SM-4331 Exercise 3

1. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$. Prove that the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ is distributed according to $\bar{X} \sim \mathrm{N}(\mu, \sigma^2/n)$. *Hint: Find the expectation and variance of* $\bar{X}$*, and use the linearity property of normal distributions.*

Solution:

$$\begin{aligned} \mathrm{E}(\bar{X}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) &= \frac{1}{n}\sum_{i=1}^n \mathrm{E}(X_i) \\ &= \frac{1}{n}\sum_{i=1}^n \mu = \mu \end{aligned}$$

$$\begin{aligned} \mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) &= \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(X_i) \\ &= \frac{1}{n^2}\sum_{i=1}^n \sigma^2 = \sigma^2/n \end{aligned}$$

By the linearity property of normal distributions, $\bar{X}$ is normal.

2. Suppose that we plan to take a random sample of size n from a normal distribution with mean μ and standard deviation $\sigma = 2$.

   (a) Suppose $\mu = 4$ and $n = 20$.

   i. What is the probability that the mean $\bar{X}$ of the sample is greater than 5?

Solution: $\bar{X} \sim \mathrm{N}(4, 2^2/20)$. Then

$$\begin{aligned} \mathrm{P}(\bar{X} > 5) &= \mathrm{P}\left(\frac{\bar{X} - 4}{2/\sqrt{20}} > \frac{5-4}{2/\sqrt{20}}\right) \\ &= \mathrm{P}(Z > 2.236) \\ &= 1 - 0.98732 \\ &= 0.01268 \end{aligned}$$

   ii. What is the probability that $\bar{X}$ is smaller than 3?

Solution: $\bar{X} \sim \mathrm{N}(4, 2^2/20)$. Then

$$\begin{aligned} \mathrm{P}(\bar{X} < 3) &= \mathrm{P}\left(\frac{\bar{X}-4}{2/\sqrt{20}} < \frac{3-4}{2/\sqrt{20}}\right) \\ &= \mathrm{P}(Z < -2.236) \\ &= \mathrm{P}(Z > 2.236) \\ &= 0.01268 \end{aligned}$$

iii. What $\mathrm{P}(|\bar{X} - \mu| \leq 1)$ in this case?

Solution: $\bar{X} - \mu \sim \mathrm{N}(0, 2^2/20)$. Then

$$\begin{aligned} \mathrm{P}(|\bar{X} - \mu| < 1) &= \mathrm{P}(-1 < \bar{X} - \mu < 1) \\ &= \mathrm{P}\left(\frac{-1}{2/\sqrt{20}} < \frac{\bar{X}-\mu}{2/\sqrt{20}} < \frac{1}{2/\sqrt{20}}\right) \\ &= \mathrm{P}(-2.236 < Z < 2.236) \\ &= 1 - 2\,\mathrm{P}(Z < -2.236) \\ &= 0.97464 \end{aligned}$$

(b) How large should n be in order that $\mathrm{P}(|\bar{X} - \mu| \leq 0.5) \geq 0.95$ for every possibly value of μ?

Solution: $\bar{X} - \mu \sim \mathrm{N}(0, 2^2/n)$. Then

$$\begin{aligned} \mathrm{P}(|\bar{X} - \mu| < 0.5) = \mathrm{P}\left(|Z| < \frac{0.5}{2/\sqrt{n}}\right) &\geq 0.95 \\ \Rightarrow 2\Phi\left(\frac{0.5}{2/\sqrt{n}}\right) - 1 &\geq 0.95 \\ \frac{0.5}{2/\sqrt{n}} &\geq \Phi^{-1}(0.975) = 1.96 \\ n &\geq (4 \times 1.96)^2 = 61.46 \end{aligned}$$

So n should be 62 or more.

(c) It is claimed that the true value of μ is 5 in a population. A random sample of size $n = 100$ is collected from this population, and the mean for this sample is $\bar{X} = 5.8$. Based on the result in (b), what would you conclude from this value of $\bar{X}$?

Solution: Here, $\bar{X} - \mu \sim \mathrm{N}(0, 2^2/100)$, and a 95% confidence interval based on the observed $\bar{X} = 5.8$ is

$$5.8 \pm 1.96 \cdot 2/10 = (5.408, 6.192),$$

which does not include $\mu = 5$. However from (b), we know that $\mathrm{P}(|\bar{X} - \mu| \leq 0.5)$ is 95% or more if $n \geq 62$. Since we collected a sample of $n = 100$, then it stands to reason that this particular sample is an anomaly (one of the 5% of the times that it is not within an error range of 0.5).

