Statistical inference
Part 5
Likelihood, and related, inference

Michael Goldstein
Durham University
APTS December 2023

Interval assessment

So far, we have been concerned with developing estimators with good properties.

We are now going to consider more general inferential questions.

We look at methods for constructing interval assessments and tests based on such estimators and otherwise.

This will allow us to explore the interplay between Bayesian and frequency based assessments.

Confidence procedures and confidence sets

We consider interval, or set, estimation.

Under the model $\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \mid \theta)\}$, for given data $X = x$, we wish to construct a set $C = C(x) \subset \Theta$

(If $\theta \in \mathbb{R}$ then the set estimate is typically an interval.)

The inference is the statement that $\theta \in C$.

Definition A random set $C(X)$ is a level $(1 - \alpha)$ **confidence procedure** when

$$\mathbb{P}(\theta \in C(X) \mid \theta) \ \geq \ 1 - \alpha$$

for all $\theta \in \Theta$.

C is an exact level $1 - \alpha$ confidence procedure if the probability equals $1 - \alpha$ for all θ.

Coverage

The value $\mathbb{P}(\theta \in C(X) \,|\, \theta)$ is termed the **coverage** of C at θ.

Exact is a special case: typically $\mathbb{P}(\theta \in C(X) \,|\, \theta)$ will depend upon θ.

The procedure is thus conservative: for a given θ_0 the coverage may be much higher than $1 - \alpha$.

Remember that X is the random quantity in the above definition, **not** θ.

The probability $1 - \alpha$ refers to the success rate of the procedure (averaged over samples).

It is **not** the probability that, for the sample that we have seen, our inference is correct.

For example, a confidence interval might be empty.

Empty confidence intervals

Here's a simple way to generate empty confidence intervals.

Suppose that you want to learn about a binomial parameter ω, but you don't have a sample from the process.

Generate a random integer between 1 and 100.

Choose the interval [0,1] if the integer is between 1 and 95.

Choose the empty interval otherwise.

The interval generated this way is a valid 95% confidence interval.

But not one that you would ever use!

Empty confidence intervals

Here's an interval that might be empty in practice.

Suppose that you want to estimate a Poisson parameter, λ.

You perform trial one giving you a sample from which you calculate confidence interval I_1.

You then perform trial two giving you a further sample which you combine with trial one and calculate a second interval I_2.

You might require intervals I_1, I_2 to be **simultaneous** confidence intervals, say at 95%.

[This means that for any λ, the probability of obtaining samples such that both $\lambda \in I_1$ and $\lambda \in I_2$ is at least 95%.]

Using this construction, the interval $I_1 \cap I_2$ is also a confidence interval at 95%.

But this could be empty.

Example: uniform distribution

Let $X_1, \ldots, X_n$ be independent and identically distributed Unif$(0, \theta)$ random variables where $\theta > 0$, so that

$$f(x|\theta) = \frac{1}{\theta}, \;\; 0 < x < \theta$$

Let $Y = \max\{X_1, \ldots, X_n\}$.

Comments

[1] Y is sufficient for θ. (check!)

[2] The uniform is a simple example of a distribution where the regularity assumptions that we have been making for the MLE evaluations do not hold, as the sample space depends on the parameter value.

Example:confidence sets

Compare two possible confidence sets:

(aY, bY) where $1 \leq a < b$

For this choice,

$\mathbb{P}(\theta \in (aY, bY) \,|\, \theta) = \left(\frac{1}{a}\right)^n - \left(\frac{1}{b}\right)^n$.

Thus, the coverage probability of the interval does not depend upon θ.

$(Y + c, Y + d)$ where $0 \leq c < d$.

For this choice,

$\mathbb{P}(\theta \in (Y + c, Y + d) \,|\, \theta) = \left(1 - \frac{c}{\theta}\right)^n - \left(1 - \frac{d}{\theta}\right)^n$.

In this case, the coverage probability of the interval does depend upon θ.

Confidence sets

We distinguish between the confidence procedure C, which is a random interval and so a function for each possible x, and the result when C is evaluated at the observation x, which is a set in Θ.

Definition (Confidence set)
The observed $C(x)$ is a level $1-\alpha$ confidence set exactly when the random $C(X)$ is a level $1-\alpha$ confidence procedure.

[If $\Theta \subset \mathbb{R}$ and $C(x)$ is an interval, then a confidence set (interval) is represented by a lower and upper value.]

A technical challenge for confidence procedures is to construct one with a specified level: to do this we start with the level and then construct a C guaranteed to have this level.

