Solutions to Final, 28/10/19, Statistics & Probability (191506103)

[The fine prints given in some of the problems indicates
how the reasoning should go in answering that question.]

1. [A standard normal r.v. is continuous, and the squared transformation is also continuous. So, to obtain the pdf, the cdf has to be calculated first and then differentiated. But, as always, we have to start with determining the range of values new random variable may take. Furthermore, while providing the final answer this range should be mentioned again.]

   Suppose Z is a standard normal r..v. with pdf given by

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < +\infty.$$

   Denoting the cdf of Z by $\Phi(z)$, we have $\Phi(z) = P(Z \le z)$ and $\Phi'(x) = \phi(z)$.

   We want to find the pdf of $X := Z^2$. Note that X can only take positive values, i.e., takes values in $(0, \infty)$. [The graph of x^2 against the x-axis will confirm this.]

   Thus, $F_X(x) := P(X \le x) = 0$, if $x \le 0$. Furthermore, for $x > 0$,

$$F_X(x) = P(X \le x) = P\left(Z^2 \le x\right) = P\left(-\sqrt{x} \le Z \le \sqrt{x}\right) = \Phi(\sqrt{x}) - \Phi(-\sqrt{x}).$$

   [Differentiating $F_X(x)$ w.r.t. x (and using the chain rule of differentiation in the process) we obtain the pdf of X.]

   Thus, the probability density function of X is given by

$$\begin{aligned} f_X(x) &= F'_X(x) = \frac{d}{dx}\left[\Phi(\sqrt{x}) - \Phi(-\sqrt{x})\right] \\ &= \Phi'(\sqrt{x}) \cdot \frac{d}{dx}(\sqrt{x}) - \Phi'(-\sqrt{x}) \cdot \frac{d}{dx}(-\sqrt{x}) \\ &= \phi(\sqrt{x}) \cdot \frac{1}{2\sqrt{x}} - \phi(-\sqrt{x}) \cdot \left(-\frac{1}{2\sqrt{x}}\right) = \frac{1}{2\sqrt{x}}\left[\phi(\sqrt{x}) + \phi(-\sqrt{x})\right] \\ &= \frac{1}{2\sqrt{x}} \cdot 2 \cdot \phi(\sqrt{x}) = \frac{1}{\sqrt{2\pi\, x}} e^{-x/2}, \quad x > 0. \end{aligned}$$

2. We are given the joint pdf: $f(x,y) = 4, \quad 0 \le y \le x; \quad x + y \le 1.$

   (a.) [To describe the marginal prob distn of Y, one must provide the range of values of Y and it should NOT depend on x. When we are talking about Y by itself (i.e., only of Y), presence of x is inappropriate/confusing.]

   Note that the poaaiblw values of the random pair (X, Y) lie in the triangle determined by the points $(0,0)$, $(1,0)$ and $(\frac{1}{2}, \frac{1}{2})$. Thus, when considered on its own, the r.v. Y can take the lowest value 0 and the highest value 1 (in combination with different X values). [Draw the triangle to see this.]

   For $0 < y < 1$, the (marginal) pdf of Y is given by $f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx$. [However, it can be seen from the drawing that "given y", the range of integration (x, in terms of y) is different for $y < \frac{1}{2}$ and $y > \frac{1}{2}$.] Integrating over the appropriate ranges, we have

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx = \begin{cases} \int_0^y 4\,dx = 4y, & 0 < y \le \frac{1}{2} \\ \int_0^{1-y} 4\,dx = 4(1-y), & \frac{1}{2} < y < 1. \end{cases}$$

2. (b.) [The conditional pdf is about the r.v. X. So, the range of values of X must be specified. Here the range MAY depend on y, because we are considering the scenario when Y is known to have taken a particular value, y.]

Note that [look at the drawn picture] given $Y = y$, the r.v. X takes values between 0 and y, if $y < \frac{1}{2}$, and between 0 and $(1 - y)$, if $y > \frac{1}{2}$. Thus, from definition we have,

$$f_{X|Y=y}(y) \;=\; \frac{f(x,y)}{f_Y(y)} \;=\; \begin{cases} \frac{4}{4y} \;=\; \frac{1}{y}, & 0 \le x \le y \qquad (\text{if } 0 < y \le \frac{1}{2}) \\ \frac{4}{4(1-y)} \;=\; \frac{1}{1-y}, & 0 \le x \le 1 - y \quad (\text{if } \frac{1}{2} < y < 1). \end{cases}$$

(c.) [To calculate $E(X|Y)$, we need to calculate $E(X|Y = y)$ first.]

$$E(X|Y = y) \;=\; \int_{-\infty}^{\infty} x\, f_{X\,|\,Y=y}(x)\, dx \;=\; \begin{cases} \int_0^y x\, \frac{1}{y}\, dx \;=\; \frac{y}{2} & \text{if } 0 < y \le \frac{1}{2} \\ \int_0^{1-y} x\, \frac{1}{1-y}\, dx \;=\; \frac{1-y}{2} & \text{if } \le \frac{1}{2} < y < 1. \end{cases}$$

Thus, $E(X|Y) = Y/2$ or $(1 - Y)/2$, depending on whether $Y \le \frac{1}{2}$ or $Y > \frac{1}{2}$.

