Part I. (20 points) Many observational studies conclude that low-fat diets protect against cardiovascular events (such as heart attacks, stroke, and so forth). Experimental results, however, prove otherwise. In an eight year long randomized controlled experiment, the percentage of people who had at least one cardiovascular event was compared between the control group (followed their usual diet) and the treatment group (followed a low fat diet). The difference between the two groups was concluded to be due to chance after conducting a test of significance.

1. (2 points) State one variable which is measured in this study and its values.

Variable: Having at least one cardiovascular event or not.
Values: 0 and 1. (Having an event = 1, not = 0)

2. (3 points) Describe an observational study in the context of the above problem (What do the investigators observe? What are the control and treatment groups, if any?) Explain in at most 3 sentences (extra sentences will be ignored!).

People who choose a more common diet (not low in fat) can be compared to people who choose a low fat diet in their daily lives. First is control, the second one is treatment group, respectively. The investigators do not assign their diets.

3. (2 points) Is the one above a longitudinal or a cross-sectional study? State which one it is, then explain how the other one would be in the context of this problem (2 sentences at most). It is cross-sectional because low-fat and usual diets are compared at the same time (even though the experiment took 8 years!). In a longitudanal study, the diet of only one group would be varied over time.

4. (3 points) Suggest a confounding variable which causes the observational studies conclude differently from the experimental study above. Why is it a confounding variable? (2 sentences at most). People who choose low-fat diet in their daily lives usually take care of their health in other ways, too, such as exercising and following regular check-ups. The latter habits are confounding variables since they affect "diet habit" as well as "fewer cardiovascular events" at the same time.

Suppose the following data set for the amount of fat (given in grams per 100 grams of food) in a low fat diet was used in another part of this study.

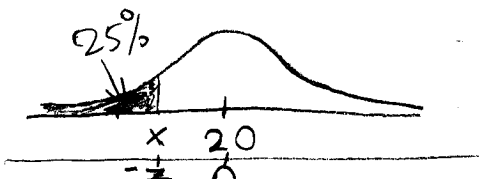
15, 30, 23, 25, 29, 18, 17, 22, 10, 14, 28, 26, 15, 11, 23, 27, 19, 18] $n = 18$

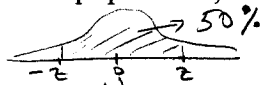
5. (3 points) Find the 75th percentile of the amount of fat using the data set.

10, 11, 14, 15, 15, 17, 18, 18, 19, 22, 23, 23, 25, **26**, 27, 28, 29, 30

$18(0.75) = 13.5 \rightarrow$ take 14th observation: 26

6. (4 points) If the amount of fat in a low-fat diet is believed to follow a normal distribution with mean 20 and standard deviation 6 in the population, what is the 25th percentile of the amount of fat in a low-fat diet?





(100 − 25 − 25)

From the normal table $z = 0.70$

$$\frac{x-20}{6} = -0.70 \Rightarrow x = 15.8 \text{ grams}$$

7. (3 points) In a *t*-test applied to this sample, the P-value is found to be 0.01. What is the value of the test-statistic *t* in this test?

d.f. $= n - 1 = 17$ $\quad 0.01 = 1\%$

$\Rightarrow t = 2.57$

Part II. (20 points) In an eight week randomized controlled experiment, 200 patients with major depression were divided into two groups, one of which (sample size 98) received St. John's wort extract (a herbal medicine) while the other (sample size 102) received a placebo. At the end of the study period, 14 of the St. John's wort patients were in remission (=feeling better), compared with 5 of the placebo patients.

1. (10 points) At $\alpha = 0.01$ level of significance, is St. John's wort effective in treating major depression? Show all steps.

$H_0: p_1 = p_2$
$H_a: p_1 > p_2$

$\hat{p}_1 = \frac{14}{98} = 0.143$

$\hat{p}_2 = \frac{5}{102} = 0.049$

$SE_{2\text{-sample}} = \sqrt{\frac{(0.143)(0.857)}{98} + \frac{(0.049)(0.951)}{102}}$

$= 0.041$

$z = \frac{0.143 - 0.049}{0.041} = 2.29$, $z_{0.01} \cong 2.35$

$\Rightarrow$ P-value $\cong \frac{100 - 97.85\%}{2} = 1.07\%$

Since $1.07\% > 1\%$ (or $2.29 < 2.35$) we do not reject H_0.
St. John's wort is not effective.

