STAT 509 Homework 2 Solution

Chong Ma

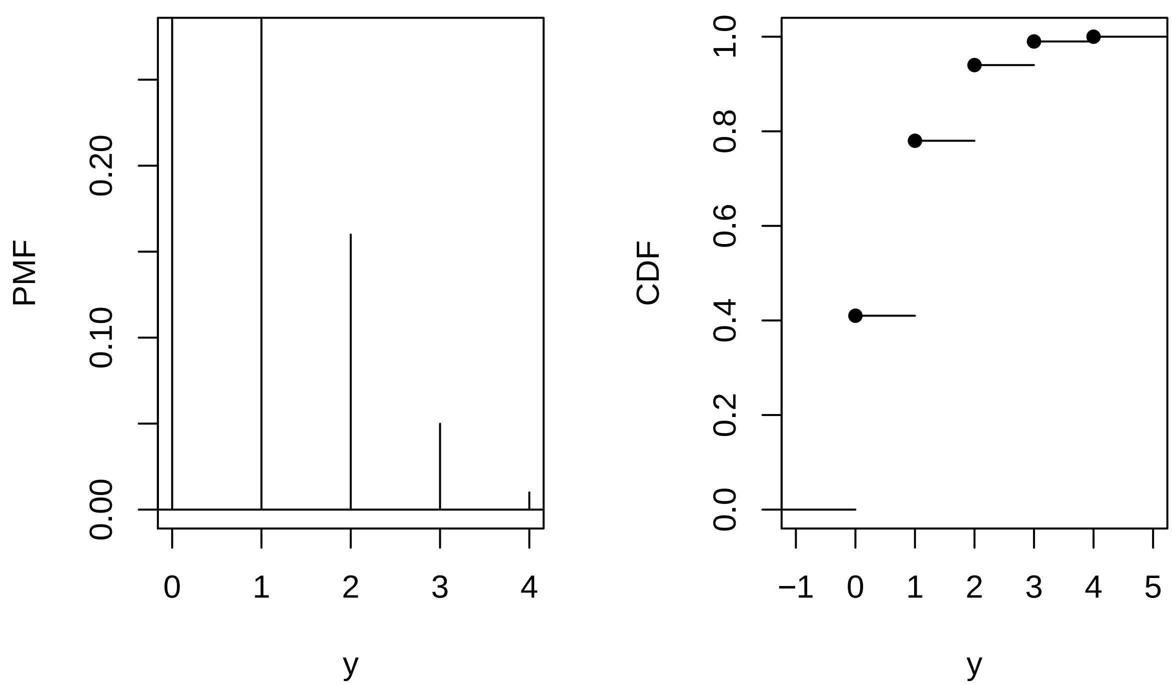
February 2, 2017

Problem 1

It tests how you understand the generic discrete distribution.

(a).

```
y=c(0,1,2,3,4); prob=c(0.41,0.37,0.16,0.05,0.01)
cdf = c(0,cumsum(prob)); cdf.plot = stepfun(y,cdf,f=0)
par(mfrow=c(1,2))
# Plot PMF
plot(y,prob,type="h",xlab="y",ylab="PMF",ylim=c(0,0.275))
abline(h=0)
# Plot CDF
plot.stepfun(cdf.plot,xlab="y",ylab="CDF",verticals=FALSE,do.points=TRUE,main="",pch=16)
```



(b). The probability that there are **at least 2** imperfections is

$$\begin{aligned} P(Y \geq 2) &= P(Y=2) + P(Y=3) + P(Y=4) \\ &= 0.16 + 0.05 + 0.01 \\ &= 0.22 \end{aligned}$$

(c). The probability that there are **at most 2** imperfections is

$$\begin{aligned} P(Y \leq 2) &= P(Y=0) + P(Y=1) + P(Y=2) \\ &= 0.41 + 0.37 + 0.16 \\ &= 0.94 \end{aligned}$$

(d). The mean of Y is given by

$$\begin{aligned} E(Y) &= \sum_y yP(Y=y) \\ &= (0)P(Y=0) + (1)P(Y=1) + (2)P(Y=2) + (3)P(Y=3) + (4)P(Y=4) \\ &= (0)0.41 + (1)0.37 + (2)0.16 + (3)0.05 + (4)(0.01) \\ &= 0.88 \end{aligned}$$