Confidence procedures

Definition (Family of confidence procedures) $C(X;\alpha)$ is a family of confidence procedures exactly when $C(X;\alpha)$ is a level-$(1-\alpha)$ confidence procedure for every $\alpha \in [0,1]$.

C is a nesting family exactly when $\alpha < \alpha'$ implies that $C(x;\alpha') \subset C(x;\alpha)$.

For $X_1, \ldots, X_n$ iid Unif$(0,\theta)$, $Y = \max\{X_1, \ldots, X_n\}$ then

$$C(Y;\alpha) = \left(\left(1 - \frac{\alpha}{2}\right)^{-1/n} Y, \left(\frac{\alpha}{2}\right)^{-1/n} Y \right)$$

is a nesting family of exact confidence procedures.

For example, if $n = 10$ then

$$C(y;0.10) = (1.0051y, 1.3493y);\ C(y;0.05) = (1.0025y, 1.4461y).$$

If we start with a family of confidence procedures for a specified model, then we can compute a confidence set for any level we choose.

Constructing confidence procedures: pivotal quantities

In the Uniform example, the coverage of the procedure (aY, bY) does not depend upon θ because the coverage probability could be expressed in terms of $T = Y/\theta$ where the distribution of T did not depend upon θ.

T is an example of a pivot and confidence procedures are straightforward to compute from a pivot.

Definition

A **pivot** for the model $\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \mid \theta)\}$ is a random variable $Q(X_1, \ldots, X_n, \theta)$ for which the distribution of Q does not depend upon θ.

For any set $\mathcal{A}$, $\mathbb{P}(Q(X_{(1:n)}, \theta) \in \mathcal{A} \mid \theta)$ does not depend on θ.

Hence, $\{\theta : Q(x_{(1:n)}, \theta) \in \mathcal{A}\}$ is an exact confidence procedure.

Example: location parameter

For any pdf $f(x)$, the family of pdfs $f(x-\theta)$, indexed by location parameter θ, is a **location family** with standard pdf $f(x)$.

Suppose that $X_1, \ldots, X_n$ are iid from $f(x-\theta)$.

Then the confidence set

$$C(x_1, \ldots, x_n) \;=\; \{\theta : \overline{x} - a < \theta < \overline{x} + b\}$$

for constants $a, b \geq 0$ has a fixed coverage for all θ.

Example: location parameter

This is because

$$\begin{aligned}\mathbb{P}(\theta \in C(X_1, \ldots, X_n) \,|\, \theta) &= \mathbb{P}(\overline{X} - a < \theta < \overline{X} + b \,|\, \theta) \\ &= \mathbb{P}\left(-b < \frac{1}{n}\sum_{i=1}^{n} Z_i < a \,\middle|\, \theta\right)\end{aligned}$$

does not depend on θ.

This is because each $Z_i = X_i - \theta$ has pdf $f(z)$ which does not depend on θ.

For location (and scale) parameters, we can easily find pivots but this is more difficult in general.

Constructing approximate confidence intervals

From our discussion of large sample properties of MLE, we have that the distribution of

$$\frac{\hat{\theta} - \theta}{\sqrt{\frac{1}{ni_1(\theta)}}}$$

is approximately a standard normal distribution.

This doesn't depend on θ so it acts as an approximate pivot, giving approximate confidence interval, at level $(1 - \alpha)$, as

$$\hat{\theta} \pm z_{\alpha/2}\sqrt{\frac{1}{ni_1(\hat{\theta})}}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ value of the standard normal distribution.

Example: Hardy Weinberg equilibrium

If gene frequencies MM, MN, NN in a population are in equilibrium, then genotypes occur in the population with frequencies

$MM : p_1 = (1-\theta)^2$

$MN : p_2 = 2\theta(1-\theta)$

$NN : p_3 = \theta^2$

(where θ is the unknown proportion of N in the population)

Suppose that in a sample from a particular population, blood types occur with the following frequencies.

$MM : x_1 = 342$

$MN : x_2 = 500$

$NN : x_3 = 187$

$Total : n = 1029$

Questions

[1] Find the MLE, $\hat{\theta}$.

[2] Find an approximate, large sample confidence interval for θ.

Multinomial distribution

Our data is an example of the multinomial distribution.

General form as follows:

X takes possible values $1, 2, ..., m$, where $\mathbb{P}(X = i) = p_i$, with $p_1 + ...p_m = 1$.

Repeat the experiment, independently, n times.

X_i is the number of times that $X = i$.

