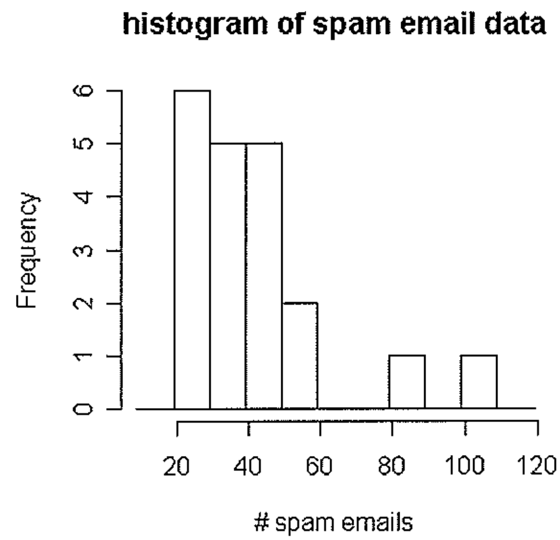
Midterm A Solution key

1. Harris recently installed a spam filter software, but he still saw spam emails in his inbox. He made a daily record of the number of spam emails that were delivered to his inbox over the past 20 days. The following is a frequency histogram for his data. The frequency refers to the number of days.



a) Harris also plotted a stemplot for the data. Which of the following is a correct stemplot for his data? Check only one answer. [2 marks]

✓ Stemplot A
___ Stemplot B
___ Stemplot C

```
A.    2 | 011355        B.    2 | 011355        C.    1 | 05
      3 | 01467               3 | 01467               2 | 011355
      4 | 12479               4 | 12479               3 | 01467
      5 | 56                  5 | 56                  4 | 12479
      6 |                     8 | 0                   5 | 56
      7 |                    10 | 5                   6 |
      8 | 0                                           7 |
      9 |                                             8 | 0
     10 | 5
```

b) What is the third quartile of the number of spam emails? Use the stemplot you have chosen in part (a) to answer this question. Check only one answer. [2 marks]
___ 23
___ 25
___ 47
_✓_ 48

2. Consider the following two studies:

Study 1: A study compared 120 patients with brain cancer to 246 healthy patients without brain cancer. The patients' cell phone use was measured using a questionnaire. The brain cancer patients used cell phones more often, on the average.

Study 2: A study exposed rats to two common types of cell phone radiation for four hours a day, five days a week, for two years. One third of the rats were randomized to be exposed to analog cell phone frequency, one third to digital cell phone frequency, and one third served as controls and received no radiation. At the end of two years, their brains were examined for cancerous tumors. No statistically significant difference in the percentage of brain cancer was found among the groups.

a) True or false? Study 1 shows that cell phone use causes brain cancer. [2 marks]
___True _✓_False

b) Identify the following elements for Study 2:
   i. the experimental unit [2 marks]: a rat
   ii. the factor [2 marks]: type of radiation
   iii. the treatments [3 marks]:
   no radiation, analog cell phone frequency, digital cell phone frequency

3. A city council was planning to turn a major street in the city from a primary traffic artery to a secondary traffic artery. It sent out a questionnaire to all the 36,589 households living in the city requesting for their input concerning the plan. Thirty four percent of the 10,375 households who returned the questionnaires opposed the plan.

   For the following statements, check all that are correct. [3 marks]

   _✓_ This survey conducted by the city council is likely to suffer from nonresponse bias.

   ____ The 10,375 households that returned the questionnaires formed a random sample of the population.

   ____ The percentage of the 10,375 households that opposed the plan, 34%, is a parameter.

4. You need to drive past two traffic lights on the way from your house to the nearest grocery store. The probability that you hit a red light is 0.5 at the first intersection and 0.4 at the second intersection. The probability that you run into a red light at both intersections is 0.25. On a random day you drive from home to that grocery store.

   Define the following events:

   E_1 = you run into a red light at the first intersection
   E_2 = you run into a red light at the second intersection
   E_3 = you run into a green light at both intersections
   E_4 = you run into a red light at both intersections

   Which of the following statements is (are) true about the above events? Check all that are correct. [6 marks]

   ____ E_1 and E_2 are independent events.

   ____ E_1 and E_2 are disjoint events.

   _✓_ E_3 and E_4 are disjoint events.

