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.

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 Eqn (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 Eqn (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

$$\frac{\partial}{\partial p_i}\mathcal{L}(\vec{p}, \lambda) = \frac{n_i}{p_i} - \lambda \tag{2}$$

$$\frac{\partial}{\partial \lambda}\mathcal{L}(\vec{p}, \lambda) = 1 - \sum p_i \tag{3}$$

Note that stationarity w.r.t λ (setting Eqn (3) to zero) enforces the constraint. Now the vanishing of derivatives w.r.t. p_i in Eqn (2) imply that $n_i/p_i = \lambda$ is constant for all i, so $p_i = n_i/\lambda$, while Eqn (3) 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 $1 \leq i \leq k$). 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 and e_i's are all large enough, we can approximate the logarithm of this by:

$$\begin{aligned}\log \Lambda(\vec{n}) &= \sum n_i \log\left(1 + \frac{n_i - e_i}{e_i}\right) \\ &\approx \sum (n_i - e_i + e_i)\left(\frac{n_i - e_i}{e_i} - \frac{(n_i - e_i)^2}{2\ e_i^2}\right)\end{aligned}$$

using the two-term Taylor series $\log(1+\epsilon) = \epsilon - \epsilon^2/2 + O(\epsilon^3)$

$$\approx \frac{1}{2}\sum \frac{(n_i - e_i)^2}{e_i} = \frac{1}{2}Q, \tag{4}$$

half the quadratic form $Q := \sum \frac{(n_i - e_i)^2}{e_i}$, since $\sum(n_i - e_i) = 0$ and since $\sum(n_i - e_i)^3 = O(1/\sqrt{n})$. The statistic Q is the so-called "Chi Squared" statistic proposed in 1900 by Karl Pearson, who found its asymptotic distribution.

Since each $n_i \sim \mathsf{Bi}(n_i, p_i)$, asymptotically each $n_i \sim \mathsf{No}\big(e_i, e_i(1-p_i^0)\big)$ and so the individual terms in the sum Eqn (4) have approximate $\mathsf{Ga}(\frac{1}{2}, \beta)$ distributions (proportional to a χ_1^2) with $\beta = 1/2(1-p_i)$, 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

$$P = 1 - \texttt{pgamma}(\texttt{Q},\ \nu/2,\ \texttt{1/2}) = \texttt{pgamma}(\texttt{Q},\ \nu/2,\ \texttt{1/2},\ \texttt{lower.tail} = \texttt{F}).$$

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

One way to compute the covariance of N_i and N_j is to use indicator variables, as follows. For $1 \le \ell \le n$ let J_ℓ be a label telling us which of the k possible outcomes happened on the ℓ'th trial— 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 $i \neq j$:

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

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

The χ^2_ν distribution is just the $\mathsf{Ga}(\alpha = \nu/2,\ \beta = 1/2)$. If the degrees of freedom parameter ν is even, it may be viewed as 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 as

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

For example, with $\nu = 2$ degrees of freedom, the P-value is simply $e^{-q/2}$, while for $\nu = 4$ and $\nu = 6$ it is $(1 + q/2)e^{-q/2}$ and $(1 + q/2 + q^2/8)e^{-q/2}$, respectively.

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

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

For any ν and q, it's available in `R` as

```
1-pchisq(q, nu)
```

or, more precisely for large q, as `pchisq(q, nu, lower.tail=FALSE)`.

2 Contingency Tables

Now consider a composite hypothesis like:

$$H_0: \ \{N_{ij}\} \sim \mathsf{MN}(n; \theta_{ij}) \text{ for some } \theta_{ij} = p_i\, q_j, \ 1 \le i \le R, \ 1 \le j \le C$$

for $R \cdot C$ counts N_{ij} summing to n. If n items are categorized separately into one of R rows and also into one of C columns, and if N_{ij} denotes the number of items in the ith row and jth column, then this hypothesis asserts that the two categorizations are *independent*. Alternately, if $N_{i+} \equiv \sum_{j=1}^{C} N_{ij}$ objects from the ith of R populations are categorized into one of C categories, then H_0 also asserts that the R populations are all *homogeneous* in the sense that they share the same distribution among the C categories.

