Chapter 11
Asymptotic Evaluations
Likelihood Ratio Asymptotics

Outline

Likelihood Ratio Tests

The likelihood ratio test is a popular choice to analyze a composite hypothesis.

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1$$

when Θ is a multidimensional parameter space and Θ_0 is a subspace. We can rewrite this as

$$H_0 : A\theta = 0 \quad \text{versus} \quad H_1 : A\theta \neq 0$$

If $\text{rank}(A) = k$, then $\dim(\Theta_0) = d - k$ where $d = \dim(\Theta)$.

Examples, For $n = 3$, we can write $H_0 : \theta_2 = 0$ using the matrix

$$A = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \qquad \text{rank}(A) = 1, \qquad \dim(\Theta_0) = 2$$

Example. $\Theta = \{\mathbf{p} = (p_1, \ldots, p_d); \sum_{i=1}^{d} p_i = 1\}$, choose a vector $\pi \in \Theta$. The choice $H_0 : \mathbf{p} = \pi$ is called goodness of fit.

Likelihood Ratio Tests

$$\Lambda(\mathbf{x}) = \frac{\sup\{L(\theta|\mathbf{x}); \theta \in \Theta_0\}}{\sup\{L(\theta|\mathbf{x}); \theta \in \Theta\}}$$

The rejection region for an α-level test is $\{\Lambda(\mathbf{x}) \leq \lambda_0\}$ where λ_0 is chosen so that

$$P_\theta\{\Lambda(\mathbf{X}) \leq \lambda_0\} \leq \alpha \text{ for all } \theta \in \Theta_0.$$

Let $\hat{\theta}_0$ be the parameter value that maximizes the likelihood for $\theta_0 \in \Theta_0$ and $\hat{\theta}$ be the parameter value that maximizes the likelihood for $\theta \in \Theta$. Then,

$$\Lambda(\mathbf{x}) = \frac{L(\hat{\theta}_0(\mathbf{x})|\mathbf{x})}{L(\hat{\theta}(\mathbf{x})|\mathbf{x})}.$$

Asymptotic Properties

Much of the attraction of maximum likelihood estimators is based on their properties for large sample sizes.

1. Consistency. If ψ_0 is the state of nature and $\hat{\theta}_n(X)$ is the maximum likelihood estimator based on n observations from a simple random sample, then

$$\hat{\theta}_n(X) \to \psi_0 \quad \text{as } n \to \infty \quad \text{in probability.}$$

2. Asymptotic normality and efficiency. Under some technical assumptions

$$\sqrt{n}(\hat{\theta}_n(X) - \psi_0).$$

converges in distribution as $n \to \infty$ to a normal random variable with mean 0 and variance $1/I(\psi_0)$, the Fisher information for one observation.

Normal Observations

Recall the two-sided hypothesis

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0.$$

n independent $N(\mu, \sigma_0)$ data with known variance σ_0^2. Set $\theta = \mu - \mu_0$. The parameter space Θ is one dimensional giving the value $\theta + \mu_0$ for the mean. $\hat{\theta} = \bar{x} + \mu_0$. Θ_0 is the single point $\{0\}$ and so $\hat{\theta}_0 = 0$. We previously showed that

$$\Lambda(\mathbf{x}) = \exp -\frac{1}{2\sigma_0^2}\left(\sum_{i=1}^{n}((x_i - \mu_0)^2 - (x_i - \bar{x})^2)\right) = \exp -\frac{n}{2\sigma_0^2}(\bar{x} - \mu_0)^2.$$

Notice that

$$-2 \ln \Lambda(\mathbf{x}) = \frac{n}{\sigma_0^2}(\bar{x} - \mu_0)^2 = \left(\frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}}\right)^2.$$

This is a χ_1^2- random variable.

Multiple Dimensions

Let's change bases $\eta = (\eta_1, \ldots, \eta_d)$, orthonormal vectors, on the parameter space Θ so that $\Theta_0 = \{\eta; \eta_1 = \cdots = \eta_k = 0\}$. Thus $\dim(\Theta_0) = d - k$.

