STAT201A – Sec. 102
Homework #4.

Steven Pollack
24112977

#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.

Proof. Recall that for a quadratic of the form $b-(x-a)^2$, a maximum is attained at $x = a$ (the point of the vertex). Thus, if we rewrite $p(1-p)$ as

$$p(1-p) = p(1-p) - \frac{1}{4} + \frac{1}{4} = \frac{1}{4} - \left(p - \frac{1}{2}\right)^2$$

we see that our function is a concave down, quadratic with vertex at $p = 1/2$.

Tying this back to $SD(X)$, for $X \sim \text{Binomial}(n, p)$, we have that $SD(X) = \sqrt{npq}$, and thus $SD(X)^2 = \text{var}(X) = np(1-p)$, which is maximized at $p = 1/2$. And because $\sqrt{\cdot}$ is a monotonic function therefore optimizing $\sqrt{f(x)}$ is equivalent to optimizing $f(x)$, we see that $SD(X)$ is maximal at $p = 1/2$. □

#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

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

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

Show that,

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

There are many ways to do this. Here's on, 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 $|P_1(A^*) - P_2(A^*)|$.

Proof. Defining the following sets:

$$\begin{aligned} A^* &= \{x \in \mathcal{S} : P_1(x) > P_2(x)\} \\ A_* &= \{x \in \mathcal{S} : P_1(x) < P_2(x)\} \\ A_*^* &= \{x \in \mathcal{S} : P_1(x) = P_2(x)\} \end{aligned}$$

and it's clear that $\mathcal{S} = A^* \sqcup A_* \sqcup A_*^*$. Furthermore, the finite size of $\mathcal{S}$ allows us to write

$$P_i(A^*) = P_i\left(\bigsqcup_{x\in A^*} \{x\}\right) = \sum_{x\in A^*} P_i(x) \qquad (i = 1, 2)$$

Hence,

$$|P_1(A^*) - P_2(A^*)| = \left|\sum_{x\in A^*} P_1(x) - \sum_{x\in A^*} P_2(x)\right|$$
$$= \left|\sum_{x\in A^*} P_1(x) - P_2(x)\right| = \sum_{x\in A^*} |P_1(x) - P_2(x)| = \sum_{x\in A^*} (P_1(x) - P_2(x))$$

Similarly,

$$|P_1(A_*) - P_2(A_*)| = \left|\sum_{x\in A_*} P_1(x) - P_2(x)\right| = \sum_{x\in A_*} |P_1(x) - P_2(x)|$$

Since $P_1(x) - P_2(x) = 0$ for all $x \in A_*^*$, we see that

$$|P_1(A^*) - P_2(A^*)| + |P_1(A_*) - P_2(A_*)| = \sum_{x\in A^*} |P_1(x) - P_2(x)| + \sum_{x\in A_*} |P_1(x) - P_2(x)|$$
$$= \sum_{x\in\mathcal{S}} |P_1(x) - P_2(x)|$$

Now, suppose $|P_1(A^*) - P_2(A^*)| > |P_1(A_*) - P_2(A_*)|$:

$$\begin{aligned} |P_1(A^*) - P_2(A^*)| > |P_1(A_*) - P_2(A_*)| &\iff P_1(A^*) - P_2(A^*) > P_2(A_*) - P_1(A_*) \\ &\iff P_1(A^*) + P_1(A_*) > P_2(A^*) + P_2(A_*) \\ &\iff 1 - P_1(A_*^*) > 1 - P_2(A_*^*) \\ &\iff P_1(A_*^*) > P_2(A_*^*) \end{aligned}$$

