Categorical Data Analysis
Assignment I
Solution Set

1.1

(a) Nominal
(b) Ordinal
(c) Interval
(d) Nominal
(e) Ordinal
(f) Nominal
(g) Ordinal

1.2

(a) Binomial based on $n = 100$ trials with success probability $\pi = 1/4$.

(b) $E(Y) = n\pi = (100)(1/4) = 25$

$\mathrm{Var}(Y) = n\pi(1-\pi) = (100)(1/4)(1-1/4) = 75/4$

It would be surprising if the student made at least 50 correct responses. The standard deviation of Y is 4.33; thus, 50 is over 5 standard deviations above the mean.

Using the normal approximation to the binomial distribution, we have

$$P(Y \geq 50) \approx P\left(Z \geq \frac{50-25}{4.33}\right) = P(Z \geq 5.77) \approx 0.$$

(c) Multinomial based on $n = 100$ trials with category probabilities $\pi_1 = \pi_2 = \pi_3 = \pi_4 = 1/4$.

(d) $E(N_j) = n\pi_j = (100)(1/4) = 25$ for $j = 1, \ldots, 4$

$\mathrm{Var}(N_j) = n\pi_j(1-\pi_j) = (100)(1/4)(1-1/4)$
$= 75/4$ for $j = 1, \ldots, 4$

$\mathrm{cov}(N_j, N_k) = -n\pi_j\pi_k = -(100)(1/4)(1/4)$
$= -25/4$ for $1 \le j \ne k \le 4$.

$\mathrm{corr}(N_j, N_k) = \mathrm{cov}(N_j, N_k) / (\mathrm{var}(N_j)\,\mathrm{var}(N_k))^{1/2}$
$= -1/3$ for $1 \le j \ne k \le 4$.

1.3

Let Y count the number of insects that survive in a batch of size n. Let π denote the survival probability.

If the factors to which the insects are sensitive vary from batch to batch, it may be reasonable to assume that π is a random variable, not a fixed constant. In this case, the conditional distribution of $Y|\pi$ may be binomial, yet the unconditional distribution of Y may have a variance exceeding that of a binomial. (See problem 1.12 (c).)

1.5 Score Test:

$$Z_S = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{m}}} = \frac{(842/1824) - 0.5}{\sqrt{\frac{(0.5)(1-0.5)}{(1824)}}} = -3.28$$

$$P = 2P(Z \leq -3.28) = 2(.0005) = \underline{\underline{.001}}$$

95% Confidence Interval (Based on Wald Approach):

$$\hat{\pi} \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{m}}$$

$$(842/1824) \pm (1.96)\sqrt{\frac{(842/1824)(982/1824)}{(1824)}}$$

$$(0.462) \pm (0.023)$$

$$\underline{\underline{(0.439, 0.485)}}$$

Both the p-value for the test and the CI strongly suggest that $H_0: \pi = 0.5$ is not a credible hypothesis. The CI indicates that π is less than 0.5.

1.6 (a) $2y \log\left(\frac{\hat{\pi}}{\pi_0}\right) + 2(m-y)\log\left(\frac{1-\hat{\pi}}{1-\pi_0}\right)$

$$= 2(0)\log\left(\frac{0}{0.5}\right) + 2(25-0)\log\left(\frac{1-0}{1-0.5}\right)$$

$$= 2(0\log(0))^{*} + 2(25)\log(2)$$

$$= 0 + 2(25)\log(2)$$

$$= 2(25)\log(25/12.5)$$

* Note: By convention, we treat $(0\log(0))$ as 0, since $\lim_{x\to 0} x\log x = 0$.