   ____ E_3 is the complement of E_4.

5. You draw two cards without replacement from a deck of 52 cards. If the first card is not a spade, find the probability that the two cards drawn are both diamonds. [6 marks]

Define events: A = the first card drawn is a spade
D_i = the ith card drawn is a diamond, $i=1,2$

$$P(D_1 \text{ and } D_2 \mid A^c) = \frac{P(D_1 \text{ and } D_2 \text{ and } A^c)}{P(A^c)}$$

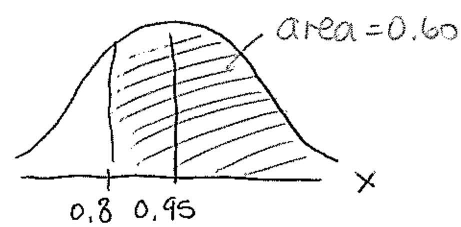
- Here $P(D_1 \text{ and } D_2 \text{ and } A^c)$ is the probability of getting a diamond on the first draw (who automatically satisfies the fact that it is not a spade) and also a diamond on the second draw, i.e. $P(D_1 \text{ and } D_2) = P(D_1) \times P(D_2 \mid D_1) = \frac{13}{52} \times \frac{12}{51}$
- $P(A^c) = 1 - P(A) = 1 - \frac{13}{52} = \frac{39}{52}$
- $P(D_1 \text{ and } D_2 \mid A^c) = \left(\frac{13}{52} \times \frac{12}{51}\right) \Big/ \frac{39}{52} = \frac{4}{51}$

6. The length of trout in a lake is normally distributed with mean $\mu = 0.95$ feet and an unknown standard deviation σ. If 60% of all trout are longer than 0.8 feet, what is the value of σ? [6 marks]

Let X = length of a trout $\sim N(\mu = 0.95 \text{ ft}, \sigma)$

60% of all trout are longer than 0.8 ft

$\Rightarrow P(X > 0.8) = 0.60$



0.8 is the 40th percentile.

z_{40} (40th percentile of $Z \sim N(0,1)$) is -0.255

$$\therefore \quad z_{40} = \frac{0.8 - \mu}{\sigma}$$

$$\sigma = \frac{0.8 - \mu}{z_{40}} = \frac{0.8 - 0.95}{-0.255} = 0.588 \text{ ft.}$$

7. In a university parking database with 5600 registered vehicles, records show that 43% of the registered vehicles are Asian makes, 23% are European makes and the remaining are American makes. Among all the 5600 cars, 20% once received a parking ticket.

a) You randomly pick three vehicles with replacement from the database. What is the probability that at most two of the three are American makes? [6 marks]

Let X = # American cars out of the 3 chosen

$X \sim Bin(n=3,\ p = 1-0.43-0.23 = 0.34)$

P(at most 2 out of 3 are American makes)

$= P(X \leq 2)$

$= P(X=0) + P(X=1) + P(X=2)$

$= \binom{3}{0} 0.34^0 (1-0.34)^3 + \binom{3}{1} 0.34^1 (1-0.34)^2 + \binom{3}{2} 0.34^2 (1-0.34)^1$

b) Consider a random sample of 100 vehicles selected from the database. The sample proportion of the 100 selected vehicles that <u>had never received</u> a parking ticket has an approximate 95% chance of falling between 0.72 and 0.88.
Fill in the blanks and show your calculation below. [4 marks]

Let p = population proportion of cars that never received a parking ticket $= 1 - 0.20 = 0.80$

$\hat{p}$ = sample proportion of cars (out of 100) that never received a parking ticket

Note: $np = 100(0.80) = 80 > 10$, $n(1-p) = 20 > 10$

$\therefore$ $\hat{p}$ follows the normal distribution approximately with mean $= 0.8$ and SD $= \sqrt{\frac{0.8(1-0.8)}{100}} = 0.04$

By 68-95-99.7 % rule, the middle 95% of coverage will be over the interval:

$0.8 \pm 2(0.04)$

$\Rightarrow (0.72, 0.88)$

8. Two stores sell watermelons. At the first store the melons weigh an average of 20 pounds with a standard deviation of 2.2 pounds. The melons are sold for 36 cents a pound. At the second store the melons are smaller, with a mean of 17 pounds and a standard deviation of 2 pounds. The store is having a sale on watermelons – only 25 cents a pound. Assume that the weights are normally distributed. Jenny selects a melon at random at each store. Find the mean and the variance of the difference in the prices Jenny pays for the two melons. [6 marks]

