MAT 2379A
Midterm Examination (with solutions)

November 2, 2016 **Professor Raluca Balan**
Time: 80 minutes

Student Number:_____

Family Name: _____ **First Name:** _____

This is a closed book examination. A formula sheet and some statistical tables are included with the exam. Only Faculty standard calculators are permitted. Record your answer to each question in the table below.

Question	Answer
1	E
2	C
3	C
4	A
5	E
6	A
7	E
8	D
9	B
10	A
11	B
12	C

NOTE: At the end of the examination, hand in only this page. You may keep the questionnaire.

.**********.

1. The vanilly mandelic acid test is a urine test which was developed to detect the presence of neuroblastoma, a rare serious disease. This test gives a positive result in 70% of cases of neuroblastoma. 8 children with neuroblastoma are administered this test. We want to calculate the probability that the test will give an incorrect (negative) result for at least 2 children in this group. Which one of the following R commands gives this probability?

   A) pbinom(2,8,0.7) B) 1-pbinom(1,8,0.7) C) pbinom(1,8,0.7)
   D) pbinom(2,8,0.3) E) 1-pbinom(1,8,0.3)

   Solution: Let X be the number of children (in the group of 8) who will receive a negative test result. X is a binomial random variable with $n = 8$ trials and probability of success $1 - 0.7 = 0.3$. The desired probability is:

   $$P(X \geq 2) = 1 - P(X \leq 1) = 1 - \text{pbinom}(1, 8, 0.3).$$

   The answer is E.

2. In a city, 6% of the population smokes cigars, 14% of the population smokes cigarettes, and 84% of the population smokes neither cigarettes nor cigars. Let A be the event that a randomly selected resident in this city smokes cigars and B the event that this person smokes cigarettes. Which one of the following statements is **incorrect**?

   A) $P(A \cup B) = 0.16$ B) $P(A' \cup B') = 0.96$
   C) $P(A \cap B) = P(A)P(B)$ D) $P(A \cup B') = 0.90$
   E) $P(A \cap B') < P(A' \cap B)$

   Solution: We know that $P(A) = 0.06$, $P(B) = 0.14$, and $P(A' \cap B') = 0.84$. Hence, $P(A \cup B) = 1 - P(A' \cap B') = 1 - 0.84 = 0.16$, so option A is correct. By addition rule, we have

   $$P(A \cap B) = P(A) + P(B) - P(A \cup B) = 0.06 + 0.14 - 0.16 = 0.04.$$

   Therefore, we can further compute

   $$\begin{aligned} P(A' \cup B') &= 1 - P(A \cap B) = 1 - 0.04 = 0.96, \\ P(A \cap B') &= P(A) - P(A \cap B) = 0.06 - 0.04 = 0.02, \\ P(A' \cap B) &= P(B) - P(A \cap B) = 0.14 - 0.04 = 0.10, \\ P(A \cup B') &= 1 - P(A' \cap B) = 1 - 0.10 = 0.90. \end{aligned}$$

Thus, options B, D, E are correct. Note that

$$P(A \cap B) = 0.04 \neq P(A)P(B) = (0.06)(0.14) = 0.0084,$$

so that C is incorrect. The answer is C.

3. There are two male rats and two female rats living in a cage. We select randomly two rats from this cage. The selection is done without replacement, i.e. the first selected rat is not returned to the cage after it was selected. Let X be the number of males among the two selected rats. Let $F(x)$ be the cumulative distribution function of X and $f(x)$ be the probability mass function of X. Which one of the following statements is **incorrect**?

   A) $F(x) > 0$ when $x = 0$
   B) $f(x) > 0$ when $x = 0$
   C) $F(x) = 0$ when $x = 3$
   D) $f(x) = 0$ when $x = 3$
   E) $f(x) = F(x)$ when $x < 0$

   Solution: X is a discrete random variable, which takes values in $\{0, 1, 2\}$ with positive probabilities. By definition of the functions F and f, we have

$$F(0) = P(X \leq 0) = P(X = 0) = f(0) > 0,$$

$$F(3) = P(X \leq 3) = 1, \quad f(3) = P(X = 3) = 0.$$

   Since $X \geq 0$, we have $F(x) = P(X \leq x) = 0$ and $f(x) = P(X = x) = 0$ for $x < 0$. So A,B,D and E are correct. The only incorrect statement is C since $F(3) = P(X \leq 3) = 1$. The answer is C.

