Part I: Background on the Binomial Distribution

A random variable is said to have a Bernoulli distribution if it takes on the value 1 with probability "p" and the value 0 with probability "1 - p".

The number of "successes" in "n" independent Bernoulli trials is said to have a binomial distribution. The probability mass function for the binomial distribution is:

```
probability(observing k successes | n, p) = choose(n, k) * p^k * (1-p)^(n-k)
```

Let's consider a concrete example. Suppose we want to know the probability of observing "k" heads in "n" tosses of a coin, where the probability of landing on heads is "p" = 1/2. The coin tosses are independent, because the outcome for the previous coin toss does not affect the outcome for the next coin toss. The coin tosses are identically distributed, because the probability of landing on heads is the same for all coin tosses. There are eight possible outcomes:

```
tails, tails, tails:    probability of this outcome = (1 - p) * (1 - p) * (1 - p) = 1/8

tails, tails, heads:    probability of this outcome = (1 - p) * (1 - p) *    p    = 1/8

tails, heads, tails:    probability of this outcome = (1 - p) *    p    * (1 - p) = 1/8

tails, heads, heads:    probability of this outcome = (1 - p) *    p    *    p    = 1/8

heads, tails, tails:    probability of this outcome =    p    * (1 - p) * (1 - p) = 1/8

heads, tails, heads:    probability of this outcome =    p    * (1 - p) *    p    = 1/8

heads, heads, tails:    probability of this outcome =    p    *    p    * (1 - p) = 1/8

heads, heads, heads:    probability of this outcome =    p    *    p    *    p    = 1/8
```

```
> # Note: there are ...
> # 3 outcomes where the number of "successes" is 1 [choose(3, 1) == 3]
> # 3 outcomes where the number of "successes" is 2 [choose(3, 2) == 3]
> n = 3
> p = 1/2
> k = c(0, 1, 2, 3)
> choose(n, k) * p^k * (1-p)^(n-k)    # probability of observing "k" successes
[1] 0.125 0.375 0.375 0.125
```

Part II: Logistic Regression Using Simulated Data

As you may recall, the binomial distribution is our source of uncertainty for logistic regression. This is analogous to using the Gaussian (normal) distribution as the source of uncertainty for linear regression.

```
> n = c(100000, 100000)
> p = c(1/3, 2/3)
> x = c(rep(0, n[1]), rep(1, n[2]))
> y = rep(0, sum(n))
> y[sample(which(x == 0), rbinom(1, n[1], p[1]))] = 1
> y[sample(which(x == 1), rbinom(1, n[2], p[2]))] = 1
> glm(y ~ x, family = binomial)

Call:  glm(formula = y ~ x, family = binomial)

Coefficients:
(Intercept)            x
    -0.6924       1.3947

> round(c(log((1/3)/(2/3)), log((2/3)/(1/3)) - log((1/3)/(2/3))), 2)
[1] -0.69  1.39
```

Part III: Estimating a Binomial Proportion

Suppose we have a set of Bernoulli outcomes, and we want to estimate the binomial proportion from data (i.e. we want to estimate "p"). We can generate a "maximum likelihood estimate" for the binomial proportion by setting the derivative of the log likelihood to zero and solving for "p".

$$\begin{aligned} probability\left(k \mid n,\theta\right) &= \binom{n}{k}\theta^{k}\left(1-\theta\right)^{\left(n-k\right)} \\ &= \binom{n}{\sum_{i=1}^{n} y_i}\theta^{\sum_{i=1}^{n} y_i}\left(1-\theta\right)^{\left(n-\sum_{i=1}^{n} y_i\right)} \end{aligned}$$

… so taking the log of both sides gives us …

$$\begin{aligned} \log\left(probability\left(k \mid n,\theta\right)\right) &= \log\left(\binom{n}{\sum_{i=1}^{n} y_i}\theta^{\sum_{i=1}^{n} y_i}\left(1-\theta\right)^{\left(n-\sum_{i=1}^{n} y_i\right)}\right) \\ &= \log\binom{n}{\sum_{i=1}^{n} y_i} + \left(\sum_{i=1}^{n} y_i\right)\log\left(\theta\right) + \left(n-\sum_{i=1}^{n} y_i\right)\log\left(1-\theta\right) \end{aligned}$$

… and taking the derivative with respect to theta gives us …

$$\begin{aligned} \frac{\partial}{\partial\theta}\log\left(probability\left(k \mid n,\theta\right)\right) &= \frac{\partial}{\partial\theta}\left[\log\binom{n}{\sum_{i=1}^{n} y_i} + \left(\sum_{i=1}^{n} y_i\right)\log\left(\theta\right) + \left(n-\sum_{i=1}^{n} y_i\right)\log\left(1-\theta\right)\right] \\ &= \frac{\partial}{\partial\theta}\left[\log\binom{n}{\sum_{i=1}^{n} y_i}\right] + \frac{\partial}{\partial\theta}\left[\left(\sum_{i=1}^{n} y_i\right)\log\left(\theta\right)\right] + \frac{\partial}{\partial\theta}\left[\left(n-\sum_{i=1}^{n} y_i\right)\log\left(1-\theta\right)\right] \\ &= 0 + \frac{\sum_{i=1}^{n} y_i}{\theta} - \frac{n-\sum_{i=1}^{n} y_i}{1-\theta} \end{aligned}$$

… and setting the derivative equal to zero and solving for theta gives us …

$$\begin{aligned} \frac{\sum_{i=1}^{n} y_i}{\theta} - \frac{n-\sum_{i=1}^{n} y_i}{1-\theta} &= 0 \\ \theta\left(1-\theta\right)\frac{\sum_{i=1}^{n} y_i}{\theta} &= \theta\left(1-\theta\right)\frac{n-\sum_{i=1}^{n} y_i}{1-\theta} \\ \left(1-\theta\right)\sum_{i=1}^{n} y_i &= \theta\left(n-\sum_{i=1}^{n} y_i\right) \\ \sum_{i=1}^{n} y_i - \theta\sum_{i=1}^{n} y_i &= \theta n - \theta\sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} y_i &= \theta n \\ \theta &= \frac{\sum_{i=1}^{n} y_i}{n} \end{aligned}$$

This estimate is sometimes called a "frequentist" estimate of the binomial proportion. The problem with this estimate occurs when the actual binomial proportion is either very small or very large, or when “n” is small.