Write $\eta^0 = (\eta_1, \ldots, \eta_k)$ and $\eta^1 = (\eta_{k+1}, \ldots, \eta_d)$. Then, the hypothesis becomes

$$H_0 : \eta^0 = 0 \quad \text{versus} \quad H_1 : \eta^0 \neq 0.$$

To prepare for the maximization problems, set

- ∇_0 be the gradient with respect to the first k coordinates.
- ∇_1 be the gradient with respect to the last $d - k$ coordinates.

Multiple Dimensions

The maximization over Θ_0 and Θ satisfies

$$\nabla_1 \ln L(\eta_0|\mathbf{x}) = \mathbf{0} \quad \text{and} \quad \begin{pmatrix} \nabla_0 \\ \nabla_1 \end{pmatrix} \ln L(\eta_0, \eta_1|\mathbf{x}) = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}$$

We can also divide the Fisher information matrix

$$I(\eta_0, \eta_1) = \begin{pmatrix} I_0(\eta_0) & \vdots & I_{0,1}(\eta_0, \eta_1) \\ \cdots & & \cdots \\ I_{0,1}(\eta_0, \eta_1)^T & \vdots & I_1(\eta_1) \end{pmatrix}$$

Multiple Dimensions

Write the linear approximation of the score function about $\psi_0 \in \Theta_0$, the true state of nature, and $\eta = (\eta_0, \eta_1)$

$$\nabla \ln L(\eta|\mathbf{X}) \approx \nabla \ln L(\psi_0|\mathbf{X}) + H \ln L(\psi_0|X)(\eta - \psi_0).$$

Here, H is the Hessian matrix of second order partial derivatives.

Now substitute $\eta = \hat{\eta}_n(\mathbf{X})$ where $\nabla \ln L(\hat{\eta}_n(\mathbf{X})|\mathbf{X}) = 0$. Then

$$\nabla \ln L(\psi_0|\mathbf{X}) \approx -H \ln L(\psi_0|X)(\hat{\eta}_n(\mathbf{X}) - \psi_0).$$

$$\frac{1}{\sqrt{n}} \nabla \ln L(\psi_0|\mathbf{X}) \approx -\frac{1}{n} H \ln L(\psi_0|X) \cdot \sqrt{n}(\hat{\eta}_n(\mathbf{X}) - \psi_0).$$

Asymptotic Properties

The random variables $\nabla \ln f(X_i|\psi_0)$ are independent with mean 0 and variance $I(\psi_0)$, the Fisher information matrix. Thus, for the term on the left

$$\frac{1}{\sqrt{n}} \nabla \ln L(\psi_0|\mathbf{X}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \nabla \ln f(X_i|\psi_0)$$

converges in distribution, by the central limit theorem, to a normal random variable with mean 0 and variance $I(\psi_0)$.

Wilks Theorem

For the term on the right $-H \ln f(X_i|\psi_0),\ i = 1, \ldots, n$ are independent with mean $I(\psi_0)$. Thus,

$$-\frac{1}{n} H \ln L(\psi_0|X) = -\frac{1}{n} \sum_{i=1}^{n} H \ln f(X_i|\psi_0)$$

converges, by the law of large numbers, to $I(\psi_0)$. Thus, by Slutsky's theorem, the distribution of $W = I(\psi_0)\sqrt{n}(\hat{\eta}_n(X) - \psi_0)$, converges to a normal random variable with variance $I(\psi_0)$.

$$\begin{aligned} \mathrm{Var}_{\psi_0}(\sqrt{n}(\hat{\eta}(\mathbf{X}) - \psi_0)) &= \mathrm{Var}_{\psi_0}(I(\psi_0)^{-1} W) = (I(\psi_0)^{-1})^T \mathrm{Var}_{\psi_0}(W) I(\psi_0)^{-1} \\ &= I(\psi_0)^{-1} I(\psi_0) I(\psi_0)^{-1} = I(\psi_0)^{-1} \end{aligned}$$

Similarly, for estimating under the restriction to Θ_0

$$\begin{aligned} \mathrm{Var}_{\psi_0}(\sqrt{n}(\hat{\eta}_0(\mathbf{X}) - \psi_0)) &= \mathrm{Var}_{\psi_0}(I_1(\psi_0)^{-1} W_0) = (I_1(\psi_0)^{-1})^T \mathrm{Var}_{\psi_0}(W_0) I_1(\psi_0)^{-1} \\ &= I_1(\psi_0)^{-1} I_1(\psi_0) I_1(\psi_0)^{-1} = I_1(\psi_0)^{-1} \end{aligned}$$