Note that $var(Y) = E(Y^2) - [E(Y)]^2$, and

$$\begin{aligned} E(Y^2) &= \sum_y y^2 P(Y=y) \\ &= (0^2)P(Y=0) + (1^2)P(Y=1) + (2^2)P(Y=2) + (3^2)P(Y=3) + (4^2)P(Y=4) \\ &= (0)0.41 + (1)0.37 + (4)0.16 + (9)0.05 + (16)(0.01) \\ &= 1.62 \end{aligned}$$

Therefore, the variance of Y is obtained by

$$var(Y) = E(Y^2) - [E(Y)]^2 = 1.62 - (0.88)^2 = 0.8456$$

(e). Given model of Y, we can calculate the probability of observing such 10 observations by

$$\begin{aligned} P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_{10} = y_{10}) &= P(Y_1 = y_1) \times P(Y_2 = y_2) \times \cdots \times P(Y_{10} = y_{10}) \\ (since Y_1, Y_2, \ldots, Y_{10} \overset{i.i.d}{\sim} Y) &= P(Y=0)^5 P(Y=1)^3 P(Y=2)^2 P(Y=4) \\ &= (0.41)^5 (0.37)^3 (0.16)^2 (0.01) \\ &= 1.5 \times 10^{-7} \approx 0 \end{aligned}$$

Because the probability of observing such 10 consecutive obervations approximates 0, thus those observations are not consistent with the model. Another set of 10 observations that are grossly inconsistent with the model is {4,4,4,4,4,4,4,4,4,4}, since the probability of observing 10 4's in a row is 10^{-20}.

Problem 2

This problem is an exercise for binomial distribution. (a).$Y \sim Binomial(20, 0.02)$ since

- drawer = `"trial"`
- drawer get stuck = `"success"`
- $p = P(\text{“}success\text{”}) = 0.02$

(b). The probability that the lot will be shipped is $P(Y = 0) = \binom{20}{0} 0.02^0 (1 - 0.02)^(20 - 0) \approx 0.668$

```
## R code for P(Y=0)
dbinom(0,20,0.02)
```

