STA261: Week 9

Likelihood Ratio Tests

Alex Stringer

March 12th - 16th, 2018

Disclaimer

The materials in these slides are intended to be a companion to the course textbook, *Mathematical Statistics and Data Analysis, Third Edition*, by John A Rice. Material in the slides may or may not be taken directly from this source. These slides were organized and typeset by Alex Stringer.

A big thanks to Jerry Brunner as well for providing inspiration for assignment questions.

License

Recap

So far, we have talked about confidence intervals and hypothesis tests.

We showed how to derive such intervals and tests using normal-theory, and the central limit theorem.

We developed tools to make inferences about whether $\mu = \mu_0$, and to find a range of plausible values for μ, given the observed data.

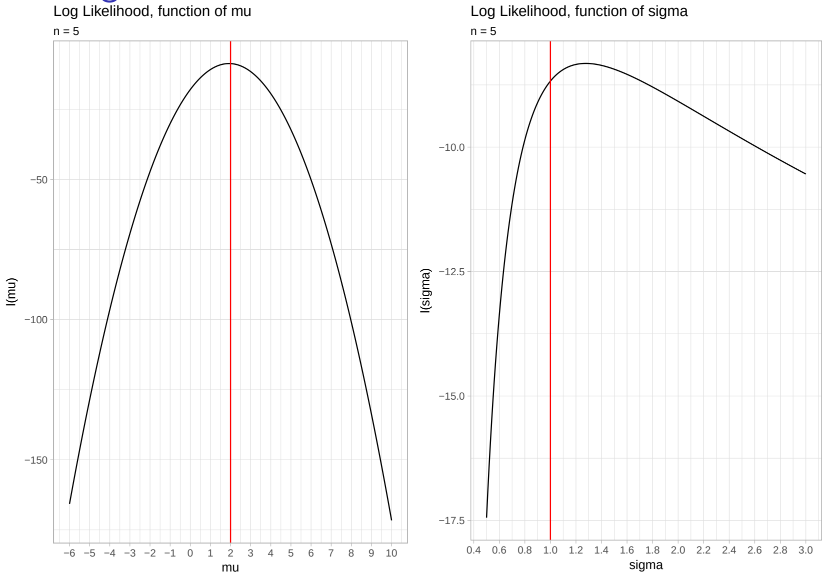
Going Forward

Today, we are going to generalize this a bit, and talk about **Likelihood Ratios**.

Recall in lecture 4, when we said that the shape of the likelihood function seemed to imply a range of values for the parameter that gave "similar" (log) likelihoods.

We used the CLT for the MLE to formalize this notion, and find a $1 - \alpha$ confidence interval for θ based off the MLE.

Recall: log-likelihood for the normal distribution



Compare two values

Suppose we wish to compare 2 candidate values of μ: μ_0 vs μ_1. We want to tell which is better supported by the data.

Look at their likelihoods: $L(\mu_0)$ and $L(\mu_1)$. The value with the higher likelihood is better supported by the data.

But by how much?

Comparing Likelihoods

Remember, absolute values of the likelihood aren't directly intepretable. The likelihood is a *relative* quantity.

To compare the likelihood of two values μ_0 and μ_1, we look at their **ratio**:

$$\Lambda = \frac{L(\mu_0)}{L(\mu_1)}$$

If this ratio is > 1, then μ_0 is better supported by the data. If it is < 1, then μ_1 is better supported by the data.

Let's look at an example.

Example: coin tossing (textbook, section 9.1, page 329)

This is a similar, but simplified, version of the coin tossing example in the textbook, page 329.

Suppose I have two coins, A and B, with respective probability of heads:

$$P_A(X = 1) = 0.3$$
$$P_B(X = 1) = 0.7$$

The corresponding likelihoods for a single flip, $x \in \{0, 1\}$, is

$$L_A(x) = 0.3^x 0.7^{1-x}$$
$$L_B(x) = 0.7^x 0.3^{1-x}$$

Example: coin tossing (textbook, section 9.1, page 329)

The likelihood ratio is

$$\Lambda = \frac{L_A(x)}{L_B(x)} = \left(\frac{0.3}{0.7}\right)^x \left(\frac{0.7}{0.3}\right)^{1-x}$$

I then throw a coin and it comes up tails. The question is: which coin did I throw?

