**Categorical Data Analysis: HW 1**

1. In class, we mentioned that when sampling **without** replacement, the independence assumption breaks down, but, somehow counterintuitively, the identical distributed assumption still holds. Let's illustrate using the drawing from a deck of 52 cards example. Let $Y_i$ be equal to 1 if the $i$-th trial (i.e. card drawn) results in a success (i.e., a heart) and equal to 0 otherwise. We sample without replacement.

   (a) Find $P(Y_1 = 1)$.

   (b) Find $P(Y_2 = 1 \mid Y_1 = 1)$ and $P(Y_2 = 1 \mid Y_1 = 0)$.

   (c) Find $P(Y_2 = 1)$. Use the law of total probability (or simply draw a tree!).

   (d) (nothing to do) Using similar arguments, one can show that $P(Y_1 = 1) = P(Y_2 = 1) = P(Y_3 = 1) = \ldots = P(Y_n = 1)$, i.e., the probability of success stays the same from trial to trial! So the random variables $Y_1, Y_2, \ldots, Y_n$ have an **identical** distribution.

   (e) Show that $P(Y_2 = 1 \mid Y_1 = 1) \neq P(Y_2 = 1)$ and hence $Y_1$ and $Y_2$ are **not** independent. The $Y_i$'s are therefore **not** Bernoulli trials, and $\sum_i Y_i$ does **not** follow a binomial but rather a hypergeometric distribution.

   (f) What are the chances that with 5 cards drawn without replacement from a deck of 52 cards, we get exactly 3 aces? Compare this probability to the one we get when sampling with replacement.

2. Using R, compute the exact coverage probability of the Wald and score confidence interval (CI) for the binomial proportion $\pi$ when $n = 15$. The Wald interval has formula $\hat{\pi} \pm z_{1-\alpha/2}\sqrt{\hat{\pi}(1-\hat{\pi})/n}$, where $\hat{\pi}$ is the sample proportion. The score interval is the one that R spits out when you use `prop.test()` with the option `correct=FALSE`. For the formula, see our class notes.

   To compute the exact coverage probability, find the confidence interval for each of the possible $n+1$ outcomes $y = 0, 1, \ldots, n$. Then, see which of the $n+1$ CIs contain $\pi$. To find the coverage probability, you simply add up the binomial probabilities (R command dbinom()) for those intervals that contain $\pi$. In other words, the exact coverage probability for a confidence interval is given by

   $$\sum_{y=0}^{n} \binom{n}{y} \pi^y (1-\pi)^{n-y} I_{CI}(y),$$

   where $I_{CI}(y)$ is an indicator function that is 1 when the confidence interval based on $y$ contains $\pi$ and 0 otherwise.

   (a) First, compute the coverage probability for $\pi = 0.05, 0.1, 0.25$ and $0.5$, and use $\alpha = 0.05$.

(b) Repeat part a for many more $\pi \in (0, 1)$ and turn in a nice graph that has $\pi$ on the x-axis (use lot's of different values so that you get a nice plot, e.g., if you are using a for loop: `for (pi in seq(0.01,0.99,0.01))` $\{\ldots\}$) and the coverage probability for the score and Wald interval (in a different color, but on the same plot) on the y-axis. You can add a line to an existing plot via the `lines()` statement, e.g., `plot(y ~ x, col="blue", type="l")` and then `lines(z ~ x, col="red")`. Remark: It is better (and good programming practice) to avoid for loops and use the powerful `apply` command in R, so check it out!

(c) Repeat part b but now using $n = 25$ and $n = 100$.

(d) Plot these charts also for $\alpha = 10\%$ and $1\%$.

3. Derive the formula for $\text{Var}[N_j]$ where $N_j$ is the multinomial count in category $j$.