Log likelihood ratio test for standardize data

Nitay Alon

August 21, 2018

In this paper we'll explore the effect of using standardize data on the computation of log-likelihood ratio test. This effect is crucial for our research since we use KLD to partition the parameter set into potential discovery region, this is doe with non standardize data. However the llrt is done with standardize and we must be sure that we can map the results from the standardize llrt to the non standardize.

1 The model

In this section we present the underline model of the data, and decompose the log-likelihood. Our null model assumes the following:

1. Both men and women are pure types
2. The variance of both populations is equal

thus, the likelihood unde the null model is simply:

$$L(x;\Theta) = \prod_{i=1}^{N_{men}} f_0(x_i) * \prod_{j=1}^{N_{women}} f_0(y_j) \tag{1}$$

$$f_0(x) = \frac{1}{\sqrt{2\pi}\sigma_0} exp^{\frac{-(x-\mu_{men})^2}{2\sigma_0^2}} \tag{2}$$

And $f_0(y)$ is the same but with the women's mean instead. Our alternative model assumes that the genders distribution is a mixture model of the sexes distributions. It assumes the following:

1. Both male and female are pure types

2. Men select the male distribution with probability p and female with probability $1-p$
3. Women select the male distribution with probability p and female with probability $1-q$
4. The variance of both sexes is equal

the likelihood of one gender under the mixture model is:

$$f_X(x) = p * \frac{1}{\sqrt{2\pi}\sigma_1} exp^{\frac{-(x-\mu_M)^2}{2\sigma_0^2}} + (1-p) * \frac{1}{\sqrt{2\pi}\sigma_0} exp^{\frac{-(x-\mu_F)^2}{2\sigma_0^2}} \quad (3)$$

(this is the men's likelihood, the women's is the same but with different mixing probabilities)

2 Log-likelihood Ratio

Now we can compute the log-likelihood ratio for the hypothesis:

$$log(\frac{f_1(x)}{f_0(x)} = log\big(\frac{p * \frac{1}{\sqrt{2\pi}\sigma_1} exp^{\frac{-(x-\mu_M)^2}{2\sigma_1^2}} + (1-p) * \frac{1}{\sqrt{2\pi}\sigma_1} exp^{\frac{-(x-\mu_F)^2}{2\sigma_1^2}}}{\frac{1}{\sqrt{2\pi}\sigma_0} exp^{\frac{-(x-\mu_{men})^2}{2\sigma_0^2}}}\big) \quad (4)$$

Re-organizing the expression:

$$log(\frac{\sigma_0}{\sigma_1}) + log\big(p * \frac{exp^{\frac{-(x-\mu_M)^2}{2\sigma_1^2}}}{exp^{\frac{-(x-\mu_{men})^2}{2\sigma_0^2}}} + (1-p)\frac{exp^{\frac{-(x-\mu_F)^2}{2\sigma_1^2}}}{exp^{\frac{-(x-\mu_{men})^2}{2\sigma_0^2}}}\big)$$

$$log(\frac{\sigma_0}{\sigma_1}) + log\big(p * exp^{\frac{-(x-\mu_M)^2}{2\sigma_1^2} - \frac{-(x-\mu_{men})^2}{2\sigma_0^2}} + (1-p) exp^{\frac{-(x-\mu_F)^2}{2\sigma_1^2} - \frac{-(x-\mu_{men})^2}{2\sigma_0^2}}\big)$$

since the variance ratio is independent of the data we can focus in the right argument. Let's re-write the exponents in a more "pleasant" way:

$$exp^{\frac{-(x-\mu_M)^2}{2\sigma_1^2} - \frac{-(x-\mu_{men})^2}{2\sigma_0^2}} = exp^{\frac{-x^2(2\sigma_0^2-2\sigma_1^2)+2x(2\sigma_1^2\mu_{men}-2\sigma_0^2\mu_M)+\mu_M^2 2\sigma_0^2-\mu_{men}^2 2\sigma_1^2}{4\sigma_1^2\sigma_0^2}}$$

$$exp^{\frac{-(x-\mu_F)^2}{2\sigma_1^2} - \frac{-(x-\mu_{men})^2}{2\sigma_0^2}} = exp^{\frac{-x^2(2\sigma_0^2-2\sigma_1^2)+2x(2\sigma_1^2\mu_{men}-2\sigma_0^2\mu_F)+\mu_F^2 2\sigma_0^2-\mu_{men}^2 2\sigma_1^2}{4\sigma_1^2\sigma_0^2}}$$