Example: coin tossing (textbook, section 9.1, page 329)

The likelihood ratio for the observed data of $x = 0$ is

$$\Lambda = \left(\frac{0.7}{0.3}\right) \approx 2.3333$$

Given that I observed tails $(x = 0)$, it is about 2.3 times more likely that the coin was coin A than coin B.

Example: coin tossing (textbook, section 9.1, page 329)

Let's restate the problem as a hypothesis test. Suppose I throw a coin once, and it has some unknown probability of heads θ. I am interested in assessing whether $\theta = 0.3$ or $\theta = 0.7$ based on the results of a single toss.

The likelihood is

$$L(\theta|x) = \theta^x (1-\theta)^{1-x}$$

and the likelihood ratio is

$$\Lambda = \frac{L(\theta_0)}{L(\theta_1)} = \left(\frac{\theta_0}{\theta_1}\right)^x \left(\frac{1-\theta_0}{1-\theta_1}\right)^{1-x}$$

where $\theta_0 = 0.3$ and $\theta_1 = 0.7$.

Note that which one is labelled θ_0 and which one is labelled θ_1 is arbitrary (for now).

Example: coin tossing (textbook, section 9.1, page 329)

So we find that $\Lambda = 2.33333$. What do we conclude?

The General Case

This idea of comparing likelihoods under two hypotheses leads to the **Likelihood Ratio Test** (LRT).

Likelihood Ratio Tests are extremely general, and have nice optimality properties.

They essentially are to hypothesis testing what the MLE was to estimation.

The General Case

Recall the most general statement of our testing problem: we have a parameter $\theta \in \Omega$, and we have two hypotheses corresponding to disjoint subsets of the parameter space:

$$H_0 : \theta \in \Omega_0$$
$$H_1 : \theta \in \Omega_1$$

We wish to see whether the observed data supports rejecting H_0 in favour of H_1.

The General Case

Definiton: the **Likelihood Ratio Statistic** for testing $H_0 : \theta \in \Omega_0$ against $H_1 : \theta \in \Omega_1$ is

$$\Lambda = \frac{\sup_{\theta \in \Omega_0} L(\theta)}{\sup_{\theta \in \Omega_1} L(\theta)}$$

Small values of Λ indicate that H_1 is better supported by the data than H_0.

We reject H_0 if Λ is "small enough"

The General Case

In general, we don't have a method for deciding whether Λ is small enough.

We do have a method, though, for a very important special case: the case where we test $H_0 : \theta \in \Omega_0$ against the alternative $H_1 : \theta \in \Omega - \Omega_0$; that is, when we are testing whether $\theta \in \Omega_0$ vs whether it is not.

In this case,

$$\Lambda = \frac{\sup_{\theta \in \Omega_0} L(\theta)}{\sup_{\theta \in \Omega} L(\theta)} = \frac{\sup_{\theta \in \Omega_0} L(\theta)}{L(\hat{\theta})}$$

where $\hat{\theta}$ is the MLE.

The General Case

Note that $0 < \Lambda \leq 1$, with $\Lambda = 1$ occurring when $\hat{\theta} \in \Omega_0$.

If the MLE is part of the null parameter space, then we of course wouldn't want to reject H_0.

If the MLE is not part of the null parameter space, then we look at whether the most likely value of θ within Ω_0 is "good enough", in the sense that it gives a likelihood that is almost as high as the maximum possible in Ω.

How good is "good enough"?

Free Parameters

Let $p = \text{dim}\Omega$, and let $d = \text{dim}\Omega_0$ be the number of *free parameters* in the whole parameter space, and under the null hypothesis.