In either case, a Generalized Likelihood Ratio test will be based on

$$\begin{aligned}\Lambda &= \frac{\sup_\theta \left\{ \prod \theta_{ij}^{N_{ij}} : \ \sum \theta_{ij} = 1 \right\}}{\sup_{p,q} \left\{ \prod (p_i q_j)^{N_{ij}} : \ \sum p_i = 1, \ \sum q_j = 1 \right\}} \\ &= \prod \left\{ \frac{\hat\theta_{ij}}{\hat p_i \hat q_j} \right\}^{N_{ij}}\end{aligned}$$

where $\hat\theta_{ij} = N_{ij}/n$, $\hat p_i = N_{i+}/n$, and $\hat q_j = N_{+j}/n$. Upon setting $\hat e_{ij} \equiv n \hat p_i \hat q_j$,

$$\begin{aligned}\log \Lambda &= \sum N_{ij} \log \left\{ \frac{N_{ij}}{\hat e_{ij}} \right\} \\ &= \sum \{ (N_{ij} - \hat e_{ij}) + \hat e_{ij} \} \log \left\{ 1 + \frac{N_{ij} - \hat e_{ij}}{\hat e_{ij}} \right\} \\ &\approx \frac{1}{2} \sum \frac{(N_{ij} - \hat e_{ij})^2}{\hat e_{ij}} = Q/2, \text{ where} \\ Q &= \sum \frac{(N_{ij} - \hat e_{ij})^2}{\hat e_{ij}}\end{aligned}$$

has approximately a χ^2_ν distribution with $\nu = RC - 1 - (R-1) - (C-1) = (R-1)(C-1)$ degrees of freedom. More generally, Q will have approximately a χ^2_ν distribution with $\nu = k - 1 - s$ degrees of freedom if there are k categories and we must estimate an s-dimensional aspect of θ from the data. The same idea may be used to test independence for three-way (or more) classifications, in which H_0 asserts that $\theta_{ijk} = p_i q_j r_k$ for some $\vec p$, $\vec q$, $\vec r$.

2.1 A Numerical Example

A 1986 study of a treatment for Hodgkins disease (Dunsmore *et al*, 1986) studied the response rates (classified into three levels: Positive, Partial, and None) for patients of four different histological types. The results are summarized in this table:

Type	Pos	Part	Neg	
LP	74	18	12	104
NS	68	16	12	96
MC	154	54	58	266
LD	18	10	44	72
	314	98	126	538

Denote by X_{ij} the entry in the ith row and jth column, and by X_{i+} and X_{+j} the row and column sums (shown in the table). The *expected* count under H_0 in cell (i, j) is $E_{ij} := X_{i+}X_{+j}/n$– $E_{11} = 104 \times 314/538 = 60.70$ for $(1, 1)$, for example, so the ξ^2 statistic is $Q = \sum(X_{ij} - E_{ij})^2/E_{ij} = 75.89$. Under the null hypothesis this would have a χ^2_ν distribution with $\nu = (R-1)(C-1) = 6$ degrees of freedom. The P-value is $P =$ `pchisq(Q, 6, low=F)` $= 2.52 \cdot 10^{-14}$, so H_0 would be rejected.

In `R` this calculation could be performed as follows:

```
Xij <- matrix( c(74,68,154,18, 18,16,54,10, 12,12,58,44), ncol=3);
row <- apply(Xij,1,sum);       # Row sums
col <- apply(Xij,2,sum);       # Column sums
Eij <- row %o% col / sum(Xij); # Expected counts
Q   <- sum( (Xij-Eij)^2/Eij ); # Chi-square statistic
P   <- pchisq(Q, 6, low=F);    # P-value
```

using the "`apply()`" function and the outer product operator "`%o%`".

