

## Goodness of fit and Wilks' theorem

Suppose we model data  $\mathbf{y}$  with a likelihood  $L(\boldsymbol{\mu})$  that depends on a set of  $N$  parameters  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_N)$ . Define the statistic

$$t_{\boldsymbol{\mu}} = -2 \ln \frac{L(\boldsymbol{\mu})}{L(\hat{\boldsymbol{\mu}})}, \quad (1)$$

where  $\hat{\boldsymbol{\mu}}$  are the ML estimators for  $\boldsymbol{\mu}$ . The value of  $t_{\boldsymbol{\mu}}$  is a measure of how well the hypothesized set of parameters  $\boldsymbol{\mu}$  stand in agreement with the data. If the agreement is poor, then  $\hat{\boldsymbol{\mu}}$  will be far from  $\boldsymbol{\mu}$ , the ratio of likelihoods will be low and  $t_{\boldsymbol{\mu}}$  will be large. Larger values of  $t_{\boldsymbol{\mu}}$  thus indicate increasing incompatibility between the data and the hypothesized  $\boldsymbol{\mu}$ .

According to Wilks' theorem, if the parameter values  $\boldsymbol{\mu}$  are true, then in the asymptotic limit of a large data sample, the pdf of  $t_{\boldsymbol{\mu}}$  is a chi-square distribution for  $N$  degrees of freedom. We will write this as

$$f(t_{\boldsymbol{\mu}}|\boldsymbol{\mu}) \sim \chi_N^2. \quad (2)$$

Suppose we have a data set that gives us an observed value of the statistic  $t_{\boldsymbol{\mu},\text{obs}}$ . We can quantify the level of compatibility between  $\boldsymbol{\mu}$  and the observed data by computing the  $p$ -value

$$p_{\boldsymbol{\mu}} = \int_{t_{\boldsymbol{\mu},\text{obs}}}^{\infty} f_{\chi_N^2}(t_{\boldsymbol{\mu}}|\boldsymbol{\mu}) dt_{\boldsymbol{\mu}}. \quad (3)$$

Now suppose that the set of parameters  $\boldsymbol{\mu}$  can be expressed as  $\boldsymbol{\mu}(\boldsymbol{\theta})$  where  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_M)$  is a set of  $M$  parameters with  $M < N$ . Now define

$$q_{\boldsymbol{\mu}} = -2 \ln \frac{L(\boldsymbol{\mu}(\hat{\boldsymbol{\theta}}))}{L(\hat{\boldsymbol{\mu}})}. \quad (4)$$

That is, in the numerator we adjust  $M$  parameters and in the denominator  $N$ . In this case, Wilks' theorem states

$$f(q_{\boldsymbol{\mu}}|\boldsymbol{\mu}(\boldsymbol{\theta})) \sim \chi_{N-M}^2 \quad (5)$$

Provided certain regularity conditions are satisfied, this holds regardless of the value of  $\boldsymbol{\theta}$ . This is a very useful property that allows one to compute  $p$ -values without needing to assume particular values for the parameters  $\boldsymbol{\theta}$ . In this case the  $p$ -value reflects the compatibility of the assumed functional form  $\boldsymbol{\mu}(\boldsymbol{\theta})$ .

# 1 Gaussian data

Suppose that the data are a set of  $N$  independent Gaussian distributed values,

$$y_i \sim \text{Gauss}(\mu_i, \sigma_i), \quad i = 1, \dots, N, \quad (6)$$

where the standard deviations  $\sigma_i$  are known but the  $\mu_i$  must be determined from the data. The likelihood is

$$L(\boldsymbol{\mu}) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(y_i - \mu_i)^2 / 2\sigma_i^2}, \quad (7)$$

so that the log-likelihood is

$$\ln L(\boldsymbol{\mu}) = -\frac{1}{2} \sum_{i=1}^N \frac{(y_i - \mu_i)^2}{\sigma_i^2} + C, \quad (8)$$

where  $C$  does not depend on  $\boldsymbol{\mu}$ . By setting the derivatives of  $\ln L(\boldsymbol{\mu})$  with respect to the  $\mu_i$  to zero we find the ML estimators to be

$$\hat{\mu}_i = y_i, \quad (9)$$

and from this we find

$$t_{\boldsymbol{\mu}} = -2 \ln \frac{L(\boldsymbol{\mu})}{L(\hat{\boldsymbol{\mu}})} = \sum_{i=1}^N \frac{(y_i - \mu_i)^2}{\sigma_i^2}. \quad (10)$$

