CATEGORICAL MODELS AND CHI-SQUARE TESTS

Categorical data are given by counts within each of a finite number of mutually exclusive categories, when a finite number of individuals are distributed by some random mechanism amongst those categories. There are several good reasons for studying the asymptotics for categorical data models.

• Pearson's χ^2 test was one of the first examples of a statistical test for goodness-of-fit. Even though Pearson (1900) got the story on degrees of freedom wrong, the χ^2 test has become a standard part of statistical methodology. Several key statistical ideas, such as maximum likelihood estimation and efficiency, were developed by Fisher during the period of his running battle with Pearson over the correct choice for the degrees of freedom of the approximating χ^2 distribution. Some of his original arguments for the goodness-of-fit problem still offer helpful insights into those general ideas.

• The χ^2 test is much used, but not always well understood. What difference does it make if one uses Poisson rather than multinomial models? Why does conditioning on various marginal totals in cross-tabulated data have the same effect on degrees of freedom as the estimation of certain parameters? It is not too hard to answer these questions in an asymptotic sense.

• As an introduction to asymptotic methods, the theory for categorical models has several technical advantages. Finiteness and monotonicity properties simplify calculations. Regularity conditions are clean and easy to understand. Pathologies that complicate general theory are easier to eliminate from categorical models. The technicalities don't obscure the ideas.

1. The multinomial model

Suppose each of n objects is placed independently into one of k mutually exclusive categories (or *cells*), labelled $1, 2, \ldots, k$, according to the distribution

$$\mathbb{P}\{\text{object placed in cell } i\} = p_i \qquad \text{for } i = 1, \ldots, k.$$

The $\{p_i\}$ are nonnegative and sum to 1. Define the cell counts

$$S_{ni} = \text{ total number in cell } i.$$

Then the random vector $(S_{n1}, \ldots, S_{nk})$ has a multinomial distribution, denoted by $\mathcal{M}(n, p_1, \ldots, p_k)$:

$$\mathbb{P}\{S_{n1} = x_1, \ldots, S_{nk} = x_k\} = \frac{n!}{x_1! \ldots x_k!}\, p_1^{x_1} \ldots p_k^{x_k},$$

where $x_1, \ldots, x_k$ range over all choices of nonegative integers summing to n.

When the vector $\mathbf{p} = (p_1, \ldots, p_k)$ of probabilities is known, the goodness-of-fit statistic

$$X_n^2(\mathbf{p}) = \sum\nolimits_\alpha \frac{(S_{n\alpha} - np_\alpha)^2}{np_\alpha}$$

has approximately a χ^2 distribution; it converges in distribution to χ^2_{k-1} as n tends to ∞. If $\mathbf{p}$ is unknown, but is modelled as a member of some set $\mathcal{P}$, replacement of the unknown p_i by estimates $\widehat{p}_{ni}$ defines an analogous goodness-of-fit statistic $X_n^2(\widehat{\mathbf{p}}_n)$. Under simple assumptions on $\mathcal{P}$, the new statistic has a limiting χ^2 distribution, but with a reduced number of degrees of freedom, for appropriately efficient estimators $\widehat{\mathbf{p}}_n$. Pearson got it wrong; Fisher (1922, 1923, 1928) got it right. Birch (1964) proved it elegantly.

Two appropriate $\widehat{\mathbf{p}}_n$ choices are (the method of minimum χ^2) the value that minimizes $X_n^2(\mathbf{p})$, and the maximum likelihood estimator (MLE), which is defined to maximize $\sum_i S_{ni} \log p_i$ over $\mathcal{P}$. Equivalently, the MLE minimizes

$$G_n^2(\mathbf{p}) = 2n \sum_i f_{ni} \log(f_{ni}/p_i)$$

over $\mathcal{P}$, where $f_{ni} = S_{ni}/n$. Both $X_n^2(\mathbf{p})$ and $G_n^2(\mathbf{p})$ provide a measure of fit between a proposed $\mathbf{p}$ and the observed vector $\mathbf{f}_n$ of proportions. Cressie & Read (1984, 1988) treated both functions as members of a one-parameter family of *power-divergence* functions:

$$J_\lambda(\mathbf{f}_n, \mathbf{p}) = \frac{2}{\lambda(\lambda+1)} \sum_i \left(f_{ni}^{1+\lambda}/p_i^\lambda - f_{ni} \right).$$

For $\lambda = 0$ and $\lambda = -1$ the functions are defined by continuity:

$$J_0(\mathbf{f}_n, \mathbf{p}) = \lim_{\lambda \to 0} J_\lambda(\mathbf{f}_n, \mathbf{p}) = 2 \sum_i f_{ni} \log(f_{ni}/p_i)$$

and similarly for J_{-1}. Using the fact that both $\{f_{ni}\}$ and $\{p_i\}$ sum to 1, it is easy to expand out the quadratic to get

$$X_n^2(\mathbf{p}) = n J_1(\mathbf{f}_n, \mathbf{p}).$$

For each real λ an estimator is defined by minimization of the J_λ function over the parameter set $\mathcal{P}$:

$$\widehat{\mathbf{p}}_n(\lambda) = \operatorname*{argmin}_{\mathbf{p} \in \mathcal{P}} J_\lambda(\mathbf{f}_n, \mathbf{p}).$$

In the next few sections I will derive limit results for these estimators and the corresponding goodness-of-fit statistics. Under regularity conditions close to those of Birch (1964), for fixed λ and λ', I will show that $\widehat{\mathbf{p}}_n(\lambda) = \widehat{\mathbf{p}}_n(\lambda') + o_p(1/\sqrt{n})$ and

$$n J_\lambda(\mathbf{f}_n, \widehat{\mathbf{p}}_n(\lambda')) \rightsquigarrow \chi^2_{k-1-s},$$

where s is a dimension defined by $\mathcal{P}$. My presentation draws many ideas from Dudley (1976).

Properties of power-divergence functions

The function $J_\lambda(\mathbf{f}_n, \mathbf{p})$ is a sum of k terms $H_\lambda(f_{ni}, p_i)$, where

$$H_\lambda(x, y) = \frac{2}{\lambda(1+\lambda)} \left(\frac{x^{1+\lambda}}{y^\lambda} + \lambda(y - x) - x \right) \qquad \text{for } x > 0, y > 0.$$

As before, the definition for $\lambda = 0$ or $\lambda = -1$ is by continuity:

$$H_0(x, y) = 2x \log(x/y) = H_{-1}(y, x) \qquad \text{for } x > 0, y > 0.$$

Similarly, the limits as $x \to 0$ or $y \to 0$, when they exist, define the values $H_\lambda(0, y)$ and $H_\lambda(x, 0)$. The extra linear term $\lambda(f_{ni} - p_i)$ has no direct effect on J_λ; it contributes zero to the sum. But it does make H_λ a more convenient function.

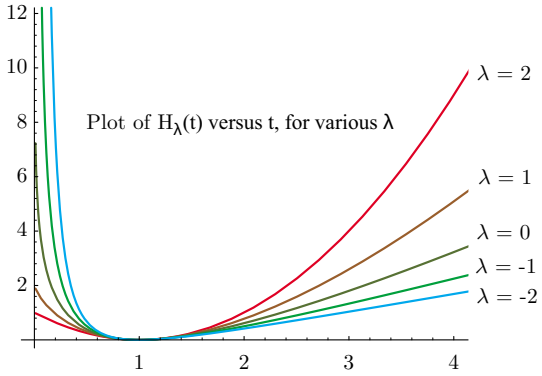
Simple algebra shows that

<1> $$H_{-1-\lambda}(y, x) = H_\lambda(x, y) = x H_\lambda(1, y/x),$$

at least when $x > 0$. The partial derivates,

$$\frac{\partial}{\partial y} H_\lambda(x, y) = \frac{2}{1+\lambda} \left(- \frac{x^{1+\lambda}}{y^{1+\lambda}} + 1 \right),$$
$$\frac{\partial^2}{\partial y^2} H_\lambda(x, y) = \frac{2x^{1+\lambda}}{y^{2+\lambda}},$$

identify $H_\lambda(x,\cdot)$ as a convex function that achieves its minimum of zero at x.



For t near 1 a Taylor expansion gives

$$H_\lambda(1,t) = (t-1)^2 - 2(2+\lambda)(t-1)^3 s^{-3-\lambda},$$

with s between 1 and t. It follows via equality <1> that for each $x_0 > 0$ there exists a neighborhood U and a constant C such that

$$\left| H_\lambda(x,y) - \frac{(y-x)^2}{x} \right| \le C|y-x|^3 \qquad \text{for } x, y \in U.$$

Of course both C and U depend on x_0 and λ. Putting $x = q_i$ and $y = p_i$ then summing over i, we get for each vector of cell probabilities $\boldsymbol{\pi}_0$ with strictly positive components a neighborhood $\mathcal{V}$ such that

<2> $$\left| J_\lambda(\mathbf{q},\mathbf{p}) - \sum_i \frac{(q_i-p_i)^2}{q_i} \right| \le C|\mathbf{q}-\mathbf{p}|^3 \qquad \text{for } \mathbf{q},\mathbf{p} \in \mathcal{V}.$$

Notice that the approximating sum does not depend on λ. The fact that $H_\lambda(x,y)$ increases as y moves away from x in either direction greatly simplifies the asymptotic theory for power-divergence estimators. It implies that $J_\lambda(\mathbf{f}_n, \mathbf{f}_n + t\mathbf{v})$ for $t \ge 0$ is an increasing function of t for each fixed $\mathbf{v}$; the function $J_\lambda(\mathbf{f}_n,\cdot)$ increases along each ray emanating from $\mathbf{f}_n$.

2. Asymptotics for power-divergence estimators

Until further notice, λ will be held at some fixed real value.

Suppose the cell counts are generated from an $\mathcal{M}(n,\boldsymbol{\pi}_0)$ distribution, where each component of the $\boldsymbol{\pi}_0$ (the *true* value of the parameter) is strictly positive. Suppose also that $\boldsymbol{\pi}_0$ belongs to some specified set of probability vectors $\mathcal{P}$.

