NAME SOLUTIONS

TEST 1
201-BZS-05
Probability & Statistics
Instructor: Emilie Richer
Date: February 13th 2009

Question 1 (3 marks) Classify the following variables by completing the table.

Variable	Qualitative/Quantitative	Nominal/Ordinal/Interval/Ratio
Weight of a bear (kg)	QUANTITATIVE	RATIO
Size of a computer file (KB)	QUANTITATIVE	RATIO
Temperature (in °K)	QUANTITATIVE	RATIO
World ranking of tennis players	QUALITATIVE	ordinal
Hair colour	QUALITATIVE	NOMINAL

Question 2 (4 marks) Briefly define the following:
(a) Chebyshev's Rule
(b) The z-score of a population element x
(c) An outlier
(d) The first quartile Q_1

(a) THIS RULE states that the proportion OF DATA within y standard deviations FROM the MEAN is given by $1-\frac{1}{y^2}$ (AS A lower bound estimate) $y>1$

(b) $Z_x = \frac{x-\bar{x}}{s}$; measures the number of standard deviations from mean

(c) DATA POINT THAT IS NOT WITHIN 3 s.d FROM MEAN

(d) $Q_1 = P_{25}$; the point for which at most 25% of data is below it

Question 3 (15 marks) Twenty CEGEP students were asked how many hours of television they watch per week. The following answers were collected:

31	15	10	20	10	2	22	23	18	1
35	7	4	29	13	7	15	19	11	9

(a) Construct a stem & leaf display of the data, then sort the data
(b) Find the sample mean $\bar{x}$
(c) Find the sample standard deviation S
(d) Find P_{42} & Q_3
(e) Build a grouped frequency distribution table summarizing the data (make sure you use an appropriate number of classes).

(a)

0	2 1 7 4 7 9
1	5 0 0 8 3 5 9 1
2	0 2 3 9
3	1 5

RANKED DATA

1 2 4 7 7 9 10 10 11 13
15 15 18 19 20 22 23 29 31 35

(b) $\bar{x} = \frac{301}{20} = 15.05$

(c) $S = 9.56$

(d) $(0.42)20 = 8.4$ depth 9 $P_{42} = 11$

$Q_3 = P_{75} \Rightarrow (0.75)(20) = 15$ depth 16 $Q_3 = 22$

(e) $\sqrt{20} \approx 4.47$ we will use 4 classes

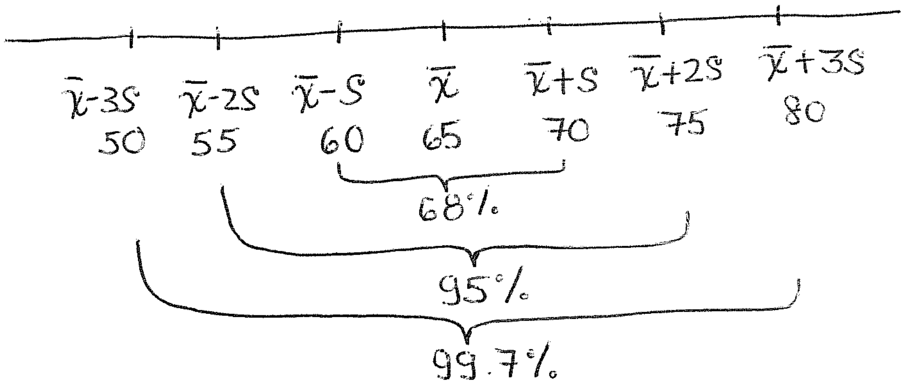
$\frac{\text{range}}{4} = \frac{35-1}{4} = 8.5$ ROUND UP TO 9

CLASS LIMITS	f
1 – 9	6
10 – 18	7
19 – 27	4
28 – 36	3

Question 4 (7 marks) One thousand students wrote an aptitude test, yielding $\bar{x} = 65$ and S = 5. If the test scores were normally distributed, then (approximately) how many of the students:

a) scored less than 60?
b) scored between 60 and 75?
c) scored more than 80?
d) What rule did you use to approximate these values?

NORMAL DISTRIBUTION = EMPIRICAL RULE



(a) 50 − 34 = 16% OF 1000 students (160)

(b) $95/2 + 34 = 81.5\%$ (815 students)

(c) $\frac{0.03}{2} = 0.015\%$ → (1.5 students between 1 & 2)

(d) EMPIRICAL RULE

Question 5 (3 marks) Let x represent the number of years of education and let y represent annual salary. Give an approximate linear correlation coefficient for these two variables. Justify your answer (there are many possible answers depending on your explanation).

We can expect a value that is positive, because salary generally rises if you have more education;

Other factors also influence salary so we can approximate r at 0.7 for example

(many many possible answers)

Question 5 (10 marks) A sample is collected amongst people having admitted to marijuana use in the last year. Data is collected regarding their first incidence of marijuana use, summarized in the grouped cumulative frequency table below:

Reported Age of First Use of Marijuana

Class Limits (in years)	*cf*
7 – 11	1
12 – 16	10
17 – 21	20
22 – 26	26
27 – 31	30

Approximate the sample mean $\bar{x}$ and the sample standard deviation S.

	M	f	M·f	$M^2 \cdot f$
	9	1	9	81
	14	9	126	1764
	19	10	190	3610
	24	6	144	3456
	29	4	116	3364
Σ		30	585	12275

$$\bar{x} \approx \frac{\sum M_i f_i}{n} = \frac{585}{30} = 19.5$$

$$S \approx \sqrt{\frac{\sum M_i^2 f_i - \frac{\left(\sum M_i f_i\right)^2}{n}}{n-1}} = \sqrt{\frac{12275 - \frac{(585)^2}{30}}{29}} = 5.47$$

Question 6 (8 marks) Ten people between the ages of 14-25 were asked the following question: "What percentage of your weekly expenses is spent on entertainment or leisure related activities?" The following table summarizes the answers collected.

x (Age in Years)	17	25	14	22	24	15	21	17	20	18	Σ 193
y (% spent)	35	5	45	20	25	50	25	35	25	50	315
$x \cdot y$	595	125	630	440	600	750	525	595	500	900	5660

(a) Find the line of best fit
(b) Find r the coefficient of linear correlation.
(c) Predict the percentage of weekly expenses spent on entertainment or leisure related activities of a 16 year old. Explain whether or not this is an good prediction based on the data on hand.

$$SS(xy) = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$$

$$= 5660 - \frac{(193)(315)}{10} = -419.5$$

$$SS(x) = \sum x_i^2 - \frac{(\sum x_i)^2}{10} = 3849 - 3724.9 = 124.1$$

$$SS(y) = 11775 - \frac{(315)^2}{10} = 1852.5$$

(a) $b_1 = \frac{SS(xy)}{SS(x)} = -3.38$

$b_0 = \bar{y} - b_1 \bar{x} = 96.74$

$$\boxed{\hat{y} = 96.74 - 3.38x}$$

(b) $r = \frac{SS(xy)}{\sqrt{SS(x) SS(y)}} = \frac{-419.5}{\sqrt{(124.1)(1852.5)}} = \boxed{-0.875}$

(c) $\hat{y} = 96.74 - 3.38(16)$
$= 42.66$

42.66% of expenses of 16 year olds spent on leisure; based on the given data, the correlation is strong (-0.875) so the prediction is good.