Comments on questions/answers to Assignment 1

1.1 This CI is 'SE-based'. It substitutes the point estimate (3/20) into the SD formula SD(statistic) = sqrt[P(1-P)/20], and uses the same Gaussian approximation i.e. the same (estimated) SD, for both limits.

You may not realize it when you write the CI as point estimate +/- times the SE, but what you are really doing is obtaining the lower limit P[L] such that 3/20 is a 'high' estimate

i.e. 3/20 = P[L] + 1.96 SD and P[U] likewise 3/20 = P[U] – 1.96 SD

If you use the same SD for both, and substitute in 3/20 as the P, then rearranging and solving these 2 equations for P[L[and P[U] yields the familiar
3/20 +/- 1.96 Sqrt[(3/20)(17/20)/20].

1.2 The reason this is a 'first principles' CI is that it does focus separately, and correctly, on each limit, and uses the correct distribution (binomial) at each limit. It makes no approximations. So the shape of the distribution at each limit, which may be quite un-Gaussian, especially at a limit which is near P = 0 or P = 1, is taken care of.

1.3 Even though it may not be obvious from the computational formula, this CI goes 'part way' towards first principles in that it recognizes the 2 separate distributions at the 2 separate limits… but it still approximates each binomial by a Gaussian distribution.

1.4 The main reason for introducing these at this early stage was as a prelude to logit regression (and the log link when dealing with risk ratios (relative risks)).

We are dealing with the null model here; if we directly use the proportion P, our model is P = beta_0

If we use the logit link (scale) our model is logit[P] = beta_0

If we use the log link (scale) our model is log[P] = beta_0

[I am using upper case P for parameter]

1.5 I created come confusion here with the meaning of 'close'. Close in absolute terms? In relative terms? All I wanted to get across is that the binomial is less well approximated by Gaussian the further one moves away from P = 0.5 towards P = 0.0 or P = 1.0.

So it is possible that even if sample size is small, but one limit is near P = 0.5, one could use a Gaussian approximation. But if a limit is near P = 0, or P=1, don't expect to be able to derive that limit using a Gaussian approximation.

2.1 The one which substitutes the point estimate P = 0 into SD = sqrt [P (1-P) / n] is in trouble.

Wilson gets a sensible upper bound, since he uses P = P[upper] as the SD to use with his Gaussian approximation.

The logit and log based CI, both substitute the point estimates, and so 'bomb out' as well.

2.2 The point about Bayesian inference for a proportion is that if one uses a Beta prior, the posterior is also Beta. Its parameters are the sum of the a's and the sum of the b's.

The reason for introducing this question in a chapter is as a prelude to an analysis where 41 was the total number of cases (exposed and unexposed) in a RCT of a HPV vaccine. When we are at the 'extreme' in such 2 sample problems involving 2 proportions, we can CONDITION on the sum of the numbers of cases in the 2 arms.

3.1 For a risk difference, usually have to use large sample (Gaussian) methods for the CI.

3.2/3 Practice in changing scales (reciprocal) and also converting the CI with you to the new scale – by the exact same method as for the point estimate.

3.4 This is now the beginning of a logistic regression with a single "X" variable with I values (0 = conventional, 1 = * surgery).

Likewise, using log instead of logit link, leads to the risk ratio. A prelude to binary regression methods.

3.5 Test-based CI's not as bad as made out to be … unless very far from null. The key idea in the test based CI is that it calculates the SD for the "difference statistic" at the null value of the comparative parameter.
i.e. at RD = 0 i.e. at RR = 1 i.e. at OR = 1.