September 15, 2017

STAT 545: Solutions to Homework 1

Dennis D. Cox
Department of Statistics
Rice University

September 15, 2017

1 Exercise 1.9

Under the model the expected proportion of green plants is 3/4 Of course, the book recommends using the Pearson χ^2 test. Here is the R session:

```
> pearsonstat = (854-.75*1103)^2/(.75*1103) + (249-.25*1103)^2/(.25*1103)
> pearsonstat
[1] 3.459958
> # df=1
> 1-pchisq(pearsonstat,1)
[1] 0.06287192
```

So, we can reject at any level α bigger than this p-value. Obviously we cannot reject $H_0 : \pi = .75$ at the $\alpha = 0.05$ level of significance, but we could reject at the 0.10 level.

The student will get full credit for doing the calculation above. Recall that this Pearson χ^2 test is the score test for this setup. Let's explore some alternatives, like the Wald test. We use the z-statistic form, which is the test most often taught in elementar statistics courses. Computing the observed proportion using R:

```
> phat = 854/1103
> phat
[1] 0.774252
```

The commonly used z-statistic and associated p-value are

```
> z = (phat-.75)/sqrt(phat*(1-phat)/1103)
> z
[1] 1.926562
> # p-value for 2 sided test:
> pvalue = 2*(1-pnorm(z))
> pvalue
[1] 0.05403427
```

The square of the z statistic would have a χ^2_1 distribution of course, but the z statistic also allows us to do one sided tests (like $H_0 : \pi \leq .75$) and then we would use a one tailed rejection region.

Now let's try the LRT and see what we get. The kernel of the likelihood is $\pi^y(1-\pi)^{n-y}$, so the LRT test statistic is

$$\lambda \;=\; 2\left(y \log \hat{\pi} + (n-y)\log(1-\hat{\pi}) - y \log .75 - (n-y)\log .25\right).$$

The R-session:

```
> lrt = 2*(854*log(phat) + 249*log(1-phat) - 854*log(.75) - 249*log(.25))
> lrt
[1] 3.539017
> 1-pchisq(lrt,1)
[1] 0.05994099
```

Hmm, still different. Well, as discussed in class, the χ^2 distribution is only an approximation.

Just to assess which of these approximate tests is most accurate, let's try the corrected and exact tests:

```
> # try the z-test with Yates correction (the default; see help file)
> stuff = prop.test(854,1103,p=.75)
> stuff$p.value
[1] 0.06795155
> # now Fisher's exact test (to be discussed in class later)
> binom.test(854,1103,p=.75)

Exact binomial test

data:  854 and 1103
number of successes = 854, number of trials = 1103, p-value = 0.06531
alternative hypothesis: true probability of success is not equal to 0.75
```

Both the corrected z-test and the exact test give p-values greater than any of the asymptotic (approximate) tests, suggesting that there may be some problems, but they are all around .06, so they're not too far apart. The score test seems to be closest to the more exact values. There are rumors that the LRT is generally better, but for discrete data the score tests are better. The results here suggest the latter rumor is probably correct, at least in this case.

2 Exercise 1.10

We have to estimate the mean (parameter) for the Poisson, then compute the Pearson χ^2 test statistic, and finally convert to a p-value. Here is the R-session:

```
> y = c(109,65,22,3,1)
> ybar = sum(y*(0:4))/sum(y)
> ybar
[1] 0.61
> sum(y)
[1] 200
> yhat = 200*dpois(0:4,ybar)
> yhat[5] = 200*(1-ppois(3,ybar)) # last class is all values >= 4
> sum(yhat) # a check
[1] 200
> pearsonstat = sum((y-yhat)^2/yhat)
> pearsonstat
[1] 0.599929
> # not going to be significant
> # for Chi-squared dist., df = no. classes (5) minus no. params. estimated (1)
> 1-pchisq(pearsonstat,4) # p-value;
[1] 0.9630716
> # definitely not significant
```

3 Exercise 1.12

(a) For a given value t of the test statistic, the p-value transformation is $P[T \geq t]$. (For this problem, all probabilites are computed using the given null distribution.) Now the largest possible value of T is 2, so the smallest possible p-value is $P[T \geq 2] = 0.08$, so if we insist on using non-randomized p-values, we never have a p-value ≤ 0.05, so we never reject H_0, so the $P[$ type I error $] = 0$, since we can never make a Type I error at the 0.05 level of significance.