2. (5 points) Another aspect of this study was about the frequency of usage of St. John's wort extract among depression patients. The 200 patients were asked if they used it on a regular basis or not, and 80 of them answered positively. Construct a 95% confidence interval for the percentage of patients who use this herbal medicine.

$\hat{p} = \frac{80}{200} = 40\%$ $\quad SE = \sqrt{\frac{(0.4)(0.6)}{200}} \cong 0.035$

$\hat{p} \mp 2SE \Rightarrow 40\% \mp 2(3.5\%)$

$\Rightarrow [33\%, 47\%]$

3. (5 points) Give an estimate of the population percentage of patients who use St. John's wort using the statistics in question 2 above. Then, use this estimate to answer the following: In a random sample of 12 patients what is the probability that at least 3 of them use St. John's wort?

$\hat{p} = 40\%$

$p(0) = (0.60)^{12} = (2.18)10^{-3}$

$p(1) = \binom{12}{1}(0.40)(0.60)^{11} = 0.0174$

$p(2) = \binom{12}{2}(0.40)^2(0.60)^{10} = 0.064$

$p(0) + p(1) + p(2) = 0.0836$

Prob. that at least 3 use $\Rightarrow 1 - 0.0836 = 0.9164$
$= 91.64\%$

Part III. (15 points) A neuroscientist conducted a field experiment for measuring the effect of high altitude (=yükseklik) on a person's ability to think critically. For each subject, the time (in seconds) to match a picture with a related sentence was recorded. A group of randomly selected 30 climbers is tested before they climbed Mount Everest and another group of 36 climbers is tested at a certain altitude in the mountain. The summary statistics is as follows:

	Mean time in seconds	Standard deviation	
Low altitude group	4.2	0.8	$n=30$
High altitude group	6.5	1.3	$n=36$

1. Find a 99% confidence interval for the difference in the time to complete the described task in low and high altitudes.

$\bar{X}_1 - \bar{X}_2 \mp z\,SE$ 99% CI $\Rightarrow z = 2.60$

$$SE = \sqrt{\frac{(0.8)^2}{30} + \frac{(1.3)^2}{36}} = 0.26$$

$$\Rightarrow (4.2 - 6.5) \mp 2.60\,(0.26)$$

$$-2.3 \mp 0.676$$

CI for $\mu_1 - \mu_2$: $[-2.976, -1.624]$ (OR for $\mu_2 - \mu_1$: $[1.624, 2.976]$)

2. Conduct a test of significance to decide if the true difference is larger than 2 seconds.

$H_0 : \mu_2 - \mu_1 = 2$

$H_a : \mu_2 - \mu_1 > 2$

$\bar{X}_2 - \bar{X}_1 = 2.3$

$$z = \frac{2.3 - 2}{0.26} = 1.15$$

$$p\text{-value} = \frac{100 - 74.99\%}{2} \cong 12.5\% > 5\%$$

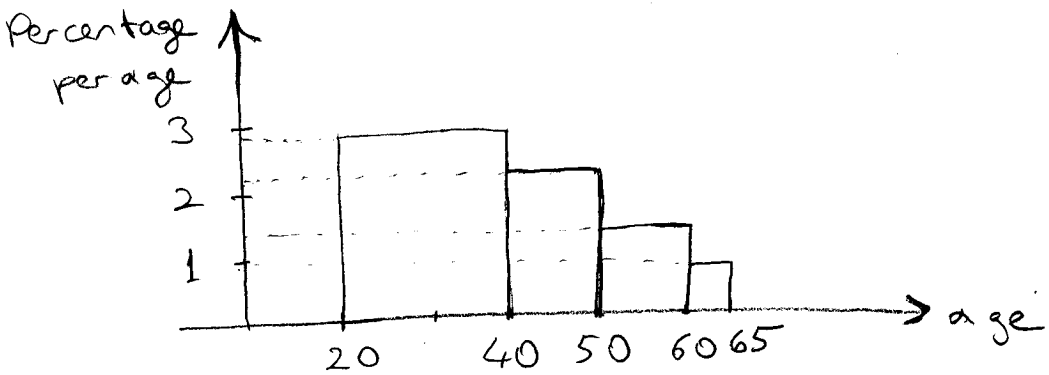
$\Rightarrow$ do not reject H_0.

The difference is not significantly larger than 2 seconds.

Part IV. (15 points) The following is the age distribution of the elementary school teachers in a country based on historical data.