Joint frequency of $X_1, ..., X_m$ is

$$f(x_1, ..., x_m | p_1, ...p_m) = \frac{n!}{x_1!...x_m!} p_1^{x_1} ...p_m^{x_m}$$

(with $x_1 + ... + x_m = n$)

Example:log likelihood

In our example, the likelihood and log likelihood functions are

$$L(\theta) = f(\underline{x}|\theta) = cp_1^{x_1} p_2^{x_2} p3^{x_3} = c'\theta^{x_2+2x_3}(1-\theta)^{2x_1+x_2}$$

$$l(\theta) = log(L(\theta)) = c'' + (2x_1 + x_2)log(1-\theta) + (x_2 + 2x_3)log(\theta)$$

Example:MLE

Therefore the derivative of the log likelihood is

$$l'(\theta) = -\frac{2x_1 + x_2}{1-\theta} + \frac{x_2 + 2x_3}{\theta} = 0$$

when

$$\hat{\theta} = \frac{x_2 + 2x_3}{2n} = 0.4247$$

Check:

$$l''(\theta) = -\frac{2x_1 + x_2}{(1-\theta)^2} - \frac{x_2 + 2x_3}{\theta^2} < 0$$

Example: Information

Fisher's information for a sample of size n=1 is

$$i_1(\theta) = -\mathbb{E}(l''(\theta)) = \mathbb{E}(\frac{2X_1 + X_2}{(1-\theta)^2}) + \frac{X_2 + 2X_3}{\theta^2}$$

As $n = 1$, $X_i \sim Bi(1, p_i)$, so $\mathbb{E}(X_i) = p_i$.

Therefore

$$i_1(\theta) = \frac{2p_1 + p_2}{(1-\theta)^2} + \frac{p_2 + 2p_3}{\theta^2} = \frac{2}{\theta(1-\theta)}$$

Example: Confidence interval

Therefore, approximately,

$$\hat{\theta} \sim N(\theta, \frac{1}{ni_1(\theta)}) = N(\theta, \frac{\theta(1-\theta)}{2n})$$

Therefore, an approximate 95% confidence interval for θ is

$$\hat{\theta} \pm 1.96\sqrt{\frac{1}{ni_1(\hat{\theta})}} = \hat{\theta} \pm 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta}}{2n}} = \hat{\theta} \pm 1.96\sqrt{0.01089^2} \approx (0.403, 0.447)$$

Example: observed information

We can approximate $ni_1(\theta)$ by the observed information $-l''(\hat{\theta})$

In our example,

$$l''(\theta) = -\frac{2x_1 + x_2}{(1-\theta)^2} - \frac{x_2 + 2x_3}{\theta^2}$$

The data is $x_1 = 342, x_2 = 500, x_3 = 187$ and $\hat{\theta} = 0.4247$.

Therefore, $-l''(\hat{\theta}) = -8356.425$ and

$$\text{Var}(\hat{\theta}) \approx -\frac{1}{l''(\hat{\theta})} = (0.01094)^2$$

So our interval is $\hat{\theta} \pm 1.96 \times 0.01094$.

Compare this to the interval based on expected information, namely $\hat{\theta} \pm 1.96 \times 0.01089$

Bayesian credible intervals

Confidence intervals describe the probabilistic properties of the process used to generate the intervals.

They do not answer the question as to what is the probability that the actual interval that you have obtained contains the parameter of interest.

Bayesian credible intervals do describe the probability that the actual interval that you have obtained contains the parameter of interest.

This probability relates to the combination of prior judgements and data which have been used to generate the interval.

Typically, we will use level sets type arguments to construct posterior credible intervals, i.e. choosing all values of θ for which $\pi(\theta|\underline{x}) > k$, for some choice of k.

Limiting posterior distributions

Large sample properties of Bayesian estimators are described by the following theorem.

Theorem

Suppose that the prior pdf, $\pi(\theta)$ for the parameter θ is positive and differentiable over Θ.

Suppose we have a sample $\underline{x} = (x_1, \ldots, x_n)$, where n is large.

Then the posterior distribution $\pi(\theta|\underline{x})$ is approximately a normal distribution.

The mean of the posterior distribution is $\hat{\theta}$.

The variance is $\frac{1}{ni_1(\hat{\theta})}$

[Note, therefore, that, for large samples, the Bayesian inference will be approximately the same, whatever the choice of prior.]