But, this is impossible, since $P_1(x) = P_2(x)$ for all $x \in A_*^*$. Thus, $|P_1(A^*) - P_2(A^*)| \leq |P_1(A_*) - P_2(A_*)|$. However, a symmetric argument shows that

$$|P_1(A_*) - P_2(A_*)| \leq |P_1(A^*) - P_2(A^*)|$$

and thus,

$$|P_1(A^*) - P_2(A^*)| = |P_1(A_*) - P_2(A_*)|$$

Finally, we establish that $d(P_1, P_2) = |P_1(A^*) - P_2(A^*)|$; Let $B \in \mathcal{F} \setminus \{A_*, A^*\}$ and suppose

$$|P_1(B) - P_2(B)| > |P_1(A^*) - P_2(A^*)| = |P_1(A_*) - P_2(A_*)|$$

Then,

$$2\,|P_1(B) - P_2(B)| > (|P_1(A^*) - P_2(A^*)| + |P_1(A_*) - P_2(A_*)|) = \sum_{x\in\mathcal{S}} |P_1(x) - P_2(x)|$$

and hence[1]

$$|P_1(B) - P_2(B)| > \sum_{x\in\mathcal{S}\setminus B} |P_1(x) - P_2(x)| \geq |P_1(B^c) - P_2(B^c)|$$

[1]note: $|P_1(B) - P_2(B)| = \left|\sum_{x\in B}(P_1(x) - P_2(x))\right| \leq \sum_{x\in B} |P_1(x) - P_2(x)|$

However,

$$|P_1(B^c) - P_2(B^c)| = |(1 - P_1(B)) - (1 - P_2(B))| = |P_2(B) - P_1(B)|$$

Thus, our assumption of strict inequality led us to the (nonsensical) conclusion that

$$|P_1(B) - P_2(B)| > |P_2(B) - P_1(B)|$$

so our max must be attained at A^* and A_*. Thus,

$$d(P_1, P_2) = \frac{1}{2} \times 2d(P_1, P_2) = \frac{1}{2}\left(|P_1(A^*) - P_2(A^*)| + |P_1(A_*) - P_2(A_*)|\right) = \frac{1}{2}\sum_{x \in \mathcal{S}} |P_1(x) - P_2(x)|$$

□

#3. So, a first look at the L_1 distance comparisons in figure 1, immediately indicates that there is a convex region in the (half) (n,p)-plane where the normal approximation is far-superior to the Poisson. Doing some rough linear regression, we can use the function $p = 0.126 \exp(-2.32 \times 10^{-3} n)$ (for $n \geq 20$) as one boundary curve of this region.

Furthermore, the graphics in figure 2 indicate that this convex region is good, not only for gaging when the normal approximation is L_1-superior, but for when it is L_∞ superior. Hence, I feel comfortable employing the following "rule of thumb": if $(n, p) \in [20, 10^3] \times [0, 1]$, and $p \leq 0.126 \exp\{-2.32 \times 10^{-3} n\}$, then go with the Poisson approximation. Otherwise, use the normal.

Note: $e^{-.00232} \approx 0.998$ and $0.126 \approx 1/8$, so we can roughly peg our rule of thumb to something like

