According to the above 3 tables of summary statistics, answer questions in the highlighted Box 1.

Box 1
1. Review the data description in the beginning of this lab. What are the variable types for HT, UI, FTV, and BWT? Dichotomous, discrete, continuous, etc.? (1pt*4=4pts) HT: dichotomous, UI: dichotomous, FTV: discrete, BWT: continuous 2. The means and standard deviations of the variables other than HT are similar for samples of sizes 50 and 100. These values are also close to that of the original birth data. Can you give a short explanation? (5pts) Sample mean and sample standard deviation are unbiased estimation for the population mean and standard deviation. Therefore, they should be consistent among the samples and approximate the true values when the sample is representative with large enough sample size. (Note that for too small sample, even if the sampling method is random, the estimates can deviate from the population parameters.) 3. What trend do you find in standard error for the variables other than HT when sample size increases from 10, 50, to 100? Please explain the reason. (5pts) The SEs decreases with increasing sample size according to the CLT $SE_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$, so when n increases, SE decreases. 4. The mean, standard deviation and standard error of HT are all 0 for the sample of size 10. What do you think is the reason? Does this make your reasoning in above questions 2 and 3 invalid? Please explain As mentioned question 2, the random sample must be representative with large enough sample size so that sample estimates can approximate the parameters. The variable HT is dichotomous and must be imbalanced with rare value 1 so it is unlikely to obtain 1 in a very small sample of size 10. Therefore, the mean and standard deviation of HT for the sample of size 10 are different from that for other two larger samples. Accordingly, $SE_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$ will be 0 because of a zero standard deviation. The reasoning in questions 2 and 3 are still valid because CLT can be applied as long as the sample is representative with sufficient sample size. (8pts)

Read above output and answer questions in Box 2.

Box 2
1. Please write the hypotheses for the question brought up by the researchers. What are the point estimate and CI (focus on the Wald CI) for the proportion of babies of not low birth weight (BWT=0)? Interpret the CI in context and explain whether it leads us to reject the null hypothesis. H_0: $p = .7$ i.e., 70% of babies are born with normal weight (2pts) H_A: $p \neq .7$ i.e., the proportion of babies of normal birth weight is not 70% (2pts) The point estimate $\hat{p} = 0.6878$ and the CI is (0.6218, 0.7539) (focus on the Wald CI). (3pts+4pts) We are 95% confident that the true proportion of babies of normal birth weight is between 0.6218 and 0.7539 which includes 0.7 so 0.7 is a plausible true proportion. Therefore, we fail to reject the null hypothesis that 70% of babies are born with normal weight. (6pts) 2. Prior to constructing the CI, can you predict whether 99% CI will be more precise or less precise than the 95% one? Why? Validate your answer with implementation in SAS studio and paste the output here. 99% CI will be less precise than 95% CI because the higher confidence level, the wider interval. We gain confidence at cost of precision. (4pts) The 99% CI is as below. (6pts)

Binomial Proportion	
LOW = 0	
Proportion	0.6878
ASE	0.0337

Confidence Limits for the Binomial Proportion		
Proportion = 0.6878		
Type	99% Confidence Limits	
Agresti-Coull	0.5957	0.7673
Clopper-Pearson (Exact)	0.5944	0.7716
Jeffreys	0.5971	0.7692
Wald	0.6010	0.7747
Wilson	0.5959	0.7670

Test of H0: Proportion = 0.7	
ASE under H0	0.0333
Z	-0.3651
One-sided Pr < Z	0.3575
Two-sided Pr > |Z|	0.7151

6 pts for attempting the lab