Let X = weight of a melon from first store
Y = " " " " " second store

$E(X) = 20$, $SD(X) = 2.2$
$E(Y) = 17$, $SD(Y) = 2$

Difference in the prices between the 2 melons
$= D = 0.36X - 0.25Y$ (in dollars)

$$E(D) = E(0.36X - 0.25Y)$$
$$= 0.36\,E(X) - 0.25\,E(Y)$$
$$= 0.36(20) - 0.25(17) = \$2.95$$

$$V(D) = V(0.36X - 0.25Y)$$
$$= 0.36^2\,V(X) + 0.25^2\,V(Y)$$

[It makes sense to assume the weights of the 2 melons are independent]

$$= 0.36^2(2.2^2) + 0.25^2(2^2)$$
$$= 0.8773 \text{ dollars}^2$$

9. Each day the value of a particular stock goes up one unit with probability 0.3, stays the same with probability 0.5 or else goes down one unit with probability 0.2. Taking changes over consecutive days to be independent of each other, estimate the probability that the stock will have increased by a value of at least five units over 500 days. [10 marks]

Let X= amount of change on a day

probability dist'n of X is

x	-1 (down)	0 (stays the same)	$+1$ (up)
$P(X=x)$	0.2	0.5	0.3

The overall change over 500 days :

$Y = X_1 + X_2 + \dots + X_{500}$

Want to find $P(Y \geqslant 5)$

$P(Y \geqslant 5)$

$= P(\bar{X} \geqslant \frac{5}{100} = 0.01)$

How is $\bar{X}$ distributed?

$\bar{X}$ will be approximately normal by CLT ($n = 500 > 30$)

$\bar{X} \overset{\text{approx.}}{\sim} N(\mu_{\bar{X}} = E(X),\ \sigma_{\bar{X}} = \frac{SD(X)}{\sqrt{500}})$

$= P\left(\frac{\bar{X} - 0.1}{0.7/\sqrt{500}} \geqslant \frac{0.01 - 0.1}{0.7/\sqrt{500}}\right)$

$= P(Z \geqslant -2.875)$

$= 1 - 0.0020$

$= 0.998$

$E(X) = \sum x\, P(X=x)$

$= -1(0.2) + 0(0.5) + 1(0.3)$

$= +0.1$

$V(X) = E(X^2) - (E(X))^2$

$= \sum x^2 P(X=x) - (0.1)^2$

$= (-1)^2(0.2) + 0^2(0.5) + 1^2(0.3) - 0.01$

$= 0.49$

$SD(X) = \sqrt{0.49} = 0.7$

Midterm B Solution Key

1. Circle the most appropriate answer: [2 marks each]

a) Form a data set that consists of four integer numbers from 1 to 10 (inclusive, without repeats).

Among all the possible data sets that can be formed (as described in the above), which of the following statements is **NOT** correct?

i. The set of numbers 1, 2, 3, 4 gives the smallest possible standard deviation.
ii. The set of numbers 4, 5, 6, 7 gives the smallest possible standard deviation.
iii. The set of numbers 1, 5, 6, 10 gives the largest possible standard deviation.
iv. The set of numbers 1, 2, 9, 10 gives the largest possible standard deviation.

b) The length of a ball of yarn is a random variable with mean 150 ft and standard deviation 2 ft. The variance of the total length of three randomly chosen balls of yarn will be

i. 6 ft^2
ii. 12 ft^2
iii. 18 ft^2
iv. 36 ft^2

c) Sixty-four percent of the teenager population is nearsighted. Consider random samples of 4 teenagers drawn from this population. The following are three statements about the sampling distribution of the sample proportion (of 4 teenagers) who do not suffer from nearsightedness.

(I) The sampling distribution has a mean of 36%.
(II) The sampling distribution has a standard deviation of 24%.
(III) The sampling distribution follows the normal distribution approximately.

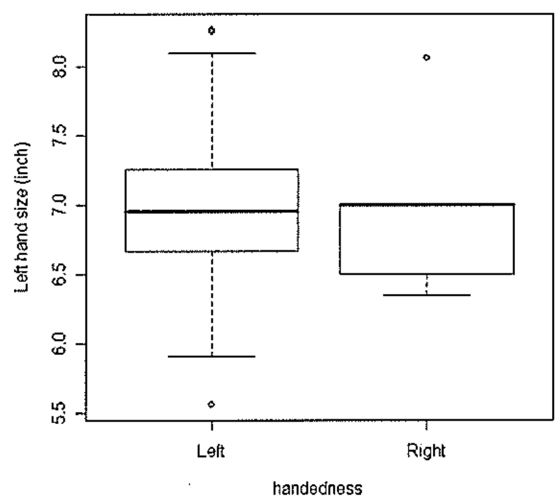
Which of the above statements is (are) correct?

