1. In a study the calcium content in wheat from a certain area for four different storage times was investigated.

   The data is included in the following table:

Storage Period	**Observations**						mean	st.dev.
0 month	58.75	57.94	58.91	56.85	55.21	57.30	57.49	1.37
1 month	58.87	56.43	56.51	57.67	59.75	58.48	57.95	1.33
2 month	59.13	60.38	58.01	59.95	59.51	60.34	59.55	0.90
4 month	62.32	58.76	60.03	59.36	59.61	61.95	60.34	1.46

   (a) (6 marks)
   Present a side–by–side boxplot for the calcium content in wheat stored 0 month and 4 month.

   You can do this! Comment and compare center, spread and shape

   (b) (7 marks)
   The incomplete ANOVA table is:

Source	**df**	**SS**	**MS**	**F**	**P-value**
Period	?	32.14	?	?	0.003
Error	?	32.90	?		
Total	?	?			
.					

   Give a complete ANOVA table for the data.

Source	**df**	**SS**	**MS**	**F**	**P-value**
Period	3	32.14	10.71	6.510	0.003
Error	20	32.90	1.645		
Total	23	65.04			
.					

   (c) (6 marks)
   Is there sufficient evidence to conclude that the mean calcium content is not the same for the different storage times? Test at a significance level of $\alpha = 0.05$.

   $H_0 : \mu_0 = \mu_1 = \mu_2 = \mu_4$ vs. $H_a : H_0$ is not true. $\alpha = 0.05$
   test statistic: $F_0 = 6.510, df_1 = 3, df_2 = 20$
   P-value=0.003 (from the ANOVA table)
   Decision: reject H_0, because the p-value is smaller than α.
   Conclusion: At significance level of 5% the data provide sufficient evidence to conclude that the mean calcium content is not the same for the different storage times.

(d) (6 marks)
Use a 95% confidence intervals to compare the mean calcium content in wheat stored for 0 month with the mean calcium content in wheat stored for 4 month. Comment on your result.
df=20.

$$(57.49 - 60.34) \pm 2.086 \cdot \sqrt{1.645}\sqrt{\frac{1}{6} + \frac{1}{6}}$$

If zero does not fall within the confidence interval, we can not conclude that we are 95% confident that μ_0 and μ_4 are not the same.

2. The following table gives the size of the living area (in square feet), x, and the selling price, y, of 10 residential properties.

x (sq. ft.)	y (thousand)
1360	178.5
1940	275.7
1750	239.5
1550	229.8
1790	195.6
1750	210.3
2230	360.5
1600	205.2
1450	188.6
1870	265.7

It is $\sum x_i = 17,290$
$\sum y_i = 2,349.4$
$\sum x_i y_i = 4,173,646$
$\sum x_i^2 = 30,477,100$
$\sum y_i^2 = 578,760$

(a) (4 marks)
Sketch a scatterplot for the two variables (use the back of the paper). Comment on the relationship between x and y.

The graph shows a strong positive linear relationship between the two variables.

(b) (4 marks)
Calculate Pearson Correlation Coefficient for x and y.

$r = 0.893$

What can be learned from this number?

that x and y are related in a strong, positive, linear fashion.

(c) (4 marks)
Give an estimate for the least squares line for x and y.

y = - 96.0 + 0.191 x

(d) (2 marks)
Give an interpretation of the slope in the context of this problem.

For every additional square foot in living space the cost go up in average by 0.191 thousand dollars.

(e) (2 marks)
Estimate the selling price of a residential property with a living area of 1500 sq.ft.