For example, for testing $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$ as before, $p = 1$ (because there is one free parameter under H_1, namely μ) and $d = 0$ (because under H_0, all parameters are fixed).

We have the following distributional result.

Distribution of $-2 \log \Lambda$

Theorem: under all the same regularity conditions as in lecture 4,

$$-2 \log \Lambda \xrightarrow{d} \chi^2_{p-d}$$

if H_0 is true, i.e. if $\theta \in \Omega_0$.

The proof is "out of scope" for our textbook.

We will prove this for the case where $\Omega_0 = \{\theta_0\}$, so $p = 1$ and $d = 0$, and the null hypothesis is that θ_0 is the true value of θ.

The result does, though, hold in a much more general setting.

Distribution of $-2\log\Lambda$

Proof: Note that $-2\log\Lambda = 2(\ell(\hat{\theta}) - \ell(\theta_0))$. Take a second-order Taylor expansion of $\ell(\theta_0)$ about the point $\hat{\theta}$ to obtain

$$\begin{aligned}
-2\log\Lambda &= 2(\ell(\hat{\theta}) - \ell(\theta_0)) \\
&\approx 2(\ell(\hat{\theta}) - (\ell(\hat{\theta}) + \ell'(\hat{\theta})(\theta_0 - \hat{\theta}) + \frac{1}{2}\ell''(\hat{\theta})(\theta_0 - \hat{\theta})^2)) \\
&= J(\hat{\theta})(\theta_0 - \hat{\theta})^2
\end{aligned}$$

because $\ell'(\hat{\theta}) = 0$ by definition, and $J(\hat{\theta}) = -\ell''(\hat{\theta})$

Distribution of $-2\log\Lambda$

Now because $J(\hat{\theta})$ is a consistent estimator of $I(\theta_0)$,

$$
\begin{aligned}
-2\log\Lambda &\approx J(\hat{\theta})(\theta_0 - \hat{\theta})^2 \\
&\xrightarrow{p} I(\theta_0)(\theta_0 - \hat{\theta})^2 \\
&= \left(\frac{\hat{\theta} - \theta_0}{1/\sqrt{I(\theta_0)}}\right)^2 \\
&\xrightarrow{p} Z^2
\end{aligned}
$$

where $Z \sim N(0,1)$, and hence $Z^2 \sim \chi^2_1$.

Example

Let's look at some examples of the likelihood ratio.

Suppose $X_i \sim N(\mu, \sigma_0^2)$ where σ_0^2 is known, and we wish to test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$.

We have $\Omega_0 = \{\mu_0\}$ and $\Omega_1 = \mathbb{R} - \{\mu_0\}$.

Since Ω_0 is a singleton set, $\sup_{\mu \in \Omega_0} L(\mu) = L(\mu_0)$.

And $\sup_{\mu \in \Omega_1} L(\mu) = L(\hat{\mu}) = L(\bar{X})$, the likelihood evaluated at the MLE.

Example

Evaluate the respective likelihoods. You don't need to worry about terms not involving μ, since they will cancel in the ratio.

$$L(\mu_0|\mathbf{x}) = c \times \exp\left(-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i - \mu_0)^2\right)$$

$$L(\bar{x}|\mathbf{x}) = c \times \exp\left(-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)$$

The likelihood ratio test statistic is then

$$\begin{aligned} -2\log\Lambda &= 2(\ell(\bar{x}) - \ell(\mu_0)) \\ &= \frac{1}{\sigma_0^2}\left(\sum_{i=1}^{n}(x_i - \mu_0)^2 - \sum_{i=1}^{n}(x_i - \bar{x})^2\right) \end{aligned}$$

Example

We then compare $-2\log\Lambda$ to the critical region of a χ^2_1 distribution.

Because random variables with a χ^2 distribution are strictly > 0, we use the critical region

$$R_\alpha = (\chi^2_{1,1-\alpha}, \infty)$$

that is, reject H_0 at the α significance level when

$$-2\log\Lambda > \chi^2_{1,1-\alpha}$$

the $1-\alpha$ quantile of the χ^2_1 distribution.

