INSTRUCTIONS

This is an open book exam.

This exam consists of 10 pages, including this one. Please check that your copy contains all 10 pages.

Multiple choice questions: If you wish, indicate how you reasoned.

The answers to the other questions are to be written in the spaces provided. USE THE AMOUNT OF "WHITE SPACE" AS A GUIDE TO THE MAXIMUM EXPECTED LENGTH OF YOUR RESPONSE.

When writing, be brief and W R I T E C L E A R L Y.

Unless specifically asked for, complete calculations are not needed. To avoid writing out formulae, indicate which table or formula would be appropriate and give a reference; explain each of the inputs to the formula.

Points for each question are shown in parentheses. The points add up to 125; a 5-point question deserves about 5 minutes' effort.

Your ID number

Do not write your name

Q	points	Q 12	points	Q 13	points
1	/5	a	/5	a	/12
2	/5	b	/5	b	/12
3	/5	c	/5		
4	/5	d	/5		
5	/8	e	/5		
6	/8	f	/5		
7	/5				
8	/10				
9	/4				
10	/8				
11	/8				
Total	/71	Total	/30	Total	/24	/125

1

1[5] In a class of 100 students, the grades on a test are summarized in the following frequency table.

Circle the interval which contains the median grade:

```
Grade        Frequency
91 -100         11
81 - 90         31
71 - 80         42
61 - 70         16
```

2[5] Consider this box plot of the systolic blood pressures (SBP) of 39 adult males.
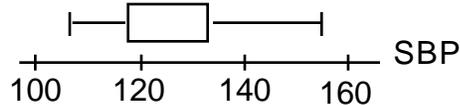
Based on this box plot...

(i) the interquartile range is approximately (circle one):   50   15   125   70

(ii) which of these statements is true?

    (circle one)

    a   The minimum blood pressure is less than 100.

    b   The maximum blood pressure is about 155.

    c   There is an outlier with a value of about 170.

    d   The median blood pressure is about 115.

3[5] A sample was taken of the salaries of 20 employees of a large company. The following are the salaries (in thousands of dollars) for this year. For convenience, the data are ordered.

```
28 31 34 35 37 41 42 42 42 47 47 51 52 52 60 61 67 72 75 80
```

(i) Suppose each employee in the company receives a $3,000 raise for next year (each employee's salary is increased by $3,000). The standard deviation of the salaries for the employees will (circle one) :

    a   be unchanged.

    b   increase by $3,000.

    c   be multiplied by $3,000.

    d   increase by 3,000 .

(ii) Suppose each employee in the company receives a 4% raise for next year. The standard deviation of the salaries for the employees will (circle one) :

    a   be unchanged.

    b   increase by 16%

    c   increase by 4%

    d   increase by 2%

4[5] Suppose that the probability that HIV will be passed from an infected person to an uninfected person during a single sexual contact is 0.01. Suppose that there are 50 such contacts.

Show how to calculate (or obtain via software) the probability that HIV will be passed on in at least one of the 50 contacts.

5[8]   The following are data collected by Statistics Canada at the 1996 census:

Montreal Metropolitan Population by knowledge of official language

English only	French only	Both English and French	Neither English nor French
8%	40%	50%	2%

Consider a person randomly chosen from this population. Let E be the variable denoting knowledge(Yes/No) of English, and F knowledge (Yes/No) of French.

a)   *Display the joint distribution of E and F as a 2 x 2 table or as a probability tree.*

b)   *What is the probability that :*

     (i) a person speaks French

     (i) a person who knows French also knows English

     (ii) a person who knows English also knows French

c)   *Are E and F independent random variables? Why or why not?*

6[8]   An auto insurance company notes whether drivers under 25 years old have had a driver's education course. Some 40% of its policyholders under 25 have had a driver's education course and 5% of this subset have an accident in a one-year period. Of those under 25 who have not had a driver's education course, 10% have an accident within a one-year period.

A 20-year-old takes out a policy with this company and within one year has an accident. What is the probability that the person did not have a driver's education course? [a probability tree may help]

7[5]   Suppose that 10% of McGill students are left-handed.

Find the probability that in a class of 20 students, at least 2 will be left-handed. State any assumptions made.

8[10] To allow for the 10% (on average) of people who make reservations but do not show up for their flights, an airline company "over-books" i.e., it takes reservations for up to 145 seats when it only has space on the plane for 140 passengers.

a)   *If the airline takes 145 reservations, what is the probability that it will not have a seat for every passenger who shows up? Actual calculations are not required, but explain the steps in sufficient detail that your research assistant could complete them.*

b)   *Would the normal distribution be appropriate here? Why or why not?*

c)   *Suppose that some of the 145 individuals are related. Does this affect the appropriateness of the probability model you base your calculations on? Why/why not?*

9[4]   An airplane has a front and a rear door which are both opened to allow passengers to exit when the plane lands. The plane has 100 passengers seated.

(i) *The number of passengers exiting through the front door should have*

   a   a binomial distribution with mean 50.

   b   a binomial distribution with 100 trials but success probability not equal to 0.5.

   c   a normal distribution with a standard deviation of 5.

   d   none of the above.

(circle one):

(ii) *justify your choice*

10[8] Suppose we want a 95% confidence interval for the average amount spent on books by students in their first year at university. The interval is to have a margin of error of $5.