$$8p \leq 0.998^n$$

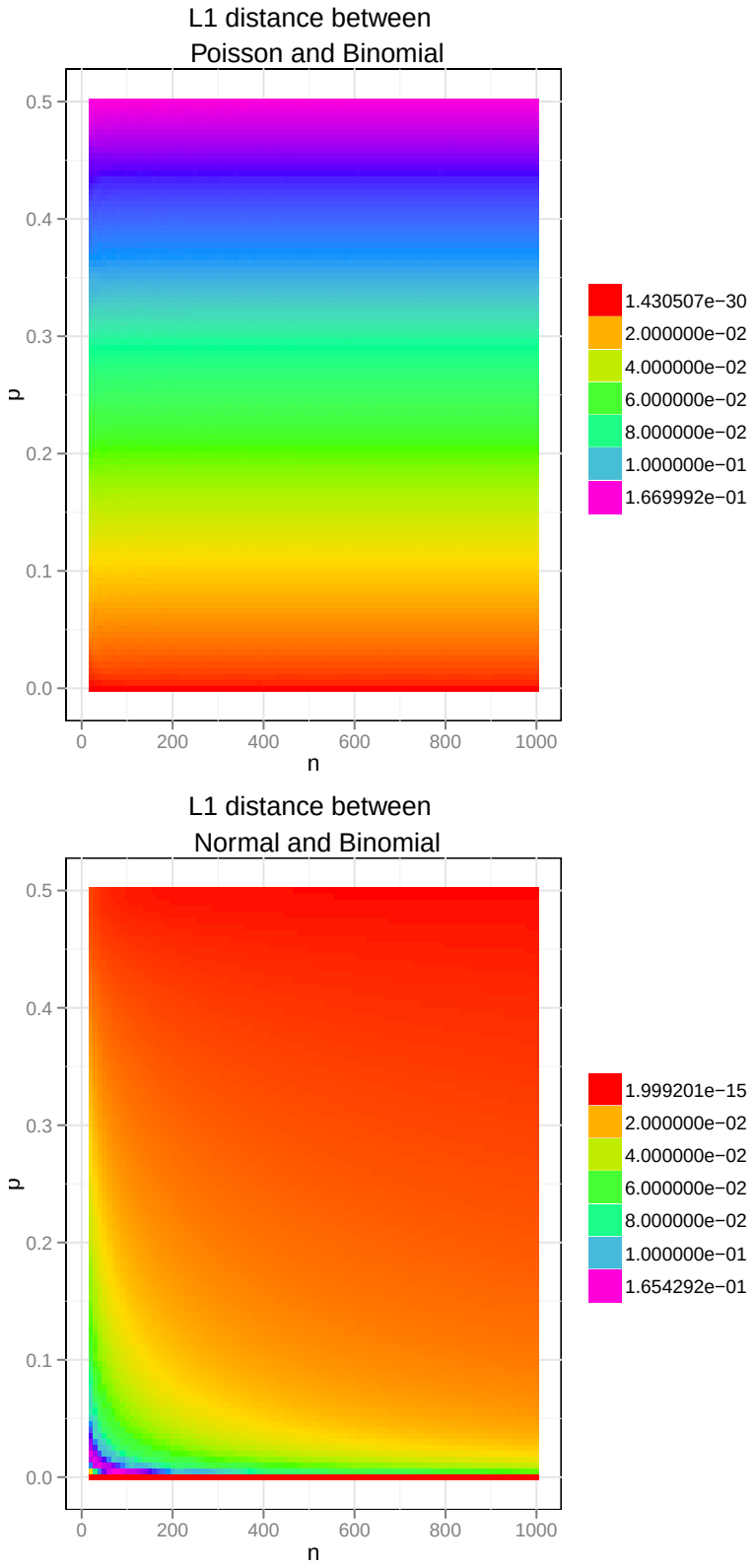

Figure 1: A graphical comparison of the L_1 distance between the two distributions as a function of (n, p). Note: red is lower in magnitude than violet.

