Social Studies 201
Winter 2001
Answers to Problem Set No. 3
March 2, 2001

1. Explain which concept of probability (subjective, theoretical, or frequency) appears to be used in each of the following statements.

> In Switzerland, a man was roughly twice as likely to be murdered in the 1880s as he is today. ... Finland had a particularly violent past: the chance of being murdered at the turn of the 19th century was higher than it is in many American cities today. (*The Economist*, October, 15, 1994, p. 21).

Answer. This appears to be a **frequency** interpretation of probability. While the data are not provided in the quote, the implication of the quote is that records of the number of murders in Switzerland and Finland were obtained and compared with the incidence of murders in American cities today. These data must be based on actual data concerning the incidence of murder.

> A professor allocates the grades of 50 students in a class according to a normal distribution with mean 65 and standard deviation 10. If a student is randomly selected from this class, the probability that the student's grade is above 85 is 0.0227.

Answer. This is a **theoretical** interpretation of probability. If the grades are allocated on this normal probability distribution, the probability of obtaining a grade of above 85 can be determined using the normal table. This table comes from a theoretical probability distribution obtained from the equation describing the normal curve.

> A new scientific study of nearly 100,000 women found that, contrary to some earlier research, hair dye does not raise a woman's **chance** of contracting certain blood and lymph cancers. (*The Leader-Post*, October 6, 1994).

Answer. This is a study using a large sample of women, and the chance is obtained by observing the incidence of various cancers among these women. This is thus a **frequency** interpretation of probability.

> Health care will **likely** not dominate the debate, in part because of Mr. Klein's effective management of the issues, where he adressed every concern thrown at him point by point (*National Post*, February 13, 2001, p. A8).

Answer. This is a **subjective** interpretation of probability since it cannot be reasoned theoretically and the circumstances surrounding the debate cannot be reproduced under the same conditions more than once. As a result, this is no more than someone's subjective judgment.

> Mr. Rock's father died of prostate cancer several years ago – a fact that doubled his own **risk** of getting the disease. (*National Post*, February 13, 2001, p. A1).

Answer. This is a **frequency** interpretation of probability, presumably based on experimental data or administrative records that show that males whose fathers die of prostate cancer are more likely than other males to die of prostate cancer.

2. Identify (i) one set of independent events and (ii) and one set of dependent events in the following quote from *The Economist*, October 15, 1994, p. 114. Explain your reasoning.

> ... the damage that smoking does to health is even larger than previously thought: the habit reduces life expectancy, on average, by eight years rather than five. The good news is that quitting smoking at any age improves life expectancy. Quitting before you reach 35 reduces the risk of death to the level of the life-long abstinent.

Answer. The set of independent events are dying as a result of smoking and quitting before reaching age 35. That is, the overall probability of dying as a result of smoking is the same as the conditional probability of dying of smoking given that one quits before age 35. The chance of

dying is no higher or lower among those who quit than among those who were life-long abstainers.

Dependent events are long life expectancy and smoking. Since smokers live eight years less, on average, the probability of a long life given that one is a smoker is considerably lower than the overall probability of a long life.

3. Use the tables from problem 6 of Computer Problem 2 to answer the following. The tables are:

```
 ROWS: v8     COLUMNS: sex
        Male    Female
           1         2      ALL

  1       59        96      155
  2       36        63       99
  3       31        26       57
  4       22        18       40
  5       13         8       21
 ALL     161       211      372
```

```
MTB > Table 'v8'  'fedvote';
SUBC>   Counts.

 ROWS: v8     COLUMNS: fedvote
        Liberal   NDP       PC    Reform    Sask     None
           1        2        3         4        5        6      ALL

  1       46       36        4        16        0       53      155
  2       42       18        3         9        1       26       99
  3       18        9        7         7        0       16       57
  4       15        3        3        10        0        9       40
  5       12        0        2         3        1        3       21
 ALL     133       66       19        45        2      107      372
```

(a) Questions from cross-classification table of V8 by SEX. If an individual is randomly selected from the Table, what is the probability of

i. Agreeing (response 4 or 5)?
$(40 + 21)/372 = 61/372 = 0.164$.

ii. Agreeing or being male?
$P\text{Agree} + P\text{Male} - P\text{Agree and Male} = 61/372 + 161/372 - (22 + 13)/372 = 187/372 = 0.503$.

iii. Being female given a neutral response (3)?
$P(\text{Female/Neutral}) = 26/57 = 0.456$.

iv. Disagreeing (response 1 or 2), given that the respondent is male?
$P(\text{Disagree/Male}) = (59 + 36)/161 = 95/161 = 0.590$.
Disagreeing given that the respondent is female?
$P(\text{Disagree/Female}) = (96 + 63)/211 = 159/211 = 0.753$.

v. From (iv), comment on the independence or dependence of events and how responses differ by sex of the individual.
In this sample, females are considerably more likely to disagree than are male and the overall probability of diagreeing must be between these. That is,
$P\text{Disagree} = (155 + 99)/372 = 0.683$.
$P(\text{Disagree/Female}) > P\text{Disagree} > P(\text{Disagree/Male})$
As a result, it can be seen that the events of disagreeing and being female, or disagreeing and being male are dependent.

