Solutions to Final, 30/10/17, 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. $1/(X+1)$ 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 $F'_X(x) = f_X(x)$.

   Want to find the pdf of $Y := \frac{1}{X+1}$. Note that Y takes values in $(0, 1)$.

   [Draw a graph to see that for small values of x, $1/(x+1)$ is close to 1, and for large values it is close to 0]

   Thus, $\quad F_Y(y) := P(Y \le y) = 0, \quad$ if $y \le 0, \quad$ and $\quad F_Y(y) = 1, \quad$ if $y \ge 1$

   For $0 < y < 1$,

$$\begin{aligned} F_Y(y) &= P(Y \le y) = P\left(\frac{1}{X+1} \le y\right) = P\left(X + 1 \ge \frac{1}{y}\right) \\ &= P\left(X \ge \frac{1}{y} - 1\right) = 1 - F_X\left(\frac{1}{y} - 1\right). \end{aligned}$$

   [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

$$\begin{aligned} f_Y(y) &= F'_Y(y) = -F'_X\left(\frac{1}{y} - 1\right) \cdot \frac{d}{dy}\left(\frac{1}{y} - 1\right) = -f_X\left(\frac{1}{y} - 1\right) \cdot \left(-\frac{1}{y^2}\right) \\ &= e^{-\left(\frac{1}{y} - 1\right)} \cdot \frac{1}{y^2} = \frac{1}{y^2} e^{1 - \frac{1}{y}}, \quad 0 < y < 1. \end{aligned}$$

2. We are given the joint pdf: $\quad f(x, y) = e^{-x}, \quad$ for $0 \le y \le x$.

   (a.) We want to find the marginal pdf of X. [So, we should provide the range of values that X could possibly take and also, it should NOT depend on y.]

   Note that if we consider the random variable X by itself, it may take any positive value (in combination with different Y values). The (marginal) pdf is then given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy = \int_0^x e^{-x}\, dy = e^{-x}\, y\Big|_{y=0}^{y=x} = x\, e^{-x}, \qquad x \ge 0.$$

   (b.) The required (conditional) pdf of Y conditional on $X = x$ (> 0) is given by

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

   [It is a density about Y; so, its range of values must be specified. However, the range MAY now depend on the particular value (x) of X, because given the value of X, the values of Y may be restricted.]

(c.) To calculate $E(Y \mid X)$, we need to calculate $E(Y \mid X = x)$ first.

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

Hence, $E(Y \mid X) = \frac{1}{2}X$.

3. Let us define X to be the number of houses the salesman needs to visit to make his first sale. From the given assumptions, it follows that X is a geometric r.v. with probability of success 0.1, where success means making a sale.

(a.) Note that the salesman will visit at least 25 houses before he stops working if and only if he does not make any sale in the <u>first</u> 24 house-visits. It does not matter how many more houses he needs to visit to make the first sale. Since making a sale on different house-visits are independent, the probability that on a given day the salesman will visit at least 25 houses before he stops working is given by

$$P(X \geq 25) = (1 - 0.1)^{24} = 0.9^{24} = 0.0798.$$

Alternatively, (since $X \sim \text{Geometric}(p = 0.1)$)

$$\begin{aligned} P(X \geq 25) \;&=\; \sum_{x=25}^{\infty} P(X = x) \;=\; \sum_{x=25}^{\infty} (1 - 0.1)^{x-1} \times 0.1 \;=\; 0.1 \sum_{x=25}^{\infty} (0.9)^{x-1} \\ &=\; 0.1 \sum_{x=24}^{\infty} (0.9)^{x} \;=\; 0.1\, \frac{0.9^{24}}{1 - 0.9} \;=\; 0.9^{24} \;=\; 0.0798. \end{aligned}$$

(b.) Let W be the number of days in a year (i.e., out of 220 days) that he visits at least 25 houses for first sale. Note that each day can be considered as a Bernoulli experiment with success being at least 25 house-visits before the salesman stops working on that day. It then follows that W is a Binomial(n, p) r.v. with $n = 220$ and the probability of success $p = 0.0798$ from part (a.) The required probability is $P(W \leq 25)$.