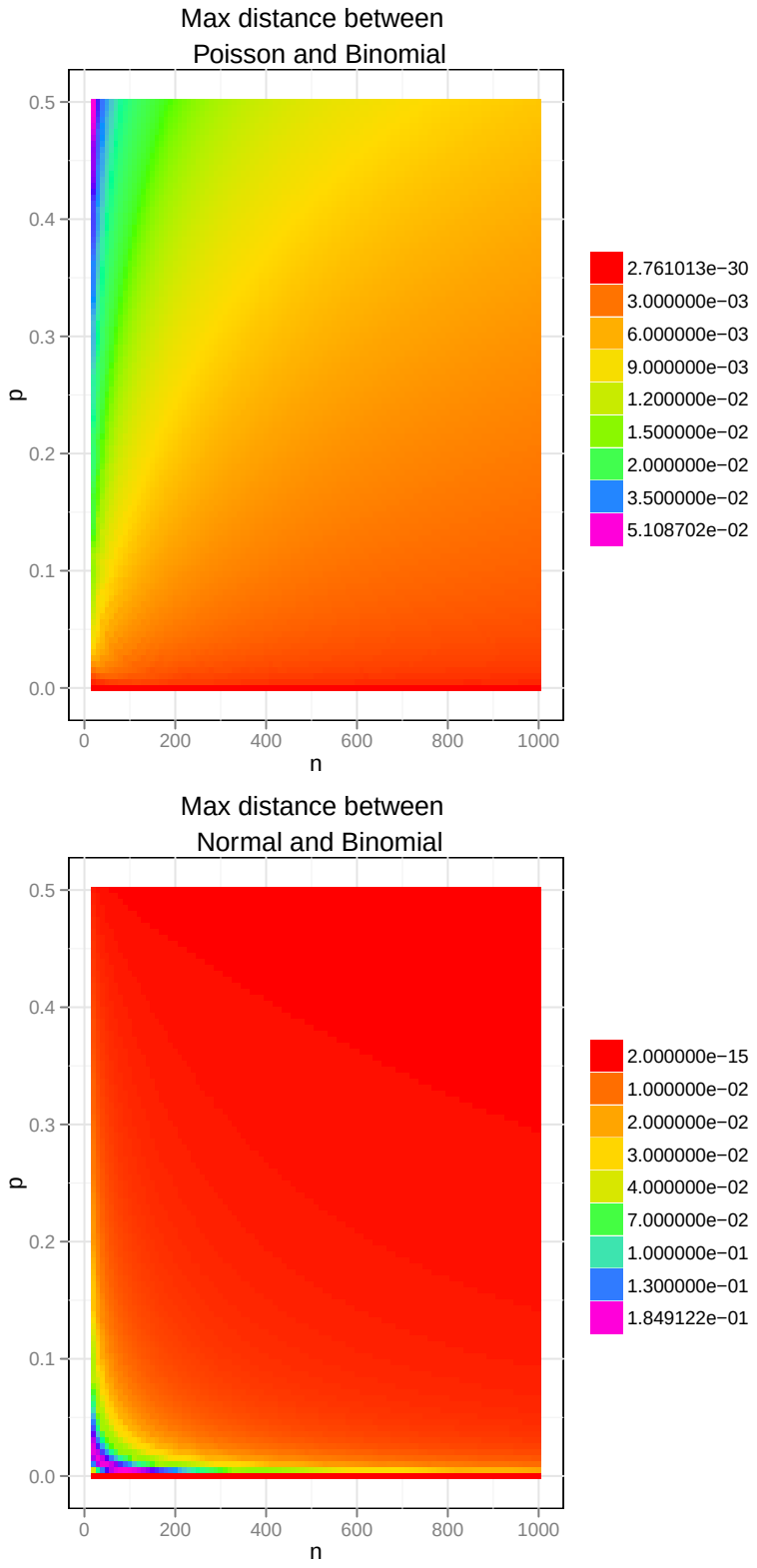

Figure 2: A graphical comparison of the L_∞ distance between the two distributions as a function of (n, p).

