STA 6351, Report.1.21

Carson Slater *Baylor University*

1.21

Fill in the details and present Pawitan's Section 4.3, pp. 79–81 — up to but not including establishing the profile likelihood. This includes working Exercise 4.11 (p. 110 of Pawitan), which establishes

$$\text{se}(\hat{\theta}) = \left(\frac{1}{x} + \frac{1}{m-x} + \frac{1}{y} + \frac{1}{n-y}\right)^{1/2}.$$

4.3 Comparing Two Proportions

Comparing two binomial proportions is probably the most important statistical problem in epidemiology and biostatistics.

Example 4.4: A geneticist believes she has located a gene that controls the spread or metastasis of breast cancer. She analyzed the presence of such gene in the cells of 15 patients whose cancer had spread (metastasized) and 10 patients with localized cancer. The first group had 5 patients with the gene, while one patient in the second group had the gene. Such data are usually presented in a 2×2 table:

	Spread	Localized	Total
Present	5 (33%)	1 (10%)	6
Absent	10	9	19
Total	15	10	25

Is the evidence strong enough to justify her belief?

It is instructive to start with the (large-sample) frequentist solution. First, assign names to the elements of the 2×2 table:

	Spread	Localized	Total
Present	x	y	$t = x + y$
Absent	$m - x$	$n - y$	$N - t$
Total	m	n	$N = m + n$

The standard test of equality of proportions is the famous χ^2 **test**, given by

$$\chi^2 = \frac{N\{x(n-y) - y(m-x)\}^2}{mnt(N-t)},$$

which, under the null hypothesis, has a χ^2_1 distribution. For the observed data,

$$\chi^2 = \frac{25(5 \times 9 - 1 \times 10)^2}{15 \times 10 \times 6 \times 19} = 1.79,$$

producing a p-value of 0.18. This is a two-sided p-value for the hypothesis of equal proportions.

This method has an appealing simplicity, but in small samples its validity is doubtful, and it does not give full information about the parameter of interest (e.g., it is not clear how to get a confidence interval). In small-sample situations we commonly use the **Fisher's exact test**. Under the null hypothesis and conditional on the observed margins, the probability of an observed table is a **hypergeometric probability**:

$$p(x) = \frac{\binom{m}{x}\binom{n}{y}}{\binom{m+n}{x+y}}.$$

Fisher's exact p-value is then computed as the probability of the observed or more-extreme tables. For the above example, there is only one more-extreme table — namely, when we get 0 "present" out of 10 localized cases. The one-sided p-value is

$$p\text{-value} = \frac{\binom{15}{5}\binom{10}{1}}{\binom{25}{6}} + \frac{\binom{15}{6}\binom{10}{0}}{\binom{25}{6}} = 0.17 + 0.03 = 0.20.$$

To proceed with a likelihood analysis, suppose the number of successes X in the first group is binomial $B(m, \pi_x)$, and independently, Y in the second group is $B(n, \pi_y)$. On observing x and y, the joint likelihood of (π_x, π_y) is

$$L(\pi_x, \pi_y) = \pi_x^x (1-\pi_x)^{m-x} \pi_y^y (1-\pi_y)^{n-y}.$$

The comparison of two proportions can be expressed in various ways — for example, using the difference $\pi_x - \pi_y$, the relative risk π_x/π_y, or the **log odds ratio** θ defined by

$$\theta = \log \frac{\pi_x/(1-\pi_x)}{\pi_y/(1-\pi_y)}.$$

In terms of θ, the null hypothesis of interest $H_0 : \pi_x = \pi_y$ is equivalent to $H_0 : \theta = 0$. Each parameterization has its own advantages and disadvantages in terms of interpretation and statistical properties. In small samples, the likelihood of the log odds ratio is more regular than that of the other parameters, so we consider the log odds ratio θ as the parameter of interest. Any other parameter can be treated as a nuisance parameter; for convenience, we use the **log odds** η defined by

$$\eta = \log \frac{\pi_y}{1-\pi_y}.$$

Some simple algebra shows that

$$\begin{aligned}\pi_y &= \frac{e^{\eta}}{1+e^{\eta}},\\ \pi_x &= \frac{e^{\theta+\eta}}{1+e^{\theta+\eta}}.\end{aligned}$$

Therefore, we get the joint likelihood

$$\begin{aligned}L(\theta,\eta) &= \left(\frac{\pi_x}{1-\pi_x}\right)^x (1-\pi_x)^m \left(\frac{\pi_y}{1-\pi_y}\right)^y (1-\pi_y)^n \\ &= \left(\frac{\pi_x/(1-\pi_x)}{\pi_y/(1-\pi_y)}\right)^x \left(\frac{\pi_y}{1-\pi_y}\right)^{x+y} (1-\pi_x)^m (1-\pi_y)^n \\ &= e^{\theta x} e^{\eta(x+y)} (1+e^{\theta+\eta})^{-m} (1+e^{\eta})^{-n}.\end{aligned}$$

The MLE of θ is available directly from the invariance property:

$$\hat{\theta} = \log \frac{x/(m-x)}{y/(n-y)}.$$

The standard error has an interesting formula:

$$\begin{aligned}\mathrm{se}(\hat{\theta}) &= \sqrt{\mathrm{Var}(\hat{\theta})} \\ &\approx \sqrt{\left[I^{-1}(\hat{\theta},\hat{\eta})\right]_{1,1}} \quad \text{(Asymptotic variance from inverse Fisher information)} \\ &\overset{\text{Delta}}{\approx} \sqrt{\mathrm{Var}\left[\log\left(\frac{\hat{\pi}_x}{1-\hat{\pi}_x}\right)\right] + \mathrm{Var}\left[\log\left(\frac{\hat{\pi}_y}{1-\hat{\pi}_y}\right)\right]} \\ &= \sqrt{\frac{1}{m\hat{\pi}_x(1-\hat{\pi}_x)} + \frac{1}{n\hat{\pi}_y(1-\hat{\pi}_y)}} \\ &= \left(\frac{1}{x} + \frac{1}{m-x} + \frac{1}{y} + \frac{1}{n-y}\right)^{1/2}.\end{aligned}$$