GROUND RULES:

- This exam contains 5 questions. Each question is worth 20 points. This exam is worth 100 points. An extra credit problem is given (written by me) that is worth 10 extra points.
- This is an open-book and open-notes exam. This means you can use anything on our course web site in addition to the textbook. You can also use anything on my course web site for STAT 511. Do not use anything else. This means no communication with other individuals (except me) and do not use any other books/web sites.
- Each question contains subparts. On each part, there is opportunity for partial credit, so show all of your work and explain all of your reasoning. **Translation:** No work/no explanation means no credit.
- If you use R to answer any part or to check your work, you must include all code and output as attachments. Do not just write out the code you used.
- Print your name legibly at the top of this page. Write your solution to each problem on its own page. You can use the back of the page if needed; you can append extra pages if needed. Collect your pages when done and staple in upper left corner. Keep all pages/solutions in order.
- I prefer you write out your solutions "by hand." In the past, I have noticed students who type their solutions (in LaTeX or Word) often do not provide enough detail and/or do not explain their reasoning very well. Please write legibly!
- Your solutions should be turned in to me no later than **April 1 at 9:40am**. Your solutions should be delivered to me in person.

HONOR PLEDGE FOR THIS EXAM:

After you have finished the exam, please read the following statement and sign your name below it.

> *I promise that I did not discuss any aspect of this exam with anyone other than the instructor, that I neither gave nor received any unauthorized assistance on this exam, and that the work presented herein is entirely my own.*

1. A researcher is interested in comparing males and females, aged 6-19, in South Carolina in terms of the number of dental caries (the number of teeth with decay or cavities). He observes independent random samples and makes the following modeling assumptions:

$$\begin{aligned} \text{Males:} \quad & Y_{11}, Y_{12}, ..., Y_{1n} \sim \text{iid Poisson with mean } \lambda_1 \\ \text{Females:} \quad & Y_{21}, Y_{22}, ..., Y_{2n} \sim \text{iid Poisson with mean } \lambda_2. \end{aligned}$$

Note that equal sample sizes are used (i.e., equal numbers of males and females are sampled). The random variable

$$Y_{ij} = \text{number of dental caries for the } j\text{th subject in the } i\text{th sample,}$$

for $i = 1, 2$ and $j = 1, 2, ..., n$.

(a) Define the sample mean of the male sample by

$$\overline{Y}_{1+} = \frac{1}{n} \sum_{j=1}^{n_1} Y_{1j}.$$

Show that

$$E(\overline{Y}_{1+}) = \lambda_1 \quad \text{and} \quad V(\overline{Y}_{1+}) = \frac{\lambda_1}{n}.$$

Similar results hold for the sample mean $\overline{Y}_{2+}$ of the female sample.

(b) What is the standard error of $\overline{Y}_{1+}$? of $\overline{Y}_{1+} - \overline{Y}_{2+}$?

(c) Argue that

$$Z_n = \frac{\overline{Y}_{1+} - \overline{Y}_{2+} - (\lambda_1 - \lambda_2)}{\sqrt{\frac{\lambda_1 + \lambda_2}{n}}} \sim \mathcal{AN}(0, 1)$$

when the (common) sample size n is large.

(d) For this part only, suppose $n = 100$ and $\lambda_1 = \lambda_2 = 2$. Where would you expect $\overline{Y}_{1+} - \overline{Y}_{2+}$ to fall with probability close to 0.95? Explain your answer.

2. Suppose Y_1, Y_2, Y_3 are mutually independent random variables with

$$\begin{aligned} Y_1 &\sim \text{exponential}(\theta) \\ Y_2 &\sim \text{exponential}(2\theta) \\ Y_3 &\sim \text{exponential}(3\theta). \end{aligned}$$

(a) Why aren't Y_1, Y_2, Y_3 iid?

(b) Define $T = Y_1 + Y_2 + Y_3$. Does T have a gamma distribution? If so, prove it. If not, explain why not.

(c) Show that

$$\widehat{\theta} = \frac{Y_1 + Y_2 + Y_3}{6}$$

is an unbiased estimator of θ and determine its standard error. Can you find an unbiased estimator of the standard error?

(d) Show that

$$Q = \frac{2}{\theta}\left(Y_1 + \frac{Y_2}{2} + \frac{Y_3}{3}\right) \sim \chi^2(6).$$

Explain why Q is a pivotal quantity. Then use Q to derive a $100(1-\alpha)\%$ confidence interval for θ. Define all notation you use in doing so.

