Math 442 – Homework Set 2

p.312[5,6,8], p.319[11,12,15], p.320[8], p.235[8,9,11]

1. (p.312 #5) Suppose that 16 percent of the students in a certain high school are freshmen, 14 percent are sophomores, 38 percent are juniors, and 32 percent are seniors. If 15 students are selected at random from the school, what is the probability that at least eight will be either freshmen or sophomores?

   Solution. First we shall assume that the population of school is large enough so that the selections are independent, or equivalent to selection with replacement. Let X_1, X_2 and X_3 be the numbers of freshmen, sophomore and juniors in the sample. Then (X_1, X_2, X_3) has a multinomial distribution $n = 15$ and $(p_1, p_2, p_3) = (0.16, 0.14, 0.38)$. The random variable $X = X_1 + X_2$ is the total number of freshmen or sophomore. Then from the property of multinomial distribution we know that X has a binomial distribution with $n = 15$ and $p = p_1 + p_2 = 0.16 + 0.14 = 0.30$, i.e. $X \sim \text{Binomial}(15, 0.30)$. Now,

   $$P(X \geq 8) = 1 - P(X \leq 7) = 1 - F_X(7) = 1 - \sum_{x=0}^{7} \binom{15}{x} (0.30)^x (0.70)^{15-x} = 0.05001254.$$

   In R the following commands are equivalent (give the same answer):

   ```
   1-pbinom(7,15,0.30) or pbinom(7,15,0.70) or dbinom(c(8:15),15,0.30))
   ```

   They all result in 0.05001254. □

2. (p.312 #6) In Exercise 5, let X_3 denote the number of juniors in the random sample of 15 students, and let X_4 denote the number of seniors in the sample. Find the value of $E(X_3 - X_4)$ and the value of $\text{Var}(X_3 - X_4)$.

   Solution. Clearly, $X_3 \sim \text{Binomial}(15, 0.38)$ and $X_4 \sim \text{Binomial}(15, 0.32)$. Hence

   $$E(X_3 - X_4) = E(X_3) - E(X_4) = 15(0.38) - 15(0.32) = 0.9,$$

   and

   $$\begin{aligned}\text{Var}(X_3 - X_4) = \text{Var}(X_3) + \text{Var}(X_4) - 2\text{Cov}(X_3, X_4) &= np_3(1 - p_3) + np_4(1 - p_4) - 2(-p_3 p_4) \\ &= 15(0.38)(0.62) + 15(0.32)(0.68) - 2(-0.38 \times 0.32) = 7.0412.\end{aligned}$$

   □

3. (p.312 #8) Suppose that the parts produced by a machine can have three different levels of functionality: working, impaired, defective. Let p_1, p_2, and $p_3 = 1 - p_1 - p_2$ be the probabilities that a part is working, impaired, and defective, respectively. Suppose that the vector $\mathbf{p} = (p_1, p_2)$ is unknown but has a joint distribution with probability density function

   $$f_{p_1,p_2}(p_1, p_2) = \begin{cases} 12{p_1}^2 & \text{for } 0 < p_1, p_2 < 1 \text{ and } p_1 + p_2 < 1, \\ 0 & \text{otherwise.} \end{cases}$$

   Suppose that we observe 10 parts that are conditionally independent given $\mathbf{p}$, and among those 10 parts, eight are working and two are impaired. Find the conditional probability density function of $\mathbf{p}$ given the observed parts.

Solution. Let $\mathbf{X} = (X_1, X_2)$, where X_1 is the total number of parts which are working, X_2 the number of impaired and $X_3 = 10 - X_1 - X_2$ the number of defective. $\mathbf{X}|\mathbf{p} \sim$ trinomial$(n = 10, \mathbf{p})$; that is

$$f_{\mathbf{X}|\mathbf{p}}(x_1, x_2|p_1, p_2) = \binom{10}{x_1 x_2} {p_1}^{x_1} {p_2}^{x_2} (1 - p_1 - p_2)^{10-x_1-x_2}.$$

By Bayes' theorem the conditional probability density function of $\mathbf{p}$ given the observed parts has the form

$$f_{\mathbf{p}|\mathbf{X}}(p_1, p_2|x_1, x_2) = \frac{f_{\mathbf{X}|\mathbf{p}}(x_1, x_2|p_1, p_2) f_{\mathbf{p}}(p_1, p_2)}{f_{\mathbf{X}}(x_1, x_2)}$$

where $f_{\mathbf{X}}$ is the marginal p.f. of $\mathbf{X}$ at $(8, 2)$ (the given parts), which can be obtained by the law of total probability for random variables:

$$\begin{aligned} f_{\mathbf{X}}(8, 2) &= \int_0^1 \binom{10}{8, 2} {p_1}^8 {p_2}^2 (1 - p_1 - p_2)^{10-8-2}\ 12{p_1}^2\ d\mathbf{p} \\ &= 12 \int_0^1 {p_1}^{10} (1 - p_1)^2\ dp_1 \\ &= 12 \times \text{Beta}(11, 3) = 12\ \frac{\Gamma(11)\Gamma(3)}{\Gamma(14)} = \frac{40}{143}. \end{aligned}$$