(b) $$S = \frac{(\hat{\pi}-\pi_0)^2}{\left(\frac{\pi_0(1-\pi_0)}{m}\right)} = \frac{(0-0.5)^2}{\left(\frac{(0.5)(1-0.5)}{25}\right)} = 25$$

(c) $$Z_\omega = \frac{\hat{\pi}-\pi_0}{\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{m}}} = \frac{0-0.5}{\sqrt{\frac{0(1-0)}{25}}} = \frac{-0.5}{0} = -\infty$$

1.8 Let π_1 = the probability of a green seedling, π_2 = the probability of a yellow seedling.

H_0: $\pi_1 = 3/4$, $\pi_2 = 1/4$
H_a: $\pi_1 \neq 3/4$ or $\pi_2 \neq 1/4$

$$X^2 = \sum_{i=1}^{2} \frac{(m_i - m\pi_{i0})^2}{m\pi_{i0}}$$

$$= \frac{(854-(1103)(3/4))^2}{(1103)(3/4)} + \frac{(249-(1103)(1/4))^2}{(1103)(1/4)}$$

$$= 0.865 + 2.595$$

$$= 3.460$$

$$P = P(\chi^2_1 \geq 3.460) = 0.0629$$

There is moderate evidence to negate the null hypothesis that the ratio of green to yellow seedlings is 3:1.

1.9

$$\hat{\mu} = \frac{(0)(109) + (1)(65) + \ldots + (4)(1)}{200} = 0.61$$

From MINITAB, we have the following probabilities for a Poisson distribution with $\mu = 0.61$:

$p(0) = 0.5434$, $p(1) = 0.3314$, $p(2) = 0.1011$,
$p(3) = 0.0206$, $p(4) = 0.0031$.

$H_0: \pi_1 = 0.5434, \pi_2 = 0.3314, \ldots, \pi_5 = 0.0031$
H_a: Preceding does not hold

$$X^2 = \sum_{i=1}^{5} \frac{(m_i - m\pi_{i0})^2}{m\pi_{i0}}$$

$$= \frac{(109 - (200)(0.5434))^2}{(200)(0.5434)} + \ldots$$

$$\ldots + \frac{(1 - (200)(0.0031))^2}{(200)(0.0031)}$$

$$= 0.001 + 0.025 + 0.157 + 0.300 + 0.222$$

$$= \underline{\underline{0.705}}$$

If we ignore the fact that we estimated the mean of the distribution before we obtained our Poisson probabilities, the degrees of freedom for X^2 would be $c-1 = 4$, and our p-value would be

$$P = P(\chi^2_4 \geq 0.705) = 0.951.$$

If we take into account the estimation of μ, the degrees of freedom for X^2 would be $(c-1)-1 = 3$, and our p-value would be

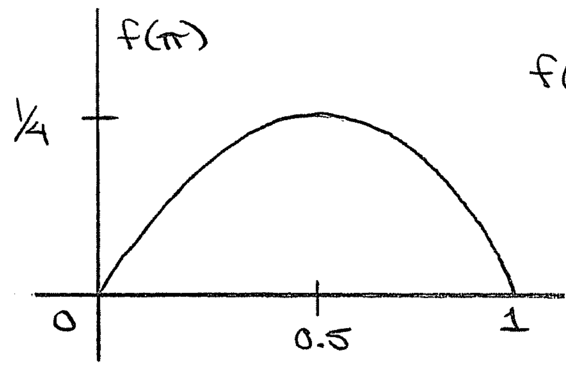
$$P = P(\chi^2_3 = 0.705) = 0.892.$$

In either case, the p-value is quite large and does not provide evidence against the probability specifications under H_0.

1.11 Recall that var$(\hat{\pi})$ is given by $\pi(1-\pi)/n$.

The function $f(\pi) = \pi(1-\pi)$, graphed on the next page for $0 < \pi < 1$, is maximized at $\pi = 1/2$. Furthermore, the function approaches zero as π approaches 0 or 1.

Thus, the estimator $\hat{\pi}$ is least precise (i.e., most variable) when $\pi = 1/2$, and becomes more precise (i.e., less variable) as π approaches 0 or 1.



$f(\pi) = \pi(1-\pi)$

1.12

(a) The distribution of Y is binomial based on m trials with success probability π.