3. (a) If Z is a random variable with a standard normal distribution, what is $\mathrm{P}(Z^2 < 3.841)$?

Solution: Using standard normal distribution,

$$\begin{aligned}\mathrm{P}(Z^2 < 3.841) &= \mathrm{P}(|Z| < \sqrt{3.841} = 1.9598) \\ &= 2\Phi(1.9598) - 1 = 0.95.\end{aligned}$$

Alternatively, we know that $Z^2 \equiv \chi^2_1$, so

$$\mathrm{P}(Z^2 < 3.841) = \mathrm{P}(\chi^2_1 < 3.841) = 0.95.$$

(b) Suppose that X_1 and X_2 are independent $\mathrm{N}(0, 4)$ random variables. Compute $\mathrm{P}(X_1^2 < 36.84 - X_2^2)$.

Solution: Since $X_i \overset{\text{iid}}{\sim} \mathrm{N}(0, 4)$, then $X_i^2/4 \overset{\text{iid}}{\sim} \chi^2_1$.

$$\begin{aligned}\mathrm{P}(X_1^2 < 36.84 - X_2^2) &= \mathrm{P}\left(\frac{X_1^2}{4} + \frac{X_2^2}{4} < 36.84/4 = 9.21\right) \\ &= \mathrm{P}\left(\chi^2_2 < 9.21\right) = 0.99.\end{aligned}$$

(c) Suppose that $X_1, X_2, X_3 \overset{\text{iid}}{\sim} \mathrm{N}(0, 1)$, while Y independently follows a χ^2_5 distribution. Compute $\mathrm{P}(X_1^2 + X_2^2 < 7.236Y - X_3^2)$.

Solution: Since $X_i \overset{\text{iid}}{\sim} \mathrm{N}(0, 1)$, then $X_i^2 \overset{\text{iid}}{\sim} \chi^2_1$.

$$\begin{aligned}\mathrm{P}(X_1^2 + X_2^2 < 7.236Y - X_3^2) &= \mathrm{P}\left(\frac{X_1^2 + X_2^2 + X_3^2}{Y} < 7.236\right) \\ &= \mathrm{P}\left(\frac{\chi^2_3/3}{\chi^2_5/5} < 7.236 \times 5/3 = 12.060\right) \\ &= \mathrm{P}\left(F_{3,5} < 12.060\right) = 0.99.\end{aligned}$$

4. Let X_i, $i = 1, 2, 3$ be independent with $\mathrm{N}(i, i^2)$ distributions. For each of the following situations, use the X_is to construct a statistic with the indicated distribution:

   (a) χ^2-distribution with 3 degrees of freedom;

   Solution: $(X_i - i)/i \overset{\text{iid}}{\sim} \mathrm{N}(0, 1)$, thus $Y = \sum_{i=1}^{3}(X_i - i)^2/i^2 \sim \chi^2_3$.

   (b) t-distribution with 2 degrees of freedom; and

   Solution: Let $Z = (X_1 - 1) \sim \mathrm{N}(0, 1)$, and $Y = \sum_{i=2}^{3}(X_i - i)^2/i^2 \sim \chi^2_2$. Then $Z/\sqrt{Y/2} \sim t_2$.

   (c) F-distribution with 1 and 2 degrees of freedom.

   Solution: Let $W = (X_1 - 1)^2 \sim \chi^2_1$, and $Y = \sum_{i=2}^{3}(X_i - i)^2/i^2 \sim \chi^2_2$. Then $W/(Y/2) \sim F_{1,2}$.

5. Imagine rolling an r-sided die n number of times independently. Define the indicator function

$$\mathbb{1}_{[k=i]}(k) = \begin{cases} 1 & \text{if roll } k \text{ is equal to } i \\ 0 & \text{otherwise} \end{cases}$$

   Suppose further that $\mathrm{P}(\mathbb{1}_{[k=i]}(k) = 1) = p_i$.

   (a) What is $\mathrm{E}\left[\,\mathbb{1}_{[k=i]}(k)\right]$ and $\mathrm{Var}\left[\,\mathbb{1}_{[k=i]}(k)\right]$?

   Solution: Since this is a Bernoulli random variable, $\mathrm{E}\left[\,\mathbb{1}_{[k=i]}(k)\right] = p_i$ and $\mathrm{Var}\left[\,\mathbb{1}_{[k=i]}(k)\right] = p_i(1 - p_i)$.