3. An ecologist samples 25 regions of land for an experiment to study plant disease. Although the radius R of each region is supposed to be 10 meters, R is better regarded as a continuous random variable with probability density function

$$f_R(r) = \begin{cases} \frac{3}{4}[1-(10-r)^2], & 9 \leq r \leq 11 \\ 0, & \text{otherwise.} \end{cases}$$

Let $R_1, R_2, ..., R_{25}$ denote the radii for the 25 regions she samples. Assume $R_1, R_2, ..., R_{25}$ are iid from $f_R(r)$.

(a) Use the Central Limit Theorem to find the approximate sampling distributions of

$$\overline{R} = \frac{1}{25}\sum_{i=1}^{25} R_i \quad \text{and} \quad T = R_1 + R_2 + \cdots + R_{25}.$$

Hint: First determine $\mu = E(R)$ and $\sigma^2 = V(R)$.

(b) For the 25 regions sampled, approximate the probability $\overline{R}$ will exceed 10.1 meters. Why can't we calculate this probability exactly?

4. I have five mutually independent statistics T_1, T_2, T_3, T_4, and T_5. I have determined these statistics have the following sampling distributions:

- $T_1 \sim \mathcal{N}(0, 1)$
- $T_2 \sim \mathcal{N}(2, 4)$, where $V(T_2) = 4$.
- $T_3 \sim \chi^2(6)$
- $T_4 \sim \chi^2(8)$
- $T_5 \sim t(14)$

On each part below, find a function of T_1, T_2, T_3, T_4, and T_5 that has the stated distribution. You don't have to use all five statistics on each part. You might only use one, two, or three of them. You don't need to be overly mathematical, but you do need to make a convincing argument your answer is correct. Make sure you pay attention to independence assumptions needed!

(a) $\chi^2(14)$
(b) $\mathcal{N}(4, 7)$
(c) $\chi^2(2)$
(d) $F(14, 1)$
(e) $F(7, 9)$
(f) $t(6)$
(g) standard Cauchy distribution (with pdf on pp 14, notes)
(h) gamma(α, β), where $\alpha = \beta = 12$.

5. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a shifted exponential distribution

$$f_Y(y) = \begin{cases} \frac{2}{5}e^{-2(y-\theta)/5} & y > \theta \\ 0, & \text{otherwise.} \end{cases}$$

The parameter θ satisfies $-\infty < \theta < \infty$ and is unknown.

(a) Find a function of $\overline{Y}$ that is an unbiased estimator of θ. Prove that your function is unbiased.

(b) Find a function of $Y_{(1)}$ that is an unbiased estimator of θ. Prove that your function is unbiased.

(c) Which unbiased estimator has a smaller variance, the one you derived in part (a) or in part (b)? Prove your claim.

6. (Extra Credit). When I wrote this question initially, the NCAA Men's Basketball Tournament had not yet started. However, before each game, gambling experts publish their "lines" or "spreads;" these describe which team is favored and by how much. Even though I don't really understand how lines/spreads are created, I can define a random variable whose value is observed for each game. For Team A favored against Team B, define

$$D = \text{game outcome for Team A minus the point spread for Team A.}$$

For example,

- suppose Team A is favored by 6 points (point spread), but Team A wins by 2 points (game outcome). The difference is $D = 2 - 6 = -4$. If you had bet on Team A to "cover the spread," then you lost the bet.
- suppose Team A is favored by 6 points, but Team A wins by 10 points. The difference is $D = 10 - 6 = 4$. If you had bet on Team A to "cover the spread," then you won the bet.
- suppose Team A is favored by 6 points, but Team A loses by 2 points. The difference is $D = (-2) - 6 = -8$. If you had bet on Team A to "cover the spread," then you certainly lost the bet.

We get to observe a sample of differences, say, $D_1, D_2, ..., D_n$ for n games played. In this problem, we will assume $D_1, D_2, ..., D_n$ is an iid sample from a $\mathcal{N}(\mu_D, \sigma_D^2)$ population distribution, where both μ_D and σ_D^2 are unknown.

In Table 1 (next page), you will see the spreads and outcomes for the first round of this year's tournament (I am ignoring the "play-in games"). There were $n = 32$ games. Point spreads were provided in advance by ESPN BET two days before the games were played.