To calculate the probability we can use normal approximation to binomial distribution, due to the CLT / De Moivre's Law, because $n = 220$ is reasonably large. Note that here $np \pm 3\sqrt{np(1-p)} = (5.498, 29.614) \subset [0, 220]$.

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

The required probability is then

$$\begin{aligned} P(W \leq 25) \;&=\; P(W \leq 25.5) \qquad\qquad \text{(with continuity correction)} \\ &=\; P\left(\frac{W - np}{\sqrt{np(1-p)}} \leq \frac{25.5 - 220 \times 0.0798}{\sqrt{220 \times 0.0798 \times 0.9202}} \right) \\ &\approx\; P(Z \leq 1.98), \quad \text{where } Z \sim N(0, 1) \qquad \text{(CLT)} \\ &=\; \Phi(1.98) \;=\; 0.9762. \qquad \text{(from table)} \end{aligned}$$

4. Let $x_1,\ x_2,\ \ldots,\ x_n$ be a random sample drawn from a geometric population with parameter p $(0 < p < 1)$. To find the maximum likelihood estimator of p, we first obtain the likelihood of the data $\{x_1, \ldots, x_n\}$ as a function of p. This is given by

$$L(p; x_1, \ldots, x_n) = \prod_{i=1}^{n} P(X_i = x_i) = \prod_{i=1}^{n} \left((1-p)^{x_i - 1}\, p\right) = p^n\, (1-p)^{\left(\sum_{i=1}^{n} x_i\right) - n}.$$

Then the log-likelihood is given by

$$l(p; x_1, \ldots, x_n) = \ln(L(p; x_1, \ldots, x_n)) = n \ln(p) + \left(\sum_{i=1}^{n} x_i \ - \ n\right) \ln(1-p).$$

By taking derivative of $l(p)$ with respect to p and setting it equal to zero we get

$$\frac{n}{p} + \frac{\sum_{i=1}^{n} x_i \ - \ n}{1-p}\,(-1) = 0 \qquad \Rightarrow \qquad \frac{n}{p} = \frac{\sum_{i=1}^{n} x_i \ - \ n}{1-p}$$

$$\Rightarrow \qquad n(1-p) = p\left(\sum_{i=1}^{n} x_i \ - \ n\right) \qquad \Rightarrow \qquad p = \frac{n}{\sum_{i=1}^{n} x_i} = 1/\overline{x}$$

[Must also check 2nd derivative]

Note further that $\quad l''(p) \ = \ -\dfrac{n}{p^2} - \dfrac{\sum_{i=1}^{n} x_i \ - \ n}{(1-p)^2} \ < \ 0, \quad$ since $\quad \displaystyle\sum_{i=1}^{n} x_i \ \geq \ n \quad$ and $n,\ p^2,\ (1-p)^2$ are all positive.

Hence the above solution is indeed a maximizer of the log-likelihood function.

Hence the MLE for p is given by $\quad \hat{p} = \dfrac{1}{\overline{X}}$.

5. (a.) Let $X \sim$ Binomial(m, p), $Y \sim$ Binomial(n, p) and $W = X + Y$. We want to calculate the moment generating function (mgf) of the r.v. W and see if it is of some known form. We can then use the uniqueness property of mgf to identify the probability distribution of W.

Note that the mgf's of X and Y are

$$m_X(t) = (1 - p + pe^t)^m \quad \text{and} \quad m_Y(t) = (1 - p + pe^t)^n.$$

The mgf of W is then given by

$$\begin{aligned} m_W(t) \ &= \ m_{X+Y}(t) \ = \ m_X(t) \cdot m_Y(t), \qquad \text{(using independence of } X \text{ and } Y\text{)} \\ &= \ (1 - p + pe^t)^m \cdot (1 - p + pe^t)^n \ = \ (1 - p + pe^t)^{m+n} \\ &= \ (1 - p + pe^t)^{n^*}, \qquad \text{where } n^* = m + n. \end{aligned}$$