3. Let us consider a box of 10 flower bulbs and let $X =$ the number of bulbs (out of the 10) that will germinate. It is known that a bulb germinates with probability $1 - 0.03 = 0.97$. We can consider checking each bulb to be a Bernoulli experiment and the germination of a bulb to be a success. Then we can model X as $X \sim \text{Binomial}(10, 0.97)$.

(a.) Guarantee is that at least 9 out of 10 (in the box) will germinate. Hence, the required probability that a box will *not* satisfy the guarantee is given by

$$\begin{aligned} P(X < 9) \;&=\; 1 - [P(X = 9) + P(X = 10)] \;=\; 1 - \left[10 \cdot 0.97^9 \cdot 0.03 + 0.97^{10}\right] \\ &=\; 1 - 0.9655 \;=\; 0.0345. \end{aligned}$$

(b.) Let us now consider the big shipment of 1000 boxes. Let W be the number of boxes (out of the 1000) that will not satisfy the guarantee. Now checking each box can be considered as a Bernoulli experiment with success being the failure to satisfy the guarantee. It then follows that W is a Binomial(n, p) r.v. with $n = 1000$ and the probability of success $p = 0.0345$ from part (a.). The required probability is $P(W \le 30)$.

To calculate the probability we can use normal approximation to binomial distribution, due to the CLT / De Moivre's Law, because $n = 1000$ is quite large. Note that here $np \pm 3\sqrt{np(1-p)} = [17.19, 51.81] \subset [0, 1000]$.

To make the approximation <u>*better*</u> we should apply a continuity correction.

The required probability is then

$$\begin{aligned} P(W \le 30) \;&=\; P(W \le 30.5) \qquad\qquad (\text{with continuity correction}) \\ &=\; P\left(\frac{W - np}{\sqrt{np(1-p)}} \le \frac{30.5 - 1000 \times 0.0345}{\sqrt{1000 \times 0.0345 \times 0.9655}}\right) \\ &\approx\; P(Z \le -0.69), \quad \text{where } Z \sim N(0,1) \qquad (\text{CLT}) \\ &=\; \Phi(-0.69) \;=\; 0.2451. \qquad (\text{From table.}) \end{aligned}$$

4. Let $X \sim N(0,1)$, and $Y \sim N(2,4)$ and $Z = 2X + 3Y$. We are required to find the probability distribution of Z using its moment generating function (mgf). So, we shall first calculate the mgf of Z, and see if it is of some known form. Subsequently, we can use the uniqueness property of mgf to identify the probability distribution of Z.

   Since the mgf for $N(\mu, \sigma^2)$ is given by $e^{\mu t + \frac{1}{2}\sigma^2 t^2}$, the mgf's of X and Y are

   $$m_X(t) = E\left[e^{tX}\right] = e^{0 \cdot t + \frac{1}{2} \cdot 1 \cdot t^2} = e^{\frac{1}{2}t^2} \quad \text{and} \quad m_Y(t) = E\left[e^{tY}\right] = e^{2 \cdot t + \frac{1}{2} \cdot 4 \cdot t^2} = e^{2t + 2t^2}.$$

   The mgf of Z is then given by

   $$\begin{aligned} m_Z(t) &= E\left[e^{tZ}\right] = E\left[e^{t(2X+3Y)}\right] = E\left[e^{t2X} \cdot e^{t3Y)}\right] \\ &= E\left[e^{2tX}\right] \cdot E\left[e^{3tY}\right] \qquad \text{(using independence of } X \text{ and } Y) \\ &= m_X(2t) \cdot m_Y(3t) = e^{\frac{1}{2}(2t)^2} \cdot e^{2(3t)+2(3t)^2} = e^{2t^2 + 6t + 18t^2} = e^{6t + 20t^2} \\ &= e^{6 \cdot t + \frac{1}{2} \cdot 40 \cdot t^2}. \end{aligned}$$

   The above mgf is the same as that of a normal r.v. with $\mu = 6$ and $\sigma^2 = 40$. Hence, from the uniqueness property of mgf we can conclude that $Z \sim N(6, 40)$.

5. X and Y are independent with $X \sim Bin(n_1, p_1)$ and $Y \sim Bin(n_2, p_2)$. We then have

   $$E(X) = n_1 p_1, \; E(Y) = n_2 p_2, \; \text{Var}(X) = n_1 p_1 (1 - p_1) \text{ and } \text{Var}(Y) = n_2 p_2 (1 - p_2).$$

   (a.) An estimator $\hat{\theta}$ is unbiased for the parameter θ, if $E(\hat{\theta}) = \theta$.