In the case where  $M$  parameters  $\theta_1, \dots, \theta_M$  are fitted, the statistic  $q_{\boldsymbol{\mu}}$  is

$$q_{\boldsymbol{\mu}} = -2 \ln \frac{L(\boldsymbol{\mu}(\hat{\boldsymbol{\theta}}))}{L(\hat{\boldsymbol{\mu}})} = \sum_{i=1}^N \frac{(y_i - \mu_i(\hat{\boldsymbol{\theta}}))^2}{\sigma_i^2}. \quad (11)$$

Thus we can use the minimized value of the sum of squares from an LS fit to test the goodness of fit. In such a case the values of  $\mu_i$  are obtained by assuming a functional relation between  $\mu$  and a control variable  $x$ , whose value is fixed for each measurement of  $y$ . That is,

$$\mu_i(\boldsymbol{\theta}) = \mu(x_i; \boldsymbol{\theta}), \quad i = 1, \dots, N. \quad (12)$$

The  $p$ -value therefore reflects the degree of compatibility between the data and the functional form  $\mu(x; \boldsymbol{\theta})$ .

# 2 Histogram of Poisson or multinomial data

Consider now a set of data values  $\mathbf{n} = (n_1, \dots, n_N)$  which we may think of as a histogram with  $N$  bins. Suppose the values  $n_i$  are independent and Poisson distributed with mean values  $\nu_i$ , so that the joint probability for the vector  $\mathbf{n}$  is

$$P(\mathbf{n}; \boldsymbol{\nu}) = \prod_{i=1}^N \frac{\nu_i^{n_i}}{n_i!} e^{-\nu_i} . \quad (13)$$

The log-likelihood is therefore

$$\ln L(\boldsymbol{\nu}) = \sum_{i=1}^N [n_i \ln \nu_i - \nu_i] + C , \quad (14)$$

where  $C$  represents terms that do not depend on  $\boldsymbol{\nu}$ .

If we regard each of the  $\nu_i$  as adjustable, then by setting the derivatives of  $\ln L(\boldsymbol{\nu})$  with respect to all of the  $\nu_i$  to zero we find the ML estimators

$$\hat{\nu}_i = n_i , \quad i = 1, \dots, N . \quad (15)$$

Using this we can write down the statistic analogous to Eq. (1),

$$t_{\boldsymbol{\nu}} = -2 \ln \frac{L(\boldsymbol{\nu})}{L(\hat{\boldsymbol{\nu}})} \quad (16)$$

$$= -2 \sum_{i=1}^N \left[ n_i \ln \frac{\nu_i}{\hat{\nu}_i} - \nu_i + \hat{\nu}_i \right] \quad (17)$$

$$= -2 \sum_{i=1}^N \left[ n_i \ln \frac{\nu_i}{n_i} - \nu_i + n_i \right] , \quad (18)$$

where in the final line we used  $\hat{\nu}_i = n_i$ . By going back to the original Poisson probabilities one can see that if  $n_i = 0$ , then the logarithmic term in Eq. (16) is in fact absent. As with the statistic  $t_{\mu}$  from above, Wilks' theorem says that the distribution of  $t_{\boldsymbol{\nu}}$  approaches a chi-square distribution for  $N$  degrees of freedom in the limit of a large data sample. Here one can see the role of the large sample limit, since then the estimators  $\hat{\nu}_i = n_i$  become approximately Gaussian distributed.

Now suppose that the set of  $N$  mean values  $\boldsymbol{\nu}$  can be determined through a set of  $M$  parameters  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_M)$ . We can then define the statistic

$$q_{\boldsymbol{\nu}} = -2 \ln \frac{L(\boldsymbol{\nu}(\hat{\boldsymbol{\theta}}))}{L(\hat{\boldsymbol{\nu}})} = -2 \sum_{i=1}^N \left[ n_i \ln \frac{\nu_i(\hat{\boldsymbol{\theta}})}{n_i} - \nu_i(\hat{\boldsymbol{\theta}}) + n_i \right] . \quad (19)$$

As with the statistic  $q_{\mu}$  above, this will follow a chi-square distribution for  $N - M$  degrees of freedom.