Substituting, the conditional probability density function of $\mathbf{p}$ given the observed parts is

$$f_{\mathbf{p}|\mathbf{X}}(p_1, p_2|X_1 = 8, X_2 = 2) = \frac{143\binom{10}{8,2} {p_1}^8 (1 - p_1)^2\ 12{p_1}^2}{40} = \frac{429\ {p_1}^{10} (1 - p_2)^2}{10}.$$

□

4. (p.319 #11) Suppose that two random variables X_1 and X_2 have a bivariate normal distribution, and $\text{Var}(X_1) = \text{Var}(X_2)$. Show that the sum $X_1 + X_2$ and the difference $X_1 - X_2$ are independent random variables.

Solution. We can write

$$X_1 = \sigma Z_1 + \mu_1, \quad \text{and} \quad X_2 = \sigma[\rho Z_1 + (1 - \rho^2)^{1/2} Z_2] + \mu_2$$

where Z_1 and Z_2 are independent random variables each having a standard normal distribution. Since $X_1 + X_2$ and $X_1 - X_2$ can each be expressed as a linear combination of Z_1 and Z_2, we know that $X_1 + X_2$ and $X_1 - X_2$ have bivariate normal distribution. So to show the independence of $X_1 + X_2$ and $X_1 - X_2$, it suffices to show that the correlation $\rho(X_1 + X_2, X_1 - X_2) = 0$. The covariance between two linear combinations of random variables can be computed using the following formula derived in Homework Set 1:

$$\text{Cov}\Big(\sum_{i=1}^m a_i X_i, \sum_{j=1}^n b_j Y_j\Big) = \sum_{i=1}^m \sum_{j=1}^n a_i b_j \text{Cov}(X_i, Y_j).$$

Accordingly,

$$\begin{aligned} \text{Cov}(X_1 + X_2, X_1 - X_2) &= \text{Cov}(X_1, X_1) - \text{Cov}(X_1, X_2) + \text{Cov}(X_2, X_1) - \text{Cov}(X_2, X_2) \\ &= \text{Var}(X_1) - 0 + 0 - \text{Var}(X_2) = 0. \end{aligned}$$

Therefore $\rho(X_1 + X_2, X_1 - X_2) = 0$ and, consequently, $X_1 + X_2$ and $X_1 - X_2$ are independent. □

5. (p.319 #12) Suppose that the two measurements from flea beetles in Examples 5.12.1 have a bivariate normal distribution with $\mu_1 = 201, \mu_2 = 118, \sigma_1 = 15.2, \sigma_2 = 6.6$, and $\rho = 0.64$. Suppose that the same two measurements from a second species also have a bivariate normal distribution with $\mu_1 = 187, \mu_2 = 131, \sigma_1 = 15.2, \sigma_2 = 6.6$, and $\rho = 0.64$. Let (X_1, X_2) be a pair of measurements on a flea beetle from one of these two species. Let a_1, a_2 be constants.

(a) For each of the two species, find the mean and standard deviation of $a_1X_1 + a_2X_2$. (Note that the variances for the two species will be the same. How do you know that?)

(b) Find a_1 and a_2 to maximize the ratio of the difference between the two means found in part (a) to the standard deviation found in part (a). There is a sense in which this linear combination $a_1X_1 + a_2X_2$ does the best job of distinguishing the two species of among all possible linear combinations.

Solution. (a) The mean of $a_1X_1 + a_2X_2$ is $E(a_1X_1 + a_2X_2) = a_1E(X_1) + a_2E(X_2)$. The variance is $\mathrm{Var}(a_1X_1 + a_2X_2) = {a_1}^2 \mathrm{Var}(X_1) + {a_2}^2 \mathrm{Var}(X_2) + 2a_1a_2\rho\sigma_1\sigma_2$, and the standard deviation is the square root of the variance.

For the first species the linear combination of the two measurements has mean $201a_1 + 118a_2$, and standard deviation

$$\sqrt{(231.04){a_1}^2 + (43.56){a_2}^2 + (128.4096)a_1a_2}.$$

For the second species the linear combination of the two measurements has mean $187a_1 + 131a_2$, and standard deviation

$$\sqrt{(231.04){a_1}^2 + (43.56){a_2}^2 + (128.4096)a_1a_2}.$$

The variances of each species were the same, as the species shared the same variances and correlation.