This mgf is the same as that of a Binomial r.v. with parameters n^* and p. Hence, using the uniqueness property of mgf we can conclude that

$$W := X + Y \ \sim \ \text{Binomial}(n^* = m + n, p).$$

(b.) Note that $m_V(t) = \left[\frac{1}{4}(3+e^t)\right]^5 = \left(\frac{3}{4}+\frac{1}{4}e^t\right)^5 = \left(1-\frac{1}{4}+\frac{1}{4}e^t\right)^5$, which is same as the mgf of a Binomial(m,p) r.v. with $m=5$ and $p=1/4$. Hence, from the uniqueness of mgf we conclude that $V \sim$ Binomial$(5, 0.25)$. Then

$$\begin{aligned} P(V<2) &= P(V\le 1) = P(V=0)+P(V+1) \\ &= (0.75)^5 + 5\times 0.25\times(0.75)^4 = 0.6328. \end{aligned}$$

6. [Clearly, this problem is about calculating a confidence interval for a population mean μ. It is given that the population (the 10-year annualized return for the stocks selected by the analyst) 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 (10-year annualized) returns from the 24 stocks, i.e., there are 24 data points.]

(a.) Let X_i be the (10-year annualized) return of the i-th stock, $i = 1, 2, \ldots n(=24)$. We assume that X_i's are i.i.d. $N(\mu, \sigma^2)$ with both μ and σ^2 unknown. For this model, normal population with unknown population variance, we know that 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: $\bar{x} = 0.1150$ and $s = 0.1017$. Furthermore, we have from the table, $t_{0.025;23} = 2.0687$. Thus, the required 95% confidence interval becomes

$$\left(0.1150 \pm 2.0687 \times \frac{0.1017}{\sqrt{24}}\right) = (0.0721\,,\, 0.1579).$$

(b.) If σ^2 was known, then in the formula for the confidence interval in part (a.), we would have used the known value of σ, instead of the estimate s, the sample standard deviation. Subsequently, the required tabular value would have come from the standard normal table and not from the t-table. The final formula would have then become: $\bar{X} \pm z_{0.025}\,\frac{\sigma}{\sqrt{n}}$.

Also, in this case, one would expect to obtain a narrower interval, but whether that will actually happen cannot be said.

7. Let X be the number of skin improvements (among the 33 studied women). By considering the studying of one woman as a Bernoulli trial and calling an improved skin to be a success, we can conclude that $X \sim$ Binomial$(33\,,\,p)$, where p, the true probability of the cream improving the skin, is unknown.

(a.) We want to know whether p is larger than 0.6. So we follow the steps of a statistical hypothesis testing.

1. Model : Let X be as defined above. $X \sim$ Binomial$(33, p)$, $\quad p$ unknown.
2. Hypotheses : $H_0 : p = 0.6$ $\quad$ Vs. $\quad H_1 : p > 0.6$
3. Test statistic : $Z = \frac{X-33\times 0.6}{\sqrt{33\times 0.6\times 0.4}}$

4. Under H_0, Z is approximately $N(0,1)$ (we assume here $n = 33 > 30$ to be large).
5. Decision rule : Reject H_0 if $Z \geq z_\alpha = 1.645$, with $\alpha = 0.05$.
6. $Z_{\text{obs}} = \frac{24-19.8}{\sqrt{7.92}} = 1.49$
7. Since $Z_{\text{obs}} = 1.49 < 1.645$, we do not reject H_0
8. At 5% level of significance the given data does not show conclusive evidence that the cream will improve the skin of more than 60% of middle-aged women.

(b.) P-value of this test is given by

$$P_{H_0}(Z \geq Z_{obs}) = P_{H_0}(Z \geq 1.49) = 1 - \Phi(1.49) = 0.0681.$$

Interpretation: If one wants to test the hypotheses with significance level $\alpha \geq 0.0681$ then H_0 will be rejected, otherwise, it will not be rejected.