Outline proof of limiting posterior form

Suppose a uniform prior $\pi(\theta) = 1$. Then

$$\pi(\theta|\underline{x}) \propto f(\underline{x}|\theta)$$

We take a Taylor expansion of $l(\theta; \underline{x})$ around $\hat{\theta}$.

$$l(\theta; \underline{x}) \approx l(\hat{\theta}; \underline{x}) + \frac{1}{2}(\theta - \hat{\theta})^2 l''(\hat{\theta}; \underline{x})$$

as $l'(\hat{\theta}; \underline{x}) = 0$.
As

$$l''(\hat{\theta}; \underline{x}) = \sum_i l''(\hat{\theta}; x_i)$$

by the SLLN

$$l''(\hat{\theta}; \underline{x}) \approx n\mathbb{E}_{\hat{\theta}}(l''(\hat{\theta}; x_1)) = -ni_1(\hat{\theta}) = -i(\hat{\theta})$$

Proof continued

Therefore

$$\pi(\theta|\underline{x}) \propto f(\underline{x}|\theta) \approx c(\underline{x})exp(-\frac{1}{2}(\theta - \hat{\theta})^2 i(\hat{\theta}))$$

(where $c(.)$ is a function which does not depend on θ).

Therefore, if $\pi(\theta)$ is uniform, then $\pi(\theta|\underline{x})$ must be, approximately, a normal density, mean $\hat{\theta}$, variance $1/i(\hat{\theta})$.

For large n, under fairly general conditions, the likelihood is sharply peaked around $\hat{\theta}$ falling off rapidly as the distance between θ and $\hat{\theta}$ increases.

Therefore, the likelihood dominates the prior and we can effectively suppose that $\pi(\theta)$ is constant so the result follows. □

Large sample credible intervals

As the large sample posterior distribution for θ is, approximately,

$$N(\hat{\theta}, \frac{1}{ni_1(\hat{\theta})})$$

it follows that an approximate credible interval, at level $(1-\alpha)$, is

$$\hat{\theta} \pm z_{\alpha/2}\sqrt{\frac{1}{ni_1(\hat{\theta})}}$$

Note that this is exactly the same as the large sample confidence interval we derived earlier.

Therefore, for large samples, Fisher's information based intervals inherit both good frequentist and Bayesian properties which explains their potential reliability.

Comment

Simulation experiments are helpful for exploring the quality of the approximation for medium sample sizes.

There are three basic issues.

Firstly, is the likelihood sufficiently peaked that using different priors gives roughly the same answer as using the uniform prior?

[Explore the effect of varying the prior away from the uniform.]

Secondly, using the uniform prior, how good is the approximation to the limiting form?

[Do this exactly or through sampling and comparing the updates.]

Thirdly, how robust are the conclusions to modifications to the likelihood?

[Add extra parameters, to explore this exactly, or add some form of random noise to your sampling experiments.]

Pre-posterior credible intervals

For moderate size samples, if the prior has been carefully chosen then Bayesian credible intervals are directly meaningful. However, it is harder to give a corresponding constructive meaning to confidence intervals.

There is one relationship between confidence and credible intervals which holds in all cases, and may be helpful in certain problems, which is as follows.

A $(1-\alpha)$ level confidence interval is a **pre-posterior** credible interval at the same level. The result is as follows.

Confidence intervals as prior credible intervals

Theorem Suppose that $I(X)$ is a $(1-\alpha)$ level confidence interval for some parameter θ. For any prior distribution for θ, the prior probability that $\theta \in I(X)$ is $(1-\alpha)$.

Proof

$$\begin{aligned}
\mathbb{P}(\theta \in I(X)) &= \int \mathbb{P}((\theta \in I(X)|\theta)\pi(\theta)d\theta \\
&= \int (1-\alpha)\pi(\theta)d\theta \\
&= 1-\alpha
\end{aligned}$$

and the result follows. □

Comment

Every confidence interval is a prior credible interval.

Whether it can function as a posterior credible interval, when you have seen the interval, depends on how much relevant prior information you have that would cause you to modify the observed interval.

In particular, in cases where you are simply using a given result of a standard statistical analysis for some other purpose, using the preposterior credible value may be a reasonable default.

Intervals and test

We are now going to turn our attention to hypothesis tests.

In particular, how such tests relate to interval estimates.

Again, we will pay particular attention to tests built around the likelihood function.

And we will develop their large sample properties.

Likelihood ratio tests

Consider a hypothesis test where we have to decide either to accept that an hypothesis H_0 is true or to reject H_0 in favour of an alternative hypothesis H_1 based on a sample $x \in \mathcal{X}$.

The set of x for which H_0 is rejected is called the **rejection region**.

The complement, where H_0 is accepted, is the **acceptance region**.

Comments on notation

Depending on context, we may replace reject/accept by reject/not reject

Also, we might construct a third region, where we are undecided.

And in a Bayesian formulation, we would report a posterior probability for the truth of H_0.