(b) Questions from cross-classification table of V8 by FEDVOTE. If an individual is randomly selected from the Table, what is the probability of

i. Disagree (response 1 or 2) and support PC party?
$P(\text{Disagree and PC}) = (4 + 3)/372 = 7/372 = 0.019$.

ii. Disagree or support PC party?
$P(\text{Disagree}) + P(\text{PC}) - P(\text{Disagree and PC}) = (155+99)/372 + 19/372 - 7/372 = 266/372 = 0.715$.

iii. Disagree given NDP? Disagree given PC?

$P(\text{Disagree/NDP}) = (36 + 18)/66 = 54/66 = 0.818.$
$P(\text{Disagree/PC}) = (4 + 3)/19 = 7/19 = 0.368.$

iv. Somewhat disagreeing (response 2), given NDP support?
$P(\text{Somewhat Disagree/NDP}) = 18/66 = 0.273.$

v. From (iii) and (iv), comment on the independence or dependence of the events.
In order to determine independence, compare the overall probability of disagreeing
$P(\text{Disagree}) = (155 + 99)/372 = 254/372 = 0.683$
with the conditional probability of disagreeing given NDP (0.818 from iii) and the conditional probability of disagreeing given PC (0.368 from iii). The events of disagreeing and supporting the PCs are very dependent, since the probability of disagreeing given PC is so much less than the overall probability of disagreeing. The events of disagreeing and supporting the NDP are also dependent, but less so than in the case of PCs.
The overall probability of somewhat disagreeing (category 2) is
$P(\text{Somewhat Disagree}) = 99/372 = 0.266$
and this is very close to the probability of somewhat disgreeing given NDP (from iv) of 0.273. The events of somewhat disagreeing and supporting the NDP are very close to being independent of each other.

4. Obtain the following using the standardized normal distribution:

(a) The area between $Z = 0$ and $Z = 1.85$ is 0.4678.

(b) The area between $Z = -2.23$ and $Z = 0.57$ is $0.4871 + 0.2157 = 0.7028$.

(c) The proportion of cases above $Z = -1.09$ is $0.3621 + 0.5000 = 0.8621$.

(d) The proportion of cases between $Z = 1.12$ and $Z = 2.12$ is $0.4830 - 0.3686 = 0.1144$.

(e) The percentage of cases that are within 1.5 standard deviations of the mean is $(0.4332 + 0.4332) \times 100\% = 86.64\%$.

The percentage of cases that are more than 2.1 standard deviations from the mean is $(0.0179 + 0.0179) \times 100\% = 3.58\%$.

(f) The 35th percentile is $Z = -0.38$ or $Z = -0.39$. The 72nd percentile is $Z = 0.58$ or $Z = 0.59$.

(g) The value of Z such that 61% of the cases are less than this is $Z = 0.28$.

(h) The Z-values so that 0.03 of the area is in each tail of the distribution are $Z = -1.88$ and $Z = +1.88$.

5. The 944 Saskatchewan respondents in Statistics Canada's General Social Survey, Cycle 11, 1996, reported a mean household income of $38 thousand, with a standard deviation of $27 thousand.

(a) If incomes are normally distributed with $\mu = 38$ and $\sigma = 27$,

i. For $X = 80$ thousand, $Z = (X - \mu)/\sigma = (80 - 38)/27 = 42/27 = 1.56$. The proportion of households with incomes above this is the B area for $Z = 1.56$, and this is 0.0594.

ii. For $X = 20$ thousand, $Z = (X - \mu)/\sigma = (20 - 38)/27 = -18/27 = -0.67$. The probability of selecting a households with incomes below this is the B area for $Z = -0.67$, and this is 0.2514.

iii. For $X = 30$, $Z = (30 - 38)/27 = -0.30$ and for $X = 50$, $Z = (50 - 38)/27 = 12/27 = 0.44$. The associated A areas for each of these are 0.1179 and 0.1700 for a total of 0.2879. Since there are 944 households, the number of households in the sample with incomes between $30,000 and $50,000 is $944 \times 0.2879 = 271.8$ or 272.

iv. The 75th percentile on the normal distribution is at $Z = 0.68$. The income that corresponds to this is $X = \mu + Z\sigma = 38 + (0.68 \times 27) = 38 + 18.36 = 56.36$ or $56 thousand.

v. For the standardized normal distribution the interquartile range is from $Z = -0.68$ to $Z = +0.68$. This is from $X = \mu + Z\sigma = 38 + (-0.68 \times 27) = 38 - 18.36 = 19.64$ to 56.36, or from $20,000 to $56,000.

Table 1: Summary of Normal and Actual Proportions of Households at Various Income Levels

Proportion with Heights	Normal	Actual
Below 20	0.251	0.303
30 - 50	0.288	0.275
Above 80	0.059	0.083

(b) The proportions from (a) and from the table showing the actual distribution of heights is in Table 1.

From Table 1, there are more households at each of the lowest and highest incomes than what would be the case if there were an exact normal distribution of incomes. Incomes are not generally normally distributed, and there are usually larger proportions of the population at lower incomes than what would be predicted by the normal distribution.