i. Statement (I) only.
ii. Statements (I) and (II) only.
iii. All the three statements.
iv. None of the three statements.

d) There are two urns. Each urn contains 1 black marble and 2 white marbles. You randomly draw one marble from each urn. Let X be the number of white marbles out of the two being drawn. Then

i. X is a Bin($2,\frac{2}{3}$) random variable.
ii. X is a Bin($3,\frac{2}{3}$) random variable.
iii. X is a Bin($6,\frac{1}{2}$) random variable.
iv. X is not a Binomial random variable.

2. The following boxplots show left hand lengths. The boxplot on the left shows data from left handers. The boxplot on the right shows data from right handers.



Which of the following statements is (are) correct about the distributions of the left hand size data for the two handedness groups? Check all that apply. [4 marks]

_✓_ The distribution among the right handers is skewed to the right.
____ For the distribution among the right handers, the first quartile is the same as the third quartile.
____ The left and right handers have roughly the same mean left hand size.
____ More than 25% of the left handers have left hand sizes that are shorter than the minimum left hand size of the right handers.

3. A farmer wants to study the effect of plant density on tomato yields. He buys two varieties of tomato plants: A and B – 24 each. He then randomizes the 24 plants of variety A to one of the three density choices: 4 plants/m^2, 6 plants/m^2 and 8 plants/m^2. He does the same to the 24 plants of variety B. At harvest, he measures the yield (in kilograms) of each tomato plant.

   Identify the following components of the farmer's experiment to study the effect of plant density: [12 marks]

   a) experimental units: tomato plants

   b) factor(s) (how many and list it / them all):

      1 factor : plant density

   c) levels of each factor (how many and what are they?):

      3 levels of plant density : 4, 6, 8 plants/m^2

   d) treatments (how many and what are they?):

      3 treatments : 4, 6, 8 plants/m^2

   e) response variable:

      yield of a tomato plant

   f) design type (check only one):

      ____ completely randomized design
      _✓_ randomized block design
      ____ matched pairs design

   Which of the following is (are) involved in the experiment? Check all that apply.

   _✓_ Replication.
   ____ Confouding.
   ____ A control group.

4. Your friend, Carl, has been doing a research project about the coffee drinking habit of UBC students. Last Wednesday during 7:30-8:30 am, he stood next to the Starbucks coffee counter in the SUB and interviewed all students who bought coffee there.

   Carl asked each student two questions:
   1. How many cups of coffee do you drink daily?
   2. Do you have an early morning class (an 8am or 9am class)?

   Carl found that the average number of cups of coffee these interviewed students drink daily is 1.78.

   a) The students interviewed by Carl formed (check only one): [2 marks]

      ____ a simple random sample.
      _✓_ a convenience sample.
      ____ a stratified random sample.
      ____ a multistage sample.

   b) The average number of cups of coffee consumed daily by the interviewed students, 1.78, is (check only one): [2 marks]

      ____ a population.
      ____ a sample.
      ____ a parameter.
      _✓_ a statistic.

   c) Do you think the average number of cups of coffee, 1.78, provides a reliable estimate of the true mean daily coffee consumption of all students in UBC? Briefly explain why or why not. [2 marks]

      No, Carl's sample suffers from the problem of undercoverage. Students who don't drink coffee were not interviewed at all. The average, 1.78, overestimates the true mean daily coffee consumption.

5. The following table shows the breakdown of the annual salaries of university graduates and the proportion of graduates falling in each income category.

Annual salary range	Proportion of university graduates
under \$20k	0.15
\$20k to \$40k	0.45
\$40k to \$60k	0.30
\$60k to \$100k	0.09
above \$100k	0.01

a) A university graduate is randomly chosen. Here are two events:

Event A : The chosen graduate has an annual salary under \$20k.

Event B : The chosen graduate has an annual salary under \$40k.

Are the two events independent? ___Yes ✓ No

Explain your reasoning. [4 marks]

$P(A) = 0.15$ $P(B) = 0.15 + 0.45 = 0.60$

If A has occurred, B must have occurred, In other words, the occurrence of A has altered the probability of B [$P(B) = 0.60$, while $P(B|A) = 1$, and $P(B) \neq P(B|A)$]

A, B are not independent.