(a) Under our model assumptions, report unbiased point estimates for μ_D and σ_D based on the 32 games played. Note I am asking for an unbiased point estimate of the population standard deviation; not the population variance.

(b) Under our model assumptions, calculate 95% confidence intervals for μ_D and σ_D based on the 32 games played. How does ESPN BET do "on average" in terms of formulating its point spreads?

Region	Game	Teams	Spread (Line)	Outcome	d_i
SOUTH	1	(1) Hou vs (16) Long	Hou by 24.5	Hou by 40	$d_1 = 15.5$
	2	(8) Neb vs (9) TAMU	Neb by 1.5	TAMU by 15	$d_2 = -16.5$
	3	(5) Wisc vs (12) JMU	Wisc by 5.5	JMU by 11	$d_3 = -16.5$
	4	(4) Duke vs (13) Verm	Duke by 11.5	Duke by 17	$d_4 = 5.5$
	5	(6) TTU vs (11) NCSU	TTU by 5.5	NCSU by 13	$d_5 = -18.5$
	6	(3) Kent vs (14) Oak	Kent by 13.5	Oak by 4	$d_6 = -17.5$
	7	(7) Fla vs (10) Colo	Fla by 2.5	Colo by 2	$d_7 = -4.5$
	8	(2) Marq vs (15) WKU	Marq by 16.5	Marq by 18	$d_8 = 1.5$
EAST	9	(1) Conn vs (16) Stet	Conn by 26.5	Conn by 39	$d_9 = 12.5$
	10	(8) FAU vs (9) NW	FAU by 2.5	NW by 12	$d_{10} = -14.5$
	11	(5) SDSU vs (12) UAB	SDSU by 6.5	SDSU by 4	$d_{11} = -2.5$
	12	(4) Aub vs (13) Yale	Aub by 12.5	Yale by 2	$d_{12} = -14.5$
	13	(6) BYU vs (11) Duq	BYU by 7.5	Duq by 4	$d_{13} = -11.5$
	14	(3) Ill vs (14) More St	Ill by 11.5	Ill by 16	$d_{14} = 4.5$
	15	(7) WSU vs (10) Drake	Drake by 1.5	WSU by 5	$d_{15} = -6.5$
	16	(2) ISU vs (15) S Dak St	ISU by 16.5	ISU by 17	$d_{26} = 0.5$
WEST	17	(1) UNC vs (16) Wag	UNC by 24.5	UNC by 28	$d_{17} = 3.5$
	18	(8) Ms St vs (9) Mi St	Mi St by 1.5	Mi St by 18	$d_{18} = 16.5$
	19	(5) SMC vs (12) GCU	SMC by 5.5	GCU by 9	$d_{19} = -14.5$
	20	(4) Ala vs (13) CofC	Ala by 9.5	Ala by 13	$d_{20} = 3.5$
	21	(6) Clem vs (11) UNM	UNM by 1.5	Clem by 19	$d_{21} = -20.5$
	22	(3) Bay vs (14) Colg	Bay by 13.5	Bay by 25	$d_{22} = 11.5$
	23	(7) Day vs (10) Nev	Nev by 1.5	Day by 3	$d_{23} = -4.5$
	24	(2) Ariz vs (15) LBSU	Ariz by 20.5	Ariz by 20	$d_{24} = -0.5$
MIDWEST	25	(1) Pur vs (16) Gramb	Pur by 26.5	Pur by 28	$d_{25} = 1.5$
	26	(8) USU vs (9) TCU	TCU by 3.5	USU by 16	$d_{26} = -19.5$
	27	(5) Gonz vs (12) McNeese	Gonz by 6.5	Gonz by 21	$d_{27} = 14.5$
	28	(4) Kan vs (13) Sam	Kan by 8.5	Kan by 4	$d_{28} = -4.5$
	29	(6) USC vs (11) Ore	USC by 1.5	Ore by 14	$d_{29} = -15.5$
	30	(3) Crei vs (14) Akron	Crei by 12.5	Crei by 16	$d_{30} = 3.5$
	31	(7) Tex vs (10) CSU	Tex by 2.5	Tex by 12	$d_{31} = 9.5$
	32	(2) Tenn vs (15) SPU	Tenn by 21.5	Tenn by 34	$d_{32} = 12.5$

Table 1: 2024 NCAA Tournament data. Match-ups, point spreads, and outcomes for first round games.