Math 181B Worksheet Week 3

Two Sample Binomial: Let $X \sim Binom(n, p_x)$ and $Y \sim Binom(m, p_y)$.

a) What is the approximate $(1 - \alpha) \times 100\%$ confidence interval of $p_x - p_y$? Compare it to the test statistic from lecture for two sample proportion test.

b) Suppose your solution for a) is $[L, R]$. By the definition of the confidence interval, what is the approximate $P(L \leq 0 \leq R)$ if $p_x = p_y$?

c) Can you propose an approximate level $\alpha = 0.05$ test based on b) for $H_0 : p_x - p_y = \delta$ vs $H_1 : p_x - p_y \neq \delta$?

d) For the test in c), can you use the two sample proportion test when $\delta \neq 0$? Please explain your reason and suggest proper adjustment if any.

e) In the case of $\delta = 0$, describe how you can compare your proposed test in c) with the two population proportions test. What do you expect?

Try the Code:

```r
px = 0.5; npy = 101; py1 = seq(0.1, length.out = npy); py2 = seq(0.4, 0.6, length.out = npy)
n1 = m1 = 20; n2 = m2 = 1000; N = 100000;
rr1 = rr2 = rr3 = rr4 = rep(0, npy)
for(j in 1:npy)
{
  x = rbinom(N, n1, px); y = rbinom(N, m1, py1[j]); pe = (x+y)/(n1+m1);
  rr1[j] = mean(abs(x/n1-y/m1)/sqrt(pe*(1-pe)*(1/n1+1/m1))>qnorm(.975));
  rr2[j] = mean(abs(x/n1-y/m1)/sqrt(x*(n1-x)/n1^3 + y*(m1-y)/m1^3)>qnorm(.975));

  x = rbinom(N, n2, px); y = rbinom(N, m2, py2[j]); pe = (x+y)/(n2+m2);
  rr3[j] = mean(abs(x/n2-y/m2)/sqrt(pe*(1-pe)*(1/n2+1/m2))>qnorm(.975));
  rr4[j] = mean(abs(x/n2-y/m2)/sqrt(x*(n2-x)/n2^3 + y*(m2-y)/m2^3)>qnorm(.975));
}
par(mfrow = c(1,2))
plot(py1, rr1, type = 'l', col = 2); lines(py1, rr2, col = 3, lty=2); abline(h=0.05, lty=2)
legend("topright", c("Test with pe", "Test with CI", "0.05"), lty=c(1:2, 1), col = c(2:3,1))
plot(py2, rr3, type = 'l', col = 2); lines(py2, rr4, col = 3, lty=2); abline(h=0.05, lty=2)
legend("topright", c("Test with pe", "Test with CI", "0.05"), lty=c(1:2, 1), col = c(2:3,1))
```
**Binning Data into Multinomial:** A professor proposes the following model for a biomaker reading in the micro-RNA:

\[ f(x) = 2x, \quad 0 \leq x \leq 1. \]

1. Under the model, what are the probabilities of \( X \) fall into the bins \([0, 0.25), \ [0.25, 0.5), \ [0.5, 0.75) \) and \([0.75, 1]\), respectively?

2. Design four bins with equal probability. What is the advantage of the design compared to the bins with equal length?

3. Under the model, give without calculation the pmf of \((k_1, k_2, k_3, k_4)\) where \( k_i \) is the count of mi-RNA readings fall into your i-th bin in b) out of 20 samples.

4. The professor recruits an undergraduate researcher to write a program that generates Monte Carlo samples from the model distribution. In a test of size 1000, he has got \((k_1, k_2, k_3, k_4) = (254, 252, 253, 241)\). Does the sample seem to match the model? Does the sample look too good such that the professor may suspect a fabrication in the program? Answer the two questions separately at level \( \alpha = 0.1 \), and explain their difference.

5. To investigate the matter, the professor uses the bins of equal widths in a) and gets \((k_1, k_2, k_3, k_4) = (132, 122, 318, 428)\). Redo the tests in d). Can you explain the results after seeing the codes?

```r
bin1 = sqrt(seq(0,1,by=0.25)); x.bin = sample(4,1000,T)
x.a = bin1[x.bin]; x.b = bin1[x.bin+1]-bin1[x.bin]
x=x.a + x.b*runif(1000)
O1 = table(cut(x,c(-Inf,bin1[-1])))
bin2 = c(-Inf,1:4/4); O2 = table(cut(x,c(-Inf,bin2[-1])))
```