So we can write the log likelihood ratio as:

$$log\big(exp^{\frac{-x^2(2\sigma_0^2-2\sigma_1^2)-\mu_{men}^2 2\sigma_1^2(x-1)}{4\sigma_1^2\sigma_0^2}}(p*exp^{\frac{2\mu_M 2\sigma_0^2(2x+\mu_M)}{4\sigma_1^2\sigma_0^2}}+(1-p)*exp^{\frac{2\mu_F 2\sigma_0^2(2x+\mu_F)}{4\sigma_1^2\sigma_0^2}})\big) \tag{5}$$

This is a critical equation. Now we'll express x in terms of the standardize RV z. First let's define $z = \frac{x-\mu}{\sigma}$, where μ is the population mean and σ is the population standard deviation.

$$log\big(exp^{\frac{-(z\sigma+\mu)^2(2\sigma_0^2-2\sigma_1^2)-\mu_{men}^2 2\sigma_1^2(z\sigma+\mu-1)}{4\sigma_1^2\sigma_0^2}}(p*exp^{\frac{2\mu_M 2\sigma_0^2(2(z\sigma+\mu)+\mu_M)}{4\sigma_1^2\sigma_0^2}}+(1-p)*exp^{\frac{2\mu_F 2\sigma_0^2(2(z\sigma+\mu)+\mu_F)}{4\sigma_1^2\sigma_0^2}})\big) \tag{6}$$

organizing:

$$log\big(exp^{\frac{-(z\sigma+\mu)^2(2\sigma_0^2-2\sigma_1^2)-\mu_{men}^2 2\sigma_1^2(z\sigma+\mu-1)}{4\sigma_1^2\sigma_0^2}}(p*exp^{\frac{\mu_M(2(z\sigma+\mu)+\mu_M)}{\sigma_1^2}}+(1-p)*exp^{\frac{\mu_F(2(z\sigma+\mu)+\mu_F)}{\sigma_1^2}})\big) \tag{7}$$

Now, using the laws of logs:

$$\frac{-(z\sigma+\mu)^2(2\sigma_0^2-2\sigma_1^2)-\mu_{men}^2 2\sigma_1^2(z\sigma+\mu-1)}{4\sigma_1^2\sigma_0^2}+log\big(p*exp^{\frac{\mu_M(2(z\sigma+\mu)+\mu_M)}{\sigma_1^2}}+(1-p)*exp^{\frac{\mu_F(2(z\sigma+\mu)+\mu_F)}{\sigma_1^2}}\big) \tag{8}$$

Not sure about this part Since we assume equal gender group sizes we can write: $\mu = \frac{1}{2}\mu_M + \frac{1}{2}\mu_F$ and the last equation turns to:

$$\frac{-(z\sigma+\frac{\mu_M+\mu_F}{2})^2(2\sigma_0^2-2\sigma_1^2)-\mu_{men}^2 2\sigma_1^2(z\sigma+\frac{\mu_M+\mu_F}{2}-1)}{4\sigma_1^2\sigma_0^2}+ log\big(p*exp^{\frac{\mu_M(2(z\sigma+\frac{\mu_M+\mu_F}{2})+\mu_M)}{\sigma_1^2}}+(1-p)*exp^{\frac{\mu_F(2(z\sigma+\frac{\mu_M+\mu_F}{2})+\mu_F)}{\sigma_1^2}}\big) \tag{9}$$

which can be written as:

$$\frac{-(z\sigma+\frac{\mu_M+\mu_F}{2})^2(2\sigma_0^2-2\sigma_1^2)-\mu_{men}^2 2\sigma_1^2(z\sigma+\frac{\mu_M+\mu_F}{2}-1)}{4\sigma_1^2\sigma_0^2}+ log\big(p*exp^{\frac{\mu_M(\frac{z\sigma}{2}+2\mu_M+\mu_F)}{\sigma_1^2}}+(1-p)*exp^{\frac{\mu_F(\frac{z\sigma}{2}+\mu_M+2\mu_F)}{\sigma_1^2}}\big) \tag{10}$$

So we can see that if we denote $\frac{z\sigma}{2} = y$ we get a nice llrt.