Using the Wilson Score Interval with Weighted Events

version 1

By Bryan Kerns

University of Illinois Urbana-Champaign

It is helpful to know the binomial theorem, equation 1.

$$(x+y)^n = \sum_{i=0}^{n} \binom{n}{i} x^i y^{(n-i)} \tag{1}$$

$\binom{n}{i}$ is syntax for n choose i. I will start by going over the derivation of the mean and variance for a standard binomial distribution, then move on to the derivation for a binomial distribution with weighted events.

1 Derivation of mean in binomial distribution

v is the number of successes in n trials with a probability of success p and probability of failure q. I will use $\hat{p}$ as the observed value of p, $\frac{v}{n}$. The mean is the sum of each possible value of v times the probability of that value. [1] was very helpful for this section and the next.

$$Prob(v) = \binom{n}{v} p^v q^{n-v} \tag{2}$$

$$\sum_{v=0}^{n} v \binom{n}{v} p^v q^{n-v} = \langle v \rangle \tag{3}$$

Start with the binomial theorem.

$$(p+q)^n = \sum_{v=0}^{n} \binom{n}{v} p^v q^{(n-v)} \tag{4}$$

1 is true for any p and q, not just $q = 1 - p$. Take the derivative with respect to p treating q as independent of p.

$$n(p+q)^{n-1} = \sum_{v=0}^{n} v \binom{n}{v} p^{v-1} q^{(n-v)} \tag{5}$$

Equation 5 is also true for any p and q. Multiply both sides by p.

$$np(p+q)^{n-1} = \sum_{v=0}^{n} v \binom{n}{v} p^v q^{(n-v)} \tag{6}$$

Set $p+q=1$.

$$np = \sum_{v=0}^{n} v \binom{n}{v} p^v q^{(n-v)} \tag{7}$$

$$\langle v \rangle = np \tag{8}$$

$$\langle \hat{p} \rangle = \left\langle \frac{v}{n} \right\rangle = p \tag{9}$$

2 Derivation of the variance in binomial distribution

$$\sigma_v^2 = \left\langle v^2 \right\rangle - \langle v \rangle^2 \tag{10}$$

$$\sum_{v=0}^{n} v^2 \binom{n}{v} p^n q^{n-v} = \left\langle v^2 \right\rangle \tag{11}$$

Start with the binomial theorem again.

$$(p+q)^n = \sum_{v=0}^{n} \binom{n}{v} p^v q^{(n-v)} \tag{12}$$

Take the derivative with respect to p twice.

$$n(n-1)(p+q)^{n-2} = \sum_{v=0}^{n} v(v-1) \binom{n}{v} p^{v-2} q^{(n-v)} \tag{13}$$

Multiply both sides by p^2.

$$n(n-1)p^2(p+q)^{n-2} = \sum_{v=0}^{n}(v^2-v)\binom{n}{v}p^v q^{(n-v)} \tag{14}$$

Set $p+q=1$.

$$n(n-1)p^2 = \left\langle v^2 \right\rangle - \langle v \rangle \tag{15}$$

We already know $\langle v \rangle$ from equation 8

$$n(n-1)p^2 + np = \left\langle v^2 \right\rangle \tag{16}$$

On the right hand side substitute in using equation 10

$$n^2p^2 - np^2 + np = \sigma_v^2 + \langle v \rangle^2 \tag{17}$$

$$-np^2 + np = \sigma_v^2 \tag{18}$$

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

$$\sigma_{\hat{p}}^2 = \sigma_{\frac{v}{n}}^2 = p(1-p)/n \tag{20}$$

3 Derivation of mean for binomial distribution with weighted events

w_i will be the weight of a single event. W will be the total weight of all successful events. To account for the weights, $\hat{p}$ needs a new definition, equation 21.

