Some useful links:

1. https://en.wikipedia.org/wiki/Fisher_information
2. http://thestatsgeek.com/2014/02/08/wald-vs-likelihood-ratio-test/
3. https://en.wikipedia.org/wiki/Models_of_DNA_evolution#JC69_model_(Jukes_and_Cantor_1969)

Note that (under a few regularity conditions, such as the MLE not being at the boundary of the feasible range), the Fisher Information ($\mathcal{I}$) at a point in parameter space, θ, is:

$$\mathcal{I}(\theta) = -\mathbb{E}_{\mathbb{P}(X|\theta)}\left(\frac{\partial^2 \ln \mathbb{P}(X \mid \theta)}{\partial \theta^2}\right) \quad (1)$$

In other words, it is the expected curvature (second derivative) of the log-likelihood function, where the expectation is taken with respect to the probability of the data (the likelihood function itself).

Revisiting sequence comparisons

We worked through a brief discussion of the code associated with the Kimura two-parameter model in class on Monday, 25 March. For simplicity, let's consider the special case where $\kappa = 1$, so all mutation types are occurring at the same rate (this is called the Jukes-Cantor model in molecular evolution). If we have a pair of sequences of total number of sites $= n$, then x will just be a count of the number of sites at which they differ. Simplifying the probability statements that we saw on Monday will reveal that:

$$\mathbb{P}(\text{same base in each seq. at a site} \mid \nu) = \frac{1 + 3e^{-4\nu/3}}{4} \quad (2)$$

$$\mathbb{P}(\text{a difference between seqs.} \mid \nu) = \frac{3 - 3e^{-4\nu/3}}{4} \quad (3)$$

where ν is a branch length which is interpreted as the expected number of changes per site. This looks complicated, but it really just amounts to their being some probability of a difference: $0 \leq p < 3/4$, so the probability of staying the same is $1 - p$ and the feasible range is $0 \leq 1 - p < 1/4$.

The easiest way to perform the inference is simply to use this alternative parameterization, an convert from p into the ν parameterization whenever needed:

$$p = \frac{3 - 3e^{-4\nu/3}}{4} \quad (4)$$

$$4p = 3(1 - e^{-4\nu/3}) \quad (5)$$

$$\frac{4p}{3} = 1 - e^{-4\nu/3} \quad (6)$$

$$e^{-4\nu/3} = 1 - \frac{4p}{3} \quad (7)$$

$$-\frac{4\nu}{3} = \ln\left(1 - \frac{4p}{3}\right) \quad (8)$$

$$\nu = -0.75 \ln\left(1 - \frac{4p}{3}\right) \quad (9)$$

Note that if we put in a value of $p > 0.75$, we end up trying to take the log of a negative number, so we have to pay attention to our feasible range.

0.1 Fisher Information for the binomial

Lo and behold, this means that (if we treat each site as independent), our likelihood is simply a binomial likelihood:

$$\log L(p) = x \ln p + (n-x)\ln(1-p) \tag{10}$$

So:

$$\frac{\partial \log L(p)}{p} = \frac{x}{p} - \frac{(n-x)}{(1-p)} \tag{11}$$

$$\hat{p} = \begin{cases} \frac{x}{n} & \text{if } \frac{x}{n} \leq 0.75 \\ 0.75 & \text{otherwise} \end{cases} \tag{12}$$

$$\mathcal{I}(p) = -\mathbb{E}_{\mathbb{P}(x|p)}\left(\frac{\partial^2 \ln \mathbb{P}(X \mid \theta)}{\partial \theta^2}\right) \tag{13}$$

$$\frac{\partial^2 \log L(p)}{p^2} = -\frac{x}{p^2} - \frac{(n-x)}{(1-p)^2} \tag{14}$$

Fortunately, the expected value for the binomial is very similar to what we saw with the Bernoulli distribution that JKK talked about. If the probability of a success is p, then the expected number of successes is simply: np:

$$\mathbb{E}_{\mathbb{P}(x|p)}(x) = np \tag{15}$$

$$\mathbb{E}_{\mathbb{P}(x|p)}(n-x) = n - np = n(1-p) \tag{16}$$

$$\tag{17}$$

Therefore the expected curvature at any point simplifies:

$$\mathbb{E}_{\mathbb{P}(x|p)}\left(\frac{\partial^2 \ln \mathbb{P}(X \mid \theta)}{\partial \theta^2}\right) = -\frac{\mathbb{E}_{\mathbb{P}(x|p)}(x)}{p^2} - \frac{\mathbb{E}_{\mathbb{P}(x|p)}(n-x)}{(1-p)^2} \tag{18}$$

$$= -\frac{np}{p^2} - \frac{n(1-p)}{(1-p)^2} \tag{19}$$

$$= -\frac{n}{p} - \frac{n}{(1-p)} \tag{20}$$

$$= -\frac{n - np + np}{p(1-p)} \tag{21}$$

$$= -\frac{n}{p(1-p)} \tag{22}$$

$$\mathcal{I}(p) = \frac{n}{p(1-p)} \tag{23}$$

$$\tag{24}$$

Note that the data does not appear in the generic form of Fisher's information because it is the expected value over all possible realizations of the data given the value of the parameter p.

Wald test and standard errors

As JKK pointed out, you can compute a Z^2 value for a test generically using:

$$t = Z^2 = \frac{(\theta - \theta_0)^2}{\mathrm{Var}(\hat{\theta})} \tag{25}$$

and you can look that up using a chi-square table with 1 degree of freedom (because the chi-square distribution with k degrees of freeedom is simply the distribution you get if you sum of the squares of k draws from the standard normal distribution).

Thus we can see that we can convert the distance between our point estimate of the parameter value to a Z statistic by dividing through by the standard error of θ

Interestingly (in many cases)

$$\mathrm{SE}(\hat{\theta}) = \sqrt{\frac{1}{\mathcal{I}(\hat{\theta})}} \tag{26}$$

That may seem like magic, but hopefully it makes some sense.

Fisher info is the variance of the slope of the log-likelihood surface.

The square root in the standard error formula should seem reasonable given the relationship between variances and standard deviations.

Since we compare hypotheses using the difference in their log-likelihoods, the slope of the $\ln L$ surface is saying something about how well we expect the data to be able to discriminate between different parameter values. So high Fisher Information, implies a very curved $\ln L$ around the MLE. Thus the peak is very sharp and pointed when Fisher Information. That means that we don't have to go out very far from the MLE before we have arrived at a much worse $\ln L$. This is the intuition of why the standard error should be something like 1 over the Fisher Information.

Wald test and standard errors for the binomial

Going back to our specific case. If we only consider cases in which $\frac{x}{n} < 0.75$, then:

$$\hat{p} = \frac{x}{n} \tag{27}$$

$$\mathcal{I}(\hat{p}) = \frac{n}{\hat{p}(1-\hat{p})} \tag{28}$$

$$\mathrm{SE}(\hat{p}) = \sqrt{\frac{1}{\mathcal{I}(\hat{p})}} \tag{29}$$

$$= \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{30}$$

Leading us (using the 1.96 from the Z distribution) to the argument that we are about 95% confident that:

$$\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Once again we find that we can derive a generic statistical test and CI using some general rules for working with likelihoods.