The first step in most asymptotic arguments is a proof of consistency. One needs to know that $\widehat{\mathbf{p}}_n$ is close to $\boldsymbol{\pi}_0$ before any sort of Taylor expansion can be of use. As it will later turn out, the first step is actually superfluous for the multinomial model. Nevertheless, it will be instructive as a prototype for a style of argument that typically would be required.

I will prove that $\widehat{\mathbf{p}}_n$ converges almost surely to $\boldsymbol{\pi}_0$. The weaker convergence in probability would actually suffice, but there is no added difficulty in establishing convergence in the stronger sense. We may as well establish the stronger result, even if we don't really need it.

The behaviour of $\mathbf{f}_n$ controls $\widehat{\mathbf{p}}_n$. The relevant facts follow directly from the SLLN and the MCLT.

<3> **Lemma.** *Under the multinomial model $\mathcal{M}(n, \boldsymbol{\pi}_0)$,*

(i) $\mathbf{f}_n \to \boldsymbol{\pi}_0$ *almost surely.*

(ii) $\sqrt{n}(\mathbf{f}_n - \boldsymbol{\pi}_0) \rightsquigarrow N(\mathbf{0}, V)$*, where the limiting variance matrix has* $(i, j)^{th}$ *element* $-\pi_i\pi_j$ *if* $i \neq j$*, and* $\pi_i - \pi_i^2$ *if* $i = j$*.*

Proof. The components of $\mathbf{f}_n$ are not independent; they sum to 1. There is another source of independence, though. Write $\mathbf{e}_\alpha$ for the unit vector with a 1 in position α and zeros elswhere. A single observation corresponds to a random vector $\mathbf{X}$ that takes the value $\mathbf{e}_\alpha$ with probability π_α, the position of the 1 indicating the cell into which the observation falls. Thus

$$\mathbb{P}\mathbf{X} = \pi_1\mathbf{e}_1 + \ldots + \pi_k\mathbf{e}_k = \boldsymbol{\pi}_0$$
$$\mathbb{P}\mathbf{X}\mathbf{X}' = \pi_1\mathbf{e}_1\mathbf{e}_1' + \ldots + \pi_k\mathbf{e}_k\mathbf{e}_k'.$$

Consequently,

$$\text{var}(\mathbf{X}) = \text{ diag } (\pi_1, \ldots, \pi_k) - \boldsymbol{\pi}_0\boldsymbol{\pi}_0' = V.$$

Write $\mathbf{f}_n$ as an average of n independent copies of $\mathbf{X}$, then deduce assertion (i) from the SLLN and (ii) from the MCLT. □

What do we need to show in order to establish the almost sure convergence of $\widehat{\mathbf{p}}_n$ to $\boldsymbol{\pi}_0$? To understand the problem let us temporarily make explicit the dependence of $\widehat{\mathbf{p}}_n = \widehat{\mathbf{p}}_n(\omega)$ on the point ω in the underlying sample space Ω. We need to find a negligible set $\mathcal{N}_0$, and for each $\delta > 0$ we need a finite $n_0(\omega, \delta)$, such that

<4> $$|\widehat{\mathbf{p}}_n(\omega) - \boldsymbol{\pi}_0| < \delta \qquad \text{for all } n \geq n_0(\omega, \delta) \text{ and } \omega \notin \mathcal{N}_0.$$

The same negligible set $\mathcal{N}_0$ works for each $\delta > 0$. If the restriction $\omega \notin \mathcal{N}_0$ were replaced by $\omega \in \mathcal{N}_\delta$, for a negligible set $\mathcal{N}_\delta$ that might depend on δ, it would be a simple matter of casting out a sequence of negligible sets to recover an appropriate $\mathcal{N}_0$: we could take $\mathcal{N}_0$ to be the union of the countable family $\{\mathcal{N}_{1/m} : m = 1, 2, \ldots\}$. In summary, for each $\delta > 0$ we need to show that, with probability one,

<5> $$|\widehat{\mathbf{p}}_n - \boldsymbol{\pi}_0| < \delta \qquad \text{eventually.}$$

Be certain that you understand the quantities hidden in the words *with probability one* and *eventually.* They are the negligible set $\mathcal{N}_\delta$ and the $n_0(\omega, \delta)$ corresponding to the modified form of assertion <4>. It would be most cumbersome to make the dependences explicit every time.

<6> **Theorem.** *Under the multinomial model $\mathcal{M}(n, \boldsymbol{\pi}_0)$, with $\boldsymbol{\pi}_0$ a point in $\mathcal{P}$ having strictly positive components, the minimizer $\widehat{\mathbf{p}}_n$ of $J_\lambda(\mathbf{f}_n, \mathbf{p})$ over $\mathcal{P}$ converges to $\boldsymbol{\pi}_0$ almost surely.*

Proof. Fix $\delta > 0$. Write B for the open ball with radius δ centered at $\boldsymbol{\pi}_0$. We need to show that, with probability one, the estimator $\widehat{\mathbf{p}}_n$ eventually lies in B, for each δ small enough. From now on let me omit the qualification *with probability one*. At the end of the proof you should mentally cast out a countable collection of negligible sets to make the assertions hold on a set with probability one.

The estimator $\widehat{\mathbf{p}}_n$ minimizes $J_\lambda(\mathbf{f}_n, \mathbf{p})$ over all $\mathbf{p}$ in $\mathcal{P}$. In particular, it must do better than the (unknown) $\boldsymbol{\pi}_0$:

$$J_\lambda(\mathbf{f}_n, \widehat{\mathbf{p}}_n) \leq J_\lambda(\mathbf{f}_n, \boldsymbol{\pi}_0).$$

By the almost sure convergence of $\mathbf{f}_n$ to $\boldsymbol{\pi}_0$ and the continuity of J_λ in its first argument,

$$J_\lambda(\mathbf{f}_n, \boldsymbol{\pi}_0) \to J_\lambda(\boldsymbol{\pi}_0, \boldsymbol{\pi}_0) = 0.$$

For each fixed $\epsilon > 0$, we must therefore have

$$J_\lambda(\mathbf{f}_n, \widehat{\mathbf{p}}_n) < \epsilon \qquad \text{eventually.}$$

To force $\widehat{\mathbf{p}}_n$ into the ball B we need only show that, for some $\epsilon > 0$,

$$\inf_{\mathbf{p}\in B^c} J_\lambda(\mathbf{f}_n, \mathbf{p}) \geq \epsilon \qquad \text{eventually.}$$

In a more typical asymptotic problem an infimum over a large piece of the parameter space might present unpleasant global complications. Strange things might happen as the parameter wanders off to the far corners of the parameter space. For the multinomial problem, monotonicity of $J_\lambda(\mathbf{f}_n, \cdot)$ along each ray emanating from $\mathbf{f}_n$ spares us the global complication.

When $|\mathbf{f}_n - \boldsymbol{\pi}_0| < \delta$, as must eventually happen by virtue of Lemma <3> part (i), the infimum over B^c is achieved on the boundary ∂B of the ball B. For each $\mathbf{p} \in B^c$ there is a p^* on the boundary for which $J_\lambda(\mathbf{f}_n, \mathbf{p}^*) \leq J_\lambda(\mathbf{f}_n, \mathbf{p})$. It follows that

$$\inf_{\mathbf{p}\in B^c} J_\lambda(\mathbf{f}_n, \mathbf{p}) = \inf_{\mathbf{p}\in \partial B} J_\lambda(\mathbf{f}_n, \mathbf{p}) \qquad \text{when } |\mathbf{f}_n - \boldsymbol{\pi}_0| < \delta.$$

Asymptotic problems do not always allow us to sidestep global considerations so easily.

If δ is chosen small enough, the boundary ∂B lies within the neighborhood of $\boldsymbol{\pi}_0$ in which the quadratic approximation <2> holds. In particular, eventually

$$\left| J_\lambda(\mathbf{f}_n, \mathbf{p}) - \sum\nolimits_i \frac{(f_{ni} - p_i)^2}{f_{ni}} \right| \leq C|\mathbf{f}_n - \mathbf{p}|^3 \qquad \text{for all } \mathbf{p} \in \partial B.$$

It is important that the inequality eventually holds *simultaneously* for all $\mathbf{p}$ on the boundary ∂B. Each f_{ni} converges almost surely to a positive π_i. Eventually the approximating quadratic will be larger than the sum obtained by replacing f_{ni} in the denominator by $2\pi_i$. Using the fact that eventually $2\delta > |\mathbf{f}_n - \mathbf{p}| > \delta/2$ for all $\mathbf{p}$ in ∂B, deduce that

$$\inf_{\mathbf{p}\in \partial B} J_\lambda(\mathbf{f}_n, \mathbf{p}) \geq \min_i (2\pi_i)^{-1}(\delta/2)^2 - C(2\delta)^3 \qquad \text{eventually.}$$

Call the lower bound ϵ. If δ is small enough, ϵ is positive. As explained earlier, it follows that $\widehat{\mathbf{p}}_n$ eventually lies within the ball B, as required. □

The proof of the last theorem made only feeble use of the almost sure convergence. Nowhere did we need to consider the behaviour of $\mathbf{f}_n$ at a fixed ω for more than one sample size n; nowhere did we need control over $J_\lambda(\mathbf{f}_n, \mathbf{p})$ for more than one n. The real role was to provide the *with probability one . . . eventually*, which justified the existence of several inequalities for a fixed sample size. If we had to rely on only a weak law of large numbers—that is, convergence of $\mathbf{f}_n$ to $\boldsymbol{\pi}_0$ in probability—the incantation would have been *eventually, with probability close to one*. The meaning would then be: given an $\epsilon > 0$ there exists an $n_0(\epsilon)$ such that $\mathbb{P}\{\ldots\} > 1 - \epsilon$ for $n \geq n_0(\epsilon)$. When we combine a fixed, finite number of such assertions we arrive at assertions that hold except possibly for a set of ω (depending on n) that has probability bounded by a fixed multiple of the (arbitrarily small) ϵ. The main line of the argument changes little.

It becomes more important that we do not depend on almost sure assertions when we study the finer asymptotics for $\widehat{\mathbf{p}}_n$. The asymptotic normality for $\mathbf{f}_n$ asserted by part (ii) of Lemma <3> controls behaviour for each fixed n, not behaviour at a fixed ω along a whole sequence of sample sizes. (A law of the iterated logarithm for $\mathbf{f}_n$ would give an almost sure bound, but the distributional consequences of the weaker result turn out to be more interesting.)

