Solutions to Final, 29/10/18, Statistics & Probability (191506103)

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

1. [Exponential r.v. is continuous and the transformed r.v. $\sqrt{X}$ 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 X is an exponential(1) r..v. with pdf given by

$$f_X(x) = e^{-x}, \quad x > 0.$$

   Denoting the cdf of X by $F_X(x)$, we have $F_X(x) = P(X \le x)$ and $F_X'(x) = f_X(x)$.

   We want to find the pdf of $Y := \sqrt{X}$. Note that Y takes values in $(0, \infty)$.

   [Looking at the graph of $\sqrt{x}$ against the x-axis will confirm this.]

   Thus, $\quad F_Y(y) := P(Y \le y) = 0, \quad$ if $y \le 0$.

   For $y > 0$,

$$F_Y(y) \;=\; P(Y \le y) \;=\; P\left(\sqrt{X} \le y\right) \;=\; P\left(X \le y^2\right) \;=\; F_X\left(y^2\right).$$

   [Differentiating $F_Y(y)$ w.r.t. y (and using the chain rule of differentiation in the process) we obtain the pdf of Y.]

   Thus, the probability density function of Y is given by

$$\begin{aligned} f_Y(y) \;&=\; F_Y'(y) \;=\; \frac{d}{dy} F_X\left(y^2\right) \;=\; F_X'\left(y^2\right) \cdot \frac{d}{dy}\left(y^2\right) \;=\; f_X\left(y^2\right) \cdot 2y \\ &=\; e^{-y^2} \cdot 2y \;=\; 2y\, e^{-y^2}, \quad y > 0. \end{aligned}$$

2. Let X be the proportion (of total capacity) stocked at the beginning of a week and Y be the proportion (of total capacity) sold during the week.
   We know that the joint density function is given by: $\quad f(x, y) = 3x, \quad 0 \le y \le x \le 1.$

   (a.) We are to find the (marginal) probability density function of X.

   [So, we should provide the range of *all possible* values that the (stand-alone) r.v. X may take. It should NOT depend on y. When we are talking only about X), presence of y is inappropriate and does not make sense.]

   Note that if we consider the random variable X by itself, the lowest possible value is 0 and the highest is 1 (in combination with different values of Y). [A rough plot the region where $0 \le y \le x \le 1$ will confirm this.] For a fixed value of $x \in [0, 1]$, we can then calculate the pdf as [noting that for fixed x, the joint pdf $f(x, y)$ is positive when $0 \le y \le x$]

$$f_X(x) \;=\; \int_{-\infty}^{\infty} f(x, y)\, dy \;=\; \int_0^x 3x\, dy \;=\; 3x \int_0^x 1\, dy \;=\; 3x \Big[y\Big]_{y=0}^{y=x} \;=\; 3x^2.$$

   Thus, the marginal density of X is given by

$$f_X(x) \;=\; 3x^2, \quad 0 \le x \le 1.$$

2. (b.) The required probability is $P(Y > \frac{1}{2} | X = \frac{3}{4})$. For this, we need the conditional density $f_{Y|X=\frac{3}{4}}(y)$. [It is a pdf about Y. So, the range of values of Y must be specified. Here, the range MAY depend on the KNOWN value (x) of the r.v. X.]

In genetral, the conditional distribution of Y given $X = x$ is given by

$$f_{Y|X=x}(y) = \frac{f(x,y)}{f_X(x)} = \frac{3x}{3x^2} = \frac{1}{x}, \qquad 0 \le y \le x.$$

Hence, $f_{Y|X=\frac{3}{4}}(y) = 4/3, \quad 0 \le y \le \frac{3}{4}$ and

$$P\left(Y > \frac{1}{2} \,\middle|\, X = \frac{3}{4}\right) = \int_{\frac{1}{2}}^{\infty} f_{Y|X=\frac{3}{4}}(y)\, dy = \int_{\frac{1}{2}}^{\frac{3}{4}} \frac{4}{3}\, dy = \frac{4}{3} \cdot \left(\frac{3}{4} - \frac{1}{2}\right) = \frac{1}{3}.$$

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

$$E(Y|X = x) = \int_{-\infty}^{\infty} y\, f_{Y\,|\,X=x}(y)\, dy = \int_0^x y\, \frac{1}{x}\, dy = \frac{1}{x}\, \frac{y^2}{2}\Big|_{y=0}^{y=x} = \frac{x}{2}.$$

Hence $E(Y|X) = \frac{X}{2}$.

3. (a.) The population mean (or equivalently, the expected value for the population model) is given by

$$E(X) = \int x f(x) dx = \int_{-1}^{1} x \cdot \frac{1}{2}(1 + \sqrt{\theta}\, x) dx = \left[\frac{1}{2} \cdot \frac{x^2}{2} + \frac{\sqrt{\theta}}{2} \cdot \frac{x^3}{3}\right]_{x=-1}^{x=1}$$

$$= \left[\left(\frac{1}{4} + \frac{\sqrt{\theta}}{6}\right) - \left(\frac{1}{4} - \frac{\sqrt{\theta}}{6}\right)\right] = \frac{\sqrt{\theta}}{3}.$$

(b.) In the model, there is only one unknown parameter (θ). [So, to find the MOM estimator, we need to set up only one equation, namely: population first moment = sample first moment.]

