Stat 201A, Fall 2012
HOMEWORK 4 (due Thursday 9/6)

1. Show that for fixed n, the binomial SD is largest when $p = 1/2$. Please avoid calculus; the variance is an easily understood function of p.

2. Let $\mathcal{S}$ be a finite set, and let P_1 and P_2 be two probability distributions on $\mathcal{S}$. Let $\mathcal{F}$ be the set of all subsets of $\mathcal{S}$. Define the *total variation distance* between P_1 and P_2 to be

$$d(P_1, P_2) \quad = \quad \max\{A \in \mathcal{F} : |P_1(A) - P_2(A)|\}$$

Thus d is the largest amount by which the two distributions differ, across all possible events.

Show that

$$d(P_1, P_2) \quad = \quad \frac{1}{2} \sum_{x \in \mathcal{S}} |P_1(x) - P_2(x)|$$

There are many ways to do this. Here's one, but you are free to use any other.

Let $A^* = \{x \in \mathcal{S} : P_1(x) > P_2(x)\}$, $A_* = \{x \in \mathcal{S} : P_1(x) < P_2(x)\}$, and $A_*^* = \{x \in \mathcal{S} : P_1(x) = P_2(x)\}$. The union of these disjoint sets is clearly $\mathcal{S}$. You should have a proof after you've investigated:

(i) the relation between $|P_1(A^*) - P_2(A^*)|$ and $\sum_{x \in A^*} |P_1(x) - P_2(x)|$

(ii) the relation between $|P_1(A^*) - P_2(A^*)|$ and $|P_1(A_*) - P_2(A_*)|$

(iii) whether the max in the definition of $d(P_1, P_2)$ can be greater than $|P_1(A^*) - P_2(A^*)|$

3. Let P_1 be the binomial distribution with parameters n and p.

Let P_2 be the Poisson approximation to P_1. To make P_2 a genuine probability distribution on $\{0, 1, \ldots, n\}$, define $P_2(n)$ to be the Poisson probability of "n or more." In terms of the Poisson histogram, you are sweeping all the area of the bars over $n+1, n+2 \ldots$ into the bar over n.

Let P_3 be the normal approximation to P_1. To make P_3 a genuine probability distribution on $\{0, 1, \ldots, n\}$, sweep all the area to the left of -0.5 into the bar over 0, and sweep all the area to the right of $n + 0.5$ into the bar over n.

The goal is to compare the two approximations, using total variation distance as the criterion. For a selection of n and p, compute $d(P_1, P_2)$ and $d(P_1, P_3)$. Smaller distances correspond to better approximations. Can you give a "rule of thumb" by which we can decide when it is better to use the Poisson approximation instead of the normal?

You'll have to make some decisions about which values of n and p you are going to use. Since this is about approximations for large n, I suggest starting at $n = 20$ or so, and choosing an evenly spaced set of values of p between 0 and 0.5.

You'll also have to make some decisions about how to display your results and present your conclusion. I look forward to reading your clear and concise analysis, and to studying your informative plots.

Please write your analysis (with plots) first. Then, in an Appendix, include your R code.