Part I. (20 points) A study was conducted to see the effect of screening (= tarama) on detecting (=saptamak) the breast cancer early enough. The subjects were 62000 women who had enrolled to a general insurance plan in a given region (like "sosyal güvenlik", which applies to almost all the population). The subjects were divided at random into two equal groups. In the treatment group, women were encouraged to visit the doctor for annual screening, including X-rays. Some of the women came in for the screening, but some others in the treatment group refused to do so. The control group was offered only usual health care (no screening for breast cancer). The following table shows the rates of death per 1000 women, by different causes.

	Cause of Death	
	Breast Cancer	All other
Treatment Group		
Examined	1.1	21
Refused	1.5	38
Total (Refused and Examined, together)	1.3	27
Control Group	2	28

1. (3 points) Is this a randomized controlled experiment or an observational study? Explain in three sentences at most.

It started as a randomized controlled experiment. However, since some of the subjects did not come for screening, they quit the treatment group themselves. It became an observational study as a result.

2. (3 points) Is this study double blind or not? Explain in two sentences at most.

It is not double blind since both the experimenter and the subjects (in the treatment group at least) know that they are being screened for cancer.

3. (3 points) What is a confounding variable in this study? Explain in two sentences at most.

The "patients' attitude towards their health" is a confounding variable because those who came for screening and those who refused are different in that respect. It affects both "being screened" and "death rate".

4. (3 points) Does screening save lives? Which numbers in the table prove your conclusion?

With the available information, we can say "yes, it does". We compare 1.3 (total in treatment group) and 2, the control group.

Part I continued. Extra information for questions 5 and 6 below: Statisticians who worked in the study examined the socio-economic level of the subjects as well. They found that poorer women were less likely to accept screening than richer ones.

5. (4 points) Looking at "two numbers" in the table, they concluded that breast cancer (like polio, but unlike most other diseases) affects the rich more than the poor. Which two numbers are these? Explain how you reach this conclusion, using at most three sentences.

Compare "Refused Group", which includes mostly poorer women, and the control group. In the first one the rate is lower (1.5) than the latter (2) which includes both poor and rich subjects.

6. (4 points) Compare the death rate from all other causes among women who accepted screening and those who refused. What explains the difference in the death rates?

It is 21 versus 38. So the death rate from all other causes is much less (about half) in the "Examined" group. We can explain this by their higher socio-economic level. Most other diseases affect poor rather than rich and possibly more educate

Part II. (15 points) "Cola wars" is the popular term for the intense competition between Coca-Cola and Pepsi displayed in their marketing campaigns. Suppose, as part of a market research, 1000 cola consumers are given a taste test in two different glasses labeled as A and B (the brand names were hidden). Each consumer is asked to state a preference for the cola in glass A or B.

1. (3 points) State (in a few words) the population targeted by this test.

All cola consumers

2. (3 points) State the variable of interest.

Preference for cola taste.

3. (5 points) What are the values of the variable that you stated above? Is this a qualitative or a quantitative variable?

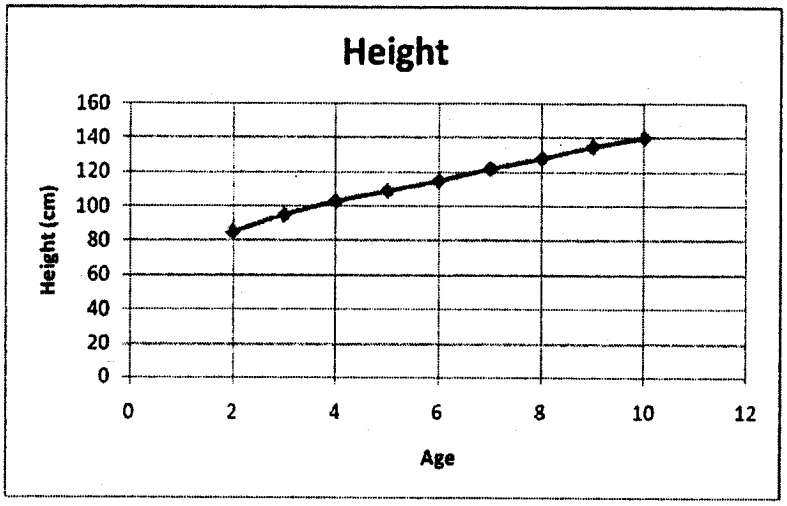
A (say Pepsi Cola) or B (say coca cola)

It is a qualitative variable, A, B or the name are just labels.

4. (4 points) Suppose 450 of cola consumers among the 1000 preferred Coca-Cola. What is the number 450, a parameter or a statistic? Why?

