Categorial Data Analysis
Solution #2

1. Just how accurate are the weather forecasts we hear every day? The following table compares the daily forecast with a city's actual weather during 10 days:

		Actual Weather		
		Rain	No rain	Total
Forecast	Rain	3	1	4
	No rain	2	4	6
	Total	5	5	10

To see an association between forecast and actual weather, we assume n_{11} ((1,1) cell count) follow hypergeometric distribution and will conduct Fisher's exact test because cell counts are too small.

(a) Note that possible values for n_{11} are 0, 1, 2, 3, 4 (do you know why?). Provide the probability $P(0)$, $P(1)$, $P(2)$, $P(3)$, and $P(4)$ using hypergeometric distribution.
Using Hypergeometric distribution, we are able to get

$$\begin{aligned}
P(0) &= \frac{\binom{5}{0}\binom{5}{4}}{\binom{10}{4}} = \frac{5}{210} = 0.0238 \\
P(1) &= \frac{\binom{5}{1}\binom{5}{3}}{\binom{10}{4}} = \frac{50}{210} = 0.2381 \\
P(2) &= \frac{\binom{5}{2}\binom{5}{2}}{\binom{10}{4}} = \frac{100}{210} = 0.4762 \\
P(3) &= \frac{\binom{5}{3}\binom{5}{1}}{\binom{10}{4}} = \frac{50}{210} = 0.2381 \\
P(4) &= \frac{\binom{5}{4}\binom{5}{0}}{\binom{10}{4}} = \frac{5}{210} = 0.0238
\end{aligned}$$

(b) Conduct Fisher's exact test whether the forecast is good or not good. (This means you should conduct one-sided test.)
Note that the forecast is better as n_{11} increases. (This means that the larger value of n_{11} supports the alternative hypothesis.) Therefore

$$p\text{-value} = P(n_{11} \geq 3) = P(3) + P(4) = 0.2619.$$

p-value is larger than .05 so we fail to reject, under the significant level $\alpha = 0.05$, the null hypothesis that the forecast is not good.

(c) conduct Fisher's exact test whether the forecast is just a random guessing. (This means you should conduct two-sided test.) Note that the forecast is far from the random guessing as n_{11} is far from $\frac{4}{2} = 2$, which is the expected number of days correctly forecasted under the random guessing. (This means that the larger difference of n_{11} from 2 supports the alternative hypothesis.) Therefore

$$p\text{-value} = P(|n_{11} - 2| \geq |3 - 2|) = P(0) + P(1) + P(3) + P(4) = 0.5238.$$

p-value is larger than .05 so we fail to reject, under the significant level $\alpha = 0.05$, the null hypothesis that the forecast is simply the random guessing.

2. The Centers for Disease Control and Prevention (CDC) has estimated that 19.8% of Americans over 15 years old are obese. The CDC conducts a survey on obesity and various behaviors. Here is a table on self-reported exercise classified by body mass index.

	Normal ($v_1 = 1$)	Overweight ($v_2 = 2$)	Obese ($v_3 = 3$)	Total
Inactive ($u_1 = 1$)	24	26	36	86
Irregularly active ($u_2 = 2$)	28	29	28	85
Regular, not intense ($u_3 = 3$)	31	31	27	89
Regular, intense ($u_4 = 4$)	17	14	9	40
Total	100	100	100	300

To quantify categories, we assign the scores as $u_1 \leq u_2 \leq u_3 \leq u_4$ for each row and $v_1 \leq v_2 \leq v_3$ for each column. So u_i is increasing as exercise activity is increasing and v_j is increasing as obesity is increasing. Suppose we are interested in the linear association between activity level and obesity level under the given scores. Try compute correlation based on the scores and test on the linear association using M^2.

Use the formula on page 11 of note 4. Note that

$$\begin{aligned}\sum_{i=1}^{4}\sum_{j=1}^{3} u_i v_j n_{ij} &= (1)(1)(24) + (1)(2)(26) + (1)(3)(36) + (2)(1)(28) + (2)(2)(29) + (2)(3)(28) \\ &\quad +(3)(1)(31) + (3)(2)(31) + (3)(3)(27) + (4)(1)(17) + (4)(2)(14) + (4)(3)(9) \\ &= 1334,\end{aligned}$$

$$\sum_{i=1}^{4} u_i n_{i+} = (1)(86) + (2)(85) + (3)(89) + (4)(40) = 683,$$

$$\sum_{i=1}^{4} u_i^2 n_{i+} = (1^2)(86) + (2^2)(85) + (3^2)(89) + (4^2)(40) = 1867,$$

$$\sum_{j=1}^{3} v_j n_{+j} = (1)(100) + (2)(100) + (3)(100) = 600,$$

and

$$\sum_{j=1}^{3} v_j^2 n_{+j} = (1^2)(100) + (2^2)(100) + (3^2)(100) = 1400.$$

Therefore,

$$r = \frac{1334 - (683)(600)/(300)}{\sqrt{(1867 - (683^2)/(300))(1400 - (600^2)/(300))}} = -0.1281$$

and $M^2 = (300-1)(-0.1281)^2 = 4.9065$. Since $\chi_{0.05}(1) = 3.84$ and $M^2 > \chi_{0.05}(1)$, we reject the null hypothesis that there is no linear association between two variables.

3. Fast food is often considered unhealthy because much of it is high in both fat and sodium. But are the two related? Here are the fat and sodium contents of several brands of burgers.

Fat (g)	19	31	34	35	39	43
Sodium (mg)	920	1500	1310	860	1180	1260

(a) In order to see the linear association, compute the Pearson's correlation and test whether or not there is an association.
$r = 0.3311$ and $t = r/\sqrt{(1-r^2)/(6-2)} = 0.7018$. Since $t_{0.025}(6-2) = 2.7764$ and $|t| < t_{0.025}(4)$, we fail to reject that there is no linear association between "Fat" and "Sodium".

(b) In order to see the general association, compute the Spearman's rank correlation and test whether or not there is an association.
Mark the ranks in the decreasing order as follow: (Of course, you can rank them in the increasing order)

r	6	5	4	3	2	1
s	5	1	2	6	4	3

Thus, $r_S = -0.0857$ and $t = r_S/\sqrt{(1-r_S^2)/(6-2)} = 0.1720$. Since $t_{0.025}(6-2) = 2.7764$ and $|t| < t_{0.025}(4)$, we fail to reject that there is no association between "Fat" and "Sodium".

(c) In order to see the general association, compute the Kendall's τ and test whether or not there is an association
Note that there are 7 concordant and 8 discordant pairs. Thus, $\tau = (7-8)/((6)(5)/(2)) = -0.0667$. Since $t = \tau/\sqrt{(2)(2\cdot 6+5)/(9\cdot 6\cdot 5)} = -0.1880$ and $|t| < z_{0.025} = 1.96$, we fail to reject that there is no association between "Fat" and "Sodium".