Likelihood ratio tests

Definition (Likelihood Ratio Test, LRT)

The likelihood ratio test (LRT) statistic for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$, where $\Theta_0 \cup \Theta_0^c = \Theta$, is

$$\lambda(x) = \frac{\sup_{\theta \in \Theta_0} L_X(\theta; x)}{\sup_{\theta \in \Theta} L_X(\theta; x)} = \frac{L(\hat{\theta}_0)}{L(\hat{\theta})}.$$

where $\hat{\theta}_0$ is the MLE if $\theta \in \Theta_0$, and $\hat{\theta}$ is the MLE if $\theta \in \Theta$.

A LRT at significance level α has a rejection region $\{x : \lambda(x) \leq c\}$ where $0 \leq c \leq 1$ is chosen so that $\mathbb{P}(\text{Reject } H_0 \,|\, \theta) \leq \alpha$ for all $\theta \in \Theta_0$.

Example

Let $X = (X_1, \ldots, X_n)$ and suppose that the X_i are independent and identically distributed $N(\theta, \sigma^2)$ random variables where σ^2 is known.

Consider the likelihood ratio test for $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.

As the maximum likelihood estimate of θ is $\overline{x}$,

$$\begin{aligned} \lambda(x) &= \frac{L_X(\theta_0; x)}{L_X(\overline{x}; x)} &&= \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left((x_i - \theta_0)^2 - (x_i - \overline{x})^2 \right) \right\} \\ &&&= \exp\left\{ -\frac{1}{2\sigma^2} n(\overline{x} - \theta_0)^2 \right\}. \end{aligned}$$

Example ctd

Notice that, under H_0,

$$\frac{\sqrt{n}(\overline{X} - \theta_0)}{\sigma} \sim N(0,1)$$

so that

$$-2\log\lambda(X) \;=\; \frac{n(\overline{X} - \theta_0)^2}{\sigma^2} \sim \chi^2_1,$$

the chi-squared distribution with one degree of freedom.

The rejection region is $\{x \,:\, \lambda(x) \leq c\} = \{x \,:\, -2\log\lambda(x) \geq k\}$.

Setting $k = \chi^2_{1,\alpha}$, where $\mathbb{P}(\chi^2_1 \geq \chi^2_{1,\alpha}) = \alpha$, gives a test at the exact significance level α.

Example ctd

The acceptance region of this test is $\{x : -2\log\lambda(x) < \chi^2_{1,\alpha}\}$ where

$$\mathbb{P}\left(\frac{n(\overline{X}-\theta_0)^2}{\sigma^2} < \chi^2_{1,\alpha} \,\middle|\, \theta=\theta_0\right) \;=\; 1-\alpha.$$

This holds for all θ_0 and so, additionally rearranging,

$$\mathbb{P}\left(\overline{X}-\sqrt{\chi^2_{1,\alpha}}\frac{\sigma}{\sqrt{n}} < \theta < \overline{X}+\sqrt{\chi^2_{1,\alpha}}\frac{\sigma}{\sqrt{n}} \,\middle|\, \theta\right) \;=\; 1-\alpha.$$

Thus,

$$C(X) = (\overline{X}-\sqrt{\chi^2_{1,\alpha}}\frac{\sigma}{\sqrt{n}}, \overline{X}+\sqrt{\chi^2_{1,\alpha}}\frac{\sigma}{\sqrt{n}})$$

is an exact level-$(1-\alpha)$ confidence procedure with $C(x)$ the corresponding confidence set.

Duality

Note that we obtained the level $(1-\alpha)$ confidence procedure by inverting the acceptance region of the level α significance test.

This correspondence, or duality, between acceptance regions of tests and confidence sets is a general property, as follows.

Theorem (Duality of Acceptance Regions and Confidence Sets)

1. For each $\theta_0 \in \Theta$, let $A(\theta_0)$ be the acceptance region of a test of $H_0 : \theta = \theta_0$ at significance level α. For each $x \in \mathcal{X}$, define $C(x) = \{\theta_0 \, : \, x \in A(\theta_0)\}$. Then $C(X)$ is a level-$(1-\alpha)$ confidence procedure.

2. Let $C(X)$ be a level-$(1-\alpha)$ confidence procedure and, for any $\theta_0 \in \Theta$, define $A(\theta_0) = \{x \, : \, \theta_0 \in C(x)\}$. Then $A(\theta_0)$ is the acceptance region of a test of $H_0 : \theta = \theta_0$ at significance level α.

