Statistics

Robert L. Wolpert
Department of Statistical Science
Duke University, Durham, NC, USA

1 Chi Square

Let's consider repeating, over and over again, an experiment with k possible outcomes. If we let n be the number of times we repeat the experiment (independently!), and count the number N_i of times the i'th outcome occurs altogether, and denote by $\vec{p} = (p_1, ..., p_k)$ the vector of probabilities of the k outcomes, then then each N_i has a binomial distribution

$$N_i \sim \mathsf{Bi}(n, p_i)$$

but they're not independent. The joint probability of the events $[N_i = n_i]$ for nonnegative integers n_i is the "multinomial" distribution, with pmf:

$$f(\vec{n} \mid \vec{p}) = \binom{n}{n_1, n_2, ..., n_k} p_1^{n_1} \cdots p_k^{n_k} \tag{1}$$

where the "multinomial coefficient" is given by

$$\binom{n}{n_1, n_2, ..., n_k} = \binom{n}{\vec{n}} = \frac{n!}{n_1!\, n_2! \cdots n_k!}$$

if each $n_i \geq 0$ and $\sum n_i = n$, otherwise zero.

$$\tag{2}$$

If we observe $\vec{N} = \vec{n}$, what is the MLE for $\vec{p}$? The answer is intuitively obvious, but *proving* it leads to something new. If we try to maximize Equation (1) using derivatives (take logs first!), we find

$$\frac{\partial}{\partial p_i} \log f(\vec{n} \mid \vec{p}) = \frac{n_i}{p_i},$$

so obviously setting these derivatives to zero won't work— they're always positive, so $f(\vec{n} \mid \vec{p})$ is increasing in each p_i. The reason is that this is really a *constrained* optimization problem— the $\{p_i\}$'s have to be non-negative and *sum to one*. As a function on $\mathbb{R}^k$, the function $f(\vec{n} \mid \vec{p})$ of Equation (1) increases without bound as we take all $p_i \to \infty$; but we're not allowed to let the sum of p_i exceed one.

An elegant solution is the method of *Lagrange Multipliers*. We introduce an additional variable λ, and replace the log likelihood with the "Lagrangian":

$$\begin{aligned}\mathcal{L}(\vec{p}, \lambda) &= \log f(\vec{n} \mid \vec{p}) + \lambda\left(1 - \sum p_i\right) \\ &= c + \sum n_i \log p_i + \lambda\left(1 - \sum p_i\right)\end{aligned}$$

with partial derivatives

$$\frac{\partial}{\partial p_i}\mathcal{L}(\vec{p}, \lambda) = \frac{n_i}{p_i} - \lambda \tag{3}$$

$$\frac{\partial}{\partial \lambda}\mathcal{L}(\vec{p}, \lambda) = 1 - \sum p_i \tag{4}$$

Note that stationarity w.r.t λ (setting Equation (4) to zero) enforces the constraint. Now the vanishing of derivatives w.r.t. p_i in Equation (3) imply that $n_i/p_i = \lambda$ is constant for all i, so $p_i = n_i/\lambda$, while Equation (4) now gives $1 = \sum n_i/\lambda = n/\lambda$, so the solutions are the ones we guessed before:

$$\hat{p}_i = n_i/n \qquad\qquad \hat{\lambda} = n.$$

1.1 Generalized Likelihood Tests

Now let's consider testing a hypothetical value $\vec{p}^0$ for the probabilities, against the omnibus alternative:

$$\begin{aligned}H_0 :&\quad \vec{p} = \vec{p}^0 = (p_1^0, \ldots, p_k^0) \\ H_1 :&\quad \vec{p} \neq \vec{p}^0\end{aligned}$$

(the alternative asserts that $p_i \neq p_i^0$ for at least one $1 \le i \le k$). The generalized likelihood ratio against H_0 is:

$$\begin{aligned}
\Lambda(\vec{n}) &= \frac{\sup_{\vec{p}} f(\vec{n} \mid \vec{p})}{f(\vec{n} \mid \vec{p}^0)} \\
&= \frac{f(\vec{n} \mid \hat{\vec{p}})}{f(\vec{n} \mid \vec{p}^0)} \\
&= \frac{\binom{n}{\vec{n}} \prod (n_i/n)^{n_i}}{\binom{n}{\vec{n}} \prod (p_i^0)^{n_i}} \\
&= \prod (n_i/np_i^0)^{n_i}
\end{aligned}$$

Introduce the notation $e_i = np_i^0$ for the "expected" number of outcomes of type i (under null hypothesis H_0) and manipulate:

$$\begin{aligned}
\Lambda(\vec{n}) &= \prod \left[\frac{n_i}{e_i}\right]^{n_i} \\
&= \prod \left[\frac{n_i - e_i + e_i}{e_i}\right]^{n_i} = \prod \left[1 + \frac{n_i - e_i}{e_i}\right]^{n_i}
\end{aligned}$$