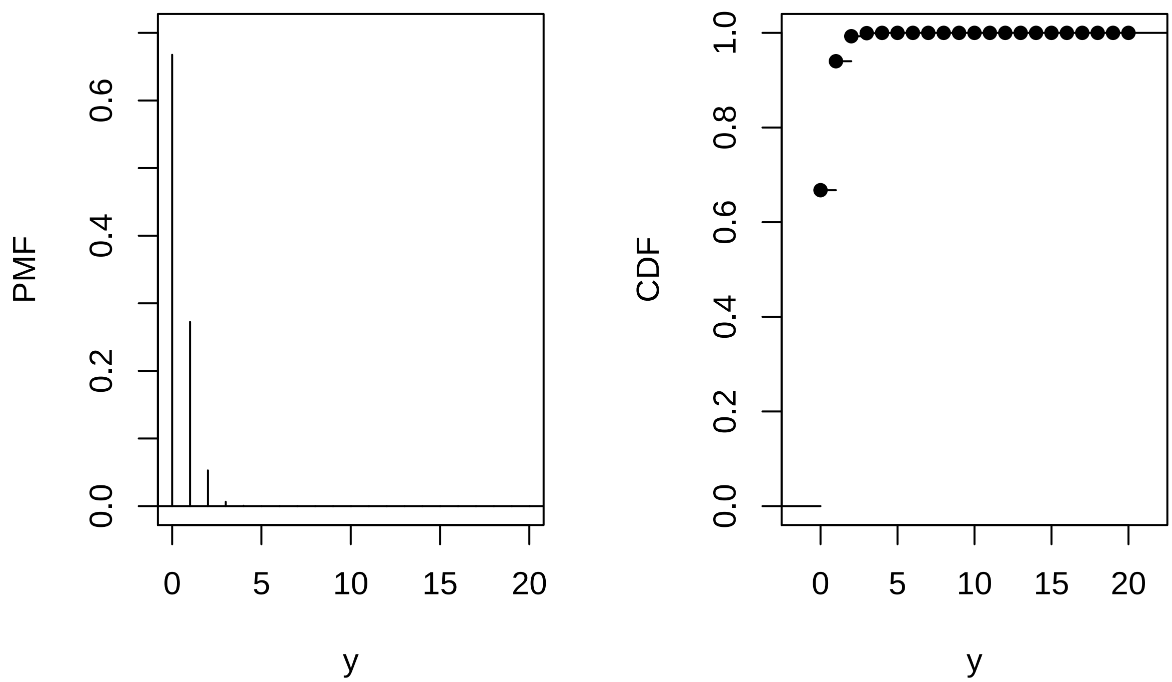
```
## [1] 0.667608
```

(c). The probability that **at least two drawers** will get stuck is

$$
\begin{aligned}
P(Y \geq 2) &= 1 - P(Y < 2) \\
&= P(Y = 0) + P(Y = 1) \\
&= \binom{20}{0} 0.02^0 (1 - 0.02)^(20 - 0) + \binom{20}{1} 0.02^1 (1 - 0.02)^(20 - 1) \\
&= 0.668 + 0.272 \approx 0.94
\end{aligned}
$$

(d). Graph the pmf of Y and the cdf of Y.

```
y = 0:20; prob = dbinom(y,20,0.02)
cdf = c(0,cumsum(prob)); cdf.plot = stepfun(y,cdf,f=0)
op<-par(mfrow=c(1,2))
# Plot PMF
plot(y,prob,type="h",xlab="y",ylab="PMF",ylim=c(0,0.7))
abline(h=0)
# Plot CDF
plot.stepfun(cdf.plot,xlab="y",ylab="CDF",verticals=FALSE,do.points=TRUE,main="",pch=16)
```



(e).

- The mean of Y: $E(Y) = np = 20(0.02) = 0.4$
- The variance of Y: $var(Y) = np(1-p) = 20(0.02)(0.98) = 0.392$
- The standard deviation of Y; $sd(Y) = \sqrt{var(Y)} = 0.626$

Problem 3

This problem is an exercise of combinations of geometric distribution, binomial distribution and hypergeometric distribution. (a). $Y \sim geom(p = 0.05)$

$$
\begin{aligned}
P(Y > 3) &= 1 - P(Y \leq 3) \\
&= 1 - (P(Y = 1) + P(Y = 2) + P(Y = 3)) \\
&= 1 - (0.05 + (0.95)(0.05) + (0.95)^2(0.05)) \\
&= 0.8574
\end{aligned}
$$

```
1-pgeom(2,0.05)
```

```
## [1] 0.857375
```

(b). $P(Y > 5|Y > 2) = P(Y > 3)$, since

$$
\begin{aligned}
P(Y>5|Y>2) &= \frac{P(Y>5,Y>2)}{P(Y>2)} = \frac{P(Y>5)}{P(Y>2)} \\
&= \frac{(1-p)^5p+(1-p)^6p+(1-p)^7p+\ldots}{(1-p)^2p+(1-p)^3p+(1-p)^7p+\ldots} \\
&= \frac{(1-p)^5(p+(1-p)p+(1-p)^2p+\ldots)}{(1-p)^2(p+(1-p)p+(1-p)^2p+\ldots)} \\
&= (1-p)^3
\end{aligned}
$$

and

$$
P(Y>3) = (1-p)^3p+(1-p)^4p+(1-p)^5p+\ldots = (1-p)^3(p+(1-p)p+(1-p)^2p+\ldots) = (1-p)^3
$$

(c). Fix 30 donations, Y turns out the number of donations that are infected. Thus, $Y \sim Binomial(30, 0.05)$.

$$
\begin{aligned}
P(Y\leq 2) &= P(Y=0)+P(Y=1)+P(Y=2) \\
&= \binom{30}{0}(0.05)^0(0.95)^{30}+\binom{30}{1}(0.05)^1(0.95)^{29}+\binom{30}{2}(0.05)^2(0.95)^{28} \\
&= 0.8122
\end{aligned}
$$

(d). $Y \sim hyper(N=30, n=3, r=10)$. The probability that exactly one is from an AA donor is

$$
P(Y=1) = \frac{\binom{10}{1}\binom{20}{2}}{\binom{30}{3}} = 0.468
$$

Problem 4

This problem is an exercise of negative binomial distribution.

(a).

- drilling wells = `"trials"`
- oil well strick = `"success"`
- $p = P(\text{“success”}) = 0.20$

(b). Note that $Y \sim geom(p=0.2)$,the probability tht the 1st successful well is found on the 4th well drilled is

$$
P(Y=4) = (1-p)^3p = (0.8)^3(0.2) = 0.1024
$$