2.2 Two by Two

An important special case of contingency table analysis is when $R = C = 2$. For example, we may study the benefit (or risk) of Exposure to some treatment (or hazard) by exploring the independence of classifications with respect to Exposure (Exposed and non-Exposed) and also to a health outcome (here, Diseased or non-Diseased). Denote the count of subjects in each class as X_{ij}, where $i \in \{0, 1\}$ indexes the exposure class (1=Exposed) and $j \in \{0, 1\}$ the disease class (1=Diseased). The object will be to test the hypothesis H_0 that exposure is unrelated to disease status, against the

two-sided alternative that there is some connection.

These data might arise from any of three possible sampling schemes, which each lead to different probability models, and somewhat different expressions for H_0:

1. **Multinomial:** For some number $n \in \mathbb{N}$ and probability vector $p = (p_{00}, p_{01}, p_{10}, p_{11})$, $\mathbf{x} = (X_{00}, X_{01}, X_{10}, X_{11}) \sim \mathsf{MN}(n, p)$. H_0 would assert that row and column classifications are independent, *i.e.*, that $p_{00}p_{11} = p_{01}p_{01}$ or, equivalently, that the ratio ψ is one, where

$$\psi = \frac{p_{00}p_{11}}{p_{01}p_{10}}$$

2. **Prospective:** For some numbers $x_{1+} \in \mathbb{N}$ of Exposed and $x_{0+} \in \mathbb{N}$ of un-Exposed subjects, we observe $X_{11} \sim \mathsf{Bi}(x_{1+}, \mathsf{P}(D \mid E))$ and $X_{01} \sim \mathsf{Bi}(x_{0+}, \mathsf{P}(D \mid E^c))$ diseased cases, respectively. H_0 would assert that $\mathsf{P}(D \mid E^c) = \mathsf{P}(D \mid E)$ or, equivalently, that the disease *odds* are equal for exposed and unexposed subjects

$$\frac{\mathsf{P}(D \mid E^c)}{\mathsf{P}(D^c \mid E^c)} = \frac{\mathsf{P}(D \mid E)}{\mathsf{P}(D^c \mid E)}$$

   This condition is satisfied if and only if the *odds ratio* is one:

$$\psi := \frac{\mathsf{P}(D \mid E)\mathsf{P}(D^c \mid E^c)}{\mathsf{P}(D \mid E^c)\mathsf{P}(D^c \mid E)} = \frac{\mathsf{P}(D \cap E)\mathsf{P}(D^c \cap E^c)}{\mathsf{P}(D \cap E^c)\mathsf{P}(D^c \cap E)} = \frac{p_{00}p_{11}}{p_{01}p_{10}}$$

3. **Retrospective:** Among some numbers $x_{+1} \in \mathbb{N}$ of Diseased and $x_{+0} \in \mathbb{N}$ of un-Diseased subjects, we discover that $X_{11} \sim \mathsf{Bi}(x_{+1}, \mathsf{P}(E \mid D))$ and $X_{10} \sim \mathsf{Bi}(x_{+0}, \mathsf{P}(E \mid D^c))$ had been exposed, respectively. H_0 would assert that $\mathsf{P}(E \mid D^c) = \mathsf{P}(E \mid D)$ or, equivalently, that the exposure *odds* are equal for diseased and undiseased subjects

$$\frac{\mathsf{P}(E \mid D^c)}{\mathsf{P}(E^c \mid D^c)} = \frac{\mathsf{P}(E \mid D)}{\mathsf{P}(D^c \mid D)}$$

   Again this is satisfied if and only if the *odds ratio* is one:

$$\psi := \frac{\mathsf{P}(E \mid D)\mathsf{P}(E^c \mid D^c)}{\mathsf{P}(E \mid D^c)\mathsf{P}(E^c \mid D)} = \frac{\mathsf{P}(E \cap D)\mathsf{P}(E^c \cap D^c)}{\mathsf{P}(E \cap D^c)\mathsf{P}(E^c \cap D)} = \frac{p_{00}p_{11}}{p_{01}p_{10}}$$