4. Individuals at risk for colon cancer are advised to have a colonoscopy beginning at age 50. The following table summarizes the results for 250 patients at risk for colon cancer:

	Colon cancer: yes	Colon cancer: no	Total
Colonoscopy: yes	x	50-x	50
Colonoscopy: no	5-x	195+x	200
Total	5	245	250

   In this table, x denotes the number of patients with colon cancer who had a colonoscopy. Find the value of x such that A and B are independent events, where A is the event that a randomly selected patient from

this group has colon cancer, and B is the event that this patient had a colonoscopy. For this value of x, can we say that A' and B' are also independent events?

A) $x = 1$. Yes, for $x = 1$, A' and B' are also independent.
B) $x = 1$. No, for $x = 1$, A' and B' are not independent.
C) $x = 5$. Yes, for $x = 5$, A' and B' are also independent.
D) $x = 5$. No, for $x = 5$, A' and B' are not independent.
E) There is no value of x for which A and B are independent.

Solution: From the table, we see that $P(A \cap B) = x/250$, $P(A) = 5/250$ and $P(B) = 50/250$. Events A and B are independent if $P(A \cap B) = P(A)P(B)$, that is

$$\frac{x}{250} = \frac{5}{250} \cdot \frac{50}{250}.$$

The unique solution to this equation is $x = 1$. In this case, events A' and B' are also independent. The answer is A.

5. In humans, the hair color is determined by a gene whose allele for dark hair (D) is dominant over the allele for red hair (d), while the eye color is determined by a gene whose allele for brown eyes (B) is dominant over the allele for blue eyes (b). A women has blue eyes and is dark-haired, being heterozygous for the hair color gene. A man is red-haired and has brown eyes being heterozygous for the eye color gene. The woman has blood type AB and the man has blood type O. Calculate the probability that their child has blue eyes, red hair and blood type A.

A) 1/4 B) 0 C) 1/2 D) 1 E) 1/8

Note: The blood type is determined by the alleles I^A, I^B, i of a gene denoted by I: I^A and I^B are dominant over i, but I^A and I^B are codominant with each other. A type A person can have genotype I^AI^A or I^Ai, a type B person can have genotype I^BI^B or I^Bi, a type O person has genotype ii, and a type AB person has genotype I^AI^B.

Solution: From the tree diagram, we see that there are 8 possible genotypes for their child: I^AiBbDd, I^AiBbdd, I^AibbDd, I^Aibbdd, I^BiBbDd, I^BiBbdd, I^BibbDd, I^Bibbdd. The probability that the child has blue eyes, red hair and blood type A is $1/8$. The answer is E.

6. Ottawa Public Health collects beach water samples every day during the summer. A non-swimming advisory is issued for a beach if bacteria level is over 200E.coli per 100 ml of water tested. Assume that the bacteria level is a normal random variable with mean $\mu = 120$ and standard deviation $\sigma = 50$. What is the probability that a non-swim advisory will be issued for Mooney's Bay beach at least twice in the next 5 years on Canada Day?

   A) 0.027 B) 0.246 C) 0.055 D) 0.163 E) 0.378

   Solution: Let X be the bacteria level for the beach water sample in a randomly chosen day. Using standardization and Table 17.3, we see that the probability that a non-swim advisory level is issued for that day is:

$$\begin{aligned} P(X > 200) &= P\left(\frac{X-120}{50} > \frac{200-120}{50}\right) \\ &= P(Z > 1.6) = 1 - 0.9452 = 0.0548. \end{aligned}$$

   Let Y be the number of times a non-swim advisory will be issued for Mooney's Bay beach in the next 5 years on Canada Day. Then Y has a binomial distribution with $n = 5$ trials and probability of success $p = 0.0548$. The desired probability is:

$$\begin{aligned} P(Y \geq 2) &= 1 - P(Y = 0) - P(Y = 1) \\ &= 1 - (0.0548)^0(1 - 0.0548)^5 - 5(0.0548)^1(1 - 0.0548)^4 \\ &= 0.027. \end{aligned}$$

   The answer is A.

7. A certain form of cancer is found in women over the age of 60 with probability 6.5%. A blood test exists for the detection of the disease, but the test is not perfect. This test has a false negative rate of 10% and a false positive rate of 4.5%. A woman over the age of 60 took the test and received a favorable (i.e. negative) test result. What is the probability that she has this form of cancer?

