Question 1 (Distribution Models, 20 points). For each of the following questions give the distribution family that best models the given situation. Note that some answers may occur more than once.

(A) Geometric
(B) Binomial
(C) Poisson
(D) Exponential
(E) Uniform
(F) Normal

(a) The exact age of a random first grader (not just their age in years).
   Answer: E
   Reason: First grades will (by and large) be people whose ages are within a particular 1-year range. Since people are born at roughly the same rate throughout the year (this is of course not exactly true in practice, but is a decent model), you would expect that the ages of first grades should be approximately uniform.

(b) The height of a random giraffe.
   Answer: F
   Reason: The height of a random giraffe is a continuous random variable whose distribution is unimodal and roughly symmetric. Furthermore, it is likely due to a conglomeration of many tiny factors adding up. Thus, normal model is the best choice.

(c) The number of years until the next time your favorite team wins the Super Bowl.
   Answer: A
   Reason: Geometric distribution models how long it will take to achieve the first success in a series of Bernoulli trials. The number of years is a discrete random variable. Thus your favorite team wins the Super Bowl this year can be viewed as a Bernoulli trial.

(d) The number of 6s you get when rolling ten dice.
   Answer: B
   Reason: This can be characterized as a Binomial model with \( n = 10 \) and \( p = 1/6 \).

(e) The number of earthquakes in San Diego in the next year.
   Answer: C
   Reason: The poisson distribution can model the probability of the occurrence of events for a variety of phenomena, which is the earthquake in this situation.
Question 2 (Survey Design, 20 points). (a) You are conducting a survey of people living in San Diego by calling random numbers from a local phone book. Which source of bias do you probably NOT need to worry about? 

(A) Undercoverage
(B) Response Bias
(C) Non-Response Bias
(D) Voluntary Response Bias

Answer: D
Reason: (D) is the answer because this is not a voluntary response sample. Individuals are not invited to respond but choose to respond. (A) is likely to happen because telephone surveys are usually conducted when you are likely to be home, such as dinnertime. If you eat out often, you may be less likely to be surveyed, a possible source of undercoverage. Similarly, a survey is likely to miss people who lack landline telephones, and so may miss certain populations such as homeless people and students. (B) also need to worry about because response biases include anything in the survey design that influences the responses. (C) is a potential source of bias because those who don’t answer the phone may differ from those who do.

(b) When computing a confidence interval, you decide to use a larger critical value. What effect does this have on your confidence level and the width of your confidence interval.

(A) Increases your confidence and widens your confidence interval.
(B) Increases your confidence and shrinks your confidence interval.
(C) Decreases your confidence and widens your confidence interval.
(D) Decreases your confidence and shrinks your confidence interval.

Answer: A
Reason: ±critical value × standad deviation of the mean characterizes the size of confidence interval. Thus the larger the critical value is, the wider your confidence interval. Your confidence level will depend on the probability that a standard normal has size no larger than your critical value. This will also increase with your critical value.

(c) When performing a hypothesis test, you get a p-value of 10%. What is the most accurate interpretation of this result.

(A) There is at most a 10% chance that the null hypothesis is true.
(B) Ninety percent of the time when you get such a result, the null hypothesis turns out to be false.
(C) If the null hypothesis were true, there is at most a 10% chance that you would have seen such extreme results.
(D) The null hypothesis is probably true.

Answer: C
Reason: This problem asks for the meaning of p-value. p-value is the greatest probability of seeing such data when the null hypothesis is true.

(d) Trying to decide whether a town is at greater risk of heart disease than the national average, you look at statistics of deaths due to heart attack in the town and compare the null hypothesis $H_0 : p = 0.25$ (the national average) to the alternative hypothesis $H_A : p > 0.25$. You reject the null hypothesis with a p-value of 0.03. If instead you had run a 2-sided alternative $H_A : p \neq 0.25$, what would you new p-value be?

(A) 0.0001
(B) 0.03
(C) 0.05
(D) 0.06
(E) 0.10

Answer: D

Reason: For two-sided alternatives, the $p$-value is the probability of deviating in either direction from the null hypothesis value, while one-sided alternatives focuses on deviations from the null hypothesis value in only one direction. Since the normal distribution is symmetric, the new $p$-value is two times the original one.

(e) When performing a survey, you decide that you would like to get confidence intervals of half the size you currently have (while maintaining the same level of confidence). How many samples would you need for this survey? 

(A) Half as many as before
(B) The same number as before
(C) Twice as many as before
(D) Four times as many as before

Answer: D

Reason:

\[
\text{size of confidence interval} = 2 \times \text{critical value} \times \text{standard error} = 2 \times \text{critical value} \times \sqrt{\frac{pq}{n}},
\]

where $n$ is the sample size. Thus in order to get confidence intervals of half the current size, the size of the new sample should be four times as large as before.
Question 3 (Continuous Distributions, 20 points). Consider a continuous random variable $X$ with probability density function

$$f(x) = \begin{cases} \frac{2}{x^2} & x \geq 1 \\ 0 & \text{otherwise.} \end{cases}$$

(a) What is $P(X \geq 3)$? [10 points]

Solution:

$$P(X \geq 3) = \int_3^\infty f(x)\,dx = \int_3^\infty \frac{2}{x^2}\,dx = -2x^{-1}\bigg|_3^\infty = 0 - \left(-\frac{1}{9}\right) = \frac{1}{9}.$$

(b) What is $E[X]$? [10 points]

Solution:

$$E[X] = \int_{-\infty}^{\infty} xf(x)\,dx = \int_1^{\infty} x \cdot \frac{2}{x^3}\,dx = \int_1^{\infty} \frac{2}{x^2}\,dx = \frac{-2}{x}\bigg|_1^{\infty} = 0 - (-2) = 2.$$
Question 4 (Expectation and Variance, 20 points). Consider four independent random variables:

- $X$ is a Binomial distribution with $n = 20$ and $p = 0.4$.
- $Y$ is an Exponential distribution with mean 5.
- $Z$ is a random variable with mean 2 and variance 4.
- $W$ is a random variable with mean $-1$ and variance 2.

(a) What are the means and variances of $X$, and $Y$? [8 points]

Solution:

$\mathbb{E}[X] = np = 20 \times 0.4 = 8.$

$\text{Var}(X) = npq = 20 \times 0.4 \times (1 - 0.4) = 4.8.$

Since $Y$ is an Exponential distribution with mean 5, the density function for $Y$ is as follows,

$$f(x) = \begin{cases} \frac{1}{5}e^{-\frac{1}{5}x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$\mathbb{E}[Y] = 5.$

$$\text{Var}(Y) = \int_{-\infty}^{\infty} (x - 5)^2 f(x) dx = \int_{0}^{\infty} (x - 5)^2 \frac{1}{5}e^{-\frac{1}{5}x} dx = 25.$$ 

(b) What are the mean and variance of $Z + 2W + 10$? [6 points]

Solution:

$\mathbb{E}[Z + 2W + 10] = \mathbb{E}[Z] + 2\mathbb{E}[W] + 10 = 2 + 2 \times (-1) + 10 = 10.$

Since $Z$ and $W$ are independent,

$$\text{Var}(Z + 2W + 10) = \text{Var}(Z) + \text{Var}(2W) = \text{Var}(Z) + 4\text{Var}(W) = 4 + 4 \times 2 = 12.$$ 

(c) Approximate the probability that the sum of five independent copies of $W$ is more than $-10$. [6 points]

Solution: Let $\bar{W}$ denote the mean of five independent copies of $W$. Then $\bar{W}$ satisfies the Central Limit Theorem and $\bar{W}$ is approximately Normal with mean $-1$ and variance $\frac{2}{5}$. This is because

$$\text{Var}(W_1 + W_2 + W_3 + W_4 + W_5) = \text{Var}(W_1) + \text{Var}(W_2) + \text{Var}(W_3) + \text{Var}(W_4) + \text{Var}(W_5) = 10.$$ 

Therefore, $\text{Var}(\bar{W})$ is $(1/5)^2 \cdot 10 = 2/5$, so the standard deviation is $\sqrt{2}/5$.

Then $\frac{\bar{W} + 1}{\sqrt{\frac{2}{5}}}$ has standard normal distribution.

$$P(\text{the sum of five independent copies of } W \text{ is more than } -10) = P(\bar{W} > -2) = P\left(\frac{\bar{W} + 1}{\sqrt{\frac{2}{5}}} > \frac{-2 + 1}{\sqrt{\frac{2}{5}}}\right)$$

$$\approx P\left(\frac{\bar{W} + 1}{\sqrt{\frac{2}{5}}} > -1.58\right) \approx 1 - 0.571 = 0.9429.$$
Question 5 (Carcinogen Testing, 20 points). You are worried that chemical X may be a carcinogen. You know that the cancer rate per decade in the population as a whole is about 5%. Surveying 1000 people who have worked significantly with chemical X over the last decade, you find that 60 have developed cancer over that time period. You would like to test whether or not this suggests that chemical X is leading to increased rates of cancer or whether this is just a coincidence.

(a) What is your null hypothesis? [4 points]
   Solution: \( H_0 : p = 0.05 \).

(b) What is your alternative hypothesis? [4 points]
   Solution: \( H_A : p > 0.05 \).

(c) What is your test statistic (please show how you computed it)? Can you reject the null hypothesis at \( \alpha = 0.05 \)? [8 points]
   Solution: The null model is a Normal distribution with a mean of 0.05 and a standard deviation of
   
   \[
   SD(\hat{p}) = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.05)(0.95)}{1000}} = 0.006892.
   \]
   
   So the z-value is
   
   \[
   z = \frac{0.06 - 0.05}{0.006892} = 1.45.
   \]
   
   The corresponding p-value is 0.0735.
   
   Conclusion: Fail to reject the null hypothesis.

(d) Explain your conclusions in common language. [4 points]
   Solution: If the true cancer rate per decade in the population as a whole is about 0.05, then the probability of observing a value of \( z \) 0.06 or greater is 0.0735. Thus I fail to reject \( H_0 \), which means that although \( H_0 \) may well be false, I do not have enough evidence to discard the hypothesis at the \( \alpha = 5\% \) level.