(b) The smallest mid-p-value is $(P[T > 2] + P[T \geq 2])/2 = (0 + .08)/2 = 0.04$, and for this case we would reject at the 0.05 level. Note that the next smallest p-value is $P[T \geq 1] = 0.62 + 0.08 = 0.70$ and is not relevant at the 0.05 level. So, we can reject with the mid-p-value if $T = 2$, and the probability of this under H_0 is 0.08, so this is our actual type I error probability with the nominal level 0.05

(c) For the ordinary p-value, the p-value map is

$$p(t) = \begin{cases} 0.04 & \text{if } t = 2; \\ 0.70 & \text{if } t = 1; \\ 1 & \text{if } t = 0. \end{cases}$$

Clearly if $\alpha = 0.05$ we reject only if $T = 2$, and then the type I error probability is 0.04. The mid-p-value map is

$$p_m(t) = \begin{cases} 0.02 & \text{if } t = 2; \\ 0.37 & \text{if } t = 1; \\ 0.85 & \text{if } t = 0. \end{cases}$$

Clearly, we only reject using the mid-p-value if we observe $T = 2$, and the type I error probability is still 0.04.

Thus, the mid-p-value approach is liberal in the setting of parts (a) and (b), i.e., the type I error probability is too large, and conservative in the setting of this part.

(d) The randomized p-value is

$$p_r(t) = (1 - U) * P[T > t] + U * P[T \geq t] = P[T > t] + U * P[T = t],$$

where U is uniform random variable on $(0, 1)$ independent of the data. If we reject when $T = 2$ and $U \leq 5/8$, the type I error probability is

$$P[T = 2 \,\&\, U \leq 5/8] = P[T = 2]P[U \leq 5/8] = 0.08 * (5/8) = 0.05.$$

One can argue if this is a sensible test or not. The most sensible approach is to report the conservative p-value $P[T \geq t]$, which most users would report, and maybe also the liberal p-value $P[T > t]$, and then the user who wishes can randomize between them. This is in the spirit of why p-values are used in the first place - they make it easy to decide whether or not to reject H_0 depending on one's own level of significance, and reporting a range (liberal, conservative) leaves it up to the reader whether he or she wishes to randomize the p-value or not.

4 Exercise 1.29

(a) The log likelihood kernel is

$$\begin{aligned}\log L(\theta) &= n_1 \log \theta^2 \;+\; n_2 \log[2\theta(1-\theta)] \;+\; n_3 \log(1-\theta)^2 \\ &= C \;+\; (2n_1 + n_2) \log \theta \;+\; (n_2 + 2n_3) \log(1-\theta).\end{aligned}$$

Here, C is a quantity that doesn't depend on θ. Taking derivative and setting equal to 0:

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

This of course is a stationary point. In fact the second derivative of the log likelihood is

$$-\frac{2n_1 + n_2}{\theta^2} - \frac{n_2 + 2n_3}{(1-\theta)^2} \;<\; 0,$$

which shows that the log likelihood is strictly concave, and so any maximizer is necessarily the unique stationary point.

(b) We derived the second derivative above, and it agrees with the formula in the book. Evaluating the expected value of the negative of this:

$$\begin{aligned}&E_\theta \left[\frac{2N_1 + N_2}{\theta^2} + \frac{N_2 + 2N_3}{(1-\theta)^2} \right] \\ &= \frac{2n\theta^2 + 2n\theta(1-\theta)}{\theta^2} + \frac{2n\theta(1-\theta) + 2n(1-\theta)^2}{(1-\theta)^2} \\ &= \frac{2n}{\theta(1-\theta)},\end{aligned}$$

where the last expression is obtained by some offline algebraic simplification.