(b) The ratio we wish to maximize is

$$\frac{(201a_1 + 118a_2) - (187a_1 + 131a_2)}{\sqrt{(231.04){a_1}^2 + (43.56){a_2}^2 + (128.4096)a_1a_2}} = \frac{14a_1 - 13a_2}{\sqrt{(231.04)a_1^2 + (43.56){a_2}^2 + (128.4096)a_1a_2}}$$

It needs to be noted that there is indeterminacy in this optimization problem. By that is meant if, for instance, global maximum occurs (δ_1, δ_2) then there will also be a global maximum at $(c\delta_1, c\delta_2)$ for any constant c because

$$\frac{14\delta_1 - 13\delta_2}{\sqrt{(231.04)\delta_1^2 + (43.56){\delta_2}^2 + (128.4096)\delta_1\delta_2}}$$

$$= \frac{14c\delta_1 - 13c\delta_2}{\sqrt{(231.04)(c\delta_1)^2 + (43.56)(c\delta_2)^2 + (128.4096)(c\delta_1)(c\delta_2)}}.$$

The main cause of the indeterminacy is division by the standard deviation and we need to constrain the standard deviation to a constant, say 1, to alleviate the indeterminacy. So the problem now becomes, maximize

$$f(a_1, a_2) = 14a_1 - 13a_2 \quad \text{subject to} \quad g(a_1, a_2) = (231.04){a_1}^2 + (43.56){a_2}^2 + (128.4096)a_1a_2 - 1 = 0.$$

for which we can use the method of Lagrangian multiplier to locate the global extrema. Setting

$$\nabla f(a_1, a_2) = \lambda \nabla g(a_1, a_2)$$

we solve for a_1, a_2 and λ where λ is the Lagrange's multiplier. That is, we set

$$\frac{\partial}{\partial a_1} f(a_1, a_2) = \lambda \frac{\partial}{\partial a_1} g(a_1, a_2) \quad \text{and} \quad \frac{\partial}{\partial a_2} f(a_1, a_2) = \lambda \frac{\partial}{\partial a_2} g(a_1, a_2)$$

and solve for a_1, a_2 and λ. On computing the partial derivatives, we get the following three equations (the last one is just the constraint)

$$\begin{aligned} 14 &= \lambda(462.08a_1 + 128.4093a_2), \\ -13 &= \lambda(87.12a_2 + 128.4093a_1) \quad \text{and} \\ 231.04a_1^2 + 43.56{a_2}^2 &+ 128.4096a_1a_2 = 1. \end{aligned}$$

Solving these three equations simultaneously, for example, in Mathematica using the following code

$$\text{Solve}[\{14 == \lambda(162.08a_1 + 128.4096a_2), -13 == \lambda(87.12a_2 + 128.4096a_1), \\ 231.04a_1^2 + 43.56a_2^2 + 128.4096 == 1\}, \{a_1, a_2, \lambda\}]$$

gives us the two extrema as $\lambda = -14.2793$, $a_1 = 0.0854177$ and $a_2 = -0.11544$, and $\lambda = 14.2793$, $a_1 = -0.0854177$ and $a_2 = 0.11544$. Obviously the first one yields a global maximum and the second one a global minimum. (Why?)

□

6. (p.319 #15) Let $X_1, \ldots, X_n$ be independent and identically distributed random variables having a normal distribution with mean μ and variance σ^2. Define $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, the sample mean. In this problem, we shall find the conditional distribution of each X_i given $\overline{X}_n$.

   (a) Show that X_i and $\overline{X}_n$ have a bivariate normal distribution with both means μ, variances σ^2 and σ^2/n, and correlation $1/\sqrt{n}$. *Hint:* Let $Y = \sum_{j\neq i} X_j$. Now show that Y and X_i are independent normals, and $\overline{X}_n$ and X_i are linear combinations of Y and X_i.

   (b) Show that the conditional distribution of X_i given $\overline{X}_n = \overline{x}_n$ is normal with mean $\overline{x}_n$ and variance $\sigma^2(1 - 1/n)$.

