2020 Probability and Statistics Examination Solution

1. a. Since the population variance is unknown, we need to use the bias-corrected variance to construct the confidence interval. Since $n = 100 > 40$, which is a siginificant amount of samples and hence a high degree of freedom, we can use Normal distribution instead of Student's t-distribution to get the z score.

   Hence the confidence interval is

$$[\bar{x} - z\frac{s}{\sqrt{n}}, \bar{x} + z\frac{s}{\sqrt{n}}] = [1.2 - 1.96 \cdot \frac{2.03}{10}, 1.2 + 1.96 \cdot \frac{2.03}{10}] = [0.80212, 1.59788]$$

   where $\bar{x} = \frac{120}{100} = 1.2$, $z = 1.96$ from the normal distribution table (two-tailed), $s = \sqrt{\frac{1}{100-1}(550 - \frac{120^2}{100})} = 2.03$

   If we draw a large number of samples, 95% percent of their mean will be captured by this interval. Central limit theorem states that the distribution of sample mean will converge to normal distribution if the number of samples drawn tends to infinity.

   b. To derive the MLE $\hat{f}$ for f, we need to first model the pmf of the allele gene. If the probability of G occuring is f, then the probability of g occuring is $1 - f$. The probability of GG is then f^2, the probability of Gg is $f(1-f)$ and then probability of gg is $(1-f)^2$.

   Then the MLE is

$$\begin{aligned}
L(f) &= (f^2)^{n_1}(f(1-f))^{n_2}(1-f)^{2n_3} \\
l(f) &= 2n_1 \log f + n_2 \log f + n_2 \log(1-f) + 2n_3 \log(1-f) \\
&= (2n_1 + n_2)\log f + (n_2 + 2n_3)\log(1-f) \\
l'(f) &= \frac{2n_1 + n_2}{f} - \frac{n_2 + 2n_3}{1-f} \\
l'(f) = 0 \Rightarrow \frac{2n_1 + n_2}{f} &= \frac{n_2 + 2n_3}{1-f} \\
2n_1 + n_2 - (2n_1 + n_2)f &= (n_2 + 2n_3)f \\
2n_1 + n_2 &= 2(n_1 + n_2 + n_3)f \\
f &= \frac{2n_1 + n_2}{2(n_1 + n_2 + n_3)}
\end{aligned}$$

   Hence $\hat{f} = \frac{2n_1 + n_2}{2(n_1 + n_2 + n_3)}$. We can check this is indeed the maximum value of $l(f)$ by taking the second derivative of it and see if it's negative:

$$l''(f) = -\frac{(2n_1 + n_2)^2}{f^2} - \frac{(n_2 + 2n_3)^2}{(1-f)^2} < 0$$

   c. We conduct a hypothesis testing against the claim made by the internet provider. The null and alternative hypothesis is shown below

$$H_0 : \mu = 30 \quad H_1 : \mu \neq 30$$

   Since the variance is bias-corrected, we do not know the true population variance and hence use the t-distribution with $21 - 1 = 10$ degrees of freedom. We calculate the test statistic T as follows

$$T = \frac{\bar{X} - \mu_0}{s_{n-1}/\sqrt{n}}$$

the observed test statistic is then

$$t = \frac{35 - 30}{\sqrt{25}/\sqrt{21}} = 4.58$$

From the table of t_{20} at significance level of 5%, we know that the rejection region is $(-\infty, -2.09) \cup (2.09, \infty)$. Our test statistic falls in this range and hence we reject the null hypothesis and the advertisement is not accurate at significance level of 5%.

d. i) The MLE $\hat{\lambda}$ is as follows

$$\begin{aligned} L(\lambda) &= p(1)^{64} p(2)^{44} p(3)^{32} p(4)^{23} p(5)^{20} p(6)^{12} \\ &= \lambda^{195} e^{-64\lambda - 2\cdot 44\lambda - 3\cdot 32\lambda - 4\cdot 23\lambda - 5\cdot 20\lambda - 6\cdot 12\lambda} \\ &= \lambda^{195} e^{-512\lambda} \\ l(\lambda) &= 195 \log \lambda - 512\lambda \\ l'(\lambda) = 0 \Rightarrow \frac{195}{\lambda} &= 512 \\ \lambda &= \frac{195}{512} \approx 0.38 \end{aligned}$$

