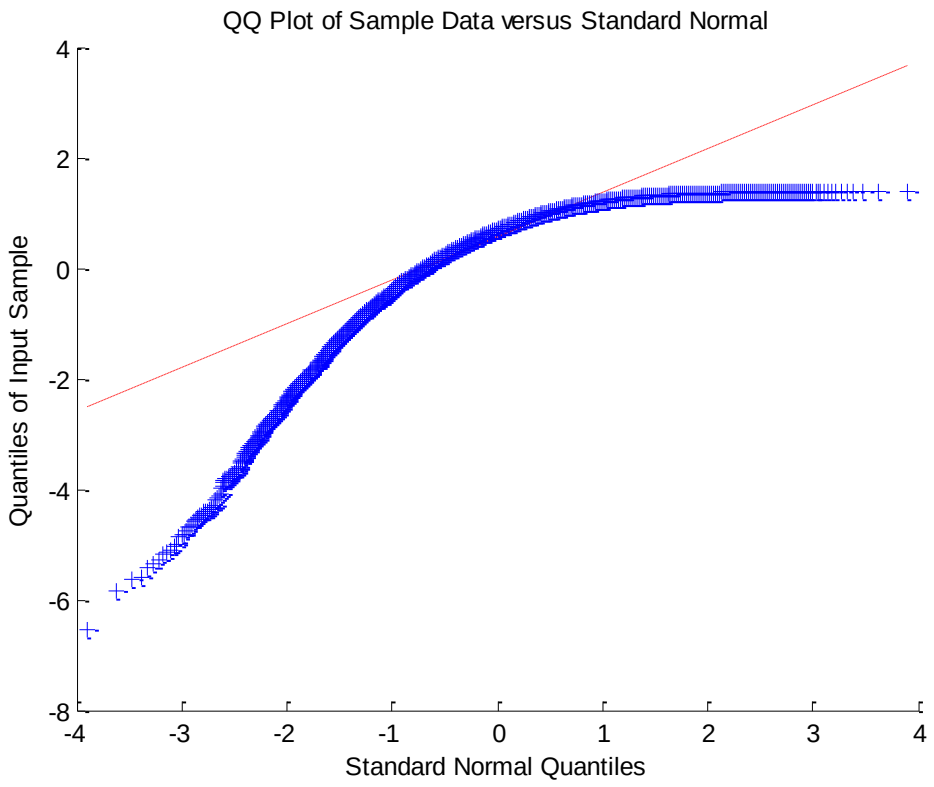


This histogram representation of the distribution is calculated from 100K particles. The x-axis units are nms. The histogram is truncated at 20 nm. 30687 particles of the 100K measured had radius ≥ 20 nm.

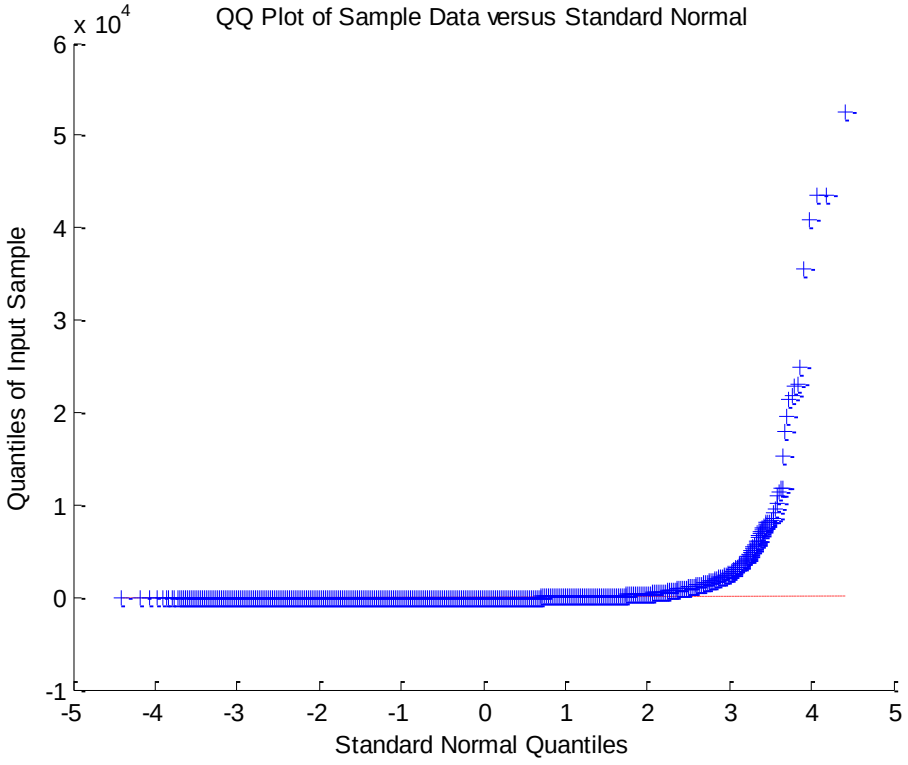
A. (5 pts)
   For the above data representing 100K particles, the scientist calculated empirical statistics.
   The empirical mean of the data is $emp - mean = e^4$ nm
   The empirical standard deviation is $emp - std = \sqrt{e^{12} - e^8}$ nm.
   The empirical median of the distribution is at e^2 nm.
   Let R denote the random variable that represents radii of the particles generated by the scientist.
   What do you think the distribution of R is? Explain your answer.