If the n_i's are all large enough, we can approximate this by:

$$\begin{aligned}
&\approx \exp\left\{\sum \frac{(n_i - e_i)}{e_i}\ n_i\right\} \\
&= \exp\left\{\sum \frac{(n_i - e_i)(n_i - e_i + e_i)}{e_i}\right\} \\
&= \exp\left\{\sum \frac{(n_i - e_i)^2}{e_i}\right\} \qquad \exp\left\{\sum \frac{(n_i - e_i)e_i}{e_i}\right\} \\
&= e^Q
\end{aligned}$$

since $\sum n_i = \sum e_i = n$ so $\sum (n_i - e_i) = 0$, where

$$Q = \sum \frac{(n_i - e_i)^2}{e_i} \tag{5}$$

is the so-called "Chi Squared" statistic proposed in 1900 by Karl Pearson. Since each $n_i \sim \mathsf{Bi}(n_i, p_i)$, asymptotically each $n_i \sim \mathsf{No}(e_i, e_i q_i^0)$ and so the individual terms in the sum Equation (5) have $\mathsf{Ga}(\frac{1}{2}, \beta)$ distributions (proportional to a χ_1^2) with $\beta = 1/2q_i$, if H_0 is true; Pearson showed that Q has approximately (and asymptotically as $n \to \infty$) a χ_ν^2 distribution with $\nu = k - 1$ degrees of freedom (we'll see why below). If H_0 is false then Q will be much bigger, of course, leading to the well-known χ^2 test for H_0, with P-value $1 -$ `pgamma(Q,` ν`/2, 1/2)`.

1.2 The Distribution of $Q(\vec{n})$

One way to compute the covariance of N_i and N_j is to use an indicator representation, as follows. For $1 \le \ell \le n$ let J_ℓ be a random integer in the range $1, ..., k$, with probability $p_j = \mathsf{P}[J_\ell = j]$ for $1 \le j \le k$. Then N_i can be represented as the sum

$$N_i = \sum_{\ell=1}^{n} \mathbf{1}_{\{J_\ell = i\}}$$

of indicator variables. This makes the following expectations easy for $i \ne j$:

$$\begin{aligned}
\mathsf{E}[N_i] &= \sum \mathsf{P}[J_\ell = i] && = np_i \\
\mathsf{E}[N_i^2] &= \mathsf{E}\left[\sum_{\ell}\sum_{\ell'} \mathbf{1}_{\{J_\ell = i\}}\mathbf{1}_{\{J_{\ell'} = i\}}\right] && = np_i + n(n-1)p_i^2 \\
& && = np_i(1-p_i) + (np_i)^2 \\
\mathsf{E}[N_i N_j] &= \mathsf{E}\left[\sum_{\ell}\sum_{\ell'} \mathbf{1}_{\{J_\ell = i\}}\mathbf{1}_{\{J_{\ell'} = j\}}\right] && = n(n-1)p_i p_j \\
\mathsf{V}(N_i) &= np_i(1-p_i) \\
\mathsf{Cov}(N_i, N_j) &= -np_i\, p_j
\end{aligned}$$

If we let $Z \sim \mathsf{No}(0,1)$ be independent of $\vec{N}$ and add $Zp_i\sqrt{n}$ to each component N_i, we will exactly cancel the negative covariance:

$$\mathsf{Cov}\big((N_i + Zp_i\sqrt{n}), (N_j + Zp_j\sqrt{n})\big) = -np_ip_j + (p_i\sqrt{n})(p_j\sqrt{n}) \qquad = 0$$

while keeping zero mean

$$\mathsf{E}\big((N_i + Zp_i\sqrt{n})\big) = 0$$

and increase the variance to

$$\mathsf{V}\big((N_i + Zp_i\sqrt{n})\big) = np_i(1-p_i) + (p_i\sqrt{n})^2 \qquad = e_i.$$

Thus the random variables $(N_i - e_i + Zp_i\sqrt{n})/\sqrt{e_i}$ are uncorrelated and have mean zero and variance one. By the Central Limit Theorem, they are approximately k independent standard normal random variables as $n \to \infty$, so the quadratic form

$$Q^+(\vec{n}) = \sum_{i=1}^{k} \frac{(N_i - e_i + Zp_i\sqrt{n})^2}{e_i}$$

has approximately a χ^2_k distribution for large n. But:

$$\begin{aligned} Q^+(\vec{n}) &= \sum \frac{(N_i - e_i)^2}{np_i} & &+ \sum \frac{2(N_i - e_i)Z\, p_i\sqrt{n}}{np_i} & &+ \sum \frac{Z^2 p_i^2 n}{np_i} \\ &= Q(\vec{n}) & &+ \frac{2Z}{\sqrt{n}} \sum (N_i - e_i) & &+ Z^2 \sum p_i \\ &= Q(\vec{n}) + Z^2, \end{aligned}$$

the sum of $Q(\vec{n})$ and a χ^2_1 random variable independent of $\vec{N}$— so $Q(\vec{n})$ itself must have approximately a χ^2_ν distribution with $\nu = (k-1)$ degrees of freedom.

1.3 P-Values

For even degrees of freedom ν the χ^2_ν distribution is just the $\mathsf{Ga}(\alpha = \nu/2, \beta = 1/2)$, the waiting time for $\nu/2$ events in a Poisson process X_t with rate $1/2$, so P-values can be computed in closed form

$$\begin{aligned} \mathsf{P}[Q > q] &= \mathsf{P}[X_q \le \nu/2] \\ &= e^{-q/2} \sum_{k=0}^{(\nu/2)-1} \frac{(q/2)^k}{k!}. \end{aligned}$$

For example, with $\nu = 2$ degrees of freedom, the P-value is simply $e^{-q/2}$. For large values of ν the χ^2_ν distribution is close to the normal $\mathsf{No}(\nu, 2\nu)$ by the Central Limit Theorem, so

$$\mathsf{P}[Q > q] \approx \Phi\left(\frac{\nu - q}{\sqrt{2\nu}}\right).$$