Likelihood and related Test statistics

Daniel Lin

Notations

$I(\theta)$	Fisher's information
$L(\theta)$	Likelihood given data x
$l(\theta)$	Log Likelihood given data x
$\hat{\theta}$	Maximum Likelihood Estimator of θ
H_0, H_1	Null and alternative Hypothesis
$\lambda(x)$	Likelihood Ratio, where x can either mean scalar or a vector of data
$W(x)$	Equivalent test function for $\lambda(x)$

I. Test statistics

Recall that given observed data x, likelihood ratio statistics is defined as

$$\lambda(x) = \frac{\sup_{\theta\in\Theta_0}(L(\theta))}{\sup_{\theta\in\Theta}(L(\theta))}$$

using a monotonic decreasing transformation $f(s) = -2\ln(s)$, one can consider $W(x) = -2\ln(\lambda(x))$, which is equivalent to likelihood ratio. Generally speaking, $Pr(\lambda(x) = 0 \text{ or } 1) = 0$, so definition of $W(x)$ is valid and it has image $(0,\infty)$. Note if $\lambda(x)$ is quite extreme(close to 0 or 1), $W(x)$ will also be extreme.

Warning. some books define likelihood ratio in the other way around. But the test statistics $W(x)$ will remain the same. i.e. if $\lambda^*(x) := \frac{\sup_{\theta\in\Theta}(L(\theta))}{\sup_{\theta\in\Theta_0}(L(\theta))}$ instead, then $W(x) := 2\ln(\lambda^*(x))$.

Consider the simple case $\Theta_0 = \{\theta_0\}$, and using Taylor expansion around θ_0,

$$\begin{aligned} W(x) &= -2(l(\theta_0) - l(\hat{\theta})) \\ &= -2\left\{(\theta_0 - \hat{\theta})l'(\hat{\theta}) + \frac{1}{2}(\theta_0 - \hat{\theta})^2 l''(\tilde{\theta})\right\} \end{aligned}$$

where $\tilde{\theta}$ is between θ_0 and $\hat{\theta}$. Using definition of $\hat{\theta}$, $l'(\hat{\theta}) = 0$. Under H_0, $\hat{\theta}$ is consistent for θ_0, so

$$\begin{aligned} W(x) &= -n(\theta_0 - \hat{\theta})^2 \frac{l''(\tilde{\theta})}{n} + o_p(1) \\ &= n(\theta_0 - \hat{\theta})^2\{I(\theta_0) + o_p(1)\} + o_p(1) \\ &= n(\theta_0 - \hat{\theta})^2 I(\theta_0) + o_p(1) \end{aligned}$$

where $o_p(1)$ means a term that converges to 0 in probability. So $W(x) = W_e(x) + o_p(1)$ where

$$W_e(x) := n(\theta_0 - \hat{\theta})^2 I(\theta_0)$$

is called **Wald test statistics**. It describes the deviation of $\hat{\theta}$ from θ after suitable normalisation. Whereas $W(x)$ measures difference between log likelihood computed at $\hat{\theta}$ and θ.

(*) With some mild regularity conditions, $W(x)$ follows χ^2_r distribution where r is the number of independent restrictions on θ we need to make in H_0. Some times we use this approximation, and if nominal level(the level of test claimed) is higher than actual test level(actual test level using asymptotic distribution), the test is said to be conservative.

We call the observed significance level or p-value to be minimum significance level s.t. null hypothesis is rejected. For example if the test statistics is $W(x)$

$$\alpha_{\text{obs}} = \sup_{\theta\in\Theta_0} P(W(x) \geq W(x); \theta)$$

II. Two sample t-test

When one needs to compare the difference of means between two groups under the assumption that two groups share the same variance, two sample t-test can be used. There is a test to whether two i.i.d. normal samples have the same variance. If this assumption cannot be made, no exact solution exists and this problem is called Behrens-Fisher Problem. But asymptotic approximation is possible.

Assume given two i.i.d. samples $(x_1,...,x_m)$ following $N(\mu_x,\sigma^2)$, $(y_1,...,y_n)$ following $N(\mu_y,\sigma^2)$.
In this case, $H_0 : \mu_x = \mu_y$, $H_1 : \mu_x \neq \mu_y$.
using the fact that under H_0, two samples have exactly the same distribution, so we can view the sample as whole,

$$\sup_{\theta\in\Theta_0}(L(\theta)) = L(\hat{\mu}, \hat{\sigma_0^2}|x,y)$$

where

$$\hat{\mu} = \frac{\sum_i x_i + \sum_j y_j}{m+n} = \frac{m\bar{x} + n\bar{y}}{m+n}$$

and

$$\hat{\sigma_0^2} = \frac{\sum_i (x_i\hat{\mu})^2 + \sum_j (y_j\hat{\mu})^2}{m+n}$$

