STATISTICS 20 **Shobhana M. Stoyanov**
FALL 2018 **MIDTERM ANSWERS**

---

1. Below is the distribution table for the average number of hours per day spent by students at a certain university on their laptops, doing things unrelated to academics. The right endpoint is included, but not the left. Assume we can continuously measure time.

Number of hours	0-2	2-4	4-8	8-12	12-20
Height (% per hour)	6	14	8		1.25

   (a) Find the missing value (2 points)
   If you find the total area of the other bins, you get (12+28+32+10)% or 82%, which means the total area of the bin from 8-12 hours 18%. The height is therefore 18/4 = 4.5% per hour.

   (b) Find the median of the distribution. (2 points)
   To find the median, we assume the data are uniformly spread within each bin. Up to the bin marked 4-8 hours, we have (12+28)% or 40%. Therefore we need 10% more for the median, out of the 32% in the bin marked 4-8 hours. 10% will need $\frac{10}{32}\times 4$ or 1.25 hours of this bin. Therefore the median is 4+1.25 or 5.25 hours. That is, you know that the height of this bin is 8% and therefore you need a width of 1.25 to get to 10%.

   (c) True or False: The median for this data is greater than the average of this data. (2 points)
   False. The data are right skewed. Therefore the median is less than the average.

---

2. Indicate whether the following statements are true or false, and give a brief explanation: (2 points each)

   (a) If in a large statistics course, the scores for the final followed the normal curve closely. The average was 70 points out of 100, and three-fourths of the class scored between 60 and 80 points, we can conclude that the sd of the scores was less than 10 points. **[T]** **F**
   The sd of the scores must be less than 10, since only about 68% of the data should be within one sd of the mean, and we have 75% of the data with 10 points of the mean.

   (b) If you add 10 to each entry on a list, that leaves the sd of the list unchanged. **[T]** **F**
   The spread is unchanged when you shift the data. If you look at the formula for sd, the mean will increase by 10 as well, so increase in the entries and the increase in the mean will cancel out when you compute the deviation from the mean.

   (c) If the sd of a list is 0, it must be true that all the numbers on the list are 0. **T** **[F]**
   If the sd is 0 it means the list has no spread, therefore all the numbers must be the *same*.

   (d) It is always possible to design a double-blind controlled experiment, provided the budget is sufficiently large. **T**
   [F] No, it is not always possible. For example, you may assign the subjects to different diets.

---

3. Suppose A and B are independent events such that $P(A) = 0.2$ and $P(B) = 0.6$.

   (a) Is it possible for A and B to be mutually exclusive? (2 points)
   No, since they are independent, the probability of their intersection is 0.12, not 0. They cannot be both independent and mutually exclusive.

   (b) What is the probability of **neither** A **nor** B occurring? (2 points)
   1-P(A or B) = 1 - (0.2+0.6-0.12) = 1 - 0.68 = 0.32.

---

4. I roll a fair **ten**-sided die **three** times.

   (a) What is the probability that all the rolls are different? (2 points)

$$\frac{10}{10}\times\frac{9}{10}\times\frac{8}{10} = 0.72$$

   (b) What is the probability that not all the rolls are the same? (2 points)

$$1 - P(\texttt{all the rolls are the same}) = 1 - \frac{10}{10}\times\frac{1}{10}\times\frac{1}{10} = 0.99$$

5. Two cards are dealt off the top of a well-shuffled standard deck (52 cards, 4 suits: hearts, clubs, spades, diamonds, 13 cards in each suit labeled 2-10, Jack, Queen, King, Ace). Compute probabilities of the following events:

   (a) The second card is a queen. (2 points)
   4/52, since each card can occupy the second position with chance 1/52.

   (b) The cards are a king and a queen. (2 points)
   The chance a king is the first card is 4/52, and the chance that a queen is the second card given that a king is the first card is 4/51. We then have to consider the reverse possibility, with queen first and king second, giving us a total of $2 \times \frac{4}{52} \times \frac{4}{51}$.