   (b) Calculate $\mathrm{E}\left[\,\mathbb{1}_{[k=i]}(k)\,\mathbb{1}_{[l=j]}(l)\right]$ when $k \neq l$.

   Solution: We note that $\mathbb{1}_{[k=i]}(k)\,\mathbb{1}_{[l=j]}(l)$ takes value 1 if and only if roll k is equal to i <u>and</u> roll l is equal to j. This happens with probability $p_i p_j$ due to independence of the rolls. Otherwise, $\mathbb{1}_{[k=i]}(k)\,\mathbb{1}_{[l=j]}(l) = 0$ with probability $1 - p_i p_j$. Thus,

$$\mathrm{E}\left[\,\mathbb{1}_{[k=i]}(k)\,\mathbb{1}_{[l=j]}(l)\right] = p_i p_j.$$

   (c) Argue that $\mathrm{E}\left[\,\mathbb{1}_{[k=i]}(k)\,\mathbb{1}_{[l=j]}(l)\right] = 0$ when $k = l$.

   Solution: It is impossible that for the same roll that the r-sided die to show faces i and j at the same time. Since this is an impossible event, its expectation is zero.

(d) Let X_i be the number of rolls that result in side i facing up. Write down the equation relating X_i and the indicator functions above. What possible values can X_i take?

Solution: As we are counting the number of occurrences that the rolls result in side i (in other words, $\mathbb{1}_{[k=i]}(k) = 1$),

$$X_i = \sum_{k=1}^{n} \mathbb{1}_{[k=i]}(k).$$

X_i can take values from 0 to n. As a remark, the vector $(X_1, \ldots, X_r)^\top$ for which $\sum_{i=1}^{r} X_i = n$ follows a multinomial distribution. Thus, we should expect X_i and X_j to be correlated (not independent).

(e) Determine $\mathrm{E}(X_i)$.

Solution: A sum of Bernoulli random variables is binomial, so $X_i \sim \mathrm{Bin}(n, p_i)$. Thus, $\mathrm{E}(X_i) = np_i$.

(f) Consider two random variables X_i and X_j defined as per (d). From your answers to (a), (b) and (c), calculate $\mathrm{E}(X_i X_j)$.

Solution: Let

$$\begin{aligned}
X_i X_j &= \left(\sum_{k=1}^{n} \mathbb{1}_{[k=i]}(k) \right) \left(\sum_{l=1}^{n} \mathbb{1}_{[l=j]}(l) \right) \\
&= \sum_{\substack{k=1 \\ k \neq l}}^{n} \sum_{l=1}^{n} \mathbb{1}_{[k=i]}(k)\, \mathbb{1}_{[l=j]}(l) + \sum_{\substack{k=1 \\ k = l}}^{n} \sum_{l=1}^{n} \mathbb{1}_{[k=i]}(k)\, \mathbb{1}_{[l=j]}(l) \\
\Rightarrow \mathrm{E}(X_i X_j) &= \sum_{\substack{k=1 \\ k \neq l}}^{n} \sum_{l=1}^{n} \mathrm{E}\left[\mathbb{1}_{[k=i]}(k)\, \mathbb{1}_{[l=j]}(l) \right] + \cancelto{0}{\sum_{\substack{k=1 \\ k = l}}^{n} \sum_{l=1}^{n} \mathrm{E}\left[\mathbb{1}_{[k=i]}(k)\, \mathbb{1}_{[l=j]}(l) \right]} \\
&= \sum_{\substack{k=1 \\ k \neq l}}^{n} \sum_{l=1}^{n} p_i p_j \\
&= (n^2 - n) p_i p_j
\end{aligned}$$

Since there are n sums each in X_i and X_j, multiplying out there are n^2 terms in $X_i X_j$. Think of a square $n \times n$ matrix. The diagonal entries are when $k = l$, and the off-diagonals are $k \neq l$. There are exactly n diagonal entries, so therefore there are $n^2 - n$ off-diagonal entries.

(g) Now calculate the covariance between X_i and X_j.