A statistic because it is obtained by observing the sample of 1000 consumers

Part III. (22 points) The following graph gives the average height for a large sample of boys, who were followed from 2 to 10 years of age.



1. (3 points) Is this a longitudinal or cross-sectional study? Explain in two sentences at most.

It is a longitudinal study because the boys were followed over time.

2. (5 points) Suppose the standard deviation at age 10 is 3cm and the height distribution can be approximated by a normal curve. Give an approximate interval for heights, which includes 95% of the observations for boys at age 10.

From the graph, average height is 140 cm for 10 year-old boys. Since average $\mp$ 2 SD includes 95% of the observations, we get $140 \mp 2(3) \Rightarrow [134, 146]$

3. (7 points) Find the standard deviation of the following sample of heights.

155, 140, 142, 150, 135

$\text{mean} = \frac{155+140+142+150+135}{5} = 144.4$ cm.

$$SD = \sqrt{\frac{(155-144.4)^2 + (140-144.4)^2 + \dots + (135-144.4)^2}{5}}$$

$\cong 7.17$

4. (4 points) Find the percentage of observations taller than 145cm in the data set of question 3.

only 150 and 155

$\Rightarrow \frac{2}{5} = 40\%$

5. (3 points) Find the 40th percentile of height distribution using the data set of question 3.

$5 \cdot \frac{40}{100} = 2 \Rightarrow 2^{nd}$ observation : 140cm

(after sorting: 135, 140, 142, 150, 155)

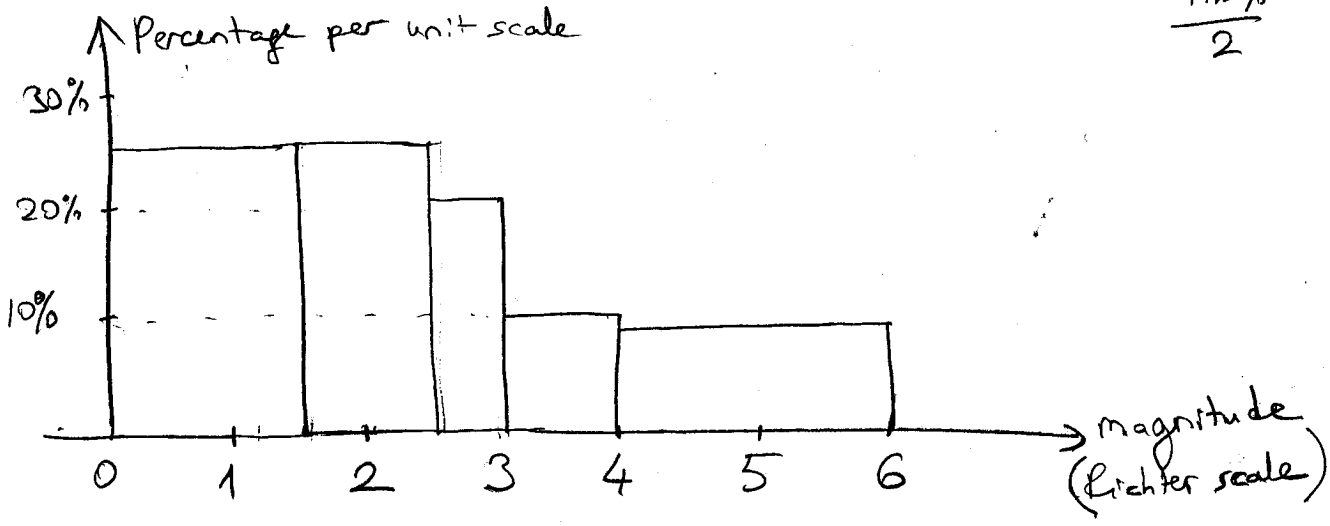
Part IV. (20 points) Consider the following frequency table for the magnitudes of aftershocks of a major earthquake, in Richter scale.

$\frac{15}{40}$ etc.

Magnitude	0- 1.5	1.5 – 2.5	2.5 - 3.0	3.0 - 4.0	4.0 - 6.0	Total
Frequency	15	10	4	4	7	40
Percentage	37.5%	25%	10%	10%	17.5%	
Perc. per scale unit	$\frac{37.5\%}{1.5}=25\%$	$\frac{25}{1}=25\%$	$\frac{10\%}{0.5}=20\%$	10%	8.75%	

1. (10 points) Draw a density scale histogram for the magnitude of aftershocks.

$\frac{17.5\%}{2}$



2. (3 points) Is the histogram left-tailed, symmetric, or right-tailed?

It's right tailed.

3. (4 points) In view of your answer to question 2, do you expect the median to be larger or smaller than, or about the same as the average?