Thus, all three sampling approaches lead to consideration of whether or not the odds ratio ψ is unity. A value of $\psi > 1$ indicates a positive association

between exposure and disease; a value $\psi < 1$ indicates a protective effect. The Maximum Likelihood Estimator for ψ in all three cases is

$$\hat{\psi} = \frac{X_{00}X_{11}}{X_{01}X_{10}},$$

and the GLRT of H_0 in all cases leads to rejection of H_0 for large values of the GLR statistic

$$\Lambda = \prod_{i,j=0,0}^{1,1} \left(\frac{n\ X_{ij}}{X_{i+}X_{+j}} \right)^{X_{ij}} = \prod_{i,j=0,0}^{1,1} (X_{ij}/E_{ij})^{X_{ij}}\ ,$$

where $E_{ij} := X_{i+}X_{+j}/n$ is the "expected" count under the hypothesis H_0 of independence. Equivalently, one would reject for large values of its logarithm

$$\begin{aligned} \log \Lambda &= \sum X_{ij} \log(X_{ij}/E_{ij}) \approx Q/2, \qquad \text{where} \\ Q &= \sum \frac{(X_{ij} - E_{ij})^2}{E_{ij}} \end{aligned}$$

has approximately a χ^2_1 distribution for large n.

2.3 A Numerical Example

But what if n is *not* large? The famous 1985 RCT test of extracorporeal membrane oxygenation (ECMO– see Ware, 1989) featured only 19 subjects. Six of ten in the control group survived, and all nine of the treated subjects survived, so the data are

$$X_{00} = 6 \qquad X_{01} = 4 \qquad X_{10} = 9 \qquad X_{11} = 0$$

and the MLE for the odds ratio is $\hat{\psi} = \infty$. Evidently this sample size is insufficient for the χ^2 approximation to hold.

Wolpert and Mengersen (2004) introduced an objective Bayesian approach using independent Jeffreys' prior distributions for the survival probabilities p and q in the Exposed (to ECMO) and un-Exposed groups, respectively, and then find the posterior probability distribution for $\psi = p(1-q)/(1-p)q$.

They found an explicit form for the pdf of $\varepsilon := \log \psi$,

$$f(\mathbf{x} \mid \epsilon) \propto e^{\varepsilon(X_{11}+1/2)}{}_2F_1\big(X_{+0}+1, X_{+1}+1; X_{++}+2; 1-e^{\varepsilon}\big) \qquad (5)$$

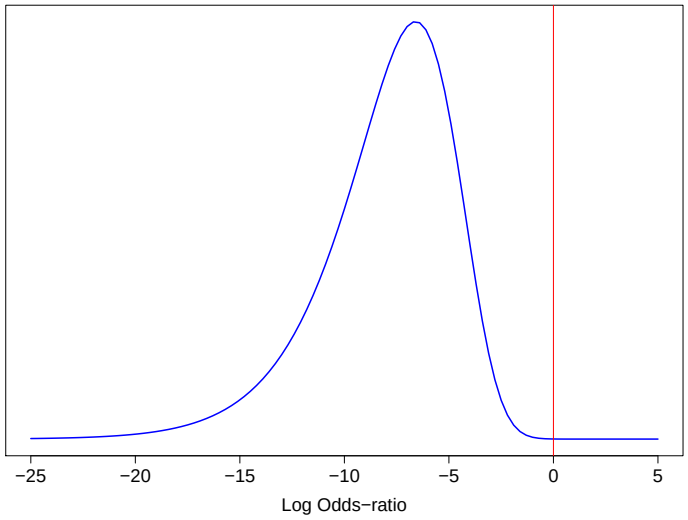

Figure 1: Reference Posterior PDF for ECMO Log Odds Ratio

in terms of the confluent hypergeometric function ${}_2F_1(a, b; c; z)$ (Abramowitz and Stegun, 1964, §15.1) and evaluated its mean and variance as