$\text{var}(Y) = m\pi(1-\pi)$

(b)

$$\begin{aligned}\text{var}(Y) &= \text{var}\left(\Sigma_i Y_i\right) \\ &= \Sigma_i \text{var}(Y_i) + \sum_{i \neq j}\Sigma_i \Sigma_j \text{cov}(Y_i, Y_j)\end{aligned}$$

Since $\text{cov}(Y_i, Y_j) = \rho\sqrt{\text{var}(Y_i)\,\text{var}(Y_j)} = \rho\,(m\pi(1-\pi)) > 0$, the preceding implies

$$\text{var}(Y) > \Sigma_i \text{var}(Y_i) = m\pi(1-\pi).$$

(c)

$$\begin{aligned}\text{var}(Y) &= E[\text{var}(Y|\pi)] + \text{var}[E(Y|\pi)] \\ &= E[m\pi(1-\pi)] + \text{var}(m\pi) \\ &= m[E(\pi) - E(\pi^2)] + m^2\,\text{var}(\pi) \\ &= m[\rho - (\text{var}(\pi) + E(\pi)^2)] + m^2\,\text{var}(\pi) \\ &= m[\rho - E(\pi)^2] + m(m-1)\,\text{var}(\pi)\end{aligned}$$

$$= m[p - p^2] + m(m-1)\,\mathrm{var}(\pi)$$
$$= mp(1-p) + m(m-1)\,\mathrm{var}(\pi)$$
$$> mp(1-p)$$

provided $m > 1$

(d) Conditional on the $\{\pi_i\}$, $Y = \sum_i Y_i$ cannot be binomial because the success probabilities for the Y_i are not constant: $P(Y_i = 1 \mid \pi_i) = \pi_i$.

Unconditionally, however, $Y = \sum_i Y_i$ is binomial because
 (i) $Y_1, \ldots, Y_m$ are independent,
 (ii) the success probabilities for the Y_i are constant.

(i) is a consequence of the independence of $\pi_1, \ldots, \pi_m$. (Note that in part (c), $Y_1, \ldots, Y_m$ are unconditionally dependent due to the common conditional success probability π.)

(ii) can be verified as follows: $P(Y_i = 1) = E(Y_i) = E[E(Y_i \mid \pi_i)] = E[P(Y_i = 1 \mid \pi_i)] = E(\pi_i) = p$.

1.14 For the multinomial distribution, we have $\mathrm{var}(N_j) = m\pi_j(1-\pi_j)$, $\mathrm{cov}(N_j, N_k) = -m\pi_j\pi_k$.

$$\mathrm{corr}(N_j, N_k) = \frac{\mathrm{cov}(N_j, N_k)}{\sqrt{\mathrm{var}(N_j)\,\mathrm{var}(N_k)}}$$

$$= \frac{-m\pi_j\pi_k}{\sqrt{(m\pi_j(1-\pi_j))\,(m\pi_k(1-\pi_k))}}$$

$$= \frac{-\pi_j \pi_k}{\sqrt{\pi_j(1-\pi_j)\pi_k(1-\pi_k)}}$$

When $c=2$, we can set $\pi_1 = \pi$, $\pi_2 = (1-\pi)$. The preceding reduces to

$$\frac{-\pi(1-\pi)}{\sqrt{\pi(1-\pi)(1-\pi)\pi}} = -1.$$

1.17 (a) The likelihood is given by the joint distribution of $Y_1, \ldots, Y_n$:

$$l(\mu) = \prod_{i=1}^{n} \frac{e^{-\mu}\mu^{y_i}}{y_i!}$$

$$= \left(\prod_{i=1}^{n} \frac{1}{y_i!}\right) \underbrace{\left(\prod_{i=1}^{n} e^{-\mu}\mu^{y_i}\right)}_{\text{Kernal}}$$

Redefining $l(\mu)$ in terms of the kernal yields

$$l(\mu) = \prod_{i=1}^{n} e^{-\mu}\mu^{y_i}$$

$$= e^{-n\mu}\mu^{\sum_{i=1}^{n} y_i}$$

$$= e^{-n\mu}\mu^{n\bar{y}}$$

For the log-likelihood, we have

$$L(\mu) = \log l(\mu)$$
$$= \log(e^{-m\mu}) + \log(\mu^{m\bar{y}})$$
$$= -m\mu + m\bar{y}\log\mu$$