The asymptotic standard error is then $2n/[\hat{\theta}(1-\hat{\theta})]$.

(c) To perform a Pearson's χ^2 test, we would compute $\hat{\theta}$ using the above formula, then the so-called expected counts:

$$\begin{aligned}
\hat{n}_1 &= n\hat{\theta}^2 \\
\hat{n}_2 &= 2n\hat{\theta}(1-\hat{\theta}) \\
\hat{n}_2 &= n(1-\hat{\theta})^2.
\end{aligned}$$

Then, the Pearson χ^2 test statistic is

$$T = \sum_{i=1}^{3} \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i}.$$

There are 2 degrees of freedom if the so-called expected counts were given, but since they are based on estimating one parameter, we lose a d.f., so the d.f. for the χ^2 distribution is 1.

Now the student will have satisfied the requirements of the problem with and answer as in the previous paragraph, but it is worthwhile to consider the alternative tests. The LRT is fairly straightforward: the MLE under H_0 is given above, and the MLE under the alternative (the general multinomial model) is the sample proportions. One can plug these into the likelihood and compute the test statistic. Coming up with a Wald test would be a challenge here. We have to figure out an estimand. Considering the full multinomial model with parameter vector $\pi = (\pi_1, \pi_2, \pi_3)$, where we have to drop one of the components because of the constraint $\sum_i \pi_i = 1$, and we will drop π_3. Under the null model,

$$\begin{aligned}
\pi_1 &= \theta^2, \\
\pi_2 &= 2\theta(1-\theta)^2 \\
&= 2\sqrt{\pi_1} - 2\pi_1,
\end{aligned}$$

so if we set

$$b(\pi_1, \pi_2) = \pi_1 - 2\sqrt{\pi_1} - 2\pi_1,$$

then the null hypothesis is equivalent to $H_0 : b(\pi_1, \pi_2) = 0$. One can then derive an asymptotic normal distribution for the estimand and construct a z-statistic for the test. Of course, other estimands could be considered, which is one of the criticisms of Wald tests: they are not unique.

5 Exercise 1.32

(a) Let Σ be the $p \times p$ matrix ($p = c - 1$) given by

$$\Sigma_{kj} = \begin{cases} -\pi_k \pi_j & \text{if } k \neq j, \\ \pi_j(1 - \pi_j) & \text{if } k = j. \end{cases}$$

We want to show that its inverse, say A, has entries

$$A_{ik} = \begin{cases} 1/\pi_c & \text{if } i \neq k, \\ 1/\pi_c + 1/\pi_i & \text{if } i = k. \end{cases}$$

Recall that $\pi_c = 1 - \sum_{i=1}^{p} \pi_i$. Letting $B = A\Sigma$, we want to show B is an identity. The seemingly straightforward approach is to slog through the algebra. If $i \neq j$, then

$$\begin{aligned} B_{ij} &= \sum_{k=1}^{p} A_{ik}\Sigma_{kj} \\ &= \left(\sum_{k=1}^{p} (-\pi_k \pi_j)/\pi_c\right) - \pi_i \pi_j/\pi_i + \pi_j/\pi_c \\ &= \left(-\frac{\pi_j}{\pi_c} \sum_{k=1}^{p} \pi_k\right) + \pi_j \left(\frac{1}{\pi_c} - 1\right) \\ &= -\frac{\pi_j}{\pi_c}(1 - \pi_c) + \frac{\pi_j}{\pi_c}(1 - \pi_c) \\ &= 0. \end{aligned}$$

If $i = j$, then

$$\begin{aligned} B_{ii} &= \sum_{k=1}^{p} A_{ik}\Sigma_{ki} \\ &= \left(\sum_{k=1}^{p} (-\pi_k \pi_i)/\pi_c\right) - \pi_i^2 \frac{1}{\pi_i} + \pi_i \left(\frac{1}{\pi_c} + \frac{1}{\pi_i}\right) \\ &= \left(-\frac{\pi_i}{\pi_c} \sum_{k=1}^{p} \pi_k\right) - \pi_i + \frac{\pi_i}{\pi_c} + 1 \\ &= -\frac{\pi_i}{\pi_c}(1 - \pi_c) - \pi_i + \frac{\pi_i}{\pi_c} + 1 \\ &= 1. \end{aligned}$$

Thus, $B = I$ as we needed to show.