We then check if the second order derivative is negative

$$l''(\lambda) = -\frac{195}{\lambda^2} < 0$$

ii) We conduct the goodness of fit test by the following procedure

We first calculate the exepcted values by using the cdf of the exponential distribution

$$\begin{aligned} E_1 &= 195 \cdot (F(1) - F(0)) = 195 \cdot (1 - e^{-0.38\cdot 1}) - (1 - e^{-0.38\cdot 0}) \approx 61.65 \\ E_2 &= 195 \cdot (F(2) - F(1)) = 195 \cdot (1 - e^{-0.38\cdot 2}) - (1 - e^{-0.38\cdot 1}) \approx 42.16 \\ E_3 &= 195 \cdot (F(3) - F(2)) = 195 \cdot (1 - e^{-0.38\cdot 2}) - (1 - e^{-0.38\cdot 1}) \approx 28.83 \\ E_4 &= 195 \cdot (F(4) - F(3)) = 195 \cdot (1 - e^{-0.38\cdot 2}) - (1 - e^{-0.38\cdot 1}) \approx 19.72 \\ E_5 &= 195 \cdot (F(5) - F(4)) = 195 \cdot (1 - e^{-0.38\cdot 2}) - (1 - e^{-0.38\cdot 1}) \approx 13.48 \\ E_6 &= 195 \cdot (F(6) - F(5)) = 195 \cdot (1 - e^{-0.38\cdot 2}) - (1 - e^{-0.38\cdot 1}) \approx 9.22 \end{aligned}$$

$$\begin{aligned} \chi^2 &= \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \\ &= \frac{(64 - 61.65)^2}{61.65} + \frac{(44 - 42.16)^2}{42.16} + \frac{(32 - 28.83)^2}{28.83} \\ &+ \frac{(23 - 19.72)^2}{19.72} + \frac{(20 - 13.48)^2}{13.48} + \frac{(12 - 9.22)^2}{9.22} \\ &= 0.0896 + 0.0803 + 0.3486 + 0.5456 + 3.1536 + 0.8382 \\ &= 5.0559 \end{aligned}$$

the degree of freedom is $6 - 1 - 1 = 4$ and hence, according to the χ^2 table we have $5.0559 < 9.49$, which indicates that the test statistic value is not in the range of rejection region. Hence, we do not reject the model of exponential distribution with parameter $\lambda = \hat{\lambda}$ at significance level 5%.

2. a. i) The probablity of $E = \text{Molly will like it}$ is as follows

$$\begin{aligned} P(E) &= P(E|F_1)P(F_1) + P(E|F_2)P(F_2) + P(E|F_3)P(F_3) \\ &= 0.95 \cdot 0.5 + 0.4 \cdot 0.2 + 0.25 \cdot 0.3 = 0.63 \end{aligned}$$

where F_n indicates the probability that the selected carton is produced at factory Lancashire, Derbyshire, and Yorkshire.

Hence the probability that Molly will like it is 0.63.

ii) Given that Molly does not like the selected carton, we can calculate the probability of production at each factory as follows

$$P(\bar{E}) = 1 - 0.63 = 0.37$$

$$P(F_1|\bar{E}) = \frac{P(\bar{E}|F_1)P(F_1)}{P(\bar{E})} = \frac{(1-0.95)\cdot 0.5}{0.37} = 0.068$$

$$P(F_2|\bar{E}) = \frac{P(\bar{E}|F_2)P(F_2)}{P(\bar{E})} = \frac{(1-0.4)\cdot 0.2}{0.37} = 0.324$$

$$P(F_3|\bar{E}) = \frac{P(\bar{E}|F_3)P(F_3)}{P(\bar{E})} = \frac{(1-0.25)\cdot 0.3}{0.37} = 0.608$$

iii) The expectation is $E(X) = 250 \cdot 0.95 + 100 \cdot 0.4 + 150 \cdot 0.25 = 315$