   A) 0.0025 B) 0.0145 C) 0.1000 D) 0.0450 E) 0.0072

   Solution: Let D be the event of having the disease and let $T-$ and $T+$ be the events that the result from the test is negative and is positive, respectively. We are given $P(D) = 0.065$, false negative rate$= P(T-|D) = 0.1$

and false positive rate$= P(T+|D') = 0.045$. We start by computing the probability that a test will be negative. By the total probability rule:

$$\begin{aligned} P(T-) &= P(T-|D)P(D) + P(T-|D')P(D') \\ &= (0.1)(0.065) + (1-0.045)(1-0.065) = 0.8994. \end{aligned}$$

We want the following probability:

$$P(D|T-) = \frac{P(D \cap T-)}{P(T-)} = \frac{P(T-|D)P(D)}{P(T-)} = \frac{(0.1)(0.065)}{0.8994} = 0.0072.$$

The answer is E.

8. A research scientist reports that mice will live a mean of 40 months when their diets are sharply restricted and then enriched with vitamins and proteins. Assuming that the lifetimes of such mice are normally distributed with a standard deviation of 6.3 months, find the probability that a given mouse will live between 37 and 49 months.

   A) 0.3920 B) 0.5845 C) 0.7205 D) 0.6080 E) 0.3888

   Solution: Let X be the lifetime of such mouse. X has a normal distribution with mean $\mu = 40$ and standard deviation $\sigma = 6.3$. We want

$$\begin{aligned} P(37 < X < 49) &= P\left(\frac{37-40}{6.3} < \frac{X-\mu}{\sigma} < \frac{49-40}{6.3}\right) \\ &= \Phi(1.43) - \Phi(-0.48) \\ &= 0.9236 - 0.3156 = 0.6080 \end{aligned}$$

   The answer is D.

9. A screening test is applied to 200 patients suffering from a certain disease. This test is also applied to 200 persons selected randomly from the community who do not have the disease, called "controls". The following table summarizes the test results:

	Patients who have the disease	Community controls	Total
Positive tests	166	18	184
Negative tests	34	182	216
Total	200	200	400

What is the specificity of this test?

A) 0.73 B) 0.91 C) 0.82 D) 0.89 E) 0.10

Solution:

$$\text{specificity} = P(\text{Test}-|\text{True}-) = \frac{P(\text{True}- \text{ and Test}-)}{P(\text{True}-)}$$

$$= \frac{182/400}{200/400} = \frac{182}{200} = 0.91.$$

The answer is B.

10. The following table gives the height measurements (in cm) of 9 female students at University of Ottawa:

160	159	164	169	176	163	161	160	178

Which of the following statements is **correct**?

A) There are no outliers. B) 178 is the only outlier.
C) 159 is the only outlier. D) 178 and 176 are two outliers.
E) 159 and 178 are two outliers.

Solution: We arrange the data in an increasing order:

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9
159	160	160	161	163	164	169	176	178

Using the formula from the book and class notes, we can compute that

$$\begin{aligned}
(1/4)(n+1) &= 10/4 = 2.5 = 2 + 0.5 \\
q_1 &= 0.5y_2 + 0.5y_3 = (160+160)/2 = 160 \\
(3/4)(n+1) &= 30/4 = 7.5 = 7 + 0.5 \\
q_3 &= 0.5y_7 + 0.5y_8 = (169+176)/2 = 172.5
\end{aligned}$$

So, $IQR = q_3 - q_1 = 172.5 - 160 = 12.5$. The upper and lower fences are given by

$$\begin{aligned}
\text{Upper fence} &= q_3 + 1.5IQR = 172.5 + 1.5 \times 12.5 = 191.25 \\
\text{Lower fence} &= q_1 - 1.5IQR = 160 - 1.5 \times 12.5 = 141.25
\end{aligned}$$

Since all measurements are within [141.25, 191.25], there are no outliers in this dataset. The answer is A.

11. The police had set up radar surveillance on a particular street and handed out a large number of tickets to drivers who drove over the speed limit. Let x be the difference between the measured speed of the vehicle and the speed limit. From a very large sample of speeding infractions on this street, we computed some descriptive statistics for x:

mean	35 km/h
standard deviation	3.9 km/h
IQR	5.2 km/h
max	53 km/h

Suppose that in this jurisdiction, the law prescribes a fine of \$100 plus \$10 per km/h over the speed limit. So the fine y (in dollars) is computed as $y = 100 + 10\,x$. Give the corresponding descriptive statistics for the fine y given out on this street.

Which of the following columns A, B, C, D, or E is **correct**?