The MLE of μ is found by solving for μ in the equation $\frac{\partial L(\mu)}{\partial\mu} = 0$.

$$\frac{\partial L(\mu)}{\partial\mu} = \frac{\partial}{\partial\mu}(-m\mu + m\bar{y}\log\mu)$$
$$= -m + \frac{m\bar{y}}{\mu}$$

Setting the preceding equal to zero and solving for μ yields $\hat{\mu} = \bar{y}$.

(b) The score function is given by

$$u(\mu) = \frac{\partial L(\mu)}{\partial\mu} = \frac{m\bar{y}}{\mu} - m$$

Thus, we have

$$\frac{\partial^2 L(\mu)}{\partial\mu^2} = \frac{\partial}{\partial\mu} u(\mu)$$

$$= \frac{\partial}{\partial \mu}\left(\frac{n\bar{y}}{\mu} - n\right)$$

$$= -\frac{n\bar{y}}{\mu^2}$$

Since the preceding is negative for all $\mu > 0$, we know $L(\mu)$ is concave and attains its maximum at $\hat{\mu} = \bar{y}$.

The Fisher information for $\hat{\mu}$ is given by

$$I(\mu) = -E\left[\frac{\partial^2 L(\mu)}{\partial \mu^2}\right]$$

$$= -E\left[-\frac{n\bar{y}}{\mu^2}\right]$$

$$= \frac{n E(\bar{y})}{\mu^2}$$

$$= \frac{n\mu}{\mu^2}$$

$$= \frac{n}{\mu}$$

Wald Test:

$$Z_w = \frac{\hat{\mu} - \mu_0}{SE} \longleftarrow SE = \sqrt{var(\hat{\mu})}, \text{ where}$$

$$var(\hat{\mu}) = 1/I(\hat{\mu})$$

$$= \frac{\hat{\mu} - \mu_0}{\sqrt{1/I(\hat{\mu})}}$$

$$= \frac{\hat{\mu} - \mu_0}{\sqrt{\hat{\mu}/m}}$$

Score Test:

$$Z_s = \frac{u(\mu_0)}{(I(\mu_0))^{1/2}}$$

$$= \frac{\left(\frac{m\bar{y}}{\mu_0} - m\right)}{\left(\frac{m}{\mu_0}\right)^{1/2}}$$

$$= \left(\frac{m\bar{y} - m\mu_0}{\mu_0}\right)\left(\frac{\mu_0}{m}\right)^{1/2}$$

$$= \frac{\sqrt{m}\,(\bar{y} - \mu_0)}{\sqrt{\mu_0}}$$

$$= \frac{\hat{\mu} - \mu_0}{\sqrt{\mu_0/m}}$$

Likelihood Ratio Test:

$$-2\left(L(\mu_0) - L(\hat{\mu})\right)$$

$$= -2\left[\left\{-m\mu_0 + m\bar{y}\log\mu_0\right\} - \left\{-m\hat{\mu} + m\bar{y}\log\hat{\mu}\right\}\right]$$

$$= 2m(\mu_0 - \hat{\mu}) + 2m\bar{y}\log(\hat{\mu}/\mu_0)$$
$$= 2m(\mu_0 - \hat{\mu}) + 2m\hat{\mu}\log(\hat{\mu}/\mu_0)$$

1.31 From page 26, the log-likelihood is given by

$$L(\pi) = m_{11}\log\pi^2 + m_{12}\log(\pi - \pi^2) + m_{22}\log(1-\pi),$$

and the first partial of the log-likelihood with respect to π is given by

$$\frac{\partial L(\pi)}{\partial \pi} = \frac{2m_{11}}{\pi} + \frac{m_{12}}{\pi} - \frac{m_{12}}{1-\pi} - \frac{m_{22}}{1-\pi}$$

