

1. (a) Based on an exact test, we are 95% confident that the true percentage of subjects suffering from nervousness is between 0.58% and 5.36%. Based on the asymptotic score interval, we are 95% confident that the true percentage lies between 0.83% and 5.34% - a slightly narrower interval. The exact confidence interval guarantees 95% coverage, while for the score interval, this is only guaranteed in the limit. In any case, for these data ( $n=188$ ), the two are pretty similar.

```
> binom.test(x=4,n=188)$conf.int*100
[1] 0.5826868 5.3577899
```

```
> prop.test(x=4,n=188,correct=FALSE)$conf.int*100
[1] 0.8304417 5.3420816
```

- (b) i. Using the asymptotic Chi-squared test (for one sample), we obtain a score statistic of  $X^2 = 34.1$  ( $df = 3$ ), and a corresponding P-value less than 0.0001. There is sufficient evidence ( $P < 0.0001$ ) that the true proportions deviate from the ones specified.

```
> chisq.test(x=c(175,5,2,6), p=c(0.98, 0.01, 0.005, 0.005))
```

Chi-squared test for given probabilities

```
data: c(175, 5, 2, 6)
X-squared = 34.0745, df = 3, p-value = 1.911e-07
```

- ii. We'll first approximate the P-value for the exact test by generating 100,000 multinomial tables based on the hypothesized multinomial probabilities. This results in a P-value of 0.00018, which is a pretty good approximation of the true exact P-value, and leads to the same conclusion as in part a.

```
> chisq.test(x=c(175,5,2,6), p=c(0.98, 0.01, 0.005, 0.005),
+ simulate.p.value=TRUE,B=100000)
```

Chi-squared test for given probabilities with simulated p-value (based on 1e+05 replicates)

```
data: c(175, 5, 2, 6)
X-squared = 34.0745, df = NA, p-value = 0.00018
```

We can also obtain the exact P-value (and not an approximation of it) by enumerating *all* multinomial tables with  $n = 188$ . The following code will do this. Note that it may take several minutes to run, since it needs to run through all  $\binom{188+4-1}{4-1} = 1,143,135$  possible multinomial tables with  $n = 188$ . The exact P-value equals 0.00016, i.e., only 0.016% of all possible tables exhibit a Chi-square statistic as large or larger than the one we observed, assuming the null hypothesis is true. This is rather exceptional, so we don't believe the null hypothesis is true. This exact P-value gives the same conclusion as in part a: There is sufficient evidence (exact  $P = 0.00016$ ) that the true proportions deviate from the ones specified.

```
> install.packages("EMT")
> require(EMT)
```

```

> X2.obs <- chisq.test(x=c(175,5,2,6),
+ p=c(0.98, 0.01, 0.005, 0.005))$statistic
> tables <- findVectors(groups=4,size=188)
> head(tables,10)
      [,1] [,2] [,3] [,4]
[1,]    0    0    0  188
[2,]    0    0    1  187
[3,]    0    1    0  187
[4,]    1    0    0  187
[5,]    0    0    2  186
[6,]    0    1    1  186
[7,]    1    0    1  186
[8,]    0    2    0  186
[9,]    1    1    0  186
[10,]   2    0    0  186
> nullprobs <- apply(tables, 1, dmultinom, size=188,
+ prob = c(0.98, 0.01, 0.005, 0.005))
> tail(nullprobs)
[1] 0.004599309 0.012248159 0.025952856 0.041022256 0.042996589 0.022413116
> X2 <- apply(tables, 1, function(x) chisq.test(x,
+ p=c(0.98, 0.01, 0.005, 0.005))$statistic)
> tail(X2)
[1] 7.066218 4.270951 2.550369 1.904472 2.333261 3.836735
> exact.P <- sum((X2>=X2.obs)*nullprobs)
> exact.P
[1] 0.0001643578

```

2. (a) Here, the multinomial probabilities are given by  $\pi_1 = \theta^2$ ,  $\pi_2 = 2\theta(1 - \theta)$ ,  $\pi_3 = (1 - \theta)^2$ , which leads to the following likelihood:

$$L(\theta|n_1, n_2, n_3) = \log \left( \frac{n!}{n_1!n_2!n_3!} \right) + 2n_1 \log(\theta) + n_2 \log(2\theta(1 - \theta)) + 2n_3 \log(1 - \theta) \quad (1)$$

Taking the derivative with respect to  $\theta$ , we obtain

$$\begin{aligned}
\frac{d}{d\theta} L(\theta|n_1, n_2, n_3) &= \frac{2n_1}{\theta} + n_2 \frac{2 - 4\theta}{2\theta(1 - \theta)} - \frac{2n_3}{1 - \theta} \\
&= \frac{2n_1}{\theta} + n_2 \frac{1 - 2\theta}{\theta(1 - \theta)} - \frac{2n_3}{1 - \theta} \\
&= \frac{2n_1}{\theta} + n_2 \left( \frac{1}{\theta} - \frac{1}{1 - \theta} \right) - \frac{2n_3}{1 - \theta} \quad (\text{partial fractions}) \\
&= \frac{2n_1 + n_2}{\theta} - \frac{n_2 + 2n_3}{1 - \theta}
\end{aligned}$$

To obtain the MLE  $\hat{\theta}$ , we set the derivative equal to 0 to obtain:

$$\begin{aligned} \frac{2n_1 + n_2}{\theta} - \frac{n_2 + 2n_3}{1 - \theta} = 0 &\implies \frac{2n_1 + n_2}{\theta} = \frac{n_2 + 2n_3}{1 - \theta} \\ &\implies \hat{\theta} = \frac{2n_1 + n_2}{2n_1 + 2n_2 + 2n_3} \end{aligned}$$

(We should check if this truly is a maximum by plugging it into the second derivative.)