There should be a more elegant way to do this. Let's note that we can write our matrices as

$$\begin{aligned} \Sigma &= D - \theta\theta^T \\ A &= D^{-1} + \frac{1}{\pi_c} U, \end{aligned}$$

where $\theta = (\pi_1, \ldots, \pi_p)$ (recall $p = c - 1$), $D = \mathrm{diag}(\theta)$, and $U = \underline{1}\underline{1}^T$ is a matrix with all entries equal to 1, and $\underline{1}$ is a p-vector with all entries equal to 1. Then

$$\begin{aligned} A\Sigma &= \left(D^{-1} + \frac{1}{\pi_c}\underline{1}\underline{1}^T\right)\left(D - \theta\theta^T\right) \\ &= I - D^{-1}\theta\theta^T + \frac{1}{\pi_c}\underline{1}\underline{1}^T D - \frac{1}{\pi_c}\underline{1}\underline{1}^T\theta\theta^T \end{aligned}$$

Now

$$\begin{aligned} D^{-1}\theta &= \begin{bmatrix} 1/\pi_1 & 0 & \cdots & 0 \\ 0 & 1/\pi_2 & \cdots & 0 \\ \vdots & \vdots & & 0 \\ 0 & \cdots & 1/\pi_{p-1} & 0 \\ 0 & \cdots & 0 & 1/\pi_p \end{bmatrix} \begin{bmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_{p-1} \\ \pi_p \end{bmatrix} \\ &= \underline{1}, \\ \underline{1}^T D &= \theta^T \\ \underline{1}^T\theta &= 1 - \pi_c. \end{aligned}$$

Substituting these into the previous calculation gives

$$\begin{aligned} A\Sigma &= I - \underline{1}\theta^T + \frac{1}{\pi_c}\underline{1}\theta^T - \frac{1}{\pi_c}\underline{1}(1 - \pi_c)\theta^T \\ &= I. \end{aligned}$$

I see that it's just about the same amount of typing as the previous solution, but it seems more elegant. It would be less typing if I didn't expand out D^{-1}.

(b) Let's try it with the slogging through, then the more elegant way. I will use π now to mean θ in the previous notation, thus to match the book's notation. I will also use δ_{ij} to denote Kronecker's δ, which are the entries of the identity matrix ($\delta_{ij} = 0$ if $i \neq j$ and $\delta_{ii} = 1$). Note also that $n\hat{\pi}_i = n_i$ and $n\pi_i = \mu_i$. Also, $\pi_c = 1 - \sum_{i=1}^{c-1} \pi_i$ and $n_c = n - \sum_{i=1}^{c-1} n_i$.

$$\begin{aligned}
n(\hat{\pi} - \pi_0)^T \Sigma_0^{-1} (\hat{\pi} - \pi_0) &= n \sum_{i=1}^{c-1} \sum_{j=1}^{c-1} (\hat{\pi}_i - \pi_i) \left(\frac{1}{\pi_c} + \delta_{ij} \frac{1}{\pi_i} \right) (\hat{\pi}_j - \pi_j) \\
&= \frac{1}{\mu_c} \sum_i \sum_j (n_i - \mu_i)(n_j - \mu_j) \; + \; \sum_i \frac{(n_i - \mu_i)^2}{\mu_i} \\
&= \frac{1}{\mu_c} \left(\sum_i (n_i - \mu_i) \right)^2 \; + \sum_{i=1}^{c-1} \frac{(n_i - \mu_i)^2}{\mu_i} \\
&= \frac{1}{\mu_c} \left(n - n_c - \sum_{i=1}^{c-1} \mu_i \right)^2 \; + \sum_{i=1}^{c-1} \frac{(n_i - \mu_i)^2}{\mu_i} \\
&= \frac{1}{\mu_c} \left(-n_c + n(1 - \sum_{i=1}^{c-1} \pi_i \right)^2 \; + \sum_{i=1}^{c-1} \frac{(n_i - \mu_i)^2}{\mu_i} \\
&= X^2.
\end{aligned}$$