Proof

1. As we have a level α test for each $\theta_0 \in \Theta$ then
   $\mathbb{P}(X \in A(\theta_0) \,|\, \theta = \theta_0) \geq 1 - \alpha$.
   Therefore, for all $\theta \in \Theta$,

$$\mathbb{P}(\theta \in C(X) \,|\, \theta) \;=\; \mathbb{P}(X \in A(\theta) \,|\, \theta) \;\geq 1 - \alpha.$$

   Hence, $C(X)$ is a level-$(1-\alpha)$ confidence procedure.

2. For a test of $H_0 : \theta = \theta_0$, the probability of a Type I error (rejecting H_0 when it is true) is

$$\mathbb{P}(X \notin A(\theta_0) \,|\, \theta = \theta_0) \;=\; \mathbb{P}(\theta_0 \notin C(X), \,|\, \theta = \theta_0) \;\leq\; \alpha$$

   since $C(X)$ is a level-$(1-\alpha)$ confidence procedure. Hence, we have a test at significance level α. □

Relationship between intervals and tests

Another way to understand the relationship between significance tests and confidence sets is as follows.

Define the set $\{(x, \theta) : (x, \theta) \in \tilde{C}\}$ in the space $\mathcal{X} \times \Theta$ where $\tilde{C}$ is also a set in $\mathcal{X} \times \Theta$.

- For fixed x, define the confidence set as $C(x) = \{\theta : (x, \theta) \in \tilde{C}\}$.
- For fixed θ, define the acceptance region as $A(\theta) = \{x : (x, \theta) \in \tilde{C}\}$.

Example revisited

Letting $x = (x_1, \ldots, x_n)$, with $z^2_{\alpha/2} = \chi^2_{1,\alpha}$, define the set

$$\{(x,\theta) : (x,\theta) \in \tilde{C}\} \quad = \quad \left\{(x,\theta) : -z_{\alpha/2}\sigma/\sqrt{n} < \overline{x} - \theta < z_{\alpha/2}\sigma/\sqrt{n}\right\}.$$

The confidence set is then

$$C(x) = \left\{\theta : \overline{x} - z_{\alpha/2}\sigma/\sqrt{n} < \theta < \overline{x} + z_{\alpha/2}\sigma/\sqrt{n}\right\}$$

and acceptance region

$$A(\theta) = \left\{x : \theta - z_{\alpha/2}\sigma/\sqrt{n} < \overline{x} < \theta + z_{\alpha/2}\sigma/\sqrt{n}\right\}.$$

Asymptotic properties

For large samples, we may carry out tests using the asymptotic distribution of the likelihood ratio test statistic

$$\lambda(x) = \frac{L(\hat{\theta}_0)}{L(\hat{\theta})}$$

We have the following theorem.

Theorem

$X_1, ..., X_n$ are an independent random sample from distribution $f(x|\theta)$.

Under suitable regularity conditions, the null distribution of $-2log(\lambda(x))$, for large n is approximately a χ^2 distribution.

The degrees of freedom is $D = dim(\Theta) - dim(\Theta_0)$,
where $dim(\Theta), dim(\Theta_0)$ are the number of free parameters under Θ, Θ_0 respectively.

Example

Let $X = (X_1, \ldots, X_n)$ and suppose that the X_i are independent and identically distributed $N(\theta, \sigma^2)$ random variables where σ^2 is known.

In this example, $dim(\Theta) = 1, dim(\Theta_0) = 0$

In that special case, we have shown that $-2log(\lambda(x)) \sim \chi_1^2$ exactly.

Proof in the case $d = m = 1$.

θ is a scalar and we may write $H_0 : \theta = \theta_0$ against $H_1 : \theta$ unrestricted, for some prescribed θ_0. A Taylor expansion of the log likelihood about the maximum likelihood estimate $\hat{\theta}$ gives

$$log(\lambda(x)) = 2(l(\hat{\theta}) - l(\theta_0)) = 2(\hat{\theta} - \theta_0)l'(\hat{\theta}) - (\hat{\theta} - \theta_0)^2 l''(\theta^*)$$

where θ^* is some other value of θ lying between θ_0 and $\hat{\theta}$.
As $l'(\hat{\theta}) = 0$, we have

$$2log(\lambda(x)) = ni_1(\theta_0)(\hat{\theta} - \theta_0)U^*V^*$$

where

$$U^* = \frac{l''(\theta^*)}{l''(\theta_0)}, \; V^* = \frac{l''(\theta_0)}{-ni_1(\theta_0)}$$

U^*, V^* both tend to one as n increases. Asymptotically $ni_1(\theta_0)(\hat{\theta} - \theta_0)$ is the square ofa standard normal random variable, hence distributed as χ^2_1. □

Example: Hardy Weinberg equilibrium

If gene frequencies MM, MN, NN in a population are in equilibrium, then genotypes occur in the population with frequencies

$MM : (1-\theta)^2 = p_1$

$MN : 2\theta(1-\theta) = p_2$

$NN : \theta^2 = p_3$

(where θ is the unknown proportion of N in the population)

Suppose that in a sample from a particular population, blood types occur with the following frequencies.