(b) Taking the second derivative of the log-likelihood, we obtain

$$\begin{aligned} \frac{d^2}{d\theta^2} L(\theta|n_1, n_2, n_3) &= -\frac{2n_1 + n_2}{\theta^2} - \frac{n_2 + 2n_3}{(1 - \theta)^2} \\ -\frac{d^2}{d\theta^2} L(\theta|n_1, n_2, n_3) &= \frac{2n_1 + n_2}{\theta^2} + \frac{n_2 + 2n_3}{(1 - \theta)^2} \end{aligned}$$

To find the information matrix, we take the expected value:

$$\begin{aligned} E\left[-\frac{d^2}{d\theta^2} L(\theta|n_1, n_2, n_3)\right] &= \frac{E(2n_1 + n_2)}{\theta^2} + \frac{E(n_2 + 2n_3)}{(1 - \theta)^2} \\ &= \frac{2n\theta^2 + 2n\theta(1 - \theta)}{\theta^2} + \frac{2n\theta(1 - \theta) + 2n(1 - \theta)^2}{(1 - \theta)^2} \\ &= 2n\left(\frac{\theta^2 + \theta(1 - \theta)}{\theta^2} + \frac{\theta(1 - \theta) + (1 - \theta)^2}{(1 - \theta)^2}\right) \\ &= 2n\left(\frac{1}{\theta} + \frac{\theta - \theta^2 + 1 - 2\theta + \theta^2}{(1 - \theta)^2}\right) \\ &= 2n\left(\frac{1}{\theta} + \frac{1}{1 - \theta}\right) \\ &= \frac{2n}{\theta(1 - \theta)} \end{aligned}$$

So we have  $I(\theta) = \frac{2n}{\theta(1-\theta)}$ . The asymptotic variance of  $\hat{\theta}$  is  $I^{-1}(\theta) = \frac{\theta(1-\theta)}{2n}$ , implying a standard deviation of  $\sqrt{\frac{\theta(1-\theta)}{2n}}$ .

If we want to estimate this, we can evaluate this at  $\hat{\theta}$ , giving  $\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{2n}}$  as the standard error.

If we have  $n_1 = 4, n_2 = 8, n_3 = 15$ , we have  $\hat{\theta} = 0.2962963$ , and our estimate of the standard error of  $\hat{\theta}$  is 0.06213856.

```
> n1 <- 4
> n2 <- 8
> n3 <- 15
> n <- sum(c(n1, n2, n3))
> theta.hat <- (2*n1 + n2)/(2*n1 + 2*n2 + 2*n3)
> se <- sqrt(theta.hat*(1-theta.hat)/(2*n))
```

```

> theta.hat - 1.96*se
[1] 0.1745047
> theta.hat + 1.96*se
[1] 0.4180879

```

- (c) In general, we will have  $k - 1 = 3 - 1 = 2$  unknown parameters, since if we know  $\pi_1$  and  $\pi_2$ , we know that  $\pi_3 = 1 - \pi_1 - \pi_2$ . Under the null hypothesis,  $\theta$  is the only unknown parameter, since once we know  $\theta$ , we know what the multinomial probabilities are. This implies one ( $= 2 - 1$ ) degree of freedom for the likelihood ratio test we're about to conduct.

Under the null, the likelihood is as given in (1) and depends on just the one parameter  $\theta$ . Clearly, the likelihood is maximized when  $\theta = \hat{\theta}$ , as we showed in 3a. So, the maximized likelihood under  $H_0$  is (1) with  $\theta$  replaced by  $\hat{\theta} = 0.2962963$ . The maximized likelihood in general is the regular multinomial likelihood with  $\pi_i$  replaced by  $n_i/n$  (i.e., the MLE's in general). Compute these two likelihoods, take their ratio, take  $-2$  times the log of that ratio and you have the value for the Likelihood Ratio Statistic  $G^2$ . Another way to get  $G^2$  is to note that under  $H_0$ , the MLE for the multinomial probabilities are  $\hat{\pi}_{10} = \hat{\theta}^2$ ,  $\hat{\pi}_{20} = 2\hat{\theta}(1 - \hat{\theta})$ ,  $\hat{\pi}_{30} = (1 - \hat{\theta})^2$ . This gives expected counts under the null of  $\mu_{i0} = n\hat{\pi}_{i0}$ . Using the fact that the likelihood ratio statistic  $G^2$  has form

$$G^2 = 2 \sum n_i \log \left( \frac{n_i}{\mu_{i0}} \right)$$

we get

```

> probs.null <- c(theta.hat^2, 2*theta.hat*(1-theta.hat), (1-theta.hat)^2)
> probs.null
[1] 0.0877915 0.4170096 0.4951989
> expected <- n*probs.null
> observed <- c(n1,n2,n3)
> G2 <- 2*sum(observed*log(observed/expected))
> G2
[1] 2.16827
> 1 - pchisq(G2, df = 1)
[1] 0.1408847

```

We get a p-value of 0.14, implying that we do not have sufficient evidence to reject the null hypothesis that the multinomial cell probabilities have the pattern described above.