$$\mu = \psi(X_{00} + \tfrac{1}{2}) \; - \psi(X_{01} + \tfrac{1}{2}) \; - \psi(X_{10} + \tfrac{1}{2}) \; + \psi(X_{11} + \tfrac{1}{2}) \tag{6a}$$

$$\sigma^2 = \psi'(X_{00} + \tfrac{1}{2}) + \psi'(X_{01} + \tfrac{1}{2}) + \psi'(X_{10} + \tfrac{1}{2}) + \psi'(X_{11} + \tfrac{1}{2}) \tag{6b}$$

where $\psi(z) = (d/dz)\log(\Gamma(z))$ and $\psi(z) = (d/dz)\psi(z)$ are the digamma and trigamma functions, respectively (Abramowitz and Stegun, 1964, §6.3, 6.4). These are included in `R` and other computing environments, but their values here can be computed easily using the identities

$$(n + \tfrac{1}{2}) - \psi(m + \tfrac{1}{2}) = \sum_{i=m}^{n-1} (i + \tfrac{1}{2})^{-1} \approx \log \frac{n}{m} \tag{7a}$$

for integers $0 \le m < n$, and

$$\psi'(n + \tfrac{1}{2}) = \frac{\pi^2}{2} - \sum_{i=0}^{n-1} (i + \tfrac{1}{2})^{-2} \approx \frac{1}{n} \tag{7b}$$

For the ECMO trial, these give $\mu = -3.75721$ and $\sigma = 2.3368$; under the normal approximation to the posterior of ε the approximate posterior probability of no effect or harmful effect would be $\mathsf{P}[\varepsilon > 0 \mid \mathbf{x}] \approx \Phi(\mu/\sigma) = 0.0539$.

In fact, due to the skewness of the pdf (see Figure (1)), it is considerably smaller— numerical integration of (5) gives $\mathsf{P}[\varepsilon > 0 \mid \mathbf{x}] \approx 9.514 \cdot 10^{-6}$, rather strong evidence in ECMO's favor despite the small sample sizes.

2.3.1 Frequentist Analysis of ECMO

Let $X \sim \mathsf{Bi}(n, p)$ and set $q := (1-p)$, the failure probability, and $\theta := \log p/q$, the log odds. The MLE for θ is

$$\begin{aligned}\hat{\theta} &= \log \frac{x/n}{1 - x/n} \\ &= \theta + \log(x/np) - \log((n-x)/nq) \\ &= \theta + \log\big(1 + \frac{x - np}{np}\big) - \log\big(1 + \frac{n - x - nq}{nq}\big) \\ &= \theta + \log\big(1 + \frac{x - np}{np}\big) - \log\big(1 - \frac{x - np}{nq}\big)\end{aligned}$$

If n is sufficiently large that $|x - np| \ll n$, then by the delta method

$$\hat{\theta} \approx \theta + \frac{x - np}{np} + \frac{x - np}{nq} = \theta + \frac{x - np}{npq} \approx \mathsf{No}(\theta, \sigma^2)$$

by the CLT, with mean θ and variance

$$\sigma^2 = \mathsf{E}\left(\frac{x - np}{npq}\right)^2 = \frac{npq}{n^2p^2q^2} = \frac{1}{npq}.$$