Reinterpreted as a manipulation of inequalities that hold eventually with probability close to one, the proof of the last theorem could establish much more than mere convergence in probability of $\widehat{\mathbf{p}}_n$ to $\boldsymbol{\pi}_0$: with δ allowed to decrease with increasing sample size, it gives an in-probability rate of convergence. Another slight improvement, which will later allow us to relate the estimators derived from different J_λ function, comes at no great cost. We do not need $\widehat{\mathbf{p}}_n$ to exactly minimize $J_\lambda(\mathbf{f}_n, \cdot)$; it has only to come within some prescribed distance from the infimum. For those who worry about such things, the slight increase in generality also eliminates technical problems related to existence, uniqueness, and measurability of an exactly minimizing value.

<7> **Theorem.** *Assume the $\mathcal{M}(n, \boldsymbol{\pi}_0)$ model with $\boldsymbol{\pi}_0$ a point in $\mathcal{P}$ having strictly positive components. Suppose $\widehat{\mathbf{p}}_n$ is an element of $\mathcal{P}$ for which*

$$J_\lambda(\mathbf{f}_n, \widehat{\mathbf{p}}_n) \le \inf_{\mathbf{p}\in\mathcal{P}} J_\lambda(\mathbf{f}_n, \mathbf{p}) + O_p(1/n).$$

Then $\widehat{\mathbf{p}}_n$ converges to $\boldsymbol{\pi}_0$ at an $O_p(1/\sqrt{n})$ rate.

Proof. Let B be the open ball with center $\boldsymbol{\pi}_0$ and (random) radius δ_n of order $O_p(1/\sqrt{n})$. The precise value for δ_n will be specified soon.

The defining inequality for $\widehat{\mathbf{p}}_n$ implies

$$J_\lambda(\mathbf{f}_n, \widehat{\mathbf{p}}_n) \le J_\lambda(\mathbf{f}_n, \boldsymbol{\pi}_0) + O_p(1/n).$$

The convergence in distribution of $\sqrt{n}(\mathbf{f}_n - \boldsymbol{\pi}_0)$ implies that $\mathbf{f}_n$ lies within a distance $O_p(1/\sqrt{n})$ of $\boldsymbol{\pi}_0$. From the quadratic approximation <2>, it follows that

$$J_\lambda(\mathbf{f}_n, \boldsymbol{\pi}_0) \le \sum\nolimits_i \frac{(f_{ni} - \pi_{0i})^2}{f_{ni}} + O(|\mathbf{f}_n - \boldsymbol{\pi}_0|^3) = O_p(1/n).$$

Thus there exists a (random) ϵ_n of order $O_p(1/\sqrt{n})$ for which

$$J_\lambda(\mathbf{f}_n, \widehat{\mathbf{p}}_n) < \epsilon_n^2.$$

To force $\widehat{\mathbf{p}}_n$ into the ball B we need to choose δ_n large enough to make (eventually, with probability close to one)

$$\inf_{\mathbf{p}\in B^c} J_\lambda(\mathbf{f}_n, p) \ge \epsilon_n^2.$$

If we ensure that $|\mathbf{f}_n - \boldsymbol{\pi}_0| < \delta_n/2$, monotonicity of $J_\lambda(\mathbf{f}_n, \cdot)$ along rays from $\mathbf{f}_n$ reduces the left-hand side to the infimum over the boundary ∂B of B, giving the lower bound

$$(\min_i (2\pi_i)^{-1})(\delta_n/2)^2 - C(2\delta_n)^3$$

when $\mathbf{f}_n$ gets close to $\boldsymbol{\pi}_0$. For n large enough (that is, eventually), with probability close to one, the lower bound is greater than $\delta_n^2/16$. To satisfy the two requirements for δ_n the value $4\epsilon_n + 2|\mathbf{f}_n - \boldsymbol{\pi}_0|$ would suffice. □

The multinomial problem is atypical in that a monotonicity property allows us to deduce a rate of convergence $O_p(1/\sqrt{n})$ directly from the behaviour of the criterion function $J_\lambda(\mathbf{f}_n, \mathbf{p})$ on the boundary of a very small ball. It is typical, however, in that no particular local structure need be imposed on $\mathcal{P}$, except that it should contain the true $\boldsymbol{\pi}_0$, in order to force the $O_p(1/\sqrt{n})$ rate. Comparisons between $J_\lambda(\mathbf{f}_n, \widehat{\mathbf{p}}_n)$ and $J_\lambda(\mathbf{f}_n, \boldsymbol{\pi}_0)$ cannot take us any further. To establish distributional results for $\sqrt{n}(\widehat{\mathbf{p}}_n - \boldsymbol{\pi}_0)$ we must impose more structure on the parameter set $\mathcal{P}$, at least in a neighborhood of $\boldsymbol{\pi}_0$. The following assumptions, essentially due to Birch (1964) with modifications by Dudley (1976), can hardly be improved upon.

Let us assume that a small piece of $\mathcal{P}$ near the true $\boldsymbol{\pi}_0$ is well approximated by a small piece of an s-dimensional hyperplane, where $s < k - 1$. The

minimization of $J_\lambda(\mathbf{f}_n, \mathbf{p})$ will then become asymptotically equivalent to the minimization of a quadratic form, derived from approximation <2>, over the whole hyperplane—an asymptotic problem of weighted least squares.

<8> **Local Smoothness Assumption.** *Say that $\mathcal{P}$ is locally s-dimensional near $\boldsymbol{\pi}_0$ if there exists a continuous, one-to-one map $\mathbf{p}(\cdot)$ from a compact neighborhood Θ_0 of the origin in $\mathbb{R}^s$ into $\mathcal{P}$ such that:*

(i) the set $\mathcal{P}_0 = \{\mathbf{p}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta_0\}$ is a neigborhood of $\boldsymbol{\pi}_0 = \mathbf{p}(\mathbf{0})$ within $\mathcal{P}$.;

(ii) the map $\mathbf{p}(\cdot)$ is differentiable at $\mathbf{0}$,

$$\mathbf{p}(\theta) = \boldsymbol{\pi}_0 + D\boldsymbol{\theta} + o(|\boldsymbol{\theta}|) \qquad \text{near } \mathbf{0},$$

with derivative matrix D of full rank s.

Property (i) means that every $\mathbf{p}$ in $\mathcal{P}$ that is close enough to $\boldsymbol{\pi}_0$ must have the form $\mathbf{p}(\boldsymbol{\theta})$ for a uniquely determined $\boldsymbol{\theta}$ in Θ_0. The assumption of full rank ensures existence of constants $0 < c_1 < c_2 < \infty$ such that

$$c_1|\mathbf{t}| \le |D\mathbf{t}| \le c_2|\mathbf{t}| \qquad \text{for all } \mathbf{t} \in \mathbb{R}^s.$$

It follows (Problem ???) that there exist other constants $0 < C_1 < C_2 < \infty$ such that

<9> $$C_1|\boldsymbol{\theta}| \le |\mathbf{p}(\boldsymbol{\theta}) - \boldsymbol{\pi}_0| \le C_2|\boldsymbol{\theta}| \qquad \text{for } \boldsymbol{\theta} \in \Theta_0$$

These inequalities will ensure that rates of convergence for estimators of $\boldsymbol{\pi}_0$ translate into the same rates of convergence for the $\boldsymbol{\theta}$ values. For example, $\mathbf{f}_n$ converges to $\boldsymbol{\pi}_0$ at $O_p(1/\sqrt{n})$ rate; with probability tending to one it has the representation $\mathbf{p}(\widehat{\boldsymbol{\theta}}_n)$, where $|\widehat{\boldsymbol{\theta}}_n| = O_p(1/\sqrt{n})$.

MAYBE BETTER TO GIVE THE ARGUMENT FOR KNOWN $\boldsymbol{\pi}_0$ FIRST. ALSO, NOTE CLASH OF NOTATION WITH H AND H_λ.

Before I state and prove a formal limit theorem, let me argue heuristically to suggest how local smoothness controls the finer behaviour of $\widehat{\mathbf{p}}_n$. The informal arguments will also establish needed notation.

From approximation <2> and the fact that $\mathbf{f}_n \approx \boldsymbol{\pi}_0$, we have

$$J_\lambda(\mathbf{f}_n, \mathbf{p}) \approx \sum\nolimits_i \frac{(f_{ni} - p_i)^2}{f_{ni}} \simeq \sum\nolimits_i \frac{(f_{ni} - p_i)^2}{\pi_{0i}} \qquad \text{for } \mathbf{p} \text{ near } \boldsymbol{\pi}_0.$$

If we define W to be the weight matrix $\operatorname{diag}(\sqrt{\pi_{01}}, \ldots, \sqrt{\pi_{0k}})$, the approximating quadratic becomes $|W^{-1}(\mathbf{f}_n - \mathbf{p})|^2$. In particular,

$$J_\lambda(\mathbf{f}_n, \mathbf{p}(\boldsymbol{\theta})) \approx Q_n(\boldsymbol{\theta}) = |\mathbf{X}_n - W^{-1}D\boldsymbol{\theta}|^2 \qquad \text{where } \mathbf{X}_n = W^{-1}(\mathbf{f}_n - \boldsymbol{\pi}_0).$$

The value $\widehat{\boldsymbol{\theta}}_n$ nearly minimizes the left-hand side over $\boldsymbol{\theta}$ in Θ_0. It should therefore also come close to minimizing $Q_n(\boldsymbol{\theta})$ over Θ_0. The global minimum of $Q_n(\boldsymbol{\theta})$ over the whole of $\mathbb{R}^s$ corresponds to the value $\boldsymbol{\theta}_n^*$ for which

<10> $$W^{-1}D\boldsymbol{\theta}_n^* = H\mathbf{X}_n,$$

where H is the matrix that projects $\mathbb{R}^k$ orthogonally onto the s-dimensional subspace $\operatorname{sp}(W^{-1}D)$ spanned by the columns of $W^{-1}D$. Because $\mathbf{f}_n - \boldsymbol{\pi}_0 = O_p(1/\sqrt{n})$ and the matrix $W^{-1}D$ is of full rank, $\boldsymbol{\theta}_n^*$ is also of order $O_p(1/\sqrt{n})$. In particular, with probability tending to one it lies in the neighborhood Θ_0.