Age	Percentage	
20 to 40	58%	$\rightarrow \frac{58}{20} = 2.9\%$
40 to 50	22%	$\rightarrow \frac{22}{10} = 2.2\%$
50 to 60	15%	$\rightarrow 15/10 = 1.5\%$
60 to 65	5%	$\rightarrow 5/5 = 1\%$

1. (6 points) Draw a histogram of the age distribution. Label the axes.



2. (9 points) A recently obtained data set suggests that the distribution has become even more concentrated on the younger age groups. Namely, the following frequency table is obtained from a random sample of 54 teachers.

	20-40	40-50	50-60	60-65
observed	36	11	5	2
expected	31.32	11.88	8.1	2.7

Has the distribution changed recently?

Expected frequencies: $58\% \times 54 = 31.32$, $22\% \times 54 = 11.88$, $15\% \times 54 = 8.1$, $(5\%) \times 54 = 2.7$

H_0: $p_1 = 58\%, p_2 = 22\%, p_3 = 15\%, p_4 = 5\%$

H_a: at least one p changed.

$$\chi^2 = \frac{(36-31.32)^2}{31.32} + \frac{(11-11.88)^2}{11.88} + \frac{(5-8.1)^2}{8.1} + \frac{(2-2.7)^2}{2.7}$$

$$= 2.13$$

d.f. $= 4-1 = 3 \Rightarrow \chi^2_{table} = 7.82$ with 5%.

Since $2.13 < 7.82$, do not reject H_0.

The distribution has not changed significantly.

Part V. (15 points). A social scientist investigates the preference of school children about their mother working outside the home, for three different groups: elementary, middle and high school students. The following is the frequency table of his findings.

	Elementary	Middle	High	Total
Prefers mother work	29 a	38 c	51 e	118
Prefers mother not work	31 b	22 d	9 f	62
Total	60	60	60	180

1. (4 points) What are the percentages of elementary, middle and high school students in this sample, respectively?

all $\frac{60}{180} = 33.33\%$

H_0: preference and educational level are independent.

H_a: they are not independent.

2. (11 points) Do the students' preferences differ significantly according to their educational levels?

Expected values: $a = \frac{60 \times 118}{180} = 39.33 = c = e$

$b = \frac{62 \times 60}{180} = 20.67$ $\quad d = \frac{60 \times 62}{180} = f = 20.67$

$$\chi^2 = \frac{(29-39.33)^2}{39.33} + \frac{(38-39.33)^2}{39.33} + \frac{(51-39.33)^2}{39.33}$$

$$+ \frac{(31-20.67)^2}{20.67} + \frac{(22-20.67)^2}{20.67} + \frac{(9-20.67)^2}{20.67}$$

$= 18.06$, d.f. $= (2-1)(3-1) = 2$

χ^2 with 5% and d.f. 2 is 5.99 from the table.

Since $18.06 > 5.99$, we reject $H_0 \Rightarrow$ preferences and educational level are not independent; the preferences differ significantly according to educational levels.

(In fact, highly significantly because P-value $< 1\%$)

Part VI. (15 points) Three different relaxation techniques are given to randomly selected subjects in an effort to reduce their stress levels. A scale is devised to measure the percentage of stress reduction in each person. The data are shown in the table.

	Technique I	Technique II	Technique III
	3	12	15
	10	12	14
	5	17	18
	6	13	14
	13	18	20
	3	9	22
	4		
Mean ($\bar{x}$)	6.2	13.5	17.2
Standard Deviation	3.8	3.4	3.4
n	7	6	6

Can the difference in the means of the three groups be explained by chance?

$H_0: \mu_1 = \mu_2 = \mu_3$

H_a: at least one mean is different

$$\bar{X}_{GM} = \frac{7(6.2) + 6(13.5) + 6(17.2)}{19} = \frac{3+10+\cdots+22}{19} = 12$$

$$S_B^2 = \frac{7(6.2-12)^2 + 6(13.5-12)^2 + 6(17.2-12)^2}{2} = 205.6$$

$$S_W^2 = \frac{6(3.8)^2 + 5(3.4)^2 + 5(3.4)^2}{16} = 12.64$$

$$F = \frac{205.61}{12.64} = 16.26$$

$F = 3.63$ for $\alpha = 0.05$ with $d.f.N = 2$ and $d.f.D = 16$.

Since $16.26 > 3.63$, reject H_0.

There is significant difference, it cannot be explained by chance.