In a prospective trial with independent treatment and control arms, it follows that for sufficiently large sample sizes the MLE $\hat{\varepsilon}$ for the log odds ratio

$$\varepsilon = \log \psi = \log \frac{\mathsf{P}(D \mid E)\mathsf{P}(D^c \mid E^c)}{\mathsf{P}(D \mid E^c)\mathsf{P}(D^c \mid E)} = \log \frac{\mathsf{P}(D \mid E)}{\mathsf{P}(D^c \mid E)} - \log \frac{\mathsf{P}(D \mid E^c)}{\mathsf{P}(D^c \mid E^c)}$$

is also approximately normally distributed with mean ε and variance

$$\begin{aligned}\sigma^2 &= \frac{1}{X_{1+}\mathsf{P}(D \mid E)\mathsf{P}(D^c \mid E)} + \frac{1}{X_{0+}\mathsf{P}(D \mid E^c)\mathsf{P}(D^c \mid E^c)} \\ &\approx \frac{1}{X_{00}} + \frac{1}{X_{01}} + \frac{1}{X_{10}} + \frac{1}{X_{11}}.\end{aligned}$$

By Eqns (6a, 7a) $\mathsf{E}[\hat{\theta}]$ is close to the reference posterior mean of ε, and by Eqns (6b, 7b) $\mathsf{Var}[\hat{\theta}]$ is close to the posterior variance of ε, for sufficiently large sample sizes. Unfortunately ECMO's sample sizes were far too small for the delta method or the CLT to apply.

3 Other Composite Hypotheses

We can also use a χ^2 test to see if data $\{X_i\}$ come from *some* unspecified member of a parametric family $f(x \mid \theta)$ of distributions. Typically we must aggregate or *bin* the data into a finite number (say, k) of categories; compute the category probabilities $p_i(\theta)$, $1 \le i \le k$; minimize Λ over all possible values of θ (or, nearly the same thing, minimize $Q(\theta)$); and approximate the distribution of $Q(\hat{\theta})$ by the χ^2_ν with $\nu = k - 1 - s$, for $\theta \in \Theta \subseteq \mathbb{R}^s$.

3.1 Poisson example

For instance, in DeGroot & Schervish (4/e) problem 5 of section 10.2, we have $n = 200$ observations $X_i \in \mathbb{Z}_+$ which may be from a $\mathsf{Po}(\theta)$ distribution:

$$\begin{aligned} X = 0: &\quad 52 \\ X = 1: &\quad 60 \\ X = 2: &\quad 55 \\ X = 3: &\quad 18 \\ X = 4: &\quad 8 \\ X \ge 5: &\quad 7 \end{aligned}$$

At any specific θ, the likelihood for the grouped data would be

$$\begin{aligned} L(\theta) &= \prod_{i=0}^{4} \left[\frac{\theta^i}{i!} e^{-\theta}\right]^{N_i} \cdot \left[1 - e^{-\theta} \sum_{i=0}^{4} \frac{\theta^i}{i!}\right]^{N_5} \\ &\propto \theta^{0\cdot 52 + 1\cdot 60 + 2\cdot 55 + 3\cdot 18 + 4\cdot 8} e^{-\theta[52+60+55+18+8]} \left[1 - e^{-\theta} \sum_{i=0}^{4} \frac{\theta^i}{i!}\right]^{7} \\ &= \theta^{256} e^{-193\theta} \left[1 - e^{-\theta} \sum_{i=0}^{4} \frac{\theta^i}{i!}\right]^{7} \end{aligned}$$

The optimal θ is $\hat{\theta} = 1.465232$ (found by a numerical search) with $Q(\hat{\theta}) = 7.696875$, for a P-value of $P = \texttt{pchisq(7.696875, df=4, low=F)} = (1 + Q/2)e^{-Q/2} = 0.1033348$. Evidently we can't reject the Poisson hypothesis at levels $\alpha \le 0.10$. Figure (2) shows a plot of the log likelihood, with $\hat{\theta}$ noted. In this example $\hat{\theta}$ is very close to the *Poisson* MLE of $\tilde{\theta} = 1.5$ (using the additional information about the "$X \ge 5$" observations offered in Problem 5 of DeGroot & Schervish, §10.2, 4/e), so the values of the log likelihood