The rate of convergence for $\boldsymbol{\theta}_n^*$ also gives

$$\begin{aligned}\sqrt{n}(\mathbf{p}(\boldsymbol{\theta}_n^*) - \boldsymbol{\pi}_0) &= \sqrt{n}D\boldsymbol{\theta}_n^* + o_p(1) \\ &= W\mathbf{X}_n + o_p(1).\end{aligned}$$

From part (ii) of Lemma <3>,

$$\mathbf{X}_n \rightsquigarrow N(\mathbf{0}, I_k - \boldsymbol{\Delta}\boldsymbol{\Delta}'),$$

where $\boldsymbol{\Delta}'$ denotes the unit row vector $(\sqrt{\pi_{01}}, \ldots, \sqrt{\pi_{0k}})$. The matrix $I_k - \boldsymbol{\Delta}\boldsymbol{\Delta}'$ represents the projection orthogonal to the one dimensional space spanned by $\boldsymbol{\Delta}$. The limiting normal distribution for $\mathbf{X}_n$ is that of $(I_k - \boldsymbol{\Delta}\boldsymbol{\Delta}')\mathbf{Z}$ for a vector $\mathbf{Z}$ of independent $N(0,1)$ random variables. Thus $\sqrt{n}(\mathbf{p}(\boldsymbol{\theta}_n^*) - \boldsymbol{\pi}_0)$ has limiting distribution $WH(I_k - \boldsymbol{\Delta}\boldsymbol{\Delta}')\mathbf{Z}$. The projection H kills the $\boldsymbol{\Delta}\boldsymbol{\Delta}'$ because $\boldsymbol{\Delta}$ is orthogonal to $\text{sp}(W^{-1}D)$:

$$\boldsymbol{\Delta}' W^{-1} D = \mathbf{1}' D = \mathbf{0}'.$$

The last equality—the orthogonality of the rows of D to the vector of ones—is a consequence of the differentiability assumption and the fact that the cell probabilities sum to one:

$$1 = \mathbf{1}'\mathbf{p}(\boldsymbol{\theta}) = 1 + \mathbf{1}' D\boldsymbol{\theta} + o(|\boldsymbol{\theta}|).$$

This gives $\mathbf{1}' D\boldsymbol{\theta} = 0$ for all $\boldsymbol{\theta}$ in $\mathbb{R}^s$, whence $\mathbf{1}' D = \mathbf{0}'$. In summary,

<11> $$\sqrt{n}(\mathbf{p}(\boldsymbol{\theta}_n^*) - \boldsymbol{\pi}_0) \rightsquigarrow H\mathbf{Z}.$$

The minimization over $\boldsymbol{\theta}$ has contributed the projection onto $\text{sp}(W^{-1}D)$.

When $\boldsymbol{\theta}_n^*$ does lie in Θ_0, it should be close to the near-minimizing value $\widehat{\boldsymbol{\theta}}_n$. This would give the approximation

$$\widehat{\mathbf{p}}_n \approx \boldsymbol{\pi}_0 + D\widehat{\boldsymbol{\theta}}_n \approx \boldsymbol{\pi}_0 + D\widehat{\boldsymbol{\theta}}_n^* = \boldsymbol{\pi}_0 + WHW^{-1}(\mathbf{f}_n - \boldsymbol{\pi}_0).$$

If these approximations are to be believed we should have

$$\sqrt{n}(\widehat{\mathbf{p}}_n - \boldsymbol{\pi}_0) \rightsquigarrow WH\mathbf{Z}$$

and

$$\begin{aligned} n J_\lambda(\mathbf{f}_n, \widehat{\mathbf{p}}_n) &\approx n\|\mathbf{X}_n - W^{-1}D\widehat{\boldsymbol{\theta}}_n^*\|^2 \\ &= \|(I - H)\mathbf{X}_n\|^2 \\ &\rightsquigarrow \|(I - H)(I - \boldsymbol{\Delta}\boldsymbol{\Delta}')\mathbf{Z}\|^2 \end{aligned}$$

The product $(I - H)(I - \boldsymbol{\Delta}\boldsymbol{\Delta}')$ is an orthogonal projection onto a subspace of dimensional $k - s - 1$: the second factor kills the component in the $\boldsymbol{\Delta}$ direction and then the first factor kills the components in the orthogonal subspace $\text{sp}(W^{-1}D)$. The second factor corresponds to the constraint that the cell counts S_{ni} sum to n; the first factor corresponds to the s parameters fitted (locally) by $\mathcal{P}$. The squared length of the projection $(I - H)(I - \boldsymbol{\Delta}\boldsymbol{\Delta}')\mathbf{Z}$ has the desired χ^2_{k-s-1} distribution.

To make the heuristic arguments rigorous we need to establish uniform probabilistic bounds on the errors of approximation. A slight extension of the stochastic order notation will save us from much tedious deatil, even if it does increase the risk of error—even more of the supporting mathematics will be hidden behind a few dangerously convenient words. We will need to make assertions of the form

$$G_n(t) = O_p(\alpha_n) \qquad \text{uniformly over } A_n,$$

for random processes $\{G_n(t) : t \in T\}$ and various subsets A_n of the parameter set T, with α_n possibly random. Such an assertion means

$$\sup_{t \in A_n} |G_n(t)| = O_p(\alpha_n).$$

That is, for each $\epsilon > 0$ there exists a constant M_ϵ and an integer $n_0(\epsilon)$ such that

$$\mathbb{P}\left\{ \sup_{t \in A_n} |G_n(t)| > M_\epsilon \alpha_n \right\} < \epsilon \qquad \text{for } n \geq n_0(\omega).$$

Ignoring an event of small probability, we may eventually assume that $|G_n(t)|$ is bounded by a fixed multiple of α_n, simultaneously for all t in A_n.

Notice that we need $\widehat{\mathbf{p}}_n$ to be slightly closer to minimizing the criterion function $J_\lambda(\mathbf{f}_n, \cdot)$ over $\mathcal{P}$—within $o_p(1/n)$ rather than the $O_p(1/n)$ from Theorem <7>. Such an estimator will be referred to, with only a slight risk of ambiguity, as a power-divergence estimator.

<12> **Theorem.** *Assume the $\mathcal{M}(n, \boldsymbol{\pi}_0)$ model with $\boldsymbol{\pi}_0$ a point of $\mathcal{P}$ having strictly positive components. Suppose $\widehat{\mathbf{p}}_n$ is an element of $\mathcal{P}$ for which*

(i) $J_\lambda(\mathbf{f}_n, \widehat{\mathbf{p}}_n) \le \inf_{\mathbf{p}\in\mathcal{P}} J_\lambda(\mathbf{f}_n, \mathbf{p}) + o_p(1/n)$.

Suppose

(ii) $\mathcal{P}$ is locally s-dimensional near $\boldsymbol{\pi}_0$, in the sense of the Local Smoothness Assumption <8>, with matrix of derivatives D.

Then

(iii) $\sqrt{n}(\widehat{\mathbf{p}}_n - \boldsymbol{\pi}_0) = \sqrt{n} W H W^{-1}(\mathbf{f}_n - \boldsymbol{\pi}_0) + o_p(1)$, *which has an asymptotic normal distribution $N(\mathbf{0}, WHW^{-1})$, where W denotes the matrix $\operatorname{diag}(\sqrt{\pi_{01}}, \ldots, \sqrt{\pi_{0k}})$, and $\boldsymbol{\Delta}'$ denotes the unit vector $(\sqrt{\pi_{01}}, \ldots, \sqrt{\pi_{0k}})$, and H denotes the matrix for orthogonal projection onto the columns of $W^{-1}D$.*

(iv) $n J_\lambda(\mathbf{f}_n, \widehat{\mathbf{p}}_n) \rightsquigarrow \chi^2_{k-s-1}$.

Proof. We need to establish approximations that hold uniformly over a ball C with center $\mathbf{0}$ in $\mathbb{R}^s$ and radius δ_n of order $O_p(1/\sqrt{n})$. With probability tending to one, such a C will be contained within the neighborhood Θ_0 where the local parametrization is valid. Choose δ_n large enough to ensure that C contains both the $\widehat{\boldsymbol{\theta}}_n$ for which $\widehat{\mathbf{p}}_n = \mathbf{p}(\widehat{\boldsymbol{\theta}}_n)$ and the $\boldsymbol{\theta}^*_n$ that minimizes the quadratic $Q_n(\boldsymbol{\theta})$ over all of $\mathbb{R}^s$. This is possible for $\widehat{\boldsymbol{\theta}}_n$ by virtue of Theorem <7> and the inequality <9>; it is possible for $\boldsymbol{\theta}^*_n$ because equality <10> has the unique solution

$$\boldsymbol{\theta}^*_n = (D'D)^{-1}D'WH\mathbf{X}_n = O_p(1/\sqrt{n}).$$

Comment further on subtlety about nonexistence of $\boldsymbol{\theta}^*_n$ with small probability?

The inverse matrix $(D'D)^{-1}$ exists because D has full rank.

Uniformly over C,

$$|\mathbf{f}_n - \mathbf{p}(\boldsymbol{\theta})| \le |\mathbf{f}_n - \boldsymbol{\pi}_0| + C_2|\boldsymbol{\theta}| = O_p(1/\sqrt{n}).$$

It follows from inequality <2> that

$$J_\lambda(\mathbf{f}_n, \mathbf{p}(\boldsymbol{\theta})) - \sum\nolimits_i \frac{(f_{ni} - p_i(\boldsymbol{\theta}))^2}{f_{ni}} = O_p(n^{-3/2}) \qquad \text{uniformly on } C.$$

Replacing the f_{ni} in the denominator by π_{0i} we perturb the summand by at most

$$O_p((|f_{ni} - p_i(\boldsymbol{\theta})|^2))O_p((|f_{ni} - \pi_{0i}|)) = o_p(1/n) \qquad \text{uniformly on } C.$$

Here I have used the weak fact that $|f_{ni} - \pi_{0i}|$ is of order $o_p(1)$, rather than the stronger $O_p(1/\sqrt{n})$, in order to stress the idea that the f_{ni} in the denominator does not play a crucial role; any quantity lying within $o_p(1)$ of π_{0i} would suffice. The numerator in the i^{th} summand has the form

$$(f_{ni} - \pi_{0i} - (D\boldsymbol{\theta})_i - O_p(1/\sqrt{n}))^2 \qquad \text{uniformly on } C.$$

The contributions from the $O_p(1/\sqrt{n})$ contribute at most $o_p(1/n)$ to the expansion of this quadratic.