$y = 96.0 - 0.191 \cdot 1500 = ...$

3. A survey was conducted to investigate the interest of middle-aged adults in physical fitness programs in Rhode Island, Colorado, California, and Florida. The objective of the investigation was to determine whether adult participation in physical fitness programs varies from one region of the US to another. A random sample of people were interviewed in each state and these data were recorded:

	Rhode Island	Colorado	California	Florida
Participate	46	63	108	121
Do not participate	149	178	192	179

(a) (3 marks)
Give a complete two way table, including marginal counts (row and column totals).
Expected counts are printed below observed counts

	C9	C11	C13	C15	Total
1	46	63	108	121	338
	63.62	78.63	97.88	97.88	
2	149	178	192	179	698
	131.38	162.37	202.12	202.12	
Total	195	241	300	300	1036

(b) (10 marks)
Do the data indicate a difference in adult participation in physical fitness programs from one state to another?
Hypotheses:
test statistic: Chi-Sq = 4.880 + 3.106 + 1.047 + 5.463 + 2.363 + 1.504 + 0.507 +

2.645 = 21.515 DF = 3,
P-Value = 0.000
Decision
Conclusion

4. Assume that the level of nitrogen oxide (NOX) in the exhaust of a particular car model, when driven in the city traffic can be modelled by a normal distribution with mean $\mu = 2.1$ grams/km and a standard deviation $\sigma = 0.5$ grams/km.

(a) (3 marks)
If the Environmental Protection Agency (EPA) mandates that a nitrogen oxide level of 2.7 grams/km cannot be exceeded, what proportion of the cars of this model would be in violation of the mandate?

$$P(X > 2.7) = 1 - P(Z \leq (2.7 - 2.1)/0.5) = 1 - P(Z \leq 1.2) = TableII$$

(b) (3 marks)
At most, 25% of the cars of this model exceed what NOX level?
Find x such that

$$P(Z \leq (x - 2.1)/0.5) = 0.25$$

Then

$$x = 0.5 \cdot 0.67 + 2.1$$

0.67 from Table II.

(c) (4 marks)
Obtain the probability that the average NOX level of 16 cars of this model is at least 2.4 grams/km.

Use $\mu_{\bar{x}} = \mu = 2.1$ and $\sigma_{\bar{x}} = \sigma/\sqrt{16}$ for standardization.

5. The cost of automobile insurance is a sore subject in California. The following table gives the 6–month premiums in 2001 for a married male, licensed for 6–8 years, who drives about 15000 miles per year, and who has no accidents or violations:

City	Allstate	21st Century
Long Beach	\$ 1050	\$ 682
Pomona	\$ 984	\$ 638
San Bernadino	\$ 900	\$ 578
Moreno Valley	\$ 964	\$ 524

The following values might help in solving this question: $s_1 = 61.8$, $s_2 = 69.1$, $s_d = 50.9$.

(a) (2 marks)
Why would you expect these pairs of observations to be dependent?

(b) (10 marks)
Do the data provide sufficient evidence to indicate that there is a difference in the average 6–month premiums for Allstate and 21st Century insurance. Test using $\alpha = 0.05$.

Paired t-test!
Hypotheses: $H_0 : \mu_A = \mu_{21}$ vs. $H_a : \mu_A \neq \mu_{21}$
teststatistic T-Value = 14.49 df=3
P-Value less than $2 \cdot 0.0005 = 0.001$
decision
conclusion

(c) (4 marks)
Find a 95% confidence interval for the true difference in the average 6–month premiums for Allstate and 21st Century insurance.

95% CI for mean difference: (288.0, 450.0)

(d) (2marks)
What does it mean in Inferential Statistics to be "95% confident"?

If you don't know this, look it up.

6. A Computer Chess Program is playing 8 games against a human champion. Assume that the two opponents are evenly matched.

(a) (2 marks)
In such a match, how many games do you expect the computer program to win?

This is a binomial distribution with $n = 8$ and $p = 0.5$
$\mu = n\ p = 4$

(b) (2 marks)
Give the probability that the computer program wins exactly 5 games.

0.219

(c) What is the probability that the computer program loses all but one game?

$P(X = 7)$

7. (a) (2 marks)
What is the difference between the standard deviation σ and the standard deviation s?

(b) (2 marks)
The data description for a qualitative variable should inclose which parts?

(c) (6 marks)
What is Inferential Statistics?

(d) (3 marks)
In ANOVA, what is the difference between the Total Sum Of Squares and the Sum of Squares for Treatments (Groups)?

(e) (3 marks)
Give an example, when we should use a significance level smaller than 0.05. Explain.