Wilks Theorem

The central limit theorem is based on comparing the data to the state of nature. Thus, we divide by $L(\psi_0|\mathbf{x})$ to relate the terms in the likelihood ratio to ψ_0.

$$\Lambda(\mathbf{x}) = \frac{L(\hat{\eta}_0(\mathbf{x})|\mathbf{x})/L(\psi_0|\mathbf{x})}{L(\hat{\eta}(\mathbf{x})|\mathbf{x})/L(\psi_0|\mathbf{x})}.$$

$$-2\ln\Lambda(\mathbf{x}) = (2\ln L(\hat{\eta}(\mathbf{x})|\mathbf{x}) - 2\ln L(\psi_0|\mathbf{x})) - (2\ln L(\hat{\eta}_0(\mathbf{x})|\mathbf{x}) - 2\ln L(\psi_0|\mathbf{x})).$$

For the first two terms, take a second order Taylor series expansion.

- The first order term $2\nabla \ln L(\hat{\eta}(\mathbf{x})|\mathbf{x})(\hat{\eta}(\mathbf{x}) - \psi_0) = 0$ because $\hat{\eta}(\mathbf{x})$ maximizes the log likelihood.
- For the second order term

$$2 \times \frac{1}{2}(\hat{\eta}(\mathbf{x}) - \psi_0)^T \cdot HL(\psi_0|\mathbf{x}) \cdot (\hat{\eta}(\mathbf{x}) - \psi_0).$$

  we will use law of large numbers on the likelihood term and the central limit theorem on $\sqrt{n}(\hat{\eta}(\mathbf{x}) - \psi_0)$ terms.

Wilks Theorem

Thus we rewrite the quadratic term in Taylor series as

$$\sqrt{n}(\hat{\eta}(\mathbf{x}) - \psi_0)^T \cdot \frac{1}{n} HL(\psi_0|\mathbf{x}) \cdot \sqrt{n}(\hat{\eta}(\mathbf{x}) - \psi_0).$$

Then, as before,

- $\frac{1}{n} HL(\hat{\eta}(\mathbf{x})|\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} Hf(X_i|\eta_0) \to I(\psi_0)$ as $n \to \infty$.
- $\sqrt{n}(\eta(\mathbf{x}) - \psi_0) \to^{\mathcal{D}} \tilde{Z}$, a normal random variable with variance matrix $I(\psi_0)^{-1}$
- Write $I(\psi_0) = U \Lambda U^T$ where U is an orthogonal matrix ($U^{-1} = U^T$) and Λ is a diagonal matrix of the (nonnegative) eigenvalues of $I(\psi_0)$.
- Write $I(\psi_0)^{-1/2} = U \Lambda^{-1/2} U^T$, where $\Lambda^{-1/2}$ is also diagonal with entries the inverse of the square root of those in Λ.
- Note that $I(\psi_0)^{-1/2} I(\psi_0)^{-1/2} = U \Lambda^{-1/2} U^T U \Lambda^{-1/2} U^T = U \Lambda^{-1/2} \Lambda^{-1/2} U^T = U \Lambda^{-1} U^T = I(\psi_0)^{-1}$

Wilks Theorem

- Let $Z = (Z_1, \ldots, Z_d)$, a vector of independent $N(0,1)$ random variables. Its covariance matrix is I_d, the $d \times d$ identity matrix.
- $\mathrm{Var}(I(\psi_0)^{-1/2} Z) = I(\psi_0)^{-1/2} I_d I(\psi_0)^{-1/2} = I(\psi_0)^{-1}$, the covariance matrix for $\tilde{Z}$.
- Putting the pieces together

$$\begin{aligned} 2(\ln L(\hat{\eta}(\mathbf{x})|\mathbf{x}) - \ln L(\psi_0)|\mathbf{x})) &\approx \sqrt{n}(\hat{\eta}(\mathbf{x}) - \psi_0)^T \cdot \frac{1}{n} HL(\psi_0|\mathbf{x}) \cdot \sqrt{n}(\hat{\eta}(\mathbf{x}) - \psi_0) \\ &\approx (I(\psi_0)^{-1/2} Z)^T I(\psi_0)(I(\psi_0)^{-1/2} Z) \\ &= Z^T (I(\psi_0)^{-1/2})^T I(\psi_0) I(\psi_0)^{-1/2} Z = Z^T Z \end{aligned}$$

This is a χ^2_d random variable.