Now, by setting up the equation and solving for θ, we obtain:

$$\frac{\sqrt{\theta}}{3} = \bar{x} \quad \Rightarrow \quad \sqrt{\theta} = 3\bar{x} \quad \Rightarrow \quad \theta = 9\bar{x}^2.$$

Thus, the MOM estimator of θ is given by $\hat{\theta} = 9\bar{X}^2$.

(c.) Estimator $\hat{\theta}$ is unbiased for parameter θ if $E(\hat{\theta}) = \theta$. Thus, we need to check whether $E(9\,\bar{X}^2) = \theta$. [To check this we must calculate $E(\bar{X}^2)$. We do not know this directly. But, it is related to the variance and expectation of $\bar{X}$, which are known.]

We recall further that if $X_1, X_2, \ldots, X_n$ are i.i.d. with population mean $E(X_i) = \mu$ and population variance $\sigma^2 = \mathrm{Var}(X_i)$, then $E(\bar{X}) = \mu$ and $\mathrm{Var}(\bar{X}) = \sigma^2/n$.
In our case, $\mu = \sqrt{\theta}/3$ (from part (a.)) and it is given that $\sigma^2 = (3-\theta)/9$. Then

$$E(\hat{\theta}) = E(9\,\bar{X}^2) = 9\,E(\bar{X}^2) = 9\left[\mathrm{Var}(\bar{X}) + \big(E(\bar{X})\big)^2\right] = 9\left[\frac{\sigma^2}{n} + \mu^2\right]$$

$$= 9\left[\frac{3-\theta}{9n} + \left(\frac{\sqrt{\theta}}{3}\right)^2\right] = \frac{3}{n} - \frac{\theta}{n} + \theta = \frac{3}{n} + \left(1 - \frac{1}{n}\right)\theta \neq \theta.$$

Thus, the MOM estimator is not unbiased for θ.

4. Let X_i be the (rounding-off) error entailed in the recorded gain of the i^{th} portfolio, where $i = 1, 2, \ldots, 48$. It is given that X_i's are i. i. d. and $X_i \sim \text{Unif}(-\frac{1}{2}, \frac{1}{2})$. [The "unit" of errors are the same as that of the recorded gains, for example, in "euros" or "thousand euros", etc.]

Denoting by Y, the (total) error in the calculated total gain, we have $Y = \sum_{i=1}^{48} X_i$. We are to find the probability that total error is within 2 unit, i.e., $P(|Y| \leq 2)$.
[To calculate this probability we need the probability distribution of Y. The exact distribution can be calcualated by doing "convolution" repeatedly. But that is time consuming. However, since Y is the sum of relatively large (≥ 30) number of i. i. d. random variables, we can approximate the distribution by using the Central Limit Theorem.]

Note that

$$\begin{aligned} E(Y) &= E\left(\sum_{i=1}^{48} X_i\right) = 48\,E(X_i) = 48 \cdot \frac{1}{2}\left(\frac{1}{2} - \frac{1}{2}\right) = 0 \quad \text{and} \\ \text{Var}(Y) &= \text{Var}\left(\sum_{i=1}^{48} X_i\right) = 48\,\text{Var}(X_i) = 48 \cdot \frac{1}{12}\left(\frac{1}{2} + \frac{1}{2}\right)^2 = 4. \end{aligned}$$

Since Y is the sum of 48 (≥ 30) i. i. d. random variables, we can use the CLT to approximate the required probability as:

$$\begin{aligned} P(|Y| \leq 2) &= P\left(\left|\frac{Y - E(Y)}{\sqrt{\text{Var}(Y)}}\right| \leq \frac{2-0}{\sqrt{4}}\right) \\ &\approx P(|Z| \leq 1), \quad \text{where } Z \sim N(0,1) \qquad \text{(CLT)} \\ &= \Phi(1) - \Phi(-1) = 0.8413 - 0.1587 = 0.6826. \qquad \text{(from table)} \end{aligned}$$

5. (a.) Suppose $X \sim \text{Exponential}(\theta)$. Let us define $Y = 2\theta X$. To show: $Y \sim \chi^2_{(2)}$.

We know that the mgf of X is given by $m_X(t) = E(e^{tX}) = \frac{\theta}{\theta - t}$ (for $t<\theta$). Then,

$$m_Y(t) = E\left(e^{tY}\right) = E\left(e^{t\,(2\theta X)}\right) = E\left(e^{(2\theta t)X}\right) = m_X(2\theta t) = \frac{\theta}{\theta - 2\theta t} = \frac{1}{1-2t}.$$

Since $\chi^2_{(n)}$ r.v. has the mgf $(1-2t)^{-n/2}$ and $m_Y(t) = (1-2t)^{-1} = (1-2t)^{-2/2}$ from the uniqueness of mgf, it follows that $2\theta X \equiv Y \sim \chi^2_{(2)}$.