$MM : 342 = x_1$

$MN : 500 = x_2$

$NN : 187 = x_3$

$Total : 1029 = n$

Question

Does data suggest that population is not in Hardy Weinberg equilibrium?

[Note this hypothesis test is a goodness of fit of the model test.]

Example:likelihood

$$f(x_1, x_2, x_3|p_1, p_2, p_3) = \frac{n!}{x_1!x_2!x_3!} p_1^{x_1} p_2^{x_2} p_3^{x_3}$$

If $\theta \in \Theta_0$, then we've already found that

$$\hat{\theta}_0 = \frac{x_2 + 2x_3}{n} = 0.4247$$

We now need to find $\hat{\theta}$.

Multinomial MLE

Theorem

$$f(x_1, ..., x_m | p_1, ...p_m) = \frac{n!}{x_1!...x_m!} p_1^{x_1} ...p_m^{x_m}$$

The MLE for each p_i is

$$\hat{p}_i = \frac{x_i}{n}$$

(Proof - exercise. Maximise log likelihood under constraint $p_1 + ... + p_m = 1$)

Example:LR statistic

$$\lambda(x) = \prod_{i=1}^{3} (\frac{p_i(\hat{\theta}_0)}{\hat{p}_i})^{x_i}$$

where

$$\hat{p}_i = \frac{x_i}{n}$$

$$p_1(\hat{\theta}_0) = (1-\hat{\theta}_0)^2, p_2(\hat{\theta}_0) = 2\hat{\theta}_0(1-\hat{\theta}_0), p_3(\hat{\theta}_0) = \hat{\theta}_0^2$$

so that

$$-2log(\lambda(x)) = -2\sum_{i=1}^{3} x_i log(\frac{np_i(\hat{\theta}_0)}{x_i}) = 0.032$$

This can be compared with an appropriate choice of upper $\alpha\%$ point of χ_1^2.
Clearly not significant.

General form

In the general version of this goodness of fit problem, we have r categories and we count the number of observations, O_i in each category.
[In our example, $r = 3$, $O_i = x_i$]
The probability of each observation falling in category i is p_i.
Under H_0 each $p_i = p_i(\theta)$.
The expected number E_i of observations falling in category i is $E_i = np_i$.
Under H_0, this is estimated as $E_i = np_i(\hat{\theta}_0)$
The LR test statistic is

$$\lambda(x) = \prod_{i=1}^{r} (\frac{p_i(\hat{\theta}_0)}{\hat{p}_i})^{O_i}$$

where $\hat{p}_i = \frac{O_i}{n}$.
Therefore, our test statistic, evaluate as Chi-square with DF $r - 1 - dim(\Theta_0)$ is

$$-2log(\lambda(x)) = 2\sum_{i=1}^{r} O_i log(\frac{O_i}{E_i})$$

Chi-square goodness of fit test

Compare the chi-square goodness of fit test for the same data set.
We have a set of r categories and we count the number of observations, O_i that we observe in each category.
We compare each O_i with the expected number E_i of observations given the model.
We then evaluate the χ^2 goodness of fit test statistic

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

whose null distribution, given the model, is a χ^2 distribution with degrees of freedom equal to $r - 1 - c$ where c is the number of parameters that we have estimated.
In our problem, $c = 1$ so $r - 1 - c = 1$. The value of χ^2 for our data is 0.0319 (check!).
Notice that this is almost exactly the same as the value we obtained for $-2log(\lambda(x)) = 0.032$

Asymptotic equivalence of tests

The two tests, with very high probability, gave about the same value.

This is generally the case for large n.

We have the following theorem.

Theorem The two tests

$$2log(\lambda(x)) = 2\sum_{i=1}^{r} O_i log(\frac{O_i}{E_i})$$

and

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

are asymptotically equivalent given H_0.

[so the chi-square goodness of fit test is approximately a likelihood ratio test]

Outline Proof

Under H_0, if n large, then $\hat{p}_i \approx p_i(\hat{\theta})$ so $\frac{x_i}{n} \approx p_i(\hat{\theta})$. Therefore

$$O_i \approx E_i$$

With $O_i = x, E_i = x_0$ carry out a Taylor expansion around x_0 of

$$f(x) = xlog(\frac{x}{x_0})$$

.