$$\hat{p} = \frac{\sum_i^{success} w_i}{\sum_i w_i} \tag{21}$$

$$\langle W \rangle = \textstyle\sum_{v=0}^{n} p^v q^{n-v} \sum_{i_1, i_2 \neq i_1, \ldots, (i_v \neq i_1, \neq i_2, \ldots)} (w_{i_1} + \cdots + w_{i_v}) \tag{22}$$

The number of different possible combinations of i_1 through i_v is n choose v, so there are $\binom{n}{v}$ parts to the right most summation. There are v weights in each part, for $\binom{n}{v}v$ weights in total. Each weight should appear the same number of times as every other weight, which means each weight appears $\binom{n}{v}\frac{v}{n}$ times.

$$\langle W \rangle = \sum_{v=0}^{n} p^v q^{n-v} \binom{n}{v} \frac{v}{n} \sum_i w_i \tag{23}$$

$$\langle W \rangle = \frac{\sum_i w_i}{n} \sum_{v=0}^{n} v p^v q^{n-v} \binom{n}{v} \tag{24}$$

We already know what the right most summation is from equation 8

$$\langle W \rangle = \frac{\sum_i w_i}{n} np \tag{25}$$

$$\langle W \rangle = p \sum_i w_i \tag{26}$$

$$\langle \hat{p} \rangle = \left\langle \frac{W}{\sum_i w_i} \right\rangle = p \tag{27}$$

As expected, the mean does not change.

4 Derivation of variance for binomial distribution with weighted events

$$\left\langle W^2 \right\rangle = \textstyle\sum_{v=0}^{n} p^v q^{n-v} \sum_{i_1, i_2 \neq i_1, \ldots, (i_v \neq i_1, \neq i_2, \ldots)} (w_{i_1} + \cdots + w_{i_v})^2 \tag{28}$$

The number of different possible combinations of i_1 through i_v is n choose v, so there are $\binom{n}{v}$ parts to the right most summation. There are v terms of type w_i^2 in each part, and $v(v-1)$

terms of type $w_i w_j$ (treating $w_1 w_2$ as different from $w_2 w_1$). Each w_i^2 term should appear the same number of times as every other similar term, and each $w_i w_j$ term should appear the same number of times as every other similar term. This means each w_i^2 term appears $\binom{n}{v}\frac{v}{n}$ times, and each $w_i w_j$ term appears $\binom{n}{v}\frac{v(v-1)}{n(n-1)}$ times.

$$\left\langle W^2 \right\rangle = \sum_{v=0}^{n} p^v q^{n-v} \binom{n}{v} \frac{v}{n} \sum_i w_i^2 + \sum_{v=0}^{n} p^v q^{n-v} \binom{n}{v} \frac{v(v-1)}{n(n-1)} \sum_i \sum_{i \neq j} w_i w_j \tag{29}$$

$$\left\langle W^2 \right\rangle = \frac{\sum_i w_i^2}{n} \sum_{v=0}^{n} v p^v q^{n-v} \binom{n}{v} + \frac{\sum_i \sum_{i \neq j} w_i w_j}{n(n-1)} \sum_{v=0}^{n} v(v-1) p^v q^{n-v} \binom{n}{v} \tag{30}$$

We know two of the summations from equations 8 and 19

$$\left\langle W^2 \right\rangle = \frac{\sum_i w_i^2}{n} np + \frac{\sum_i \sum_{j \neq i} w_i w_j}{n(n-1)} n(n-1)p^2 \tag{31}$$

$$\left\langle W^2 \right\rangle = p \sum_i w_i^2 + p^2 \sum_i \sum_{j \neq i} w_i w_j \tag{32}$$