Wilks Theorem

The Taylor series expansion for

$$2 \ln L(\hat{\eta}_0(\mathbf{x})|\mathbf{x}) - 2 \ln L(\psi_0|\mathbf{x})$$

based on $\hat{\eta}_0(\mathbf{x})$, the maximum of the log-likelihood on Θ_0 is more complicated.

The first order terms converge to an expression that involves the $I_{0,1}(\eta_0, \eta_1)$ terms in the information matrix and cancels the last $d - k$ terms in $Z^T Z$.
This results in

$$-\ln \Lambda(\mathbf{x}) \to^{\mathcal{D}} X_k,$$

a χ^2_k random variable.

Fit of a Distribution

Goodness of fit tests examine the case of a sequence of independent observations each of which can have 1 of d possible categories. For example, each of us has one of 4 possible of blood types, O, A, B, and AB. The local blood bank has good information from a national database of the fraction of individuals having each blood type,

$$\pi_O, \pi_A, \pi_B, \text{ and } \pi_{AB}.$$

The actual fraction p_O, p_A, p_B, and p_{AB} of these blood types in the community for a given blood bank may be different than what is seen in the national database. As a consequence, the local blood bank may choose to alter its distribution of blood supply to more accurately reflect local conditions.

Introduction

To place this assessment strategy in terms of formal hypothesis testing, let $\pi = (\pi_1, \ldots, \pi_d)$ be postulated values of the probability

$$P_\pi\{\text{individual is a member of } i\text{-th category}\} = \pi_i$$

and let $\mathbf{p} = (p_1, \ldots, p_d)$ denote the possible states of nature. Then, the parameter space is

$$\Theta = \{\mathbf{p} = (p_1, \ldots, p_d); p_i \geq 0 \text{ for all } i = 1, \ldots, d, \sum_{i=1}^{d} p_i = 1\}.$$

This parameter space has $d - 1$ free parameters. Once these are chosen, the remaining parameter value is determined by the requirement that the sum of the p_i equals 1. Thus, $\dim(\Theta) = d - 1$.

Fit of a Distribution

- The hypothesis is

$$H_0 : p_i = \pi_i, \text{ for all } i = 1, \ldots, d \quad \text{versus} \quad H_1 : p_i \neq \pi_i, \text{ for some } i = 1, \ldots, d$$

- The parameter space for the null hypothesis is a single point $\pi = (\pi_1, \ldots, \pi_d)$. Thus, $\dim(\Theta_0) = 0$.
- Consequently, the likelihood ratio test statistic has a distribution that is approximately chi-square with $\dim(\Theta) - \dim(\Theta_0) = d - 1$ degrees of freedom.
- The data $\mathbf{x} = (x_1, \ldots, x_n)$ are the categories for each of the n observations.

Likelihood Function

Let's use the likelihood ratio criterion to create a test for the distribution of human blood types in a given population. For the data

$$\mathbf{x} = \{O, B, O, A, A, A, A, A, O, AB\}$$

in the case of independent observations, the likelihood is

$$L(\mathbf{p}|\mathbf{x}) = p_O \cdot p_B \cdot p_O \cdot p_A \cdot p_A \cdot p_A \cdot p_A \cdot p_A \cdot p_O \cdot p_{AB} = p_O^3 p_A^5 p_B p_{AB}.$$

Notice that the likelihood has a factor of p_i whenever an observation take on the value i. In other words, a sufficient statistic for the data is

$$n_i = \#\{\text{observations from category } i\}$$

to create $\mathbf{n} = (n_1, n_2, \cdots, n_d)$, a vector that records the number of observations in each category, then, the likelihood function

$$L(\mathbf{p}|\mathbf{n}) = p_1^{n_1} \cdots p_d^{n_d}.$$

Likelihood Ratio

The likelihood ratio is the ratio of the maximum value of the likelihood under the null hypothesis and the maximum likelihood for any parameter value. In this case, the numerator is the likelihood evaluated at π.