Solution. (a) As suggested let $Y = \sum_{j\neq i} X_j$, then as Y is the sum of $n-1$ independent and identically distributed random variables each of which has a normal distribution with mean μ and variance σ^2. Thus $Y \sim N(\mu(n-1), \sigma^2(n-1))$. Of course, $X_i \sim N(\mu, \sigma^2)$. As $X_1, \ldots, X_n$ are mutually independent, we have $Y = \sum_{j\neq i} X_j \perp X_i$, that is Y and X_i are independent normals. Let Z_1 and Z_2 be independent random variables, each of which has a standard normal distribution. Then we can write $X_i = \sigma Z_1 + \mu$ and $Y = \sigma(n-1)^{1/2} Z_2 + \mu(n-1)$. As $X_i = X_i$ and $\overline{X}_n = 1/n(Y + X_i)$, we have that

$$\begin{aligned} X_i &= \sigma Z_1 + \mu \\ \overline{X}_n &= \frac{1}{n}(Y + X_i) = \frac{1}{n}\left[\sigma(n-1)^{1/2} Z_2 + \mu(n-1) + \sigma Z_1 + \mu\right] \\ &= \frac{\sigma}{\sqrt{n}}\left(\frac{1}{\sqrt{n}} Z_1 + \frac{\sqrt{n}(1-1/n)^{1/2}}{\sqrt{n}} Z_2\right) + \mu = \frac{\sigma}{\sqrt{n}}\left(\frac{1}{\sqrt{n}} Z_1 + (1-\frac{1}{n})^{1/2} Z_2\right) + \mu. \end{aligned}$$

Hence X_i and $\overline{X}_n$ have a bivariate normal distribution with parameters $\mu_1 = \mu$, $\mu_2 = \mu$, ${\sigma_1}^2 = \sigma^2$, $\sigma_2^2 = \sigma^2/n$ and correlation $\rho = 1/\sqrt{n}$.

(b) Using the property of the bivariate normal distribution, the conditional distribution of X_i given $\bar{X}_n = \bar{x}_n$ is normal distribution with mean

$$E(X_i|\bar{X}_n = \bar{x}_n) = \mu_1 + \rho\sigma_1 \left(\frac{\bar{x}_n - \mu_2}{\sigma_2}\right) = \mu + \frac{1}{\sqrt{n}}\sigma \left(\frac{\bar{x}_n - \mu}{\sigma/\sqrt{n}}\right) = \bar{x}_n$$

and variance

$$\text{Var}(X_i|\bar{X}_n = \bar{x}_n) = (1-\rho^2)\sigma_1^2 = (1-\frac{1}{n})\sigma^2.$$

□

7. (p.320 #8) Suppose that a random sample of 16 observations is drawn from a normal distribution with mean μ and standard deviation 12; and that independently another random sample of 25 observations is drawn from a normal distribution with the same mean μ and standard deviation 20. Let $\overline{X}$ and $\overline{Y}$ denote the sample means of the two samples. Evaluate $P(|\overline{X} - \overline{Y}| < 5)$.

Solution. Recall that, if the random variables $X_1, X_2, \ldots, X_n$ form a random sample from a normal distribution with mean μ and variance σ^2, then the sample mean $\overline{X}_n$ has a normal distribution with mean μ and variance σ^2/n.
So in this case

$$\overline{X}_n \sim N(\mu, 9) \quad \text{and} \quad \overline{Y}_n \sim N(\mu, 16).$$

Let $Z = (\overline{X}_n - \overline{Y}_n)/5$. Then Z will have a normal distribution with mean

$$E(Z) = \frac{1}{5}\, E(\overline{X}_n) - \frac{1}{5}\, E(\overline{Y}_n) = \frac{\mu-\mu}{5} = 0$$

and variance

$$\text{Var}(Z) = \frac{1}{25}\,\text{Var}(\overline{X}_n) + \frac{1}{25}\,\text{Var}(\overline{Y}_n) = \frac{9}{25} + \frac{16}{25} = 1.$$

Therefore,

$$P(|\overline{X} - \overline{Y}| < 5) = P(|Z| < 1) = P(-1 < Z < 1) = \Phi(1) - \Phi(-1) \approx 0.6826895$$

where $\Phi(\cdot)$ is the CDF of $N(0,1)$. In R,

```
> pnorm(1,0,1) - pnorm(-1,0,1)
[1] 0.6826895
```

□

8. (p.235 #8) Suppose that 30 percent of the items in a large manufactured lot are of poor quality. Suppose also that a random sample of n items is to be taken from the lot, and let Q_n denote the proportion of items in the sample that are of poor quality. Find a value of n such that $P(0.2 \le Q_n \le 0.4) \ge 0.75$ by using (a) the Chebyshev inequality and (b) the tables of the binomial distribution at the end of this book [1].