Here, X^2 is the Pearson statistic from equation (1.16), p. 18.

Now let's try doing it the more elegant way.

$$\begin{aligned}
n(\hat{\pi} - \pi_0)^T \Sigma_0^{-1} (\hat{\pi} - \pi_0) &= n(\hat{\pi} - \pi_0)^T \left(D^{-1} + \frac{1}{\pi_c} \underline{1}\underline{1}^T \right) (\hat{\pi} - \pi_0) \\
&= n(\hat{\pi} - \pi_0)^T D^{-1} (\hat{\pi} - \pi_0) \; + \; \frac{n}{\pi_c} (\hat{\pi} - \pi_0)^T \underline{1}\underline{1}^T (\hat{\pi} - \pi_0) \\
&= n \sum_{i=1}^{c-1} \frac{(\hat{\pi}_i - \pi_i)^2}{\pi_i} \; + \; \frac{n}{\pi_c} \left[\underline{1}^T (\hat{\pi} - \pi_0) \right]^2 \\
&= \sum_{i=1}^{c-1} \frac{n^2 (\hat{\pi}_i - \pi_i)^2}{n^2 \pi_i} \; + \; \frac{n}{\pi_c} \left[\sum_{i=1}^{c-1} (\hat{\pi}_i - \pi_i) \right]^2 \\
&= X^2.
\end{aligned}$$

This seems to be just about the same as the previous calculation.

(c) When $c = 2$, then in equation (1.11), p. 13, $\pi = \pi_1$ and

$$\begin{aligned}
Z_s^2 &= \left(\frac{\hat{\pi}_1 - \pi_1}{\sqrt{\pi_1 (1 - \pi_1)/n}} \right)^2 \\
&= n \frac{(\hat{\pi}_1 - \pi_1)^2}{\pi_1 (1 - \pi_1)}
\end{aligned}$$

$$
\begin{aligned}
speq \quad & n\frac{n^2\left(\hat{\pi}_1-\pi_1\right)^2}{n^2\pi_1(1-\pi_1)} \\
= \quad & n\frac{(n\hat{\pi}_1-n\pi_1)^2}{\mu_1(n-\mu_1)} \\
= \quad & n\frac{(n_1-\mu_1)^2}{\mu_1(n-\mu_1)}.
\end{aligned}
$$

Boy, that doesn't look like the formula for X^2, but let's work with the definition of X^2 and simplify it.

$$
\begin{aligned}
X^2 \quad &= \quad \frac{(n_1-\mu_1)^2}{\mu_1}+\frac{(n_2-\mu_2)^2}{\mu_2} \\
&= \quad \frac{(n_1-\mu_1)^2}{\mu_1}+\frac{(n-n_1-n+\mu_1)^2}{n-\mu_1} \\
&= \quad \frac{(n_1-\mu_1)^2}{\mu_1}+\frac{(n_1-\mu_1)^2}{n-\mu_1} \\
&= \quad n\frac{(n_1-\mu_1)^2}{\mu_1(n-\mu_1)}.
\end{aligned}
$$

Now we have solved the problem.

This problem suggests that in many cases, a Wald statistic can be made into a score statististic by simply substituting the constrained MLE into the formula for the standard error rather than the full MLE. Of course, that can't be true in general since the Wald test is not unique - it depends on the estimand that is constrained to give the null hypothesis.