$$\Lambda(\mathbf{n}) = \frac{L(\pi|\mathbf{n})}{L(\hat{\mathbf{p}}|\mathbf{n})} = \frac{\pi_1^{n_1}\pi_2^{n_2}\cdots\pi_d^{n_d}}{\hat{p}_1^{n_1}\hat{p}_2^{n_2}\cdots\hat{p}_d^{n_d}} = \left(\frac{\pi_1}{\hat{p}_1}\right)^{n_1}\cdots\left(\frac{\pi_d}{\hat{p}_d}\right)^{n_d}$$

To find the maximum likelihood estimator $\hat{p}$, we, as usual, begin by taking the logarithm of the likelihood,

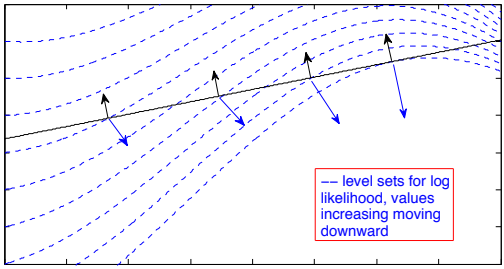
$$\ln L(\mathbf{p}|\mathbf{n}) = \sum_{i=1}^{d} n_i \ln p_i.$$

Not every set of values for p_i is admissible. So, we cannot just take derivatives, set them equal to 0 and solve. Indeed, we must find a maximum under the constraint

$$s(\mathbf{p}) = \sum_{i=1}^{d} p_i = 1.$$

Lagrange Multipliers

- Level sets of the log-likelihood function.shown in dashed blue.
- The constraint is level set $\{s(\mathbf{p}) = 1\}$ shown in black.
- The gradients are indicated by arrows.
- At the maximum, these two arrows are parallel. Their ratio λ is called the Lagrange multiplier.



$$\begin{aligned} \nabla_{\mathbf{p}} \ln L(\hat{\mathbf{p}}|\mathbf{n}) &= \lambda \nabla_{\mathbf{p}} s(\hat{\mathbf{p}}). \\ \left(\frac{\partial}{\partial p_1} \ln L(\hat{\mathbf{p}}|\mathbf{n}), \ldots, \frac{\partial}{\partial p_d} \ln L(\hat{\mathbf{p}}|\mathbf{n}) \right) &= \lambda \left(\frac{\partial}{\partial p_1} s(\mathbf{p}), \ldots, \frac{\partial}{\partial p_d} s(\mathbf{p}) \right) \\ \left(\frac{n_1}{\hat{p}_1}, \ldots, \frac{n_d}{\hat{p}_d} \right) &= \lambda (1, \ldots, 1) \end{aligned}$$

Lagrange Multipliers

$$\frac{n_i}{\hat{p}_i} = \lambda, \quad n_i = \lambda \hat{p}_i \quad \text{for all } i = 1, \ldots, d.$$

Now sum this equality for all values of i and use the constraint $s(\mathbf{p}) = 1$ to obtain

$$n = \sum_{i=1}^{d} n_i = \lambda \sum_{i=1}^{d} \hat{p}_i = \lambda s(\hat{\mathbf{p}}) = \lambda.$$

Thus, we have that

$$\frac{n_i}{\hat{p}_i} = n \quad \text{and} \quad \hat{p}_i = \frac{n_i}{n}.$$

The estimate for p_i is the fraction of observations in category i. Thus, for the blood bank example,

$$\hat{p}_O = \frac{3}{10}, \quad \hat{p}_A = \frac{5}{10}, \quad \hat{p}_B = \frac{1}{10}, \quad \text{and} \quad \hat{p}_{AB} = \frac{1}{10}.$$