In some problems one may want to model a histogram of values  $\mathbf{n} = (n_1, \dots, n_N)$  as following a multinomial distribution. This is similar to the Poisson case above except that the total number of entries,

$$n_{\text{tot}} = \sum_{i=1}^N n_i \quad (20)$$

is regarded as constant. There are in effect  $N - 1$  free parameters in the problem, which can be taken as all but one of the probabilities  $\mathbf{p} = (p_1, \dots, p_N)$  for an event to be in one of the  $N$  bins. One of the  $p_i$  is fixed from the constraint

$$\sum_{i=1}^N p_i = 1 . \quad (21)$$

The multinomial distribution for  $\mathbf{n}$  is

$$P(\mathbf{n}|\mathbf{p}, n_{\text{tot}}) = \frac{n_{\text{tot}}!}{n_1!n_2!\dots n_N!} p_1^{n_1} p_2^{n_2} \dots p_N^{n_N} . \quad (22)$$

Since  $n_{\text{tot}}$  is fixed, we can regard the parameters to be  $\nu_i = p_i n_{\text{tot}}$ . The log-likelihood function is then

$$\ln L(\boldsymbol{\nu}) = \sum_{i=1}^N n_i \ln \frac{\nu_i}{n_{\text{tot}}} + C . \quad (23)$$

As in the Poisson case the ML estimators for the  $\nu_i$  are found to be  $\hat{\nu}_i = n_i$ , so the statistic  $t_{\boldsymbol{\nu}}$  then becomes

$$t_{\boldsymbol{\nu}} = -2 \sum_{i=1}^N n_i \ln \frac{\nu_i}{n_i} . \quad (24)$$

That is, it is the same as in the Poisson case but without the terms  $-\nu_i + n_i$ . Because here there are only  $N - 1$  fitted parameters (one of the  $\hat{\nu}_i$  can be determined from  $n_{\text{tot}}$  minus the sum of the rest), Wilks' theorem says that  $t_{\boldsymbol{\nu}}$  follows a chi-square distribution for  $N - 1$  degrees of freedom.

If the  $N$  mean values  $\boldsymbol{\nu}$  are determined from  $M$  parameters  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_M)$ , then the distribution of the corresponding  $q_{\boldsymbol{\nu}}$ ,

$$q_{\boldsymbol{\nu}} = -2 \sum_{i=1}^N n_i \ln \frac{\nu_i(\hat{\boldsymbol{\theta}})}{n_i} , \quad (25)$$

is a chi-square distribution for  $N - M - 1$  degrees of freedom.

Now suppose instead of evaluating the  $\nu_i$  terms in Eqs. (19) and (25) with the ML estimators for  $\boldsymbol{\theta}$ , we write the corresponding quantities as a function of  $\boldsymbol{\theta}$ , i.e.,

$$\chi_{\text{M}}^2(\boldsymbol{\theta}) = -2 \sum_{i=1}^N n_i \ln \frac{\nu_i(\boldsymbol{\theta})}{n_i} , \quad (26)$$

$$\chi_{\text{P}}^2(\boldsymbol{\theta}) = -2 \sum_{i=1}^N \left[ n_i \ln \frac{\nu_i(\boldsymbol{\theta})}{n_i} - \nu_i(\boldsymbol{\theta}) + n_i \right] , \quad (27)$$

where the subscripts M and P refer to the multinomial or Poisson cases, respectively. These expressions are equal to the corresponding values of  $-2 \ln L(\boldsymbol{\theta})$ . So to maximize the likelihood one can simply minimize  $\chi_{\text{P}}^2(\boldsymbol{\theta})$  or  $\chi_{\text{M}}^2(\boldsymbol{\theta})$ , and the same ML estimators  $\hat{\boldsymbol{\theta}}$  will result.

As an added bonus, however, the value of the minimized function can be used directly for a test of the goodness of fit, and to the extent that Wilks' theorem is satisfied, its sampling distribution is a chi-square distribution for  $N - M$  (Poisson) or  $N - M - 1$  (multinomial) degrees of freedom.

## References

- [1] S.S. Wilks, *The large-sample distribution of the likelihood ratio for testing composite hypotheses*, Ann. Math. Statist. **9** (1938) 60-62.
- [2] G. Cowan, *Statistical Data Analysis*, Oxford University Press, 1998.
- [3] Steve Baker and Robert D. Cousins, *Clarification of the use of the chi-square and likelihood functions in fits to histograms*, NIM **221** (1984) 437.