Habib University
Dhanani School of Science and Engineering

Dr. Muntazir Abidi
Integrated Sciences and Mathematics (iSciM)

**AstroStatistics Spring 2022**
**Midterm Exam Solutions**

NOTE : Solve ANY THREE of the following four problems. All of them carry equal marks. You're allowed to use calculators. Total time for this exam is TWO HOURS.

## Problem 1 (25 Marks)

This problem is about the conditional probability. Let us consider a medical test scenario. One variable is the test results $T$ for some disease. The test can be negative (0) or positive (1). The other variable is the health state of the patient, or the presence of disease $D$ : the patient can have a disease (1) or not (0). If the patient is healthy ($D = 0$), the probability for the test being positive (a false positive) is $p(T = 1|D = 0) = \epsilon_{fP}$, where $\epsilon_{fP}$ is the small number. If the patient has the disease ($D = 1$), the probability for the test being negative (a false negative) is $p(T = 0|D = 1) = \epsilon_{fN}$, where $\epsilon_{fN}$ is also a small number. Let us assume that we also know that the prior probability (in the absence of any testing, for example, based on some large population studies unrelated to our test) for the disease in question is $p(D = 1) = \epsilon_D$ , where $\epsilon_D$ is a small number. Given these information answer the following questions :

1. Write the possible combinations in the sample space.
2. Write the expressions for $p(T = 0|D = 0)$, $p(T = 1|D = 1)$, and $p(D = 0)$?
3. Assume now that our patient took the test and it came out positive ($T = 1$). What is the probability that our patient has contracted the disease, $p(D = 1|T = 1)$ ?
4. For rare diseases, we cannot take the risk. What is the condition for $p(D = 1|T = 1) \sim 1$ ? That is whenever the test is positive we know that the patient has the disease.
5. Let us compute some numbers. Suppose the frequency of the disease in the population (base rate) is 0.5%. The test is highly accurate with a 5% false positive rate and a 10% false negative rate. What is the probability that a patient has a disease given that the test is positive, that is $p(D = 1|T = 1)$ ?

**Solution :**

1. Sample space is : $(T = 0, D = 0)$, $(T = 1, D = 0)$, $(T = 0, D = 1)$, and $(T = 1, D = 1)$.
2. If $p(T = 1|D = 0) = \epsilon_{fP}$ then $p(T = 0|D = 0) = 1 - p(T = 1|D = 0) = 1 - \epsilon_{fP}$. Similarly, if $p(T = 0|D = 1) = \epsilon_{fN}$ then $p(T = 1|D = 1) = 1 - p(T = 0|D = 1) = 1 - \epsilon_{fN}$.
   If $p(D = 1) = \epsilon_D$ then $p(D = 0) = 1 - p(D = 1) = 1 - \epsilon_D$.
3. Using Bayes rule

$$p(D = 1|T = 1) = \frac{p(T = 1|D = 1)p(D = 1)}{p(T = 1|D = 0)p(D = 0) + p(T = 1|D = 1)p(D = 1)} \tag{1}$$

   and given our assumptions,

$$p(D = 1|T = 1) = \frac{\epsilon_D - \epsilon_{fN}\epsilon_D}{\epsilon_D + \epsilon_{fP} - [\epsilon_D(\epsilon_{fP} + \epsilon_{fN})]} \tag{2}$$

For simplicity we can neglect the second-order terms since all $\epsilon$ parameters are presumably small, and thus

$$p(D=1|T=1) = \frac{\epsilon_D}{\epsilon_D + \epsilon_{fP}} \tag{3}$$

4. We can only reliably diagnose a disease (i.e., $p(D=1|T=1) \sim 1$) if $\epsilon_{fP} \ll \epsilon_D$. For rare diseases the test must have an exceedingly lo false-positive rate!
5. $p(D=1) = 0.05$ and $p(D=0) = 0.995$. The false positive and false negative are conditional probabilities :

$$p(\text{false positive}) = p(T=1|D=0) = 0.05 \tag{4}$$

and

$$p(\text{false negative}) = p(T=0|D=1) = 0.9 \tag{5}$$

We also know that

$$p(T=0|D=0) = 1 - p(T=1|D=0) = 0.95 \tag{6}$$

and

$$p(T=1|D=1) = 1 - p(T=0|D=1) = 0.9 \tag{7}$$

using

$$p(D=1|T=1) = \frac{p(T=1|D=1)p(D=1)}{p(T=1|D=0)p(D=0) + p(T=1|D=1)p(D=1)} \tag{8}$$

$$= \frac{0.9 \times 0.005}{0.995 \times 0.05 + 0.005 \times 0.9} = 0.082949 \approx 8.2\% \tag{9}$$

## Problem 2 (25 Marks)

Let $X$ have range $[0, 3]$ and the probability density function $p_X(x) = kx^2$. Let $Y = X^3$.