---

6. Can we model the following situations using binomial random variables? If yes, indicate what n and p should be. If not, indicate why. (2 points each)

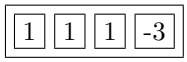
   (a) The number of aces in a 6-card hand dealt off a well-shuffled standard deck of cards.
   No. The probabilities do not stay the same from draw to draw as we draw without replacement.

   (b) If a die is rolled three times and a coin is tossed three times, the sum of the number of times the coin lands heads and the die lands with an odd number of spots.
   Yes, as the probabilities are the same for heads or odd number of spots. We can model this using a box with one ticket marked 0 and one ticket marked 1, and we draw six times from this box. The first three draws represent the coin toss, and the next three represent the die rolls. Therefore n = 6 and p = 0.5.

---

7. Let X be a discrete random variable such that X takes the values 1 and -1 with probabilities 0.4 each and the value 2 with probability 0.2. Find the expected value of X. (2 points)
   E(X) = 1× 0.4 + (-1) × 0.4 + 2 × 0.2 = 0.4.

---

8. Consider the following box of tickets: 

   (a) What is the standard deviation of the tickets in the box? (2 points)

$$\texttt{Box sd} = (\texttt{big - small}) \times \sqrt{\texttt{fraction of big} \times \texttt{fraction of small}} = (1 - (-3))\sqrt{\frac{3}{4} \times \frac{1}{4}} = \sqrt{3}.$$

   (b) If we draw 300 times from this box, what will be the standard error of the sum of these 300 draws? (2 points)
   SE(sum of 300 draws) = $\sqrt{300}\times$ box SD = $\sqrt{3} \times 10 \times \sqrt{3}$ = 30.

   (c) What is the approximate probability that the sum of 300 draws will fall between -30 and 60? (2 points)
   -30 is one SE to the left of the expected value of the sum of 300 draws, which is 0 (300 × box average), and 60 is 2 SEs to the right of the expected value, therefore using the central limit theorem, the normal approximation to this probability is 68%/2+95%/2 = 81.5%.

   (d) How would you use `pnorm()` to compute this probability? (2 points)
   We can use the values in SU and therefore we will use pnorm(2) - pnorm(-1), where we are using the default values for the arguments mean (=0) and sd (=1); OR we can put in values 0 and 30 for mean and sd respectively: pnorm(60, mean = 0, sd = 30)-pnorm(-30, mean = 0, sd = 30).

---

9. Consider a binomial random variable X with parameters $n = 100$, and $p = 0.05$.

   (a) In R, how would you create a vector that consisted of ninety-five 0's, and five 1's? (2 points)
   c(rep(0,95),rep(1,5))

   (b) How would you use R to compute the probability that $X \leq 2$? (2 points)
   pbinom(2, size=100, prob=0.05)

---

10. Consider the following data frame that is a subset of the one that you have used in a homework assignment, giving total earnings and salaries of some top athletes, in millions of dollars.

Name	Total	Salary	Sport
Cristiano Ronaldo	108.0	61.0	Soccer
Neymar	90.0	73.0	Soccer
LeBron James	85.5	33.5	Basketball
Roger Federer	77.2	12.2	Tennis
Stephen Curry	76.9	34.9	Basketball
Kevin Durant	57.3	25.3	Basketball

(a) Assume that `dplyr` is loaded in R. Write the code that you would use to create a new column called "Endorsements" that gives the difference between the player's total earnings (called Total) and his salary. (2 points)

```
mutate(x,Endorsements = total-salary)
```

(b) Write code to compute the average salary, grouped by sport. (2 points)

```
summarize(group_by(x,sport),avg_salary=mean(salary))
```

(c) What would you expect the following code to return? (2 points)

```
arrange(filter(forbes18,sport=="Soccer"),total)
```

```
We would see a data frame consisting of two rows with only the soccer players in ascending order
of their total income.  We would expect to see:
Neymar                90       73       Soccer
Cristiano Ronaldo    108       61       Soccer
```