The combined effect of replacing $\mathbf{p}(\boldsymbol{\theta})$ by its linear approximation and replace the f_{ni} by π_{0i} in the denominator is a uniform quadratic approximation,

<13> $$J_\lambda(\mathbf{f}_n, \mathbf{p}(\boldsymbol{\theta})) = Q_n(\boldsymbol{\theta}) + o_p(1/n) \qquad \text{uniformly on } C.$$

If we expand $Q_n(\boldsymbol{\theta} - \boldsymbol{\theta}^*_n + \boldsymbol{\theta}^*_n)$ as a quadratic in $\boldsymbol{\theta} - \boldsymbol{\theta}^*_n$, the cross product term disappears (otherwise $\boldsymbol{\theta}^*_n$ could not be a global minimizing value) leaving

$$Q_n(\boldsymbol{\theta}) = Q_n(\boldsymbol{\theta}^*_n) + |W^{-1}D(\boldsymbol{\theta} - \boldsymbol{\theta}^*_n)|^2 \qquad \text{for all } \boldsymbol{\theta} \in \mathbb{R}^s.$$

When $\boldsymbol{\theta}_n^*$ lies in Θ_0 it defines a member of $\mathcal{P}$ for which

$$J_\lambda(\mathbf{f}_n, \widehat{\mathbf{p}}_n) \le J_\lambda(\mathbf{f}_n, \mathbf{p}(\boldsymbol{\theta}_n^*)) + o_p(1/n).$$

Replacing both J_λ terms by means of approximation <13> then consolidating $o_p(1/n)$ terms, we get

$$Q_n(\widehat{\boldsymbol{\theta}}_n) \le Q_n(\boldsymbol{\theta}_n^*) + o_p(1/n).$$

That is,

$$|W^{-1}D(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^*)|^2 \le o_p(1/n).$$

It would perhaps be more precise to make such a set of comparisons only with the explicit stipulation that $\widehat{\mathbf{p}}_n \in \mathcal{P}_0$ and $\boldsymbol{\theta}_n^* \in \Theta_0$. The final inequality offers us a way to avoid those details. The $o_p(\cdot)$ acknowledges existence of a set where the $1/n$ rate assertion has no effect; that set covers all the realizations where the comparison argument is not strictly justified.

The rest of the argument is easy. The full rank assumption on D ensures existence of a positive constant C_3 for which $|W^{-1}D\mathbf{t}| \ge C_3|\mathbf{t}|$ for all $\mathbf{t} \in \mathbb{R}^s$. Consequently,

$$\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^* = o_p(1/\sqrt{n}).$$

The differentiability then gives

$$\widehat{\mathbf{p}}_n = \mathbf{p}(\boldsymbol{\theta}_n^*) + o_p(1/\sqrt{n}),$$

which translates into assertion (iii) of the Theorem. It also gives

$$\begin{aligned} J_\lambda(\mathbf{f}_n, \widehat{\mathbf{p}}_n) &= Q_n(\boldsymbol{\theta}_n) + o_p(1/n) \\ &= Q_n(\boldsymbol{\theta}_n^*) + o_p(1/n) \\ &= |(I - H)\mathbf{X}_n|^2 + o_p(1/n). \end{aligned}$$

As explained during the heuristic discussion leading up to the Theorem, when multiplied by n the last quadratic expression has the asserted limiting χ^2 distribution. The $o_p(1)$ perturbation does not affect that limit. □

Notice that the choice of λ had no effect on the last proof, except for the hidden control over various constants. For every λ, the estimator $\widehat{\mathbf{p}}_n(\lambda)$ has the asymptotic representation asserted by (iii) in Theorem <12>; the standardized estimators for different λ values are equal up to the $o_p(1)$ error terms. That is, for all λ and λ',

$$\widehat{\mathbf{p}}_n(\lambda) - \widehat{\mathbf{p}}_n(\lambda') = o_p(1/\sqrt{n}),$$

as asserted at the end of Section 1. Similarly, the behaviour of $nJ_\lambda(\mathbf{f}_n, \widehat{\mathbf{p}}_n(\lambda))$ is determined by the approximation

$$\widehat{\mathbf{p}}_n(\lambda) = \mathbf{p}(\boldsymbol{\theta}_n^*) + o_p(1/\sqrt{n}).$$

The $\boldsymbol{\theta}_n^*$ does not depend on λ. It follows that, under the conditions of the Theorem, $nJ_\lambda(\mathbf{f}_n, \widehat{\mathbf{p}}_n(\lambda')) \rightsquigarrow \chi^2_{k-s-1}$, for all λ and λ'.

What happens if we use an estimator for $\boldsymbol{\pi}_0$ that is not defined by near minimization of one of the $J_\lambda(\mathbf{f}_n, \cdot)$ functions? More specifically, suppose $\tilde{\mathbf{p}}_n$ has the form $\mathbf{p}(\tilde{\boldsymbol{\theta}}_n)$ with probability tending to one, where $\tilde{\boldsymbol{\theta}}_n = O_p(1/\sqrt{n})$. If $\tilde{\boldsymbol{\theta}}_n$ does not lie within $o_p(1/\sqrt{n})$ of the $\boldsymbol{\theta}_n^*$ that minimizes Q_n, the term $n|W^{-1}D(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^*)|^2$ will not converge in probability to zero; it will inflate the approximating $nQ_n(\boldsymbol{\theta}_n^*)$, which has the limiting χ^2_{k-s-1} distribution, by a quantity that does not disappear asymptotically.

As we will see in Section 5, the property $\tilde{\boldsymbol{\theta}}_n = \boldsymbol{\theta}_n^* + o_p(1/\sqrt{n})$ corresponds to the property called *efficiency*.

3. Power-divergence estimators under local alternatives

How does the $\widehat{\mathbf{p}}_n$ from Theorem <12> behave if the fixed $\boldsymbol{\pi}_0$ is replaced by a $\boldsymbol{\pi}_n$ that changes with n? Such a question arises when one considers the power of goodness-of-fit test. It also appears as part of the definition of a *regular estimator*, a restriction that will be introduced in Section 5 as part of the program to rescue the flawed concept of efficiency.

Consider first the behaviour of $\widehat{\mathbf{p}}_n$ under a model $\mathcal{M}(n, \boldsymbol{\pi}_1)$, for a fixed $\boldsymbol{\pi}_1 \notin \mathcal{P}$ for which $\inf\{\|\boldsymbol{\pi}_1 - \mathbf{p}\| : \mathbf{p} \in \mathcal{P}\} > 0$. The last assumption eliminates the awkward posibility that $\boldsymbol{\pi}_1$ might be a limit of points in $\mathcal{P}$ without actually belonging to $\mathcal{P}$ itself. Let B be a small open ball that is disjoint from $\mathcal{P}$, centered at $\boldsymbol{\pi}_1$. Then, by an argument similar to the proof of Theorem <6>, when $\mathbf{f}_n \in B$ we have

$$\begin{aligned} J_\lambda(\mathbf{f}_n, \widehat{\mathbf{p}}_n) &\geq \inf_{\mathbf{p}\in\partial B} J_\lambda(\mathbf{f}_n, \mathbf{p}) \\ &\to \inf_{\mathbf{p}\in\partial B} J_\lambda(\boldsymbol{\pi}_1, \mathbf{p}) \qquad \text{almost surely} \\ &> 0. \end{aligned}$$

The rescaled statistic $n J_\lambda(\mathbf{f}_n, \widehat{\mathbf{p}}_n)$ diverges to infinity; it no longer has a proper limiting distribution. A test of fit will reject the model $\mathcal{P}$ with probability tending to one, under $\mathcal{M}(n, \boldsymbol{\pi}_1)$. The test has power tending to one at $\boldsymbol{\pi}_1$.

Discrimination between $\boldsymbol{\pi}_1$ and $\mathcal{P}$ is to easy a task for any halfway decent estimator or testing procedure. We need to pose a more difficult task if we are to compare different procedures. We must consider power at *local alternatives*, which change with n. Replace $\boldsymbol{\pi}_1$ by a sequence $\{\boldsymbol{\pi}_n\}$ that moves towards $\mathcal{P}$ at a rate that allows good procedures to discriminate between $\mathcal{P}$ and $\boldsymbol{\pi}_n$, but not with probability tending to one. Alternatives of the form

<14> $$\boldsymbol{\pi}_n = \boldsymbol{\pi}_0 + W\boldsymbol{\delta}_n/\sqrt{n} \qquad \text{with } \boldsymbol{\delta}_n \to \boldsymbol{\delta},$$

lend themselves to easy limiting calculations. Here $\boldsymbol{\pi}_0$ and W have the same meaning as in Theorem <12>. By building the W into the perturbation we eliminate a number of W^{-1} factors in later formulae.