Solution:

$$\begin{aligned}\mathrm{Cov}(X_i, X_j) &= \mathrm{E}(X_i X_j) - \mathrm{E}(X_i)\,\mathrm{E}(X_j) \\ &= (n^2 - n)p_i p_j - np_i \cdot np_j \\ &= -np_i p_j\end{aligned}$$

6. Let $\{X_1, \ldots, X_n\}$ be a random sample from a $\mathrm{N}(\mu, \sigma^2)$ population.

   (a) Let $M = \sum_{i=1}^{n}(X_i - \bar{X})^2$, where $\bar{X}$ is the sample mean. Work out the distribution of M/σ^2.

   Solution: We know that $\bar{X} \sim \mathrm{N}(\mu, \sigma^2/n)$, and $\sqrt{n}(\bar{X} - \mu)/\sigma \sim \mathrm{N}(0, 1)$, and therefore

   $$\frac{n(\bar{X} - \mu)^2}{\sigma^2} \sim \chi_1^2.$$

   Also, $(X_i - \mu)/\sigma \sim \mathrm{N}(0, 1)$, and thus

   $$\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \sim \chi_n^2$$

   Furthermore,

   $$\overbrace{\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2}^{\chi_n^2} = \overbrace{\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2}^{M/\sigma^2} + \overbrace{\frac{n}{\sigma^2}(\bar{X} - \mu)^2}^{\chi_1^2}$$

   so we have that $M/\sigma^2 \sim \chi_{n-1}^2$.

   (b) Let $\alpha = 0.05$. Using the χ^2 probability tables, determine the values of $\chi_{14}^2(\alpha/2)$ and $\chi_{14}^2(1-\alpha/2)$, i.e. the top and bottom $\alpha/2$ point of the χ_{14}^2 distribution where $\mathrm{P}\left(Y < \chi_k^2(a)\right) = a$ when $Y \sim \chi_k^2$.

   Solution: $\chi_{14}^2(0.025) = 26.12$ and $\chi_{14}^2(0.975) = 5.63$.

   (c) Suppose $n = 15$ and the sample variance is $s^2 = 24.5$. What is a 95% confidence interval for σ^2?

   Solution: Note that $s^2 = M/(n-1) = 24.5$, so $M = 24.5 \times 14 = 343$. We also know that $\mathrm{P}(5.63 < M/\sigma^2 < 26.12) = 0.95$, therefore

   $$\begin{aligned}\{5.63 < M/\sigma^2 < 26.12\} &= \{M/26.12 < \sigma^2 < M/5.63\} \\ &= \{13.13 < \sigma^2 < 60.92\}\end{aligned}$$

   is a 95% confidence interval for σ^2.

7. Let $\{Y_{ij}\}$ be sample from $\mathrm{N}(\mu_j, \sigma^2)$, $i = 1, \ldots, n_j$ and $j = 1, \ldots, m$. In total there are $n = \sum_{j=1}^m n_j$ samples. Further, let $S = \sum_{i=1}^{n_j} \sum_{j=1}^m (Y_{ij} - \bar{Y})^2$, where $\bar{Y} = \frac{1}{n} \sum_{i=1} \sum_{j=1}^m Y_{ij}$.

(a) Define the sample group means to be $\bar{Y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{ij}$. Add and subtract the sample group mean $\bar{Y}_j$ into the squared sum in S to show that

$$\sum_{i=1}^{n_j} \sum_{j=1}^{m} (Y_{ij} - \bar{Y})^2 = \sum_{i=1}^{n_j} \sum_{j=1}^{m} (Y_{ij} - \bar{Y}_j)^2 + \sum_{j=1}^{m} n_j (\bar{Y}_j - \bar{Y})^2$$

Solution:

$$\begin{aligned}\sum_{i,j} (Y_{ij} - \bar{Y})^2 &= \sum_{i,j} (Y_{ij} - \bar{Y}_j + \bar{Y}_j - \bar{Y})^2 \\ &= \sum_{i,j} (Y_{ij} - \bar{Y}_j)^2 + \sum_{i,j} (\bar{Y}_j - \bar{Y})^2 \\ &\quad + 2 \sum_{i,j} (Y_{ij} - \bar{Y}_j)(\bar{Y}_j - \bar{Y})\end{aligned}$$

The third component of the RHS is

$$2 \sum_{i=1}^{n_j} \sum_{j=1}^{m} (Y_{ij} - \bar{Y}_j)(\bar{Y}_j - \bar{Y}) = 2 \sum_{j=1}^{m} (\bar{Y}_j - \bar{Y}) \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)$$

but $\sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j) = \frac{n_j}{n_j} \sum_{i=1}^{n_j} Y_{ij} - n_j \bar{Y}_j = 0$, so the entire ssum is zero. Also, the second component of the RHS is

$$\sum_{i=1}^{n_j} \sum_{j=1}^{m} (\bar{Y}_j - \bar{Y})^2 = \sum_{j=1}^{m} \sum_{i=1}^{n_j} (\bar{Y}_j - \bar{Y})^2 = \sum_{j=1}^{m} n_j (\bar{Y}_j - \bar{Y})^2$$

(b) What is the distribution of $\bar{Y}$ and $\bar{Y}_j$?