Suppose the actual binomial proportion is p = 0.001 and we only have 100 observations. The expected number of "successes" (n * p = 100 * 0.001 = 0.1) is less than one. This means our frequentist estimate is likely to be 0, even though the actual binomial proportion is 0.001. This can cause trouble in any domain where observed probabilities can be very small or very large; e.g. when estimating click-through rates for online advertising or purchase probabilities for items in the "tail" of an ecommerce merchant's inventory. We would like to avoid degenerate estimates that express the certainty of either zero or one.

Bayesian estimation can be used to adjust the frequentist estimate. Recall that posterior = prior * likelihood / evidence.

$$
\begin{aligned}
prior &= p\left(\theta \mid \alpha_1, \alpha_0\right) \\
&= \frac{\theta^{\alpha_1}\left(1-\theta\right)^{\alpha_0}}{\Gamma\left(\alpha_1\right)\Gamma\left(\alpha_0\right)/\Gamma\left(\alpha_1+\alpha_0\right)} \\
&= dbeta\left(\theta,\ \alpha_1,\ \alpha_0\right) \\
likelihood &= probability\left(k \mid n, \theta\right) \\
&= \binom{n}{k}\theta^k\left(1-\theta\right)^{(n-k)} \\
evidence &= p\left(n, k, \alpha_1, \alpha_0\right) \\
&= \int_{\theta=0}^{1}\binom{n}{k}\theta^k\left(1-\theta\right)^{(n-k)}\frac{\theta^{\alpha_1}\left(1-\theta\right)^{\alpha_0}}{\Gamma\left(\alpha_1\right)\Gamma\left(\alpha_0\right)/\Gamma\left(\alpha_1+\alpha_0\right)}d\theta \\
posterior &= p\left(\theta \mid n, k, \alpha_1, \alpha_0\right) \\
&= \frac{\frac{\theta^{\alpha_1}\left(1-\theta\right)^{\alpha_0}}{\Gamma\left(\alpha_1\right)\Gamma\left(\alpha_0\right)/\Gamma\left(\alpha_1+\alpha_0\right)}\binom{n}{k}\theta^k\left(1-\theta\right)^{(n-k)}}{\int_{\theta=0}^{1}\frac{\theta^{\alpha_1}\left(1-\theta\right)^{\alpha_0}}{\Gamma\left(\alpha_1\right)\Gamma\left(\alpha_0\right)/\Gamma\left(\alpha_1+\alpha_0\right)}\binom{n}{k}\theta^k\left(1-\theta\right)^{(n-k)}d\theta} \\
&= \frac{\theta^{(k+\alpha_1)}\left(1-\theta\right)^{(n-k+\alpha_0)}}{\Gamma\left(k+\alpha_1\right)\Gamma\left(n-k+\alpha_0\right)/\Gamma\left(\left(k+\alpha_1\right)+\left(n-k+\alpha_0\right)\right)} \\
&= dbeta\left(\theta,\ k+\alpha_1,\ n-k+\alpha_0\right)
\end{aligned}
$$

The expected value (mean) of the posterior distribution is simply (k + $alpha_1$) / (k + $alpha_1$ + (n-k) + $alpha_0$). We are essentially adjusting the frequentist estimate by adding $alpha_1$ and $alpha_0$ "pseudo" counts to the "k" and "n-k" observed counts. Using the posterior mean for our estimate of the binomial proportion is the same as using a weighted mean of (k / n) and ($alpha_1$ / ($alpha_1$ + $alpha_0$)), so we'll want to keep the pseudo counts relatively small to allow the observed counts to overtake the pseudo counts relatively quickly …

$$\begin{aligned}
\hat{\theta} &= \left(\frac{\alpha_1 + \alpha_0}{n + \alpha_1 + \alpha_0}\right)\left(\frac{\alpha_1}{\alpha_1 + \alpha_0}\right) + \left(\frac{n}{n + \alpha_1 + \alpha_0}\right)\left(\frac{k}{n}\right) \\
&= \left(\frac{1}{n + \alpha_1 + \alpha_0}\right)\left((\alpha_1 + \alpha_0)\left(\frac{\alpha_1}{\alpha_1 + \alpha_0}\right) + n\left(\frac{k}{n}\right)\right) \\
&= \left(\frac{1}{n + \alpha_1 + \alpha_0}\right)(\alpha_1 + k) \\
&= \frac{k + \alpha_1}{n + \alpha_1 + \alpha_0}
\end{aligned}$$

We're using a Bayesian prior to avoid the certainty of using either zero or one for our estimate of the binomial proportion. In practice, folks often use either $alpha_1$ = $alpha_0$ = 1 (a Laplace prior) or $alpha_1$ = $alpha_0$ = 0.5 (a Jeffreys prior).

We can obtain a 95% confidence interval for the binomial proportion with the following R command …

```
qbeta(c(0.025, 0.975), k + alpha.1, n - k + alpha.0)
```

Note: the R binom.test() command uses the following slightly more conservative (larger) interval …

```
c(qbeta(0.025, k, n - k + 1), qbeta(0.975, k + 1, n - k))
```