An easy application of MCLT for triangular arrays shows that

$$\sqrt{n}(\mathbf{f}_n - \boldsymbol{\pi}_n) \rightsquigarrow N(0, V) \qquad \text{under } \mathcal{M}(n, \boldsymbol{\pi}_n),$$

with the same limiting variance matrix $V = \operatorname{diag}(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0\boldsymbol{\pi}_0'$ as before. The asymptotic arguments from Section 2 were driven by the behaviour of the standardized random vector $\mathbf{X}_n = W^{-1}\sqrt{n}(\mathbf{f}_n - \boldsymbol{\pi}_0)$. Under $\mathcal{M}(n, \boldsymbol{\pi}_0)$ it has a limiting $N(\mathbf{0}, I - \boldsymbol{\Delta}\boldsymbol{\Delta}')$ distribution. The local alternative adds a shift:

$$\mathbf{X}_n \rightsquigarrow N(\boldsymbol{\delta}, I - \boldsymbol{\Delta}\boldsymbol{\Delta}') \qquad \text{under } \mathcal{M}(n, \boldsymbol{\pi}_n).$$

At this point you should reexamine the proofs of Theorems <7> and <12> to convince yourself that most of the argument required only that $\mathbf{f}_n - \boldsymbol{\pi}_0$ be of order $O_p(1/\sqrt{n})$); the limiting distribution was needed only in the last paragraph of the second theorem. Under $\mathcal{M}(n, \boldsymbol{\pi}_n)$ it is still true that

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_n &= \boldsymbol{\theta} *_n + o_p(1/\sqrt{n}), \\ \sqrt{n}W^{-1}(\widehat{\mathbf{p}}_n - \boldsymbol{\pi}_0) &= H\mathbf{X}_n + o_p(1), \\ n J_\lambda(\mathbf{f}_n, \widehat{\mathbf{p}}_n) &= \|(I - H)\mathbf{X}_n\|^2 + o_p(1). \end{aligned}$$

One should be a little careful with the interpretation of the $o_p(\cdot)$ quantities here. They represent random vectors and random variables that are probabilistically small under the model $\mathcal{M}(n, \boldsymbol{\pi}_n)$. The same assertions were established under $\mathcal{M}(n, \boldsymbol{\pi}_0)$ in the proofs of the theorems. If we could pass directly from $o_p(\cdot)$ under $\mathcal{M}(n, \boldsymbol{\pi}_0)$ to $o_p(\cdot)$ under $\mathcal{M}(n, \boldsymbol{\pi}_n)$, there would be no need to

reexamine the proofs for subtle consequences of calculations carried out under the alternative models. It will turn out that the dual interpretation of $o_p(\cdot)$ is justified by Le Cam's concept of *contiguity*, which will be discussed in more detail in Section 4.

The changes in the conclusions of Theorem <12> are due only to changes in the limiting distributions of the functions of $\mathbf{X}_n$:

$$H\mathbf{X}_n \rightsquigarrow H\boldsymbol{\delta} + H\mathbf{Z},$$
$$\|(I_k - H)\mathbf{X}_n\|^2 \rightsquigarrow \|(I_k - H)\boldsymbol{\delta} + (I_k - H)(I_k - \boldsymbol{\Delta}\boldsymbol{\Delta})\mathbf{Z}\|^2,$$

where $\mathbf{Z}$ has a $N(\mathbf{0}, I_k)$ distribution, as before.

The quadratic form in $\mathbf{Z}$ has a noncentral chi-square distribution $\chi^2_{k-s-1}(\gamma)$ with noncentrality parameter $\gamma = \|(I_k - H)\boldsymbol{\delta}\|$. That is, it has the same distribution as

$$(Y_1 + \gamma)^2 + Y_2^2 + \ldots + Y_{k-s-1}^2,$$

with the $\{Y_i\}$ independent $N(0, 1)$ random variables. The tail probabilities

$$\beta(t, \gamma) = \mathbb{P}\{\chi^2_{k-s-1}(\gamma) \geq t\}$$

increase with γ. For a formal goodness-of-fit test, one chooses t to make $\beta(t, 0)$ a prescribed small value (the size of the test). Then $\beta(t, \gamma)$ becomes the asymptotic power for the local alternatives $\{\boldsymbol{\pi}_n\}$.

Notice that the asymptotic power depends on $\boldsymbol{\delta}$ only through its component $(I_k - H)\boldsymbol{\delta}$ orthogonal to $\text{sp}(W^{-1}D)$. In particular, the test has no asymptotic power for alternatives with $\boldsymbol{\delta} \in \text{sp}(W^{-1}D)$. It would be most unfortunate if this were not so, because such $\boldsymbol{\delta}$ could correspond to $\{\boldsymbol{\pi}_n\}$ approaching $\boldsymbol{\pi}_0$ along the surface $\mathcal{P}$; the alternatives would then part of the model, and no test should be able to distinguish between $\mathcal{P}$ and such a $\{\boldsymbol{\pi}_n\}$.

The effect of the $\boldsymbol{\delta}_n$ perturbation also shows up in the asymptotic behaviour of $\widehat{\mathbf{p}}_n$: under $\mathcal{M}(n, \boldsymbol{\pi}_n)$,

$$\sqrt{n}(\widehat{\mathbf{p}}_n - \boldsymbol{\pi}_0) = WH\mathbf{X}_n + o_p(1) \rightsquigarrow WH\boldsymbol{\delta} + H\mathbf{Z}.$$

If we center at $\boldsymbol{\pi}_n$, which is the parameter that $\widehat{\mathbf{p}}_n$ really estimates, the $\boldsymbol{\delta}$ disappears: using the fact $\boldsymbol{\delta}_n = \boldsymbol{\delta} + o(1)$, we have

$$\begin{aligned}\sqrt{n}(\widehat{\mathbf{p}}_n - \boldsymbol{\pi}_n) &= WH\mathbf{X}_n - W\boldsymbol{\delta}_n + o_p(1)\\ &\rightsquigarrow WH\boldsymbol{\delta} + WH\mathbf{Z} - W\boldsymbol{\delta}.\end{aligned}$$

If $\boldsymbol{\pi}_n \in \mathcal{P}$, the vector $\boldsymbol{\delta}$ lies in $\text{sp}(W^{-1}D)$ and $H\boldsymbol{\delta} = \boldsymbol{\delta}$. Thus $\sqrt{n}(\widehat{\mathbf{p}}_n - \boldsymbol{\pi}_n)$, has the same limiting distribution for all local alternatives $\boldsymbol{\pi}_n = \boldsymbol{\pi}_0 + \boldsymbol{\delta}_n/\sqrt{n}$ that approach $\boldsymbol{\pi}_0$ through the model. An estimator with this property will be called *regular* (in Hájek's sense). Section 5 will develop the concept of efficiency for regular estimators.

4. The conditional Poisson model

This Section unedited. It contains many errors.

In the multinomial model, the constraint that the cell counts sum to k has the asymptotic effect of removing one degree of freedom from the limiting χ^2 distribution. It is also the reason for the $I_k - \boldsymbol{\Delta}\boldsymbol{\Delta}'$ factor in the limiting variance of the standardized counts.

In other applications of the χ^2-test, further linear constraints are placed on the cell counts, resulting in further reductions in the limiting degrees of freedom.

<15> **Example.** In a two-way table, the cells are arranged into a rectangular array, with the rows and columns corresponding to different partitions of the population. The table of Greenwood & Yule (1915), which cross-classified individuals as either inoculated or uninoculated against cholera and as either attacked or not attacked by the disease, was cited by Fisher (1922) as a case where degrees of freedom must be adjusted.

CHOLERA	Not Attacked	Attacked	Total
Inoculated	1625	5	1630
Not	1022	11	1033
Total	2647	16	2663

More generally, if cell counts S_{ij} for $i = 1, \ldots, r$ and $j = 1, \ldots, c$ are analyzed with the marginal totals S_{i+} and S_{+j} treated as fixed (that is, the analysis is done conditional on those marginal totals), the degrees of freedom are reduced by $r + c - 1$, not just by the 1 due to the fixed sample size. Notice that one of the marginal constraints is redundant, because $\sum_i S_{i+} = \sum_j S_{+j} = n$; there are only $r + c - 1$ linearly independent equalities involved. □

Fisher recognized that the fixed marginal totals play the same asymptotic role as the estimation of parameters, which he saw as another way to force the estimated cell probabilities to come closer to the observed frequencies. He referred to estimation as a "method for reconstructing the population". He clearly understood the (asymptotic) equivalence of conditioning and parameter estimation. His concept of degrees of freedom recognized estimation or marginal conditioning as constraints that forced the vector $\widehat{\mathbf{p}}_n$ to lie in lower dimensional subspaces (asymptotically).

> . . . in all cases linear restrictions imposed upon the frequencies of the sampled population, by our methods of reconstructing that population, have exactly the same effect upon the distributions of χ^2 as have restrictions placed upon the cell contents of the sample.
>
> [Fisher (1922), page 92]

> In cases where the population, with which the sample is compared in calculating χ^2 has been itself reconstructed from the sample, we must also take account of the number of degrees of freedom absorbed in this process of reconstruction. The two cases of widest application were (i) contingency tables in which the population is reconstructed by assigning to the margins the frequencies observed in the sample, and (ii) frequency curves constructed to agree with the sample in respect of one or more moments. The common sense of this correction lies in the fact that when the population with which the sample is compared has been artificially identified with the sample in certain respects, such as the marginal frequencies, or the moments, we shall evidently make an exaggerated estimate of the closeness of agreement between sample and population, if we regard the sample as an unselected sample of a population known *à priori* . It was possible to show that the distribution was in fact that which arises when from any population a large number of samples are taken, and only those samples chosen which agree with the population in (say) the marginal frequencies; these samples compared to the true population will give values of χ^2 distributed in the same manner as in the practical case in which we compare any sample with a population artificially constructed from it.
>
> [Fisher (1923), page 139]

Fisher also gave a geometric interpretation to explain the large sample behaviour of the goodness-of-fit statistic. (He wrote x for the difference between the observed counts and the (estimated) expected counts.)

> The most general way of proving this result consists in regarding the values of x (above) as independent co-ordinates in generalised space; then owing to the linear relations by which the deviations are restricted, for example that the marginal totals of the population should be equal to those observed, all possible sets of observations will lie relative to the centre of the distribution, specified by the assumed population, in a plane space, of the same number of dimensions as there are degrees of freedom.
>
> [Fisher (1922, page 88)]

The interesting point here is the idea of identifying dependence as a consequence of constraining a vector of independent quantities, which is the main topic of this section.