   In our case, we have

   $$E\left(\frac{X}{n_1} - \frac{Y}{n_2}\right) = \frac{E(X)}{n_1} - \frac{E(Y)}{n_2} = \frac{n_1 p_1}{n_1} - \frac{n_2 p_2}{n_2} = p_1 - p_2.$$

   Hence $\frac{X}{n_1} - \frac{Y}{n_2}$ is an unbiased estimator of $p_1 - p_2$.

   (b.) Since $\frac{X}{n_1} - \frac{Y}{n_2}$ is an unbiased estimator, i.e., bias=0, the mean squared error of the estimator is given by

   $$\begin{aligned} \text{MSE} &= \text{Var}\left(\frac{X}{n_1} - \frac{Y}{n_2}\right) \overset{\text{(indep)}}{=} \text{Var}\left(\frac{X}{n_1}\right) + \text{Var}\left(\frac{Y}{n_2}\right) \\ &= \frac{1}{n_1^2}\text{Var}(X) + \frac{1}{n_2^2}\text{Var}(Y) = \frac{1}{n_1^2} n_1 p_1 (1 - p_1) + \frac{1}{n_2^2} n_2 p_2 (1 - p_2) \\ &= \frac{p_1 (1 - p_1)}{n_1} + \frac{p_2 (1 - p_2)}{n_2}. \end{aligned}$$

6. Clearly this is a problem for calculating a confidence interval for the population mean μ. Furthermore it is given/assumed that the population (melting point of hydrogenated oil) is normally distributed. Since nothing is mentioned about standard deviation of the population we assume that it is unknown.

   Subsequently, we formulate the following model for the given data: Let X_i be the melting point of the i-th sample from the new brand, $i = 1, 2, \ldots, n(= 16)$. Then $X_1, \ldots X_{16}$ are i.i.d. $N(\mu, \sigma^2)$, where both μ and σ are unknown.

6. (a.) From theory we know that for the model mentioned above, the 95% confidence interval for μ is given by $\bar{X} \pm t_{0.025;n-1} \frac{S}{\sqrt{n}}$, where $\bar{X}$ is the sample mean and S is the sample standard deviation. For the given data observed values are: $\bar{x} = 94.32$ and $s = 1.275$. Furthermore $t_{0.025;15} = 2.1315$. Then the required 95% confidence interval becomes

$$\left(94.32 \pm 2.1315 \times \frac{1.275}{4}\right) = (93.64, 95.00).$$

(b.) If a 97% confidence interval is calculated based on the same data, the interval would be wider than the 95% confidence interval. This is because if one wants a higher confidence (probability) from the same set of observation (without adding any more data) one must make the interval larger. In the formula, this will be reflected in $t_{0.015;15} > t_{0.025;15}$.

If, on the other hand, a 97% confidence interval is calculated based on a different set of data of 16 samples, then though I would expect a wider interval than in (a.), it cannot be guaranteed. This is because, not only that the tabular values will be changed, but also the obtained sample mean $\bar{x}$ and sample standard deviation s will be different.

7. [This is a hypothesis testing problem related to the population proportion of (young) university students who uses the Tinder app. In the study, a sample of 200 university students are used.]

(a.) Let X be the number of students (out of the 200 surveyed) who uses the app. Then $X \sim \text{Binomial}(n, p)$, where $n = 200$ and p is unknown. We want to know whether p is more than 1/2. So, we follow the steps of a statistical hypothesis testing.

1. Model : $X \sim \text{Binomial}(200, p)$, $\quad p$ unknown.
2. Hypotheses : $H_0 : p = 0.5 \quad$ Vs. $\quad H_1 : p > 0.5 \qquad$ (with $\alpha = 0.05$.)
3. Test statistic : $Z = \dfrac{X - 200 \times 0.5}{\sqrt{200 \times 0.5 \times (1 - 0.5)}} = \dfrac{X - 100}{\sqrt{50}}$.
4. Under H_0, Z is approximately $N(0, 1)$-distributed $\quad$ (since $n = 200$ is large).
5. Decision rule : Reject H_0 if $Z \geq z_\alpha = z_{0.05} = 1.645$.
6. $Z_{\text{obs}} = \frac{113-100}{\sqrt{50}} = 1.84$
7. Since $Z_{\text{obs}} = 1.84 > 1.645$, we reject H_0.
8. At 5% level of significance the given data shows conclusive evidence that the more than half of the university students use the Tinder app regularly.

(b.) P-value for the test in (a.) is given by

$$P(Z \geq Z_{\text{obs}}) = P(Z \geq 1.84) = 1 - \Phi(1.84) = 1 - 0.9671 = 0.0329.$$

The P-value $0.0329 < 0.05 = \alpha$ indicates that H_0 should be rejected, as was the case in part (a.). However, if one uses a significance level $\alpha^* = 0.01$ then one will fail to reject H_0, because the P-value $0.0329 > 0.01 = \alpha^*$.