(b.) Define $Y_i = 2\theta X_i$. Since X_i's are independent, so are Y_i's. Furthermore, from (a.), we have that $m_{Y_i}(t) = (1-2t)^{-1}$ for each i. Note, also, that

$$2n\theta\bar{X} = 2n\theta\,\frac{\sum_{i=1}^n X_i}{n} = 2\theta\sum_{i=1}^n X_i = \sum_{i=1}^n (2\theta X_i) = \sum_{i=1}^n Y_i.$$

Then from the independence of Y_i's it follows that

$$m_{2n\theta\bar{X}}(t) = m_{\sum_{i=1}^n Y_i}(t) = \prod_{i=1}^n m_{Y_i}(t) = \prod_{i=1}^n (1-2t)^{-1} = (1-2t)^{-n} = (1-2t)^{-(2n)/2}.$$

Once again, from the uniqueness property of the mgf we conculde that $2n\theta\bar{X} \sim \chi^2_{(2n)}$.

5. (c.) From (b.), we have that $2n\theta\bar{X} \sim \chi^2_{(2n)}$. Denoting by $\chi^2_{\alpha;m}$ the value below which lies α probability, under the density-curve of a $\chi^2_{(m)}$ r.v., we have

$$P\left(\chi^2_{\frac{\alpha}{2};2n} \le 2n\theta\bar{X} \le \chi^2_{1-\frac{\alpha}{2};2n}\right) = 1-\alpha \quad \Leftrightarrow \quad P\left(\frac{\chi^2_{\frac{\alpha}{2};2n}}{2n\bar{X}} \le \theta \le \frac{\chi^2_{1-\frac{\alpha}{2};2n}}{2n\bar{X}}\right) = 1-\alpha.$$

Hence the formula for $100(1-\alpha)\%$ confidence interval for θ is

$$\left[\frac{\chi^2_{\frac{\alpha}{2};2n}}{2n\bar{X}}\ ,\ \frac{\chi^2_{1-\frac{\alpha}{2};2n}}{2n\bar{X}}\right].$$

6. [Clearly, this problem is about performing a hypothesis test regarding a population mean μ. It is given that the population (the return of the stocks from which the analyst makes his choices) is normally distributed. However, nothing is mentioned about the standard deviation of the population. So, we assume that it is unknown. In any case, we should start by describing the model about the data, where the data are the returns from the 24 stocks, i.e., there are 24 data points.]

(a.) Let X_i be the return of the i^{th} stock, where $i = 1, 2, \ldots 24$. We have to test whether the (true) expected value of X_i beats the benchmark value, i.e., it is higher than 10.10%. The (sample) average of the returns from the 24 stocks (in the sample) is $\bar{x} = 11.50\%$ and the standard deviation of the returns is $s = 10.17\%$.
We follow the steps of a statistical hypothesis testing.

1. Model : $X_1, X_2, \ldots X_{24}$ are i.i.d. $N(\mu, \sigma^2)$, with both μ and σ^2 unknown.
2. Hypotheses : $H_0 : \mu = 0.1010$ Vs. $H_1 : \mu > 0.1010$.
3. Test statistic : $T = \dfrac{\bar{X} - 0.1010}{S/\sqrt{24}}$, where $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is the sample variance.
4. Under H_0, T is distributed as $t_{(23)}$.
5. Decision rule : Reject H_0 if $T \ge t_{\alpha;23} = 1.7139$, with $\alpha = 0.05$.
6. $T_{\text{obs}} = \frac{\bar{x}-0.1010}{s/\sqrt{24}} = \frac{0.1150-0.1010}{0.1017/\sqrt{24}} = 0.6744$.
7. Since $T_{\text{obs}} = 0.6744 < 1.7139$, we do not reject H_0.
8. At 5% level of significance the given data do not show conclusive evidence for the claim that the analyst can choose stocks which has higher expected return than the benchmark value of 10.10%.

(b.) In the model we have assumed that the standard deviation of the returns (of the stocks that the analyst selects) are unknown. Subsequently, we used the sample standard deviation in our analysis. Since we are going to compare the analyst's performance with the benchmark values and the benchmark standard deviation is known (15.67%), we could have made the model assumption that population standard deviation is known to be $\sigma = 0.1567$.
With this model, the formula for the test statistic would use the known value of σ (instead of the estimate s): $Z := \frac{\bar{X}-0.1010}{0.1567/\sqrt{24}}$.
Subsequently, the decision rule would use the $N(0,1)$-table and not the t-table: *"Reject H_0, if observed test statistic $Z \ge z_{0.05} = 1.645$."*