For the second partial of the log-likelihood, we have

$$\frac{\partial^2 L(\pi)}{\partial \pi^2} = -\frac{2m_{11}}{\pi^2} - \frac{m_{12}}{\pi^2} - \frac{m_{12}}{(1-\pi)^2} - \frac{m_{22}}{(1-\pi)^2}$$
$$= -\frac{2m_{11}+m_{12}}{\pi^2} - \frac{m_{12}+m_{22}}{(1-\pi)^2}$$

For the Fisher information, we have

$$I(\pi) = -E\left[\frac{\partial^2 L(\pi)}{\partial \pi^2}\right]$$
$$= -E\left[-\frac{2m_{11}+m_{12}}{\pi^2} - \frac{m_{12}+m_{22}}{(1-\pi)^2}\right]$$

$$= \frac{2E(n_{11}) + E(n_{12})}{\pi^2} + \frac{E(n_{12}) + E(n_{22})}{(1-\pi)^2}$$

$$= \frac{2n\pi^2 + n\pi(1-\pi)}{\pi^2} + \frac{n\pi(1-\pi) + n(1-\pi)}{(1-\pi)^2}$$

$$= \frac{n\pi(\pi+1)}{\pi^2} + \frac{n(1-\pi^2)}{(1-\pi)^2}$$

Since $\hat{\pi} = 0.494$ and $n = 156$, we have

$I(\hat{\pi}) = 932.4$,

$SE = (1/I(\hat{\pi}))^{1/2} = \underline{\underline{0.0327}}$

134 For a discrete random variable X with probability mass function $p(x)$ and sample space S, we have

$$E(X) = \sum_{x \in S} x\, p(x).$$

Now let X be a random variable that assumes the value $x_j = \pi_{j0}/\hat{\pi}_j$ with probabilities $p(x_j) = \hat{\pi}_j$ for $j = 1, \ldots, c$.

Consider $E(X)$ and $E(-2n \log X)$.

$$E(X) = \sum_{j=1}^{c} x_j \, p(x_j)$$

$$= \sum_{j=1}^{c} (\hat{\pi}_j) \left(\frac{\pi_{j0}}{\hat{\pi}_j} \right)$$

$$= \sum_{j=1}^{c} \pi_{j0}$$

$$= 1$$

$$E(-2m \log X) = -2m \, E(\log X)$$

$$= -2m \sum_{j=1}^{c} (\log x_j) \, p(x_j)$$

$$= -2m \sum_{j=1}^{c} \left(\log \left(\frac{\pi_{j0}}{\hat{\pi}_j} \right) \right) (\hat{\pi}_j)$$

$$= G^2$$

Now by Jensen's inequality,

$$\log E(X) \geq E(\log X).$$

Thus, we have

$$G^2 = E(-2m \log X)$$

$$= -2m \, E(\log X)$$

$$\doteq -2n\{\log E(X)\}$$

$$= -2n\{\log(1)\}$$

$$= 0$$

1.35 Let $X_1, \ldots, X_k$ be a sequence of independent chi-squared random variables with respective degrees of freedom $v_1, \ldots, v_k$.

Let $Y = \sum_{i=1}^{k} X_i$ and $v = \sum_{i=1}^{k} v_i$.

Consider the m.g.f. of Y.

$$m_Y(t) = E(e^{tY})$$

$$= E\left(e^{t\sum_{i=1}^{k} X_i}\right)$$

$$= E\left(\prod_{i=1}^{k} e^{tX_i}\right)$$

By independence of the X_i

$$= \prod_{i=1}^{k} E(e^{tX_i})$$

$$= \prod_{i=1}^{k} m_{X_i}(t)$$

$$= \prod_{i=1}^{k} (1-2t)^{-v_i/2}$$

$$= (1-2t)^{-(\sum_{i=1}^{k} v_i)/2}$$

$$= (1-2t)^{-v/2}$$

By the uniqueness of m.g.f.'s, we can assert that Y is a chi-squared random variable with degrees of freedom v.