Example: Hospital Wait Times

With this theory in hand, we don't need to stick with the normal distribution.

Suppose we have patients arriving at a hospital waiting room, randomly. We can model their wait times X_i according to an exponential distribution,

$$X_i \sim Exp(\theta), E(X) = \theta$$

The hospital claims that the average waiting time is 60 minutes. We go on a randomly selected day and observe that $n = 100$ patients have an average wait time of $\bar{x} = 75$ minutes.

Is the hospital's claim supported by the data?

Example: Hospital Wait Times

The hypothesis we wish to test is

$$H_0 : \theta = 60$$
$$H_1 : \theta \neq 60$$

The likelihood is

$$L(\theta|\mathbf{x}) = \frac{1}{\theta^n} \exp\left(-\frac{1}{\theta}\sum_{i=1}^{n} x_i\right)$$

and the MLE is $\bar{X}$.

Example: Hospital Wait Times

The likelihood ratio is then

$$\Lambda = \left(\frac{\bar{x}}{\theta_0}\right)^n \exp\left(n\left(1 - \frac{\bar{x}}{\theta_0}\right)\right)$$

and the test statistic is

$$-2\log\Lambda = -2n\left(\log\bar{x} - \log\theta_0 + 1 - \frac{\bar{x}}{\theta_0}\right) \sim \chi^2_1$$

Example: Hospital Wait Times

With $\theta_0 = 60$, $n = 100$ and $\bar{x} = 75$, we evaluate

$$-2\log\Lambda = -2(100)\left(\log 75 - \log 60 + 1 - \frac{75}{60}\right) = 5.37$$

which we compare to $\chi^2_{1,0.95} = 3.84$.

Because 5.37 $>$ 3.84, we reject H_0 at the 5% significance level.

Example: Hospital Wait Times

We can also compute the p-value of this test. The p-value is the probability of observing a result with as much or greater evidence against H_0 if H_0 is true. If H_0 is true, then $-2\log\Lambda \sim \chi^2_1$, so

$$p_0 = P(\chi^2_1 > 5.37) = 0.02$$

R Code

```r
# Critical value
round(qchisq(.95,1),2)
```

```
## [1] 3.84
```

```r
# P-value
1 - round(pchisq(-2*100*(log(75) - log(60)
+ 1 - (75/60)),1),2)
```

```
## [1] 0.02
```

Example: Unknown Variance

When the variance was known, we recovered our usual normal-theory test using the likelihood ratio. What about when the variance is unknown?

Let $X_i \sim N(\mu, \sigma^2)$ with both parameters unknown. We wish to test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ using a likelihood ratio test.

We need to get the MLE of (μ, σ^2) under the null, and in general.

We know that in general,

$$\hat{\mu} = \bar{X}$$
$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$$

so $\hat{\sigma}^2 = \frac{n-1}{n}s^2$ when $\hat{\mu} = \bar{X}$.

Example: Unknown Variance

However, even though H_0 doesn't directly specify any restrictions on σ^2, it *does* restrict μ.

And $\hat{\sigma}^2$ depends on μ.

Hence under $H_0 : \mu = \mu_0$,

$$\hat{\sigma}_0^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_0)^2$$

Example: Unknown Variance

The maximized restricted likelihood under H_0 is thus

$$L(\mu_0, \hat{\sigma}_0^2) = \left(2\pi\hat{\sigma}_0^2\right)^{-n/2} \exp\left(-\frac{1}{2\hat{\sigma}_0^2}\sum_{i=1}^{n}(X_i - \mu_0)^2\right)$$
$$= \left(2\pi\hat{\sigma}_0^2\right)^{-n/2} \exp\left(-\frac{n}{2}\right)$$