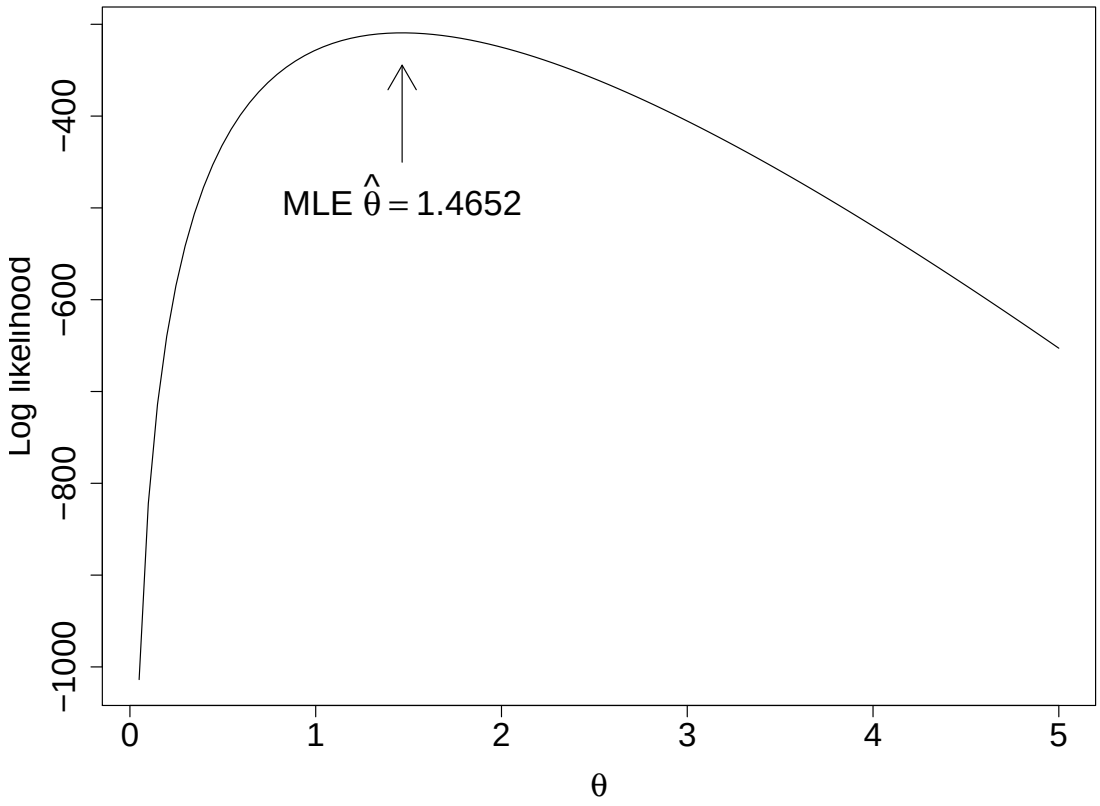

Figure 2: Multinomial log likelihood.

and of Q agree to two decimal places and the same conclusions would be drawn using either method.

3.2 Geometric example

Let's test to see if the same data come from the geometric distribution $\mathsf{Ge}(p)$ for any $p \in [0, 1]$. Setting $q = 1 - p$, the geometric probabilities are

$$\mathsf{P}[X = i] = pq^i, \quad 0 \le i \le 4 \qquad\qquad \mathsf{P}[X \ge 5] = q^5$$

so

$$L(p) = \prod_{i=0}^{4} \left[p\ q^i\right]^{N_i} \cdot \left[q^5\right]^{N_5} = p^{\sum_{i=0}^{4} N_i} q^{\sum_{i=0}^{5} i\, N_i} = p^{193} q^{291}.$$

This attains its maximum at $\hat{p} = 193/(193 + 291) = 193/484$, leading to "expected" counts of $e_i = \hat{p}\hat{q}^i$ for $0 \le i \le 4$, and $e_5 = \hat{q}^5$. The log GLR statistic and the quadratic form Q are

$$\begin{aligned} \log \Lambda &= \sum N_i \log(N_i/e_i) = 19.6416 \\ Q &= \sum (N_i - e_i)^2/e^i \ = 41.8620 \approx 2 \log \Lambda \end{aligned}$$

for a P-value of $P =$ `pchisq(41.86, 4, low=F)` of $P \approx 1.78 \cdot 10^{-8}$. This offers clear evidence that these data do *not* come from any exponential distribution.