The standard deviation is

$$\sigma = \sqrt{250 \cdot 0.95 \cdot 0.05 + 100 \cdot 0.4 \cdot 0.6 + 150 \cdot 0.25 \cdot 0.75} = \sqrt{64} = 8$$

These calculations are made based on the fact that the three factory are independent and the number of catons in the box that Molly will like is a binomial distribution.

b. i) The definitino of pgf is

$$G(z) = E\{z^X\} = \sum_x z^x p(x)$$

Assuming that $X_1, X_2, \dots X_n$ are independent random variables. Then we have

$$\begin{aligned} G_{S_n}(z) = E(z^{\sum_{i=1}^n X_i}) = E(\prod_{i=1}^n z^{X_i}) &= \prod_{i=1}^n E(z^{X_i}) \qquad \text{(By independence)} \\ &= \prod_{i=1}^n G_{X_i}(z) \end{aligned}$$

ii) The pgf of a Bernoulli random variable X with parameter p is

$$\begin{aligned} G_X(z) = E(z^X) &= \sum_x z^x (p^x (1-p)^{1-x}) \\ &= z^0 (p^0 (1 - p^{1-0})) + z^1 (p^1 (1-p)^{1-1}) \\ &= 1 - p + zp \end{aligned}$$

From the conclusion in i) we can acquire that $G_{\text{Bin}}(z) = \prod_i G_{\text{Bernoulli}_i}(z) = (1-p+zp)^n$

iii) We will show $\lim_{n\to\infty} G_{\text{Bin}}(z)$ as follows

$$\lim_{n\to\infty} G_{\text{Bin}}(z) = \lim_{n\to\infty} (1 - \frac{\lambda}{n} + \frac{\lambda z}{n})^n = \lim_{n\to\infty} (1 + \frac{\lambda(z-1)}{n})^n = e^{\lambda(z-1)} = e^{-\lambda(1-z)}$$

iv) The pgf of Poisson distribution with parameter λ is

$$\begin{aligned}G_{\text{Poisson}}(z) &= \sum_x z^x p(x) = z^0\frac{e^{-\lambda}}{0!} + z^1\frac{\lambda e^{-\lambda}}{1!} + z^2\frac{\lambda^2 e^{-\lambda}}{2!} + \ldots \\ &= e^{-\lambda}(\frac{(z\lambda)^0}{0!} + \frac{(z\lambda)^1}{1!} + \frac{(z\lambda)^2}{2!} + \ldots) \\ &= e^{-\lambda}e^{\lambda z} = e^{-\lambda(1-z)}\end{aligned}$$

which is precisely the limit calculated in iii). As shown in iii), if we put $p = \frac{\lambda}{n}$, the pgf of Binomial distribution is approaching the pgf of a Poisson distribution with $\lambda = np$.

c. i) The definition of $P_{XY}(B_X, B_Y)$ is as follows (**this is an alternative definition besides the one from lecture note**)

$$P_{XY}(B_X, B_Y) = P(X^{-1}(B_X) \cap Y^{-1}(B_Y)) \qquad B_X, B_Y \subset \mathbb{R}$$

where $X(s) \in B_X, X(s) \in B_Y, s \in S$

ii) The probability of $X < Y$ is as follows

$$\begin{aligned}P(X < Y) &= \int_{y=0}^{\infty}\int_{x=0}^{y} f_{XY}(x,y)dxdy \\ &= \int_{y=0}^{\infty}\int_{x=0}^{y} \lambda e^{-\lambda x} \cdot \mu e^{-\mu y}dxdy \\ &= \int_{y=0}^{\infty}(1 - e^{-\lambda y}) \cdot \mu e^{-\mu y}dy \\ &= 1 + (0 - \frac{\mu}{\lambda + \mu}) \\ &= 1 - \frac{\mu}{\lambda + \mu} \\ &= \frac{\lambda}{\lambda + \mu}\end{aligned}$$