1. Find $k$ and the cumulative distribution function of $X$
2. Find the 30th percentile of $X$.
3. Compute $E(Y)$.
4. Write down an explicit formula, involving an integral, for $Var(Y)$ (Do not compute the value of the integral.)
5. Find the probability density function $p_Y(y)$ for $Y$ .

**Solution :**

1. $\int_0^3 p_X(x)dx = 1$ this implies that $\int_0^3 kx^2dx = k(x^3/3)\Big|_0^3 = 1$ which is $27k = 3$ implies that $k = 1/9$. For cumulative densifty function we know that cdf $= P(X \leq x)$ can be computed as

$$F(x) = P(X \leq x) = \int_0^x \frac{1}{9}x^2dx = \frac{1}{27}x^3\Big|_0^x = \frac{1}{27}x^3$$

2. Now, to find the 30th percentile, we just need to set 0.3 equal to $F(x_{30})$

$$0.3 = F(x_{30}) = \frac{1}{27}x_{30}^3 \implies x_{30} = (0.3 \times 27)^{1/3} \approx 2.006$$

3. Using the formula of expectation

$$E(Y) = E(X^3) = \int_0^3 x^3 \frac{1}{27} x^2 dx = \frac{1}{27 \times 6} x^6 \Big|_0^3 = \frac{1}{27 \times 6}(3)^6 = \frac{243}{54}$$

4. $Var(Y) = E(Y^2) - E(Y)^2$. We know that $E(Y) = \frac{243}{54}$ so the integral for $E(Y^2)$ is given by

$$E(Y^2) = \int_0^3 \frac{1}{27} x^6 \times x^2 dx = \int_0^3 \frac{1}{27} x^8 dx$$

so

$$\text{Var}(Y) = \int_0^3 \frac{1}{27} x^8 dx - \Big(\frac{243}{54}\Big)^2$$

5. using the transformation of variables, the cumulative probability dist : $P(Y \leq y) = P(X^3 \leq y) = P(X \leq y^{1/3}) = \frac{1}{27} y^{3/3} = \frac{1}{27} y$. Differentiating this with respect to y we get the probability density function :

$$p_Y(y) = \frac{d}{dy}\left(\frac{1}{27} y\right) = \frac{1}{27}$$

One can check that $\int_0^{27} p_Y(y) dy = 1$, so $p_Y(y)$ is indeed the probability density function of $Y$. If the range of X is $[0, 3]$, the range of $Y$ is $[0, 27]$.

## Problem 3 (25 Marks)

Recall that the normal distribution $N(\mu, \sigma^2)$ has the pdf

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big), \tag{10}$$

where $\mu$ is the mean and $\sigma$ is the standard deviation. The standard normal distribution $N(0, 1)$ has mean 0 (by symmetry), variance 1, and pdf $\phi(z)$ given by setting $= 0$ and $\sigma = 1$ above. The cdf is denoted $\Phi(z)$ and does not have a nice formula. In this problem, we'll show that scaling and shifting a normal random variable gives a normal random variable. Suppose $Z \sim N(0, 1)$ and $X = aZ + b$.

1. Compute the mean $\mu$ and variance $\sigma^2$ of $X$.
2. Express the cdf $P(X \leq x)$ of $X$ in terms of $\Phi$ and then use the chain rule to find the pdf $p(x)$ of $X$.
3. Use (b) to show that $X$ follows the $N(b, a^2)$ distribution.
4. Lets assume that $a = 3$ and $b = 1$. Find $P(-1 \leq X \leq 1)$
5. The probability that $Z$ is within one standard deviation of its mean is approximately 68%. What is the probability that $X$ is within one standard deviation of its mean.

**Solution :**

1. We know that the mean of $X$ is $E(X) = aE(Z) + b = 0 + b = b$ and the variance is

$$Var(X) = Var(aZ + b) = a^2 Var(Z) = a^2 \tag{11}$$

2. Let x be any real number. We will first compute $F_X(x) = P(X \leq x)$. Since $X = aZ + b$, we get
$$F_X(x) = P(X \leq x) = P(aZ + b \leq x) = P(Z \leq \frac{x-b}{a}) = \Phi(\frac{x-b}{a}) \tag{12}$$
Differentiating this with respect to $x$ we find
$$f_X(x) = \frac{d}{dx}\Phi(\frac{x-b}{a}) = \frac{1}{a}\phi(\frac{x-b}{a}) = \frac{1}{a\sqrt{2\pi}}e^{-\frac{(x-b)^2}{2a^2}} \tag{13}$$