We compare to the maximized unrestricted likelihood

$$L(\hat{\mu}, \hat{\sigma}^2) = \left(2\pi\hat{\sigma}^2\right)^{-n/2} \exp\left(-\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}(X_i - \hat{\mu})^2\right)$$
$$= \left(2\pi\hat{\sigma}^2\right)^{-n/2} \exp\left(-\frac{n}{2}\right)$$

Example: Unknown Variance

The squared likelihood ratio is then

$$\Lambda^2 = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)^n$$

or

$$\begin{aligned} -2\log\Lambda &= n\left(\log\hat{\sigma}_0^2 - \log\hat{\sigma}^2\right) \\ &= n\log\left(1 + \frac{t^2}{n-1}\right) \end{aligned}$$

where $t^2 = \frac{(\bar{X}-\mu_0)^2}{s^2/n}$.

You will be asked to show that last part on the assignment. Hint: add and subtract $\bar{X}$ inside $\sum_{i=1}^n (X_i - \mu_0)^2$ to show that $\sum_{i=1}^n (X_i - \mu_0)^2 = \sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - \mu_0)^2$.

Example: Unknown Variance

On the assignment, you will also be asked to show (the above hint will help) that the case with the *known* variance that we discussed before gives *exactly* the same test statistic (and therefore, same decision).

In the unknown variance case, we see that the kind of decision we are making is the same: reject H_0 for large values of $|t| = \frac{|\bar{X} - \mu_0|}{s/\sqrt{n}}$.

But the two test statistics aren't identical. So which is better? For the same significance level α, which has a lower probability of Type II error?

Testing Independence

We will cover one final, and very important, example of a likelihood ratio test.

Suppose we have N individuals sampled from a population, classified into two sets of discrete categories.

For example, we could sample N Canadians and ask what province they are from (BC, Alberta,...;13 levels) and who they voted for in the last election (Liberal, Conservative, NDP, Green, BQ, Other, Didn't Vote; 7 levels).

We want to test the hypothesis that the two categories are unrelated, against the alternative that they are related in some way.

Testing Independence

More formally: we have data $y_{ij}, i = 1 \dots R, j = 1 \dots C$, corresponding to counts of individuals observed in category (i, j).

We can arrange the data in a *contingency table*:

	c_1		c_C
r_1	y_{11}	$\cdots$	y_{1C}
	$\vdots$		$\vdots$
r_R	y_{R1}	$\cdots$	y_{RC}

Testing Independence

We have the following constraints:

$$N = \sum_{i=1}^{R} \sum_{j=1}^{C} y_{ij}$$

$$r_i = \sum_{j=1}^{C} y_{ij}$$

$$c_j = \sum_{i=1}^{R} y_{ij}$$

How to test that the two categories are unrelated? We need a model for the y_{ij}.

Testing Independence

Suppose the y_{ij} are drawn randomly from a population in which the true proportion of subjects in cell (i,j) is p_{ij}. Then the joint distribution of the data is

$$(Y_{11},\ldots,Y_{RC}) \sim Multinomial(p_{11},\ldots,p_{RC})$$

Key observation: if the row and column categories are *independent*, then

$$p_{ij} = P(Y_{ij}=1) = p_{i.} \times p_{.j}$$

where $p_{i.}$ is the marginal probability of an observation being in the i^{th} row, and $p_{.j}$ is the marginal probability of an observation being in the j^{th} column.

Testing Independence

So we wish to test

$$H_0 : p_{ij} = p_{i\cdot} \times p_{\cdot j}$$

against the alternative that the p_{ij} are not restricted.