3.3 Generic example

We can construct a GLR test of the null hypothesis that observations $X_1, \ldots, X_n$ come from *any* parametric family $\mathcal{P} = \{f_\theta(x) :\ \theta \in \Theta\}$ with finite-dimensional parameter space $\Theta \subset \mathbb{R}^s$ as follows:

- Partition the outcome space $\mathcal{X} = \cup_{i=1}^{k} A_i$ into some number $k > s + 1$ of disjoint sets A_i;
- Evaluate the probabilities $p_i(\theta) = \int_{A_i} f_\theta(x)\, dx$ that X will fall into each partition element;
- Count the observed occupancies $N_i = \sum_j \mathbf{1}_{A_i}(X_j) = \#\{j :\ X_j \in A_i\}$;
- Find $\hat{\theta} = \text{argmax}_\theta \sum N_i \log p_i(\theta)$ and set $\hat{p}_i := p_i(\hat{\theta})$ and $E_i := n\hat{p}_i$;
- Evaluate

$$Q(\hat{\theta}) := \sum_{i=1}^{k} \frac{\left(N_i - np_i(\hat{\theta})\right)^2}{np_i(\hat{\theta})} = \sum_{i=1}^{k} \frac{(N_i - E_i)^2}{E_i};$$

- Report a P-value of $P =$ `pchisq(Q, k-s-1, low=F)`.

The fourth step can be replaced with "Find $\hat{\theta} = \text{argmin}_\theta \, Q(\theta)$", but *not* by "Set θ equal to its MLE under the model $\mathcal{P}$". Since the multinomial likelihood function is very nearly proportional to e^Q (that's how the χ^2 test was derived, after all), the multinomial MLE $\hat{\theta}$ is very nearly the minimizing value of Q, but other estimates $\tilde{\theta}$ of θ will lead to a heavier-tailed distribution for $Q(\tilde{\theta})$ than the χ^2_{k-s-1}. For the MLE $\tilde{\theta}$ under the model $\mathcal{P}$, Chernoff and Lehmann (1954) showed that $Q(\tilde{\theta})$ is distributed like the sum of a χ^2_{k-s-1} random variable and an independent sum $\sum_{i=1}^{2} \lambda_i Z_i^2$ for $\{Z_i\} \overset{\text{iid}}{\sim} \mathsf{No}(0,1)$ and numbers $0 \leq \lambda_i \leq 1$. It follows that its CDF lies between those of the χ^2_{k-s-1} and χ^2_{k-1}, so it is valid to reject H_0 at level α if $Q(\tilde{\theta})$ exceeds the $(1-\alpha)$th quantile of the χ^2_{k-1} distribution.

References

Abramowitz, M. and Stegun, I. A., eds. (1964), *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*, *Applied Mathematics Series*, volume 55, Washington, D.C.: National Bureau of Standards, reprinted in paperback by Dover (1974); on-line at `http://www.math.sfu.ca/~cbm/aands/`.

Chernoff, H. and Lehmann, E. L. (1954), "The Use of Maximum Likelihood Estmates in χ^2 Tests for Goodness of Fit," *Annals of Mathematical Statistics*, 21, 579–586, doi:10.1214/aoms/1177728726.

Ware, J. H. (1989), "Investigating Therapies of Potentially Great Benefit: ECMO," *Statistical Science*, 4, 298–340, (With discussion).

Wolpert, R. L. and Mengersen, K. L. (2004), "Adjusted Likelihoods for Synthesizing Empirical Evidence from Studies That Differ in Quality and Design: Effects of Environmental Tobacco Smoke," *Statistical Science*, 19, 450–471, doi:10.1214/088342304000000350.