STA 250 (Fall 13): Midterm III

Total time: 1hr 10min

The **three** questions below carry a total of 30 points. Attempt all question parts and show work to guarantee partial/full credit. Make use of the tables and basic probability facts attached at the end. **No other cheat sheets allowed**. You will be provided with sheets of white paper to write your answers on. Please remember to staple them before turning in and write your name on the top. You should be able to attempt any part of any question whether or not you have attempted the parts before it.

Tip for quick table look-up: $\boxed{z_k(2\alpha) = c \iff \alpha = 1 - \Phi_k(c)}$.

1. In the NSW study, out of $n = 185$ subjects receiving job training $X = 120$ reported a substantial increase in income (defined as their 1978 annual income being at least 25% more than their 1975 annual income). From the control arm of the study, it was determined that there is a 56% chance of a substantial income increase for people with similar socio-economic background who did not receive any training. Consider the model: $X \sim \mathsf{Binomial}(n, p), p \in (0, 1)$ and answer the following. [$4 + 2 + 4 = 10$ points]

   (a) Report a 95% ML confidence interval for p.

   (b) If 0.56 is below the left boundary of the above interval, what can you say about the p-value for testing $H_0 : p \le 0.56$ against $H_1 : p > 0.56$ based on ML tests?

   (c) Consider a formal Bayes testing of $H_0 : p = 0.56$ (treatment has no additional effect) against $H_1 : p = 0.65$ (treatment gives a 50% boost to the odds of substantial increase). From binomial pmf formula,

   $$\frac{P(X = 120|n = 185, p = 0.65)}{P(X = 120|n = 185, p = 0.54)} = 20.3$$

   Which decision will you take under the loss:

		Truth	
		$p = 0.56$	$p = 0.65$
Decision	No effect	0	1
	50% boost	25	0

   and prior: $P(H_0) = 1/2 = P(H_1)$?

2. In a high energy physics experiment, 10 energy channels are searched for a signal that may or may not exist. If the signal exists, then it must show up in exactly one of the 10 channels. Let $X_1, \ldots, X_{10}$ denote the channel measurements. If channel j has the signal then $X_j \sim \mathsf{Normal}(3, 1)$, otherwise $X_j \sim \mathsf{Normal}(0, 1)$. Consider testing

   $$H_0 : \text{the signal does not exist} \quad vs. \quad H_1 : \text{the signal exists}$$

   based on the channel measurements data and answer the following. [$4 + 4 = 8$ points]

(a) For each channel j consider the pair of hypotheses

$$H_{0j}: \text{ channel } j \text{ does not have the signal} \quad vs. \quad H_{1j}: \text{ channel } j \text{ has the signal.}$$

For this pair, any ML test rejects for large values of X_j and the ML tests based p-value is precisely $1 - \Phi(X_j)$ where Φ is the standard normal CDF. We could combine these ML tests for all channels to test the overall H_0 stated earlier. For carrying out this multiple testing, what would be the most appropriate error rate to control for: the "family-wise error rate" or the "false discovery rate"? Justify your answer.

(b) Whichever error rate you pick in part (a), control it at 10% level and decide whether to reject the overall H_0 when the recorded p-values from the 10 channels are:

```
0.30 0.91 0.32 0.72 0.92 0.58 0.013 0.54 0.56 0.003
```

3. In an online survey of $n = 60$ students randomly chosen from all students registered in all statistics courses in Fall 2013, $X_1 = 25$ reported satisfaction with homework assignments, $X_2 = 23$ reported dissatisfaction and $X_3 = 12$ did not respond. Assume there are only two types of students: 'satisfied' and 'dissatisfied' with regards to homework assignments. Also assume truthful reporting. Consider the model $X \sim \mathsf{Multinomial}(n, p), p \in \Delta_3$. We are interested in testing the null hypothesis

H_0 : a third of all students are dissatisfied with homework assignments.

Answer the following [$3 + 4 + 2 + 3 = 12$ points]

(a) Assume that chance of responding is NOT related to satisfaction level. Show that H_0 corresponds to the null subset

$$\Delta_3^0 = \{(\tfrac{2}{3}a, \tfrac{1}{3}a, 1 - a) : a \in (0, 1)\},$$

and give an interpretation of the free parameter 'a'.

(b) Argue (with logic and precise calculations) that the expected cell counts under the null is $(32, 16, 12)$. [Hint: For any two non-negative numbers r and s, the function $g(a) = a^r(1 - a)^s$ over $a \in (0, 1)$ is maximized at $\hat{a} = \frac{r}{r+s}$].

(c) From (b), the Pearson's chi-square test statistic value is 4.59 and the p-value $=$ 0.03. So in carrying out a 5% level testing, you will reject H_0. What would you conclude about the actual proportion of dissatisfied students? Does it appear larger or smaller than a third?

(d) What would have been the expected counts under H_0 and the p-value had we assumed 'dissatisfied' students were two times as likely to respond to the survey than 'satisfied' students? Select an option from below and give a short justification.

(i) $(32, 16, 12)$ and p-value $= 0.03$ (ii) $(16, 32, 12)$ and p-value < 0.03
(iii) $(24, 24, 12)$ and p-value > 0.03 (iv) $(40, 8, 12)$ and p-value < 0.03.