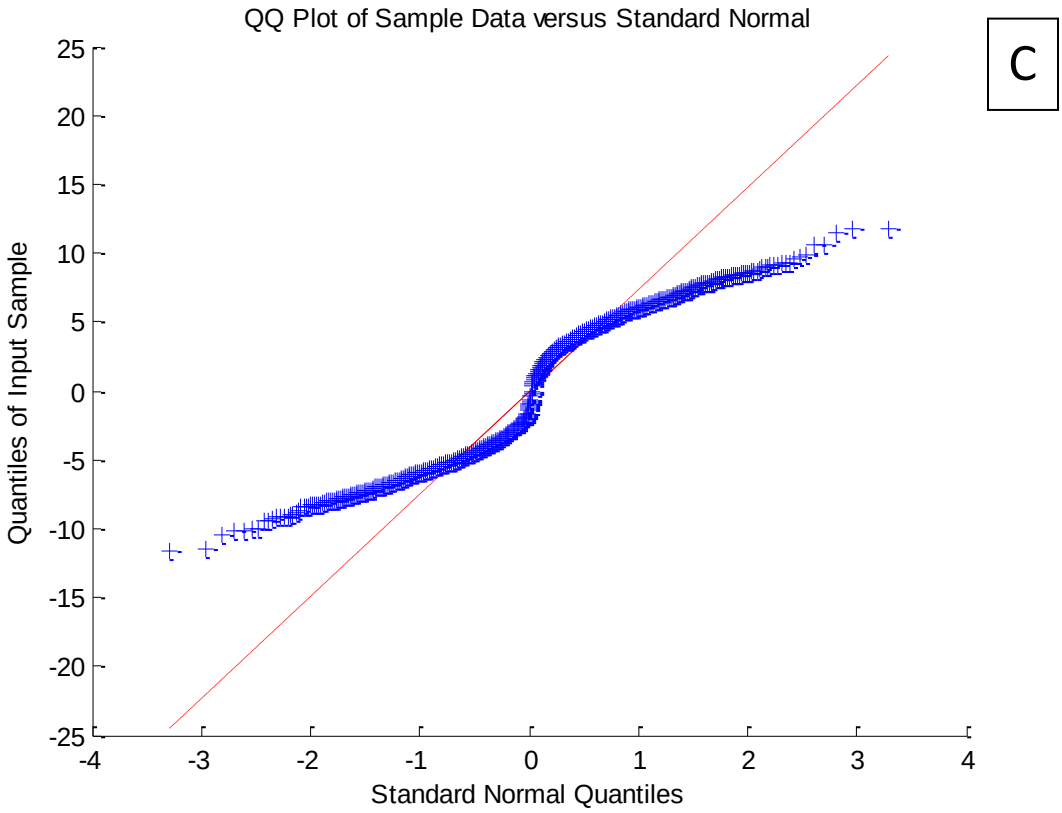
B. (5 pts)
   The scientist produced QQ plots for her data against the standard normal distribution. Amongst the 4 plots (marked A-D) in the next two pages, indicate which one (if any) corresponds to a QQ-plot of the quantiles of log(R) and which one (if any) corresponds to a QQ-plot of the quantiles of R. Explain your answer.