We need

- The MLE of $p_{i\cdot}$ and $p_{\cdot j}$ under H_0
- The MLE of p_{ij} under H_1

Testing Independence

The unrestricted likelihood is

$$L(\mathbf{p}|\mathbf{y}) = c \times \prod_{i=1}^{R} \prod_{j=1}^{C} p_{ij}^{y_{ij}}$$

Under H_0, the likelihood is

$$L_0(\mathbf{p}|\mathbf{y}) = c \times \prod_{i=1}^{R} \prod_{j=1}^{C} (p_{i\cdot} \times p_{\cdot j})^{y_{ij}}$$

Testing Independence

Maximizing these requires lagrange multipliers, due to the constraint that $\sum_{i,j} p_{ij} = 1$. The result is what we would expect though:

$$\hat{p}_{ij} = \frac{y_{ij}}{N}$$
$$\hat{p}_{i\cdot} = \frac{r_i}{N}$$
$$\hat{p}_{\cdot j} = \frac{c_j}{N}$$

i.e. the MLEs are the respective sample proportions.

Testing Independence

It follows that the test statistic for a likelihood ratio test is (exercise on assignment 9: verify this):

$$-2\log\Lambda = 2\sum_{i=1}^{R}\sum_{j=1}^{C} y_{ij}\log\left(\frac{Ny_{ij}}{r_i c_j}\right)$$

We reject H_0 if this is large, i.e. if the counts in any cell deviate strongly from what we would expect under the hypothesis of independence.

Testing Independence

How large is large enough? We know that $-2\log\Lambda$ asymptotically follows a χ^2 distribution. What are the degrees of freedom?

Under H_1, there are $RC - 1$ free parameters, because there are RC cell probabilities, which all have to sum to 1.

Under H_0, there are $(R-1)$ free row probabilities and $(C-1)$ free column probabilities.

Hence the degrees of freedom are

$$RC - 1 - ((R-1) + (C-1)) = (R-1)(C-1)$$

Example

Let's consider an example. The following is a synthetic dataset from 2 categories each with 2 levels. This could represent something like "smoking" vs "respiratory illness" or "treatment/control" vs some binary clinical state.

	40	50
43	10	33
47	30	17

Example

We wish to test whether the rows and columns are independent. We find

$$
\begin{aligned}
N &= 10 + 33 + 30 + 17 = 90 \\
\hat{p}_{1\cdot} &= 40/90 = 0.44 \\
\hat{p}_{2\cdot} &= 50/90 = 0.56 \\
\hat{p}_{\cdot 1} &= 43/90 = 0.48 \\
\hat{p}_{\cdot 2} &= 47/90 = 0.52 \\
\hat{p}_{11} &= 10/90 = 0.11 \\
\hat{p}_{12} &= 33/90 = 0.37 \\
\hat{p}_{21} &= 30/90 = 0.33 \\
\hat{p}_{22} &= 17/90 = 0.19
\end{aligned}
$$

Example

Under H_0,

$$\hat{p}_{11} = 0.44 \times 0.48 = 0.21$$
$$\hat{p}_{12} = 0.44 \times 0.52 = 0.23$$
$$\hat{p}_{21} = 0.56 \times 0.48 = 0.27$$
$$\hat{p}_{22} = 0.56 \times 0.52 = 0.29$$

Are these far enough away from their unrestricted estimates under H_1 to conclude that the observed data provides evidence against the null hypothesis of independence?

Example

Our test statistic is

$$
\begin{aligned}
-2\log\Lambda &= 2\times 10\times\log\left(\frac{90\times 10}{43\times 40}\right) \\
&\quad + 2\times 33\times\log\left(\frac{90\times 33}{43\times 50}\right) \\
&\quad + 2\times 30\times\log\left(\frac{90\times 30}{47\times 40}\right) \\
&\quad + 2\times 17\times\log\left(\frac{90\times 17}{47\times 50}\right) \\
&= 15.5
\end{aligned}
$$

Example

If our hypothesis of independence is correct, then this should be a realization of a χ_1^2 random variable.

The p-value of the test is

$$\begin{aligned} p_0 &= P(\chi_1^2 > 15.5) \\ &\approx 8.2579629 \times 10^{-5} \end{aligned}$$

If the test statistic is χ_1^2, then observing a value of 15.5 is extremely improbable.

We reject H_0 at any reasonable significance level, and conclude that the two categories are related somehow.