Where do those independent coordinates come from? Fisher (1922) introduced the device of treating the multinomial as a conditioned set of independent Poissons. Suppose $Y_1, \ldots, Y_k$ are independent, with Y_i distributed Poisson (λ_i). Then $Y_1 + \ldots + Y_k$ has a Poisson (λ) distribution, with $\lambda = \lambda_1 + \ldots + \lambda_k$, and

$$\begin{aligned}
&\mathbb{P}\{Y_1 = y_1, \ldots, Y_k = y_k \mid Y_1 + \ldots + Y_k = n\} \\
&= \prod_i \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} \Big/ \frac{e^{-\lambda}\lambda^n}{n!} \qquad \text{if } y_1 + \ldots + y_k = n \\
&= \frac{n!}{y_1! \ldots y_n!}\, p_1^{y_1} \ldots p_k^{y_k} \qquad \text{where } p_i = \lambda_i/\lambda.
\end{aligned}$$

The conditioning on the sum accounts for the one degree of freedom lost in the $\mathcal{M}(n, \mathbf{p})$ model. The vector of standardized counts $X_i = (Y_i - \lambda_i)/\sqrt{\lambda_i}$ has an asymptotic $N(\mathbf{0}, I_k)$ distribution. For the multinomial model, the limiting distribution would appear to be that of the $N(\mathbf{0}, I_k)$ conditioned on the value of a particular linear combination. If we choose the λ_i so that $\sum_i \lambda_i = n$ the constraint is simple: $\sum_i X_i = 0$. Clearly we are free to choose the λ_i in this way, because the $\{p_i\}$ depend only on the ratios of the $\{\lambda_i\}$. The idea can be taken further. In an excellent paper summarizing the state of the χ^2 art, Cochran (1952) noted:

> This approach also makes it clear that if further homogeneous linear restrictions are imposed [on the $Y_i - \lambda_i$], either by the structure of the data or in the process of fitting, the effect will merely be to reduce the degrees of freedom in χ^2.
>
> [Cochran (1952), page 319]

For example, in the two-way table, a multinomial model where marginal totals are used to estimate cell probabilities under an independence hypothesis leads to a limiting $\chi^2_{(r-1)(c-1)}$ distribution for the goodness-of-fit statistics. An analysis treating the marginal totals as fixed would lead to the same distribution; and so would an analysis with row marginals fixed but with column marginals used to estimate the remaining unknown parameters. All analyses fit into the same framework of a table of independent Poisson counts, conditioned on certain linear combinations and with possible parameter estimation.

Unfortunately, there exists quite a technical gap between what is intuitively clear and what is mathematically provable. Haberman (1974, Theorem 1.1) has established a rigorous result that exposes a number of the hidden subtleties.

The limit theory in Section 2—Theorem <12> in particular—was driven by the convergence in distribution of the random vector $W^{-1}\sqrt{n}(\mathbf{f}_n - \boldsymbol{\pi}_0)$.

Under the Poisson model with $\pi_{oi} = \lambda_i/\lambda$ and $\lambda = n$, the components of this vector are precisely the standardized counts $(Y_i - \lambda_i)/\sqrt{\lambda_i}$. If we want asymptotic theory for the conditional Poisson model, we have only to establish a conditional limit theorem for the random vector

$$\mathbf{X} = \Lambda^{-1}(\mathbf{Y} - \boldsymbol{\lambda}) \qquad \text{where } \Lambda = \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_k}).$$

The limit should be taken as the sum $n = \sum \lambda_i$ tends to infinity.

What sort of conditioning constraint should we consider? In the example with the two-way table, the various marginal totals all correspond to quantities of the form $\mathbf{A} \,.\, \mathbf{Y}$, where $\mathbf{A}$ is a vector with (nonnegative) integer components. Moreover, the Poisson means could be chosen so the constraints were

<16> $$\mathbf{A}_j \,.\, \mathbf{Y} = \mathbf{A}_j \,.\, \boldsymbol{\lambda} \qquad \text{for } j = 1, \ldots, k - s,$$

for linearly independent, integer vectors $\{\mathbf{A}_j\}$. If we apply the Gram-Schmidt procedure to the expanded collection $\mathbf{A}_1, \ldots, \mathbf{A}_{k-s}, \mathbf{e}_1, \ldots, \mathbf{e}_k$, where $\mathbf{e}_i$ is the unit vector with a 1 in the ith position, we construct an orthogonal basis for $\mathbb{R}^k$ consisting of vectors with rational coordinates. To see this, consider the first few steps in the procedure. Let $\mathbf{A}_1$ have squared length α_1, an integer. Then the component of $\mathbf{A}_2$ orthogonal to $\mathbf{A}_1$ is

$$\mathbf{B}_2 = \mathbf{A}_2 - \frac{1}{\alpha_1}(\mathbf{A}_2 \,.\, \mathbf{A}_1)\mathbf{A}_1,$$

which certainly has all coordinates rational. The squared length α_2 of $\mathbf{B}_2$ is rational. In the direction orthogonal to both $\mathbf{A}_1$ and $\mathbf{A}_2$ the vector $\mathbf{A}_3$ has component

$$\mathbf{B}_3 = \mathbf{A}_3 - \frac{1}{\alpha_1}(\mathbf{A}_3 \,.\, \mathbf{A}_1)\mathbf{A}_1 - \frac{1}{\alpha_2}(\mathbf{A}_3 \,.\, \mathbf{B}_2)\mathbf{B}_2,$$

which has rational coordinates. And so on. Multiplication of the $\{\mathbf{B}_i\}$ by suitably large integers will then produce an orthogonal basis for $\mathbb{R}^k$ with vectors having only integer coordinates.

Let $\mathbf{V}_1, \ldots, \mathbf{V}_s$ denote the basis vectors that span the subspace $\mathcal{L}$ that is orthogonal to $\operatorname{sp}(\mathbf{A}_1, \ldots, \mathbf{A}_{k-s})$. Then the linear constraints <16> have the interpretation that $\mathbf{Y} - \boldsymbol{\lambda}$ should belong to $\mathcal{L}$.

Notice that $\mathcal{L}$ must contain points from the integer lattice $\mathbb{Z}^k$; integer linear combinations of the $\{\mathbf{V}_i\}$ have only integer components. If $\boldsymbol{\lambda}$ has integer coordinates, as will be assumed from now on, the set $\boldsymbol{\lambda} \oplus \mathbb{Z}^k$ will contain lattice points to which $\mathbf{Y}$ attaches nonzero probability, when $\min \lambda_i$ is large enough. We will be able to make use of the elementary notion of conditional probability, and not need to worry about abstract methods of conditioning. To avoid nonnormal limiting behaviour (Problem??) we should also require that none of the λ_i increases much more slowly than n. Remember that Λ is the diagonal matrix $\operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_k})$.

<17> **Theorem.** *Let $\mathcal{L}$ denote the subspace of $\mathbb{R}^k$ spanned by vectors $\mathbf{V}_1, \ldots, \mathbf{V}_k$ with integer coordinates. Let $\mathbf{Y}$ be a random k-vector of independent Poisson variables whose vector of means $\boldsymbol{\lambda}$ has integer coordinates for which $\sum_i \lambda_i = n$ and $\min_i \lambda_i/n$ stays bounded away from zero. Then the conditional distribution of the vector of standardized counts $(Y_i - \lambda_i)/\sqrt{\lambda_i}$, given that $\mathbf{Y} - \boldsymbol{\lambda}$ belongs to $\mathcal{L}$, converges to a $N(\mathbf{0}, I_k)$, conditioned to lie in $\Lambda^{-1}\mathcal{L}$.*

Perhaps it is imprecise to assert convergence to a conditioned normal distribution depending on $\boldsymbol{\lambda}$. Nothing in the hypotheses of the theorem requires the subspace $\Lambda^{-1}\mathcal{L}$ to settle down in any of the usual senses. The limit

distribution is not fixed; it might change with $\boldsymbol{\lambda}$. The assertion should be interpreted to mean

$$\mathbb{P}((h(\Lambda^{-1}(\mathbf{Y}-\boldsymbol{\lambda}))\mid \mathbf{Y}-\boldsymbol{\lambda}\in\mathcal{L}))-\mathcal{N}_{\boldsymbol{\lambda}}h\to 0$$

for each bounded, uniformly continuous $h(\cdot)$ on $\mathbb{R}^k$. The approximating probability distribution $\mathcal{N}_{\boldsymbol{\lambda}}$ has density $\psi(\mathbf{x})=(2\pi)^{-s/2}\exp(-\frac{1}{2}\|\mathbf{x}\|^2)$ with respect to Lebesgue measure σ on the subspace $\Lambda^{-1}\mathcal{L}$; it is the distribution of a $N(\mathbf{0}, I_k)$ random vector conditioned to lie in $\Lambda^{-1}\mathcal{L}$.

Rather than presenting a complete, detailed proof of Theorem <17>, I will give only an outline of a simplified version of Haberman's argument, leaving the technical details to the problems.

The main ingredient is a local limit theorem for Poisson probabilities, derived from Stirling's approximation, $m!\approx\sqrt{2\pi}\,m^{m+\frac{1}{2}}e^{-m}$. If Y has a Poisson(λ) distribution then, for an integer $m=\lambda+t\sqrt{\lambda}$,

$$\mathbb{P}\{Y=m\}\approx\frac{\phi(t)}{\sqrt{\lambda}}$$

where $\phi(\cdot)$ denotes the $N(0,1)$ density. More precisely, the excellent error bounds in Stirling's approximation give

<18> $$\mathbb{P}\{Y=m\}=\frac{\phi(t)}{\sqrt{\lambda}}\big(1+o(1)\big)\qquad\text{where } t=\frac{m-\lambda}{\sqrt{\lambda}}$$

uniformly over a $o(\sqrt{\lambda})$ range for t, as λ tends to infinity. For the vector $\mathbf{Y}$ of independent Poissons, multiplication of k such approximating factors leads to the local limit theorem, which relates the Poisson probability attached to a lattice point $\mathbf{m}$ and the normal integral over a small region:

<19> $$\mathbb{P}\{\mathbf{Y}=\mathbf{m}\}=(2\pi)^{-(k-s)/2}\psi(\mathbf{t})\prod\nolimits_i\lambda_i^{-1/2}\qquad\text{for } \mathbf{m}=\boldsymbol{\lambda}+\Lambda\mathbf{t},$$

uniformly over $\|\mathbf{t}\|=o(\sqrt{n})$. The product term on the right-hand side decreases like $n^{-k/2}$:

$$\prod\nolimits_i\lambda_i^{-\frac{1}{2}}=n^{-k/2}\alpha(\boldsymbol{\lambda}),$$

where $\alpha(\boldsymbol{\lambda})$ is a factor that stays bounded away from zero and infinity under the conditions of the theorem. Absorb the $(2\pi)^{-(k-s)/2}$ into this factor.