R Code

```
### R code from vignette source './tex/stat201a_hw4.Rnw'

###################################################
### code chunk number 1: prepare_data
###################################################
calculateDistance <- function(X1,X2) { #L1 distance
  return(sum(abs(X1-X2)))
}

findL1DistanceOfBinomialAndPoisson <- function(n,p) {
  X <- dbinom(x=0:n, size=n, prob=p) # vector of P(X=x), 0 <= x <= n
  Y <- dpois(x=0:n, lambda=n*p) # vector of P(Y=x), 0 <= x <= n
  return(0.5*calculateDistance(X,Y)) #scale according to
      definition of d(P_1,P_2)
}

findL1DistanceOfBinomialAndNormal <- function(n,p) {
  X <- dbinom(x=0:n, size=n, prob=p)
  Z <- ( pnorm(q=seq(from=0.5, to=n+0.5, by=1),
               mean=n*p, sd=sqrt(n*p*(1-p))
               ) - pnorm(q=seq(from=-0.5,to=n-0.5,by=1),
                         mean=n*p, sd=sqrt(n*p*(1-p)) ) )
  return(0.5*calculateDistance(X,Z))
}

findSupNormOfBinomialAndPoisson <- function(n,p) {
  X <- dbinom(x=0:n, size=n, prob=p)
  Y <- dpois(x=0:n, lambda=n*p)
  return(max(abs((X-Y))))
}

findSupNormOfBinomialAndNormal <- function(n,p) {
  X <- dbinom(x=0:n, size=n, prob=p)
  Z <- ( pnorm(q=seq(from=0.5, to=n+0.5, by=1),
               mean=n*p, sd=sqrt(n*p*(1-p))
               ) - pnorm(q=seq(from=-0.5, to=n-0.5,by=1),
                         mean=n*p, sd=sqrt(n*p*(1-p)) ) )
  return(max(abs((X-Z))))
}


# set up grid
```

```
num_of_trials <- seq(from=20,to=10**3,by=10)
p_values <- seq(from=10**-16, to=0.5,
  length.out=length(num_of_trials))

surfaces <- expand.grid(n=num_of_trials, p=p_values,
                        KEEP.OUT.ATTRS=FALSE)

#calculate all the distances
surfaces$L1_distance_poisson <- with(surfaces,mapply(
    findL1DistanceOfBinomialAndPoisson, n=n, p=p))

surfaces$sup_norm_poisson <- with(surfaces, mapply(
  findSupNormOfBinomialAndPoisson, n=n, p=p))

surfaces$L1_distance_normal <- with(surfaces, mapply(
  findL1DistanceOfBinomialAndNormal, n=n, p=p))

surfaces$sup_norm_normal <- with(surfaces, mapply(
  findSupNormOfBinomialAndNormal, n=n, p=p))


#########################################################
### code chunk number 2: plot_L1_poisson
#########################################################
library(ggplot2)

maximum <- max(surfaces$L1_distance_poisson)
minimum <- min(surfaces$L1_distance_poisson)

d <- ggplot() + layer(data=surfaces, geom="tile",
                      mapping=aes(x=n, y=p,
                                  fill=L1_distance_poisson))

d <- d + scale_fill_gradientn(
  colours=rainbow(7), breaks=c(seq(from=minimum,to=.1,
                                   length.out=6),maximum))

d <- d + opts(
  panel.background=theme_rect(fill="white", colour="black"),
  panel.grid.major=theme_line(colour = "grey90"),
  title="L1 distance between \n Poisson and Binomial",
  legend.title=theme_blank())

print(d)
```

```
#####################################################
### code chunk number 3: plot_L1_normal
#####################################################
maximum <- max(surfaces$L1_distance_normal)
minimum <- min(surfaces$L1_distance_normal)

d <- ggplot() + layer(data=surfaces, geom="tile",
                      mapping=aes(x=n, y=p, fill=L1_distance_normal))
d <- d + scale_fill_gradientn(
   colours=rainbow(7),breaks=c(seq(from=minimum,to=.1,
                                   length.out=6),maximum))
d <- d + opts(
   panel.background=theme_rect(fill="white",colour="black"),
   panel.grid.major=theme_line(colour = "grey90"),
   title="L1 distance between \n Normal and Binomial",
   legend.title=theme_blank())

print(d)


#####################################################
### code chunk number 4: plot_sup_norm_poisson
#####################################################
maximum <- max(surfaces$sup_norm_poisson)
minimum <- min(surfaces$sup_norm_poisson)

d <- ggplot() + layer(data=surfaces, geom="tile",
                      mapping=aes(x=n, y=p, fill=sup_norm_poisson)
                         )

d <- d + scale_fill_gradientn(
   colours=rainbow(7),
   breaks=c(seq(from=minimum,to=.015,length.out=6),.02,.035,maximum
      ))

d <- d + opts(
   panel.background=theme_rect(fill="white",colour="black"),
   panel.grid.major=theme_line(colour = "grey90"),
   title="Max distance between \n Poisson and Binomial",
   legend.title=theme_blank())

print(d)
```

```
#######################################################
### code chunk number 5: plot_sup_norm_normal
#######################################################
maximum <- max(surfaces$sup_norm_normal)
minimum <- min(surfaces$sup_norm_normal)

d <- ggplot() + layer(data=surfaces, geom="tile",
                      mapping=aes(x=n, y=p, fill=sup_norm_normal))

d <- d + scale_fill_gradientn(
  colours=rainbow(7),
  breaks=c(seq(from=minimum,to=.04,length.out=5),.07,.1,.13,
    maximum))

d <- d + opts(
  panel.background=theme_rect(fill="white",colour="black"),
  panel.grid.major=theme_line(colour = "grey90"),
  legend.position="right",
  title="Max distance between \n Normal and Binomial",
  legend.title=theme_blank())

print(d)
```