Likelihood Ratio

Next, we substitute the maximum likelihood estimates $\hat{p}_i = n_i/n$ into the likelihood ratio to obtain

$$\Lambda(\mathbf{n}) = \frac{L(\pi|\mathbf{n})}{L(\hat{\mathbf{p}}|\mathbf{n})} = \left(\frac{\pi_1}{n_1/n}\right)^{n_1} \cdots \left(\frac{\pi_d}{n_d/n}\right)^{n_d} = \left(\frac{n\pi_1}{n_1}\right)^{n_1} \cdots \left(\frac{n\pi_d}{n_d}\right)^{n_d}.$$

Let $\mathbf{N} = (N_1, \ldots, N_d)$ denote the random vector of observed number of occurrences for each category i. When the null hypothesis holds true,

$$-2 \ln \Lambda(\mathbf{N}) = -2 \sum_{i=1}^{d} N_i \ln \frac{n\pi_i}{N_i} = 2 \sum_{i=1}^{d} N_i \ln \frac{N_i}{n\pi_i}$$

has approximately a χ^2_{d-1} distribution.

Likelihood Ratio

Using the notation $O_i = n_i$ for the number of observed occurrences of i and $E_i = n\pi_i$ for the number of expected occurrences of i as given by H_0, we can write the test statistic as

$$G^2 = -2 \ln \Lambda_n(\mathbf{O}) = 2 \sum_{i=1}^{d} O_i \ln \frac{O_i}{E_i}.$$

The traditional method for a test of goodness of fit, we use, instead of the G^2 statistic, an approximation

$$\chi^2 = \sum_{i=1}^{d} \frac{(E_i - O_i)^2}{E_i}.$$

In either case the p-value will be the probability that the a χ^2_{d-1} random variable takes a value greater than the test statistic.

Chi-Square Statistic

The Red Cross recommends that a blood bank maintains 44% type O, 42% type A, 10% type B, and 4% type AB. You suspect that the distribution of blood types in Tucson is not the same as the recommendation. In this case, the hypothesis is

$$H_0 : p_O = 0.44, p_A = 0.42, p_B = 0.10, p_{AB} = 0.04$$

versus

$$H_1 : \text{at least one } p_i \text{ is unequal to the given values.}$$

Based on 400 observations, we observe 228 for type O, 124 for type A, 40 for type B and 8 for type AB. We find the expected occurrences by computing $400 \times p_i$ using the values in H_0. This gives the table

type	O	A	B	AB
observed	228	124	40	8
expected	176	168	40	16

Blood Bank

Enter the observations and the proportions under H_0 in the `chisq.test` command. R computes the expected number of observations.

```
> chisq.test(c(228,124,40,8),p=c(0.44,0.42,0.10,0.04))
Chi-squared test for given probabilities
data:  c(228, 124, 40, 8)
X-squared = 30.8874, df = 3, p-value = 8.977e-07
```

The number of degrees of freedom is $4 - 1 = 3$. Note that the p-value is very low and so the distribution of blood types in Tucson is very unlikely to be the same as the national distribution.

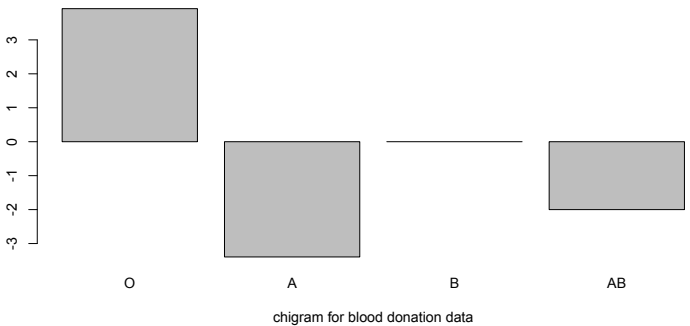
Exercise. Compute by hand the χ^2 statistic from the blood bank data. Use the `pchisq` command to determine the p-value.

Hanging Chi-Gram

To visualize the discrepancies from from the null hypothesis, we use a hanging chi-gram. This plots category i with a bar, height

$$\frac{O_i - E_i}{\sqrt{E_i}}.$$

Note that these values can be either positive or negative.




```
> resid<-(O-E)/sqrt(E)
> barplot(resid, names.arg=c("O","A","B","AB"),
  xlab="chigram for blood donation data")
```