OR $P(A \text{ and } B) = 0.15 \neq P(A) \times P(B) = 0.15 \times 0.60 = 0.09$

b) Two university graduates are chosen at random. What is the probability that one has an annual salary under \$20k and the other has an annual salary between \$20k and \$60k? [6 marks]

Define 2 events :

A_i = Salary is under \$20K for the ith individual

B_i = salary is between \$20K and \$60K for the ith individual

$i = 1, 2$

$P(A_i) = 0.15$, $P(B_i) = 0.45 + 0.30 = 0.75$

$P(\underbrace{\{A_1 \text{ and } D_2\}}_{} \text{ OR } \underbrace{\{A_2 \text{ and } D_1\}}_{})$

disjoint

$= P(A_1 \text{ and } D_2) + P(A_2 \text{ and } D_1)$ by addition rule

$= P(A_1) \times P(D_2) + P(A_2) \times P(D_1)$ by multiplication rule as the 2 individuals should be independent.

$= 0.15 \times 0.75 + 0.15 \times 0.75$

$= 0.225$

we want to approximate the probability!

c) A random sample of 400 university graduates is drawn from all university graduates. $\boxed{\text{Approximate}}$ the probability that 180 or fewer of the sampled graduates earn more than \$40k annually. State any assumption(s) you have made in the calculation. [7 marks]

Let X = # graduates out of 400 who earn > \$40K annually.

$X \sim Bin(n=400, \; p=0.30+0.09+0.01=0.40)$

We use normal approximation to the Binomial

Check conditions: (1) we have a random sample of graduates

(2) $n=400 < 5\%$ of population size (all graduates)

(3) $np = 400(0.4) = 160 > 10, \quad n(1-p) = 240 > 10$

$X \overset{approx.}{\sim} N(\mu = np = 160, \; \sigma = \sqrt{np(1-p)} = 9.797959)$

$P(X \leq 180) = P(X < 180.5)$ ← continuity correction

$= P\left(Z < \frac{180.5-160}{9.797959}\right) = P(Z < 2.09) = 0.9817$

6. At the present time the noise level per jet takeoff in one neighbourhood near the airport is believed to follow the normal distribution with mean 102 decibels and standard deviation 5 decibels. Suppose a regulation is passed that requires jet noise in this neighbourhood to be lower than 105 decibels 95% of the time. How much should the mean noise level be lowered to comply with the regulation? [6 marks]

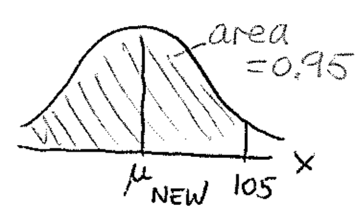
Let X = noise level under the regulation $\sim N(\mu_{NEW}, \sigma = 5)$

↑ assume σ is unchanged.

Want X be < 105 dB 95% of the time

i.e. $P(X < 105) = 0.95$

105 dB is the 95th percentile.



z_{95} = [z score for area = 0.95]

$= 1.645$

$\therefore \; z_{95} = \frac{105 - \mu_{NEW}}{\sigma}$

$\mu_{NEW} = 105 - \sigma \times z_{95} = 105 - 5 \times 1.645 = 96.775$ dB

The mean noise level should be lowered by

$(102 - 96.775) = 5.225$ dB.

7. According to a research study of fish in the Tennessee River, the weights of fish in the river follow the normal distribution with mean 950 grams and standard deviation 180 grams.

   a) True or false? The average weight of a random sample of 36 fish drawn from the Tennessee River follows the normal distribution. This is a result of the Central Limit Theorem. [2 marks]

   ___ True

   _✓_ False

   b) Find the probability that a random sample of 36 fish drawn from the Tennessee River has an average weight of at least 980 grams. [5 marks]

Let X = weight of a fish $\sim N(\mu = 950g, \sigma = 180g)$

$\bar{X}$ = average weight of 36 fish $\sim N(\mu_{\bar{X}} = 950g, \sigma_{\bar{X}} = \frac{180}{\sqrt{36}} = 30g)$

$$P(\bar{X} \geq 980)$$
$$= P\left(\frac{\bar{X} - 950}{30} \geq \frac{980-950}{30}\right)$$
$$= P(Z \geq 1)$$
$$\doteq \frac{1}{2}(1 - 0.68)$$ by 68-95-99.7 rule
$$= 0.16$$