Solution. For $i = 1, 2, \ldots, n$, let the random variable X_i represent the outcome of a selection with $X_i = 1$ if an item of poor quality was selected and $X_i = 0$ if an acceptable item was selected. Then the sample mean $\overline{X}_n$ will be equal to the proportion, Q_n, of items in the sample that are of poor quality. Let $T = \sum_{i=1}^{n} X_i$ denote the total number of items of poor quality in the sample of n items. Then T is the sum of n independent Bernoulli trials each with probability of success $p = 0.3$. Thus

[1] Morris H. DeGroot and Mark J. Schervish, *Probability and Statistics*, Third Edition, Addison-Wesley:New York, 2002.

T has a binomial distribution with parameters n and $p = 0.30$. It follows that $E(T) = np = 0.3n$ and $\mathrm{Var}(T) = np(1-p) = 0.21n$. Since $Q_n = \overline{X}_n = T/n$, from Chebyshev's inequality we have:

$$\begin{aligned} P(0.2 \le Q_n \le 0.4) &= P(0.2n \le T \le 0.4n) \\ &= P\left(|T - 0.3n| \le 0.1n\right) \ge 1 - \frac{0.21n}{(0.1n)^2} = 1 - \frac{21}{n}. \end{aligned}$$

This probability will be at least 0.75 if

$$1 - \frac{21}{n} \ge 0.75 \quad \Leftrightarrow \quad \frac{21}{n} \le 0.25 \quad \Leftrightarrow \quad n \ge 84.$$

That is, $n \ge 84$.
From the tables of binomial distribution given at the end of our book, it is found that for $n = 20$,

$$\begin{aligned} P(0.2 \le Q_n \le 0.4) = P(4 \le T \le 8) &= \sum_{k=4}^{8} P(T = k) \\ &= 0.1304 + 0.1789 + 0.1916 + 0.1643 + 0.1144 = 0.7796. \end{aligned}$$

Therefore, a sample of 20 items would actually be sufficient to satisfy the specified probability requirement. □

9. (p.235 #9) Let $Z_1, Z_2, \ldots$ be a sequence of random variables; and suppose that, for $n = 1, 2, \ldots$, the distribution of Z_n is as follows:

$$P(Z_n = n^2) = \frac{1}{n} \quad \text{and} \quad P(Z_n = 0) = 1 - \frac{1}{n}.$$

Show that

$$\lim_{n\to\infty} E(Z_n) = \infty \text{ but } Z_n \xrightarrow{p} 0.$$

Solution. Recall that, for a discrete random variable the expected value is the sum of the products of each possible value and its respective probability. Thus,

$$E(Z_n) = (n^2)P(Z_n = n^2) + (0)P(Z_n = 0) = n^2\left(\frac{1}{n}\right) = n \quad \Longrightarrow \quad \lim_{n\to\infty} E(Z_n) = \lim_{n\to\infty} n = \infty.$$

To show that $Z_n \xrightarrow{p} 0$, we compute the limit as $n \to \infty$ of $P(|Z_n| < \varepsilon)$ for every number $\varepsilon > 0$. Clearly,

$$P(|Z_n| < \varepsilon) = \begin{cases} \frac{1}{n} & \text{if} \quad \varepsilon \ge n^2 \\ 1 - \frac{1}{n} & \text{if} \quad \varepsilon < n^2 \end{cases}.$$

Therefore for every $\varepsilon \in (0, \infty)$,

$$\lim_{n\to\infty} P(|Z_n| < \varepsilon) = \lim_{n\to\infty} (1 - \frac{1}{n}) = 1$$

as to be proved. □

10. (p.236 #11) Prove that if the sequence $Z_1, Z_2, \ldots$ converges to a constant b in the quadratic mean, then the sequence also converges to b in probability.

Solution. We wish to prove that

$$\lim_{n\to\infty} E(Z_n - b)^2 = 0 \Rightarrow Z_n \xrightarrow{p} b.$$

Equivalently that,

$$\lim_{n\to\infty} E(Z_n - b)^2 = 0 \Rightarrow \lim_{n\to\infty} P(|Z_n - b| < \varepsilon) = 1$$

for every number $\varepsilon > 0$. By applying Markov's Inequality (see Math 441),

$$P(|Z_n - b| > \varepsilon) = P((Z_n - b)^2 > \varepsilon^2) \leq \frac{E(Z_n - b)^2}{\varepsilon^2}.$$

Taking the limit as n tends to infinity

$$0 \leq \lim_{n\to\infty} P(|Z_n - b| > \varepsilon) \leq \lim_{n\to\infty} \frac{E(Z_n - b)^2}{\varepsilon^2} = 0$$

which implies

$$\lim_{n\to\infty} P(|Z_n - b| > \varepsilon) = 0$$

as to be proved. □