This is

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + ...$$

which evaluates as

$$f(x) = 0 + (x - x_0) + \frac{1}{2x_0}(x - x_0)^2 + ...$$

Outline Proof ctd

Summing over i gives

$$-2log(\lambda = 2[\sum_i (O_i - E_i) + \sum_i \frac{1}{2E_i}(O_i - E_i)^2] + ...$$

As

$$\sum_i (O_i - E_i) = 0$$

this gives

$$-2log(\lambda) \approx \sum_i \frac{1}{E_i}(O_i - E_i)^2$$

□

Comparing experiments:tea tasting

Imagine carrying out Fisher's famous tea-tasting experiment.

Here an individual, Joan say, claims to be able to tell whether the milk or the tea has been added first in a cup of tea.

We perform the experiment of preparing ten cups of tea, choosing each time on a coin flip whether to add the milk or tea first.

Joan then tastes each cup and gives an opinion as to which ingredient was added first.

We count the number, X, of correct assessments. Suppose, for example, that $X = 9$.

Comparing experiments: ESP

Now compare the tea-tasting experiment to an experiment where an individual, Harry say, claims to have ESP as demonstrated by being able to forecast the outcome of fair coin flips.

We test Harry by getting forecasts for ten flips.

Let X be the number of correct forecasts.

Suppose that, again, $X = 9$.

Comparing the experiments

Within the traditional view of statistics, we might accept the same formalism for the two experiments.

For each experiment, each assessment is independent with probability p of success.

In each case, X has a binomial distribution parameters 10 and p, where $p = 1/2$ corresponds to pure guessing.

Within the traditional approach, the likelihood is the same, the point null is the same if we carry out a test for whether $p = 1/2$, and confidence intervals for p will be the same.

Comparing the experiments

However, even without carrying out formal calculations, I would be fairly convinced of Joan's tea tasting powers while remaining unconvinced that Harry has ESP.

You might decide differently, but that is because you might make different prior judgements.

Traditional statistical assessment of a positive result on a highly significant, very powerful test is not, of itself, a convincing inference.

The prior assessment of the plausibility of the new hypothesis can be converted into a posterior assessment, and without such an assessment there is no inference.

Bayes factors

The Bayesian approach for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$, where $\Theta_0 \cup \Theta_0^c = \Theta$, adds two ingredients.

Firstly, we must specify prior probabilities, $\mathbb{P}(H_0), \mathbb{P}(H_1)$.

Secondly, we must specify prior distributions $\pi_0(\theta)$, for $\theta \in \Theta_0$ and $\pi_1(\theta)$, for $\theta \in \Theta_0^c$.

Then, from Bayes theorem, we have

$$\frac{\mathbb{P}(H_0|\underline{x})}{\mathbb{P}(H_1|\underline{x})} = \frac{\mathbb{P}(\underline{x}|H_o)}{\mathbb{P}(\underline{x}|H_1)} \times \frac{\mathbb{P}(H_0)}{\mathbb{P}(H_1)}$$

where

$$\mathbb{P}(\underline{x}|H_o) = \int f(\underline{x}|\theta)\pi_0(\theta)d\theta, \quad \mathbb{P}(\underline{x}|H_1) = \int f(\underline{x}|\theta)\pi_1(\theta)d\theta$$

.

General discussion

Our Bayesian updating rule is that

posterior odds ratio = integrated likelihood ratio $\times$ prior odds ratio

The integrated likelihood ratio

$$L_\Theta = \frac{\mathbb{P}(\underline{x}|H_o)}{\mathbb{P}(\underline{x}|H_1)}$$

is often called the Bayes factor.

This formulation combines

(i) the prior judgements as to the relative plausibility of the two hypotheses [the prior odds ratio]

(ii) the evidence from the data [the Bayes factor].

Asymptotics

The asymptotic relationship between the integrated likelihood ratio L_Θ and the likelihood ratio test discussed earlier is as follows :

$$-2\log(\frac{\mathbb{P}(\underline{x}|H_o)}{\mathbb{P}(\underline{x}|H_1)}) \approx -2\log\frac{L(\hat{\theta}_0)}{L(\hat{\theta}_1)} - d\log(n)$$

(where $d = dim(\Theta) - dim(\Theta_0)$.)

This relationship allows us to move between Bayesian weight of evidence and the evidence in the test statistic.

(see Schwarz (1978) Estimating the dimension of a model, Ann Statist,
and the general discussion in Kass and Raftery (1995), Bayes Factors, JASA)