	A	B	C	D	E
mean	\$450	\$450	\$45	\$450	\$450
standard deviation	\$139	\$39	\$39	\$39	\$139
IQR	\$152	\$52	\$52	\$152	\$52
max	\$630	\$630	\$63	\$630	\$630

Solution: We have a linear function with a scaling factor of 10 and a shift of 100. We must take both the scaling and the shift, when we compute the mean and the maximum:

$$\text{mean} = 100 + 10\,(35) = \$450 \quad \text{and} \quad \text{max} = 100 + 10\,(53) = \$630.$$

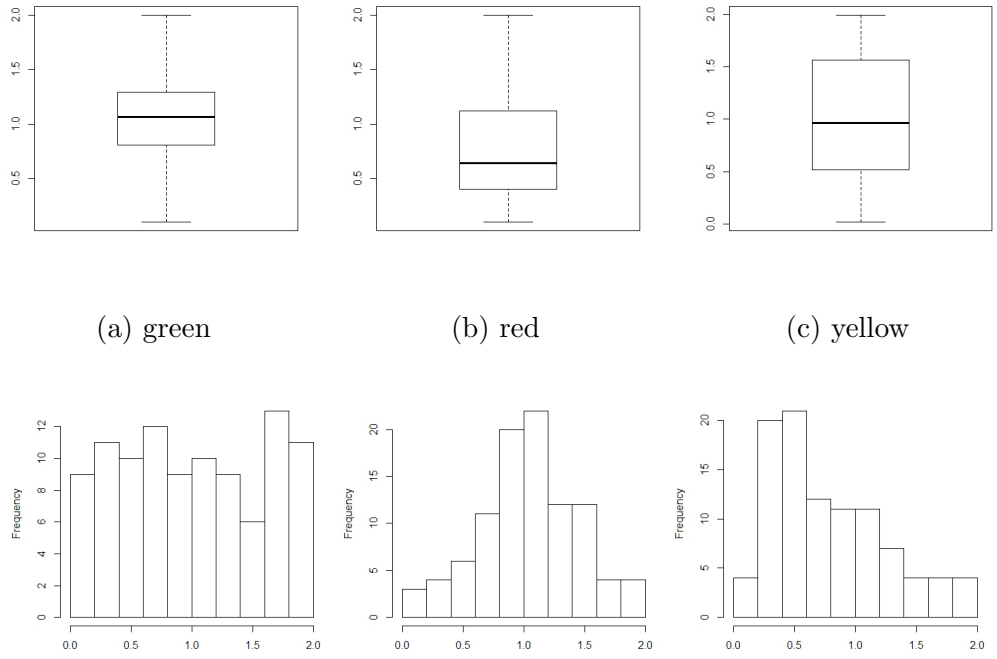
As we scale by a factor of 10, the distance between the values becomes 10 times larger. But as we then add 100 to all the values, this shift does not effect distances between values. So the measures of dispersion are not affected by this shift. This means that when considering measures of dispersion such as the standard deviation and the IQR, we only need

to consider the scaling by a factor of 10. The shift does not affect the measures of dispersion. So we get

$$\text{standard deviation} = 10\,(3.9) = \$39 \quad \text{and} \quad \text{IQR} = 10\,(5.2) = \$52.$$

The answer is B.

12. The most common species of sweat bees are green, red, and yellow. For each color, we select a sample of 100 of bees of this color and record their length (in inches). We then create 3 data sets of 100 observations each, called "green", "red" and "yellow". Below are the boxplots and histograms for these data sets. The labels of the variables are missing from the histograms, but are included in the boxplots. Our task is to identify the missing labels. Which one of the following statements is **correct**?



(a) green (b) red (c) yellow

(d) (e) (f)

A) Histogram (d) is for green, (e) is for red, and (f) is for yellow.
B) Histogram (f) is for green, (e) is for red, and (d) is for yellow.

C) Histogram (e) is for green, (f) is for red, and (d) is for yellow.
D) Histogram (f) is for green, (d) is for red, and (e) is for yellow.
E) Histogram (d) is for green, (f) is for red, and (e) is for yellow.

Solution: We have to match each boxplot with the corresponding histogram. All samples have the same range, so the whiskers cannot be used for the matching. Notice that one of the histogram is skewed to the left, while the other two histograms are approximately symmetric. Here we would have to identify a median away from the center for the skewed distribution. Therefore b matches with f. For the symmetric distributions, we can compare the dispersion for the matching. Notice that the values in histogram e are less dispersed (i.e. more concentrated in the center) compared to the values in histogram d. When comparing the boxplots a and c, we should notice that the box in a is smaller (i.e. less dispersed). Therefore, a is matches with e and c is matched with d. The answer is C.