To approximate the probability of $\mathbf{Y}$ lying in some subset we have only to sum the terms from <19> over the lattice points in the set. Two slight complications arise with this idea. First, the approximation <19> only works for lattice points within $o(\sqrt{n})$ of $\boldsymbol{\lambda}$. Second, it is no mean feat to figure out which lattice points lie within a given set.

The first complication actually causes very little trouble. The Poisson(λ) probabilities drop off rapidly as m moves away from λ:

$$\mathbb{P}\{\mid Y-\lambda\mid > t\sqrt{\lambda}\}\le 2\exp((-\tfrac{1}{2}t^2B(t/\sqrt{\lambda}))),$$

where $B(\cdot)$ is a continuous function that decreases monotonely from $B(0)=1$. (See Problem ??) In particular, it is easy to find a t or order $o(\sqrt{\lambda})$ such that the tail probability decrease faster than any fixed power of $1/\lambda$. Correspondingly, it is easy to dispose of contributions from regions where <19> fails.

The second complication was foreseen by the specification of integer coordinates for the vectors $\mathbf{V}_1,\ldots,\mathbf{V}_s$ that span $\mathcal{L}$. For each $\mathbf{w}$ in $\mathbb{Z}^s$ define a cell

$$C(\mathbf{w})=\{(w_1+t_1)\mathbf{V}_1+\ldots+(w_s+t_s)\mathbf{V}_s : -\tfrac{1}{2}\le t_i<\tfrac{1}{2}\text{ for each } i\}.$$

The cells partition $\mathcal{L}$ into disjoint sets, each of which has the same s-dimensional Lebesgue measure. Because each cell is a translation of $C(\mathbf{0})$ by integer multiples of the $\mathbf{V}_i$ vectors, it must contain the same number of lattice

points from $\mathbb{Z}^k$. It might be difficult to calculate the actual number, N, of lattice points per cell, but its constancy simplifies the approximation of Poisson probabilities:

$$\begin{aligned}\mathbb{P}\{\mathbf{Y}-\boldsymbol{\lambda}\in C(\mathbf{w})\} &\approx N\mathbb{P}\{\mathbf{Y}-\boldsymbol{\lambda}=\mathbf{w}\}\\ &\approx n^{-k/2}\alpha(\boldsymbol{\lambda})N\psi(\Lambda^{-1}\mathbf{w}).\end{aligned} \qquad <20>$$

The linear transformation Λ^{-1} maps each cell $C(\mathbf{w})$ into a smaller cell in $\Lambda^{-1}\mathcal{L}$ with Lebesgue measure $\sigma(\Lambda^{-1}C(\mathbf{0})) = n^{-s/2}\beta(\boldsymbol{\lambda})$, where $\beta(\boldsymbol{\lambda})$ is another factor that stays bounded away from zero and infinity under the conditions of the theorem. Using the approximation

$$\psi(\Lambda^{-1}\mathbf{w})\sigma(\Lambda^{-1}C(\mathbf{w})) \approx \int\{\mathbf{x}\in\Lambda^{-1}C(\mathbf{w})\}\psi(\mathbf{x})\sigma(d\mathbf{x}),$$

we deduce from <20> that

$$\mathbb{P}\{\mathbf{Y}-\boldsymbol{\lambda}\in C(\mathbf{w})\} \approx n^{-(k-s)/2}\gamma(\boldsymbol{\lambda})\int\{\mathbf{x}\in\Lambda^{-1}C(\mathbf{w})\}\psi(\mathbf{x})\sigma(d\mathbf{x}),$$

where $\gamma(\boldsymbol{\lambda}) = N\alpha(\boldsymbol{\lambda})/\beta(\boldsymbol{\lambda})$. Of course, we should worry about bounds on the error. But if we ignore that difficulty while summing over $\mathbf{w}$ in a large region we get $\mathbb{P}\{\mathbf{Y}-\boldsymbol{\lambda}\in\mathcal{L}\} \approx n^{-(k-s)/2}\gamma(\boldsymbol{\lambda})$, because $\int\{\mathbf{x}\in\Lambda^{-1}\mathcal{L}\}\psi(\mathbf{x})\sigma(d\mathbf{x}) = 1$. Some attention to errors in the approximation would show that it omits only terms of order $o(n^{-(k-s)/2})$.

A similar argument would give an approximation to the contribution

$$\mathbb{P}h(\Lambda^{-1}(\mathbf{Y}-\boldsymbol{\lambda}))\{\mathbf{Y}-\boldsymbol{\lambda}\in C(\mathbf{w})\}$$

for a bounded, uniformly continuous $h(\cdot)$. Summation over all cells in a large region would then lead to

$$\mathbb{P}h(\Lambda^{-1}(\mathbf{Y}-\boldsymbol{\lambda}))\{\mathbf{Y}-\boldsymbol{\lambda}\in\mathcal{L}\} \approx n^{-(k-2)/2}\gamma(\boldsymbol{\lambda})\int\{\mathbf{x}\in\Lambda^{-1}\mathcal{L}\}h(\mathbf{x})\psi(\mathbf{x})\sigma(d\mathbf{x}),$$

with an error again of order $o(n^{-(k-s)/2})$. The assertion of the theorem is established when we take the ratio of the last two approximating quantities.

Remark: *Things to do: explain how the conditional limit theorem fits with the goodness-of-fit test with parameter estimation.*

5. Problems

THESE PROBLEMS HAVE NOT YET BEEN EDITED TO MAKE ANY SENSE. MORE PROBLEMS WILL BE ADDED.

[1] If $P_n \rightsquigarrow P$, show that $\liminf P_n\ell \geq P\ell$ for all lower-semicontinuous functions ℓ that are bounded below. [Hint:]

[2] Suppose $\{x_1,\ldots,x_n\}$ and $\{y_1,\ldots,y_n\}$ are positive numbers with $\sum_i x_i = \sum_i y_i = 1$. Show that $\sum_i x_i \log(x_i/y_i) \geq \frac{1}{2}\sum_i |x_i - y_i|^2$. [Hint: Try a Taylor expansion of $x\log(x/y)$ about y.]

[3] Suppose n objects are independently cross-classified according to one of r possible row characteristics and one of c possible column characteristics, with probability p_{ij} of an object ending up in row i and column j. (This gives a multinomial model with $k = rc$ cells.) Let $\mathcal{P}$ be the model that specifies independence between row and column characteristics:

$$p_{ij} = \left(\sum\nolimits_\ell p_{i\ell}\right)\left(\sum\nolimits_\ell p_{\ell j}\right)/rc \qquad \text{for all } i,\ j, \text{ under } \mathcal{P}.$$

Suppose the true $\boldsymbol{\pi}_0$ has all components positive.

(i) Show that $\mathcal{P}$ is locally s-dimensional near $\boldsymbol{\pi}_0$, for an s that you should specify.

(ii) Find the maximum likelihood estimator $\widehat{\mathbf{p}}_n$.

(iii) Find the limiting distribution of $\sqrt{n}(\widehat{\mathbf{p}}_n - \boldsymbol{\pi}_0)$.

(iv) Describe the space $\text{sp}(W^{-1}D)$. (Basis vectors?)

[4] Under the conditions specified by the Local Smoothness Assumption <8>, prove that there exist constants $0 < C_1 < C_2 < \infty$ such that

$$C_1|\boldsymbol{\theta}| \le |\mathbf{p}(\boldsymbol{\theta}) - \boldsymbol{\pi}_0| \le C_2|\boldsymbol{\theta}| \qquad \text{for } \boldsymbol{\theta} \in \Theta_0$$

[Hint: First consider the bound for $\boldsymbol{\theta}$ in a neighborhood of $\mathbf{0}$ small enough to make the $o(\|\boldsymbol{\theta}\|)$ term less than a suitably small multiple of $\|\boldsymbol{\theta}\|$. Then use continuity and compactness to extend the range of the inequalities.]

[5] If $\{X_n\} = o_p(1)$, show that there exists a sequence $\{\epsilon_n\}$ that converges to zero slowly enough to ensure $\mathbb{P}\{|X_n| > \epsilon_n\} \to 0$. [Hint: Build ϵ_n using an increasing sequence $n(k)$ such that $\mathbb{P}\{|X_n| > 1/k\} < 1/k$ for $n \ge n(k)$.]

[6] Let Q_n be the $\text{Bin}(n, \delta_n)$ distribution and P_n be the $\text{Bin}(n, \frac{1}{2})$. Show that $Q_n \lhd P_n$ if and only if $\delta_n = \frac{1}{2} + O(1/\sqrt{n})$.

[7] Remove $O(1/\sqrt{n})$ rate of converegnce restriction, under extra Lipschitz on model, for convergence under laternatives.

References

Birch, M. W. (1964), 'A new proof of the Fisher-Pearson theorem', *Annals of Mathematical Statistics* pp. 817–824.

Cochran, W. G. (1952), 'The χ^2 test of goodness of fit', *Annals of Mathematical Statistics* **23**, 315–3455.

Cressie, N. & Read, T. R. C. (1984), 'Multinomial goodness-of-fit tests', *Journal of the Royal Statistical Society, Series B* **46**, 440–464.

Cressie, N. & Read, T. R. C. (1988), *Goodness-of-Fit Statistics for Discrete Multivariate Data*, Springer, New York.

Dudley, R. M. (1976), 'Convergence of laws on metric spaces, with a view to statistical testing'. Lecture Note Series No. 45, Matematisk Institut, Aarhus University.

Fisher, R. A. (1922), 'On the interpretation of χ^2 from contingency tables, and the calculation of p', *Journal of the Royal Statistical Society* **85**, 87–94.

Fisher, R. A. (1923), 'Statistical tests of agreement between observation and hypothesis', *Econometrica* **3**, 139–147.

Fisher, R. A. (1928), 'On a property connecting the χ^2 measure of discrepency with the method of maximum likelihood', *Atti. Cong. Int. Mat. Bologna* **6**, 95–100.

Greenwood, M. & Yule, G. U. (1915), 'The statistics of snti-typhoid and snti-cholera inoculations and the interpretation of such statistics in general.', *Proc. Roy. Soc. of Medicine,* Section of Epidemiology and State Medicine **??**, 113–194.

Haberman, S. J. (1974), *The Analysis of Frequency Data*, University of Chicago Press.

Pearson, K. (1900), 'On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling', *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **50**, 157–175.