A

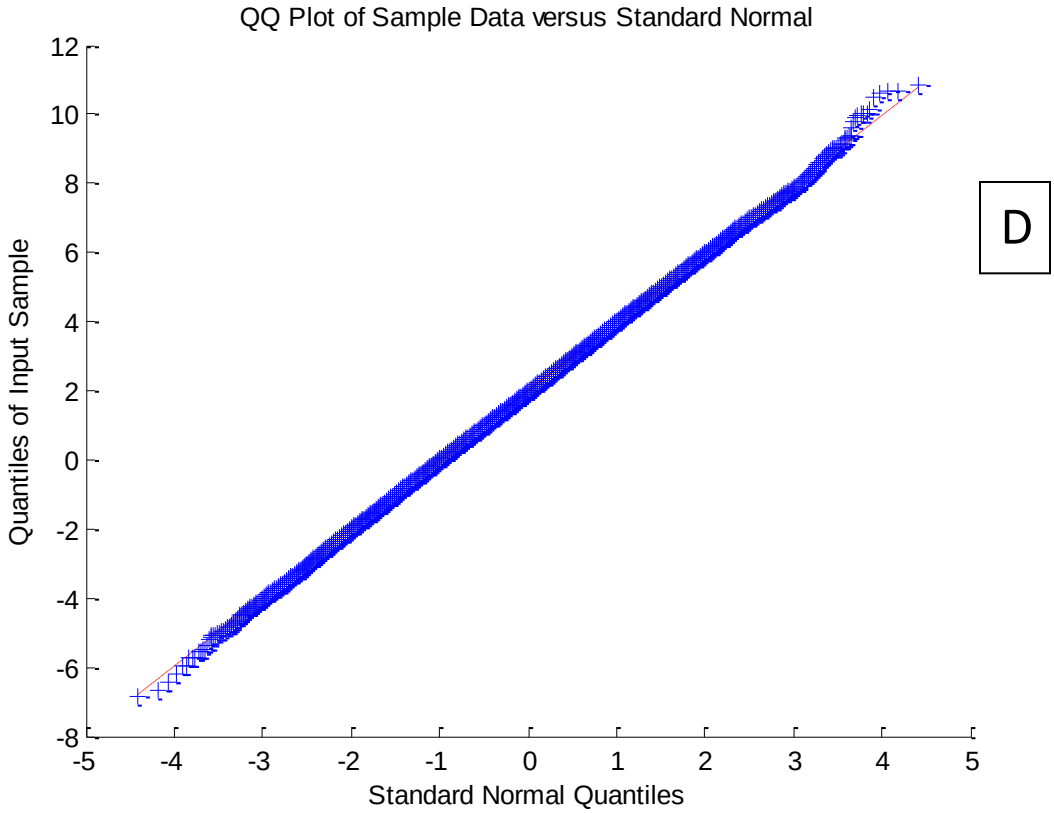

QQ Plot of Sample Data versus Standard Normal
Quantiles of Input Sample
Standard Normal Quantiles
4
2
0
-2
-4
-6
-8
-4
-3
-2
-1
0
1
2
3
4

B


QQ Plot of Sample Data versus Standard Normal
x 10^4
Quantiles of Input Sample
Standard Normal Quantiles
6
5
4
3
2
1
0
-1
-5
-4
-3
-2
-1
0
1
2
3
4
5


QQ Plot of Sample Data versus Standard Normal
Quantiles of Input Sample
Standard Normal Quantiles
25
20
15
10
5
0
-5
-10
-15
-20
-25
-4
-3
-2
-1
0
1
2
3
4

C


QQ Plot of Sample Data versus Standard Normal
Quantiles of Input Sample
Standard Normal Quantiles
12
10
8
6
4
2
0
-2
-4
-6
-8
-5
-4
-3
-2
-1
0
1
2
3
4
5

D

C. (5 pts)
According to the model you have developed what is the radius r so that
# of particles with radius < r = 20000? (leave answer in exp notation if necessary)

D. (5 pts)
The experiment requires at most 10% of particles to have a radius larger than e^4 nm.
Show, based on your model, that the population generated here is therefore not adequate for the experiment.

E. (5 pts)
The scientist can treat the particles and decrease all particle radii.
A reasonably priced process will lead to all radii decreasing exactly $\sqrt{e}$ fold (a particle with radius r will have radius $r \cdot 1/\sqrt{e}$ after the treatment).
A more expensive process will lead to all radii decreasing exactly e fold (a particle with radius r will have radius $r \cdot 1/e$ after the treatment).
She consulted with her statistician colleague as to whether either of the treatments will solve the problem and specifically as to whether the less expensive one will do it.
What advice would you give in this case? Show all your calculations.