Using the pdf of normal distribution, we can find log likelihood without restriction H_0:

$$l(\mu_x,\mu_y,\sigma^2|x,y) = -\frac{m+n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_x)^2 - \sum_{j=1}^{n}(y_j-\mu_y)^2$$

taking two derivatives with respect to σ^2 yields that

$$\hat{\sigma^2} = \frac{\sum_i (x_i-\bar{x})^2 + \sum_j (y_j-\bar{y})^2}{m+n}$$

and $\hat{\mu_x} := \frac{\sum_i x_i}{m}, \hat{\mu_y} := \frac{\sum_j y_j}{n}$ maximises $l(\mu_x,\mu_y,\sigma^2|x,y)$.
Substituting the MLE back to likelihood ratio, we get

$$\begin{aligned} \lambda(x) &= \frac{\hat{\sigma_0^2}}{\hat{\sigma^2}} = \frac{\sum_i (x_i-\hat{\mu})^2 + \sum_j (x_j-\hat{\mu})^2}{\sum_i (x_i-\hat{\mu_x})^2 + \sum_j (x_j-\hat{\mu_y})^2} \\ &= \frac{\sum_i (x_i-\hat{\mu_x})^2 + m(\hat{\mu}-\hat{\mu_x})^2 + \sum_j (x_j-\hat{\mu_y})^2 + n(\hat{\mu}-\hat{\mu_y})^2}{\sum_i (x_i-\hat{\mu_x})^2 + \sum_j (x_j-\hat{\mu_y})^2} \end{aligned}$$

We can rewrite

$$\begin{aligned} & m(\hat{\mu}-\hat{\mu_x})^2+n(\hat{\mu}-\hat{\mu_y})^2 \\ =m\left(\bar{x}-\frac{m\bar{x}+n\bar{y}}{m+n}\right)^2 & +n\left(\bar{y}-\frac{m\bar{x}+n\bar{y}}{m+n}\right)^2) \\ =\frac{m^2n}{(n+m)^2}(\bar{x}-\bar{y})^2 & +\frac{mn^2}{(n+m)^2}(\bar{y}-\bar{x})^2 \\ & =\frac{mn}{m+n}(\bar{x}-\bar{y})^2 \end{aligned}$$

So

$$\lambda=\left(1+\frac{t^2}{m+n-2}\right)^{-(m+n)/2}$$

where $t=\frac{(\bar{x}-\bar{y})/\sqrt{1/m+1/n}}{s_p}$, and the pooled sample variance s_p^2 is defined as

$$s_p^2:=\frac{\sum_i(x_i-\hat{\mu_x})^2+\sum_j(x_j-\hat{\mu_y})^2}{m+n-2}$$

since $\frac{\sum_i(x_i-\hat{\mu_x})}{\sigma}$ follows χ^2_{m-1} and $\frac{\sum_j(y_j-\hat{\mu_y})}{\sigma}$ follows χ^2_{n-1}, $\sum_i(x_i-\hat{\mu_x})^2+\sum_j(x_j-\hat{\mu_y})^2$ follows $\sigma^2\chi^2_{m+n-2}$. So by definition of t-distribution:

$$T=\frac{Z}{\sqrt{V/n}}$$

where $Z\sim N(0,1), V\sim\chi^2_n$, one obtain $t\sim t_{m+n-2}$.

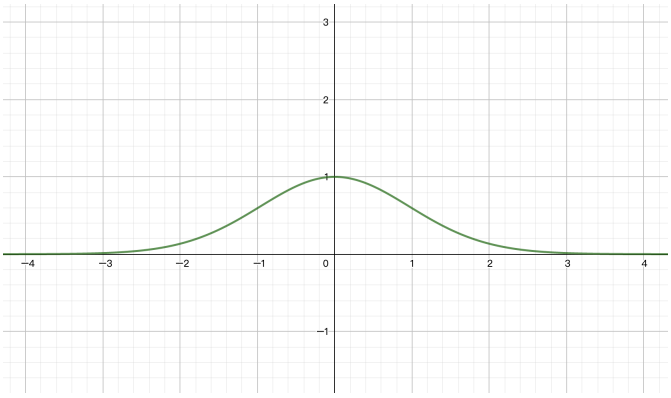

Fig. 1. Graph of λ against t, using $m+n=40$

We can see from the graph above that extreme values of t indeed implies extreme values of λ.
The distribution of λ is difficult to obtain, but we can use observed significance level instead

$$-2\ln(\lambda)=(m+n)\ln\left(1+\frac{t^2}{m+n-2}\right) \quad (1)$$

$$=\frac{m+n}{m+n-2}t^2+o_p(\frac{1}{m+n})(m+n)=\frac{m+n}{m+n-2}t^2+o_p(1) \quad (2)$$

so for large m,n, W is approximately Z^2 where $Z\sim N(0,1)$. So $W\sim\chi^2_1$. This checks (*).

Now

$$\alpha_{\text{obs}}\approx\sup_{\theta\in\Theta_0}P(|t|\geq|t_{\text{obs}}|;\theta)$$

by equation (2). So two-sided t confidence interval can be used to construct a rejection region at level α:

$$(-\infty,-t_{\alpha/2}] \text{ or } [t_{\alpha/2},\infty)$$

Reference

Statistical Inference based on the likelihood, written by Adelchi Azzalini. page 110-127