We expect it to be smaller than the average because the right tail affects by increasing the average

4. (3 points) Which interval does the 35th percentile belong to?

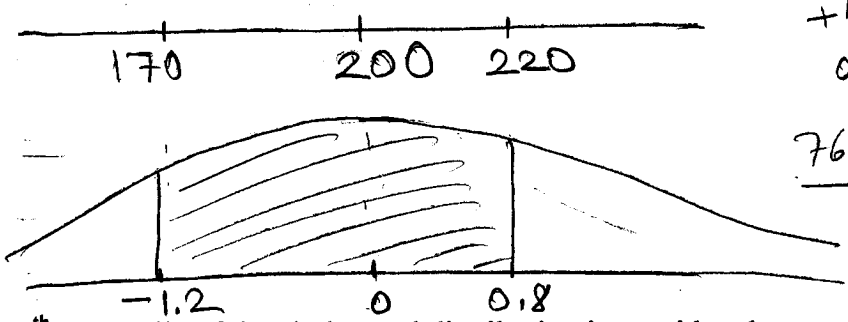
It belongs to 0–1.5 interval, because 37.5% of all observations are in that interval.

Part V. (23 points) A study of cholesterol levels in psychiatric patients showed that cholesterol level is approximately normally distributed with an average of 200 mg/dL (milligrams per deciliter) and a standard deviation of 25 mg/dL.

1. (8 points) What is the chance of observing a psychiatric patient with a cholesterol level above 170 mg/dL but below 220 mg/dL?

$\frac{170-200}{25} = -1.2$

$\frac{220-200}{25} = 0.8$

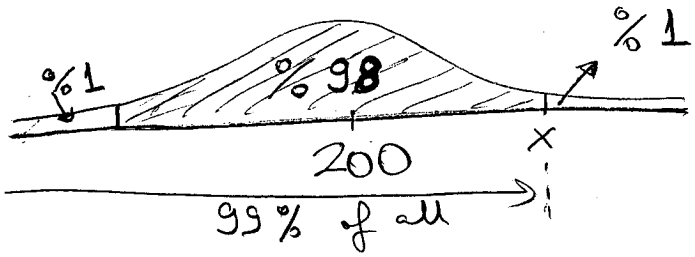


$+1.2 \Rightarrow 76.99$

$0.8 \Rightarrow 57.63$

$\frac{76.99\% + 57.63\%}{2} = 67.31\%$

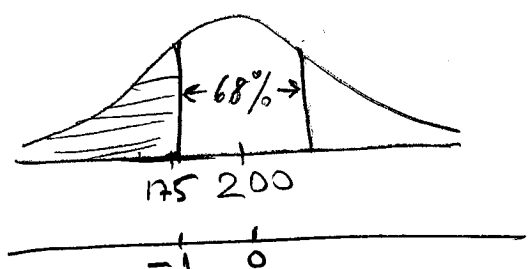
2. (5 points) 99th percentile of the cholesterol distribution is considered as an extremely high cholesterol level. Find it.



$\frac{X-200}{25} = 2.35$ (from table)

$\Rightarrow X = 258.75$ mg/dL

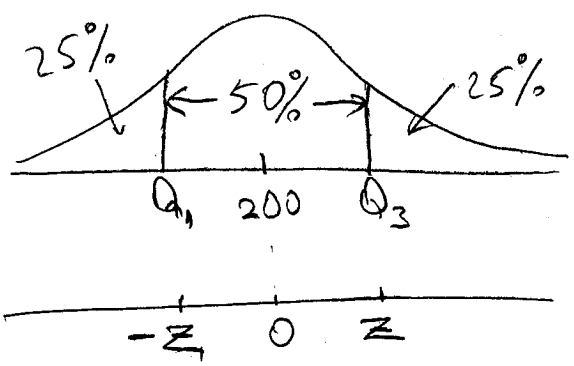
3. (5 points) According to other research, cholesterol levels lower than 175 mg is related to violent behavior. What percent of the patients are suspected to have violent behavior?



$50\% - \frac{68\%}{2} = 16\%$

4. (5 points) Find the IQR (inter-quartile range) of the cholesterol distribution.

$IQR = 75^{th}$ percentile $- 25^{th}$ percentile $\equiv Q_3 - Q_1$



$\Rightarrow z \equiv 0.70$

$\frac{Q_3-200}{25} = 0.70 \Rightarrow Q_3 = 217.5$

$\frac{Q_1-200}{25} = -0.70 \Rightarrow Q_1 = 182.5$

$\Rightarrow IQR = 217.5 - 182.5 = 35$ mg/dL