Solution: $\bar{Y} \sim \mathrm{N}(\mu, \sigma^2/n)$ and $\bar{Y}_j \sim \mathrm{N}(\mu_j, \sigma^2/n_j)$.

(c) Assuming that $\mu_j = \mu$, for all $j = 1, \ldots, m$ and using your answer to (b), determine then the following distributions

i. $\frac{1}{\sigma^2} \sum_{i=1}^{n_j} \sum_{j=1}^{m} (Y_{ij} - \mu)^2$

Solution: Since $Y_{ij} \sim \mathrm{N}(\mu, \sigma^2)$, $(Y_{ij} - \mu)/\sigma \sim \mathrm{N}(0, 1)$, so

$$\frac{1}{\sigma^2} \sum_{i=1}^{n_j} \sum_{j=1}^{m} (Y_{ij} - \mu)^2 \sim \chi^2_n$$

.

ii. $\frac{n}{\sigma^2}(\bar{Y}-\mu)^2$

> **Solution:** Since $\bar{Y} \sim \mathrm{N}(\mu, \sigma^2/n)$, $\sqrt{n}(\bar{Y}-\mu)/\sigma \sim \mathrm{N}(0,1)$, so
>
> $$\frac{n}{\sigma^2}(\bar{Y}-\mu)^2 \sim \chi_1^2$$
>
> .

iii. $\frac{1}{\sigma^2}\sum_{i=1}^{n_j}\sum_{j=1}^{m}(Y_{ij}-\bar{Y})^2$

> **Solution:** We can show that
>
> $$\frac{1}{\sigma^2}\sum_{i=1}^{n_j}\sum_{j=1}^{m}(Y_{ij}-\mu)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n_j}\sum_{j=1}^{m}(Y_{ij}-\bar{Y})^2 + \frac{n}{\sigma^2}(\bar{Y}-\mu)^2$$
>
> and therefore $\frac{1}{\sigma^2}\sum_{i=1}^{n_j}\sum_{j=1}^{m}(Y_{ij}-\bar{Y})^2 \sim \chi_{n-1}^2$.

iv. $\frac{1}{\sigma^2}\sum_{j=1}^{m} n_j(\bar{Y}_j-\mu)^2$

> **Solution:** Since $\bar{Y}_j \sim \mathrm{N}(\mu, \sigma^2/n_j)$, $\sqrt{n}_j(\bar{Y}_j-\mu)/\sigma \sim \mathrm{N}(0,1)$, so
>
> $$\frac{n_j}{\sigma^2}(\bar{Y}_j-\mu)^2 \sim \chi_1^2,$$
>
> and thus $\frac{1}{\sigma^2}\sum_{j=1}^{m} n_j(\bar{Y}_j-\mu)^2 \sim \chi_m^2$.

v. $\frac{1}{\sigma^2}\sum_{i=1}^{n_j}\sum_{j=1}^{m}(Y_{ij}-\bar{Y}_j)^2$

> **Solution:** We can also show that
>
> $$\frac{1}{\sigma^2}\sum_{i=1}^{n_j}\sum_{j=1}^{m}(Y_{ij}-\mu)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n_j}\sum_{j=1}^{m}(Y_{ij}-\bar{Y}_j)^2 + \frac{1}{\sigma^2}\sum_{j=1}^{m} n_j(\bar{Y}_j-\mu)^2$$
>
> and therefore $\frac{1}{\sigma^2}\sum_{i=1}^{n_j}\sum_{j=1}^{m}(Y_{ij}-\bar{Y}_j)^2 \sim \chi_{n-m}^2$.

Hint: Use the sum of squares decomposition with $\bar{Y}$ and $\bar{Y}_j$, and then use the properties of χ^2-distributions.

(d) Finally using the properties of χ^2 distributions, argue that $\sum_{j=1}^{m} n_j(\bar{Y}_j-\bar{Y})^2$ must follow a χ_{n-m}^2 distribution.