(i) *If the 'best estimate' is that the amounts spent have a standard deviation σ = $50, then the number of observations required (in a SRS) is closest to:*

   a   196
   b   250
   c   400
   d   475

(circle one)

(ii) *The amounts spent by individuals are unlikely to have a normal distribution. Is this critical to your calculation? Explain.*

\

10...continued

(iii) The university administrator, who fancies himself as a statistician, says that we should use a multiple from the "t" distribution rather than the z distribution in the sample size formula.

Technically, he is correct. Explain why.

Practically, in this case, it doesn't matter very much. Explain -- in a sentence or two -- why.

11[8] The nicotine content in cigarettes of a certain brand is normally distributed with mean (in milligrams) μ and standard deviation $\quad = 0.1$. The brand advertises that the mean nicotine content of its cigarettes is 1.5, but measurements on a random sample of 100 cigarettes of this brand gave a mean of $\bar{x} = 1.53$. Is this evidence that the mean nicotine content is actually higher than advertised?

(i) To answer this, test the hypotheses

H_0: $\mu = 1.5$, H_a: $\mu > 1.5$

at the 5% significance level. *SHOW YOUR WORK*

(ii) You conclude

(circle one) :

a   that H_0 should be rejected.

b   that H_0 should not be rejected.

c   that H_a should be rejected.

d   there is a 5% chance that the null hypothesis is true.

12 Refer to the article "Oral contraceptive use and bone mineral density in premenopausal women" As the authors did, concentrate on Never vs. Ever (Table 1, first two columns)

a[5]   The reported SD for alcohol consumption is larger than the mean.

(i) *Can this be, or is it a mistake? Explain, using a diagram to show what the distribution looks like.*

(ii) *What other summary measures might have been more informative for describing the shapes of the distributions of alcohol and cigarette consumption?*

b[5]   *Reconstruct, from the data given, the 95% CI of (–0.06 to 0.00) that accompanies the point estimate of –0.03 for the lumbar BMD.*

d[5]   If you were calculating a P-value from the 1.03 versus 1.06, you would have had to decide between a "pooled variance" test or a "separate sample" t-statistic.

Give two reasons why -- in this instance -- the P-value would have been just about the same either way.

e[5]   The authors do not explain what the "error bars" on the top of the rectangles in Fig. 1 represent. By eye, I estimate that the lengths of these bars for the two "Spine" BMD's are of the order of 0.08 g/cm^2.

What could they be? Or if you can't be sure, say what they could not be. Explain your reasoning.

c[5]   The authors do not report any P-values, but in at least two places (3 lines from bottom of first column of page 1025, legend to Figure 1) they do use the concept of statistical significance.

Without doing any calculations, deduce directly from what is given in the table the (approximate) 2-sided P-value associated with the 1.03 versus 1.06 for the lumbar BMD. Explain your reasoning.

f5] Suppose you were told that the Lumbar BMD of one of the 524 women was --
when rounded -- 1.35g/cm^2.

What is the probability that the woman was in the "Ever" group?

(use the information below)

		Lumbar BMD categories											
		0.65	0.75	0.85	0.95	1.05	1.15	1.25	**1.35**	1.45	1.55	1.65	(1)
Prior Prob.	OC												
0.87	**Ever**	0.00	0.03	0.12	0.25	0.30	0.20	0.07	**0.01**	0.00	0.00	0.00	(2E)
# 454													
		1.9	13.7	53.4	115	138	91.0	33.3	**6.7**	0.8	0.0	0.0	(3E)
									???				
									???				
		0.1	1.3	5.8	15.0	21.4	16.9	7.4	**1.8**	0.2	0.0	0.0	(3N)
# 70													
0.13	**Never**	0.00	0.02	0.08	0.21	0.31	0.24	0.11	**0.03**	0.00	0.00	0.00	(2N)
		2.1	14.9	59.2	130	159	108	40.7	**8.5**	1.0	0.1	0.0	(T)

Rows (1) : categories of Lumbar BMD (i.e., BMD rounded to end in a 5)
(2E) : Prob(BMD category | Ever)   (= 1)
(2N) : Prob(BMD category | Never)   (= 1)
(3E) : 454 × (2E)   (= 454)
(3N) : 70 × (2N)   (= 70)
(T)   : totals of 3E and 3N (for each Lumbar BMD category)

13   Refer to the article "The effect of working serial night shifts on the cognitive functioning of emergency physicians"

a[12]   "The mean day-shift KAIT score was 119.1 (SD=7.7), and the mean night-shift KAIT score was 107.2 (SD=10.2). This difference was significant (mean difference=11.9; 95% confidence interval 7.0 to 16.8; *P*<.001), with the day-shift scores being statistically higher than the night-shift scores" (Abstract; but see also more complete summaries in Table 1)

   (i) *Reconstruct the 95% CI 7.0 to 16.8 from the summaries given.*

   (ii) State the null and alternative hypotheses tested and verify that "P< .001"

   (iii) Why, in the last row of the Table, doesn't $\sqrt{7.7^2 + 10.2^2}$ equal   9.2 ?

13 ... continued

b[12] "Residents in group B, who were tested first after working night shifts, had a larger difference between their 2 scores than residents in group A, who were tested first on the day shift (night first: mean difference=17.1 [SD=8.6]; day first: mean difference= 6.6 [SD=6.7]; *P*=.017). On the basis of these scores, the order of testing with the KAIT (night first or day first) did make a difference" [Bottom of page 153 and top of page 154]

(i) Reconstruct the P-value (0.017) from the summaries given.

(ii) Explain in words -- to a resident who is working the day shift -- what the P-value of 0.017 is (after the night shift, don't even try!).

(iii) Why did the order of testing make a difference? What is the lesson for investigators who are attracted to the crossover design?