We know $\langle W \rangle$ from equation 26

$$\sigma_W^2 = \left\langle W^2 \right\rangle - \langle W \rangle^2 = p \sum_i w_i^2 + p^2 \sum_i \sum_{j \neq i} w_i w_j - p^2 (\sum_i w_i)^2 \tag{33}$$

$$\sigma_W^2 = p \sum_i w_i^2 + p^2 \sum_i \sum_{j \neq i} w_i w_j - p^2 \sum_i w_i^2 - p^2 \sum_i \sum_{j \neq i} w_i w_j \tag{34}$$

$$\sigma_W^2 = p(1-p) \sum_i w_i^2 \tag{35}$$

$$\sigma_{\hat{p}}^2 = \sigma^2{}_{\frac{W}{\sum_i w_i}} = p(1-p) \frac{\sum_i w_i^2}{(\sum_i w_i)^2} \tag{36}$$

Unlike the mean, the new variance for $\hat{p}$ is significantly more complicated. One can see that it reduces to equation 20 when all the weights are equal. When one of the weights is much larger than all the others combined, it approaches $p(1-p)$ as if $n = 1$ in equation 20.

5 Derivation of Wilson score interval

[2] is the source for this entire section. I am deriving the non-weighted version here. Define P_λ as the area between $-\lambda\sigma$ and $+\lambda\sigma$ under a bell curve (This is the opposite definition from [2], but I am rewording for clarity). This means we have P_λ confidence of equation 37 being true.

$$|p - \hat{p}| < \lambda\sigma_{\hat{p}} \tag{37}$$

The ends of the confidence interval can be found by changing the $<$ to a $=$.

$$|p - \hat{p}| = \lambda\sigma_{\hat{p}} \tag{38}$$

Take note that equation 38 is only true for p equal to one of the ends of the confidence interval. We can square this to get rid of the absolute value sign, then get the equation in the form of a quadratic equation.

$$(p - \hat{p})^2 = \lambda^2\sigma_{\hat{p}}^2 \tag{39}$$

$$(p - \hat{p})^2 = \lambda^2 p(1-p)/n \tag{40}$$

I define x as λ^2/n for simplicity.

$$p^2 - 2p\hat{p} + \hat{p}^2 - xp + p^2x = 0 \tag{41}$$

$$p^2(1+x) + p(-2\hat{p} - x) + \hat{p}^2 = 0 \tag{42}$$

Quadratic equation solution

$$p = \frac{2\hat{p}+x\pm\sqrt{4\hat{p}^2+4\hat{p}x+x^2-4\hat{p}^2-4\hat{p}^2x}}{2+2x} \tag{43}$$

$$p = \frac{\hat{p}+0.5x}{1+x} \pm \frac{\sqrt{\hat{p}(1-\hat{p})x+0.25x^2}}{1+x} \tag{44}$$

The lower value is the lower bound for a confidence of P_λ and the greater value is the upper bound. Again, we do not have a general equation for p, but for what p could be if it was exactly at the end of the confidence interval. Note that the middle of the interval depends on x and thus on λ. The median value of p is still $\hat{p}$, as you can deduce from the limit of x approaches 0.

6 Wilson Score Interval and Weighted Events

The logic up to equation 39 is the same. However, we have equation 45 instead of equation 40 because the variance has changed.

$$p^2 - 2p\hat{p} + \hat{p}^2 = \lambda^2 p(1-p)\frac{\sum_i w_i^2}{(\sum_i w_i)^2} \tag{45}$$

If we now define x as $\lambda^2 \frac{\sum_i w_i^2}{(\sum_i w_i)^2}$, we end up with 41, and thus equation 44 again, with a different but still known value of x. This interval for weighted events in a binomial distribution can be applied to our kTracker efficiency fits to get the proper uncertainty. I recommend using the median as the center point and using asymmetric error bars.

References

[1] John R. Taylor. *An Introduction to Error Analysis, The Study of Uncertainties in Physical Measurements, 2nd edition.* University Science Books, 1982, 1997.

[2] Edwin B. Wilson. *Probable Inference, The Law of Succession, and Statistical Inference.* Journal of the American Statistical Association, Vol. 22, No. 158 (Jun., 1927), 209-212.