3. from (2) we see that $f_X(x)$ is the normal probability density function $N(b, a^2)$. From (2) and (3), we see that if $Z$ is standard normal, then $\sigma Z + \mu$ follows a $N(\mu, \sigma^2)$ distribution. From (1) we know that $E(\mu + \sigma Z) = \mu$ and $Var(\sigma Z + \mu) = \sigma^2$.
4. We have
$$P(-1 \leq X \leq 1) = P(-\frac{2}{3} \leq X \leq 0) = \Phi(0) - \Phi(-\frac{2}{3}) \approx 0.2475 \tag{14}$$
This mean that 24.75%. The last value we found using the standard normal table.
5. Since $E(X) = 1$, $Var(X) = 9$,we want $P(2 \leq 4)$.We have
$$P(2 \leq X \leq 4) = P(3 \leq 3Z \leq 3) = P(1 \leq Z \leq 1) \approx 0.68. \tag{15}$$

## Problem 4 (25 Marks)

Suppose $X$ and $Y$ are random variables with means $\mu_X$ and $\mu_Y$ . The covariance of $X$ and $Y$ is defined as
$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] \tag{16}$$
Using this expression, proof the following properties.

1. $\text{Cov}(aX + b, cY + d) = ac\text{Cov}(X, Y)$ for constants $a$, $b$, $c$, $d$.
2. $\text{Cov}(X, Y) = E(XY) - \mu_X\mu_Y$ .
3. Var(X + Y ) = Var(X) + Var(Y ) + 2Cov(X, Y ) for any X and Y .
4. **Discrete Case :** : If $X$ and $Y$ have joint pmf $p(x_i, y_j)$ then show that
$$\text{Cov}(X, Y) = \left(\sum_i \sum_j p(x_i, y_j)x_i x_j\right) - \mu_X\mu_Y \tag{17}$$
5. **Continuous Case :** : If $X$ and $Y$ have joint pdf $f(x, y)$ over range $[a, b] \times [c, d]$, then show that
$$\text{Cov}(X, Y) = \left(\int_a^b \int_c^d xyf(x, y)\right) - \mu_X\mu_Y \tag{18}$$

**Solution :**

1. Note that by linearity $E(X + b) = E(X) + b$ and $E(aX) = aE(X)$. Now using the definition of the covraiance :
$$\text{Cov}(aX + b, cY + d) = E[(aX + b - E(aX + b))(cY + d - E(cY + d))]$$

$$\text{Cov}(aX+b, cY+d) = E[(aX+b-aE(X)-b)(cY+d-cE(Y)-d)]$$

$$\text{Cov}(aX+b, cY+d) = E[(aX-aE(X))(cY-cE(Y))] = E[ac(X-E(X))(Y-E(Y))]$$
$$\text{Cov}(aX+b, cY+d) = acE[(X-E(X))(Y-E(Y))] = acE[(X-\mu_X)(Y-\mu_Y)]$$

because $E(X) = \mu_X$ and $E(Y) = \mu_Y$. QED.

2. Using the linearity of Expectation we can simplify the above expression to
$$\text{Cov}(X,Y) = E[XY] - E[X\mu_Y] - E[Y\mu_X] + E[\mu_X\mu_Y]$$
since $\mu_X$ and $\mu_Y$ are number we can take them out of the expectations, by doing so we get
$$\text{Cov}(X,Y) = E[XY] - \mu_Y E[X] - \mu_X E[Y] + E[\mu_X\mu_Y]$$
Remember that $E[X] = \mu_X$ and $E[Y] = \mu_Y$, so we get
$$\text{Cov}(X,Y) = E[XY] - \mu_Y\mu_X - \mu_X\mu_Y + \mu_X\mu_Y$$
$$\text{Cov}(X,Y) = E[XY] - \mu_Y\mu_X$$
QED.

3. We know by definition that
$$\text{Var}(X) = E(X^2) - E(X)^2$$
so for X+Y we will get
$$\text{Var}(X+Y) = E((X+Y)^2) - E(X+Y)^2$$
Using linearity :
$$\text{Var}(X+Y) = E(X^2+Y^2+2XY) - [E(X)+E(Y)]^2$$
$$\text{Var}(X+Y) = E(X^2) + E(Y^2) + 2E(XY) - E(X)^2 - E(Y)^2 2E(X)E(Y)$$
$$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) - 2(E(XY) - E(X)E(Y))$$
We know that $\text{Cov}(X,Y) = E(XY) - \mu_X\mu_Y$, so we get
$$\text{Var}(X+Y) = \text{Var}(Y) + \text{Var}(Y) + 2\text{Cov}(X,Y)$$
QED.

4. Given that
$$\text{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$
for the discrete case, the probability density function is $p(x,y)$ so the $E[XZ]$ is
$$E[XZ] = \sum_i \sum_j p(x_i, y_j) x_i y_j$$
so,
$$\text{Cov}(X,Y) = \sum_i \sum_j p(x_i, y_j) x_i y_j - \mu_X\mu_Y.$$

5. Similarly for the continuous case, if the density function is $f(x,y)$, $E[XY]$ is equal to
$$E[XY] = \int_a^b \int_c^d f(x,y) xy dx dy$$
so
$$\text{Cov}(X,Y) = \int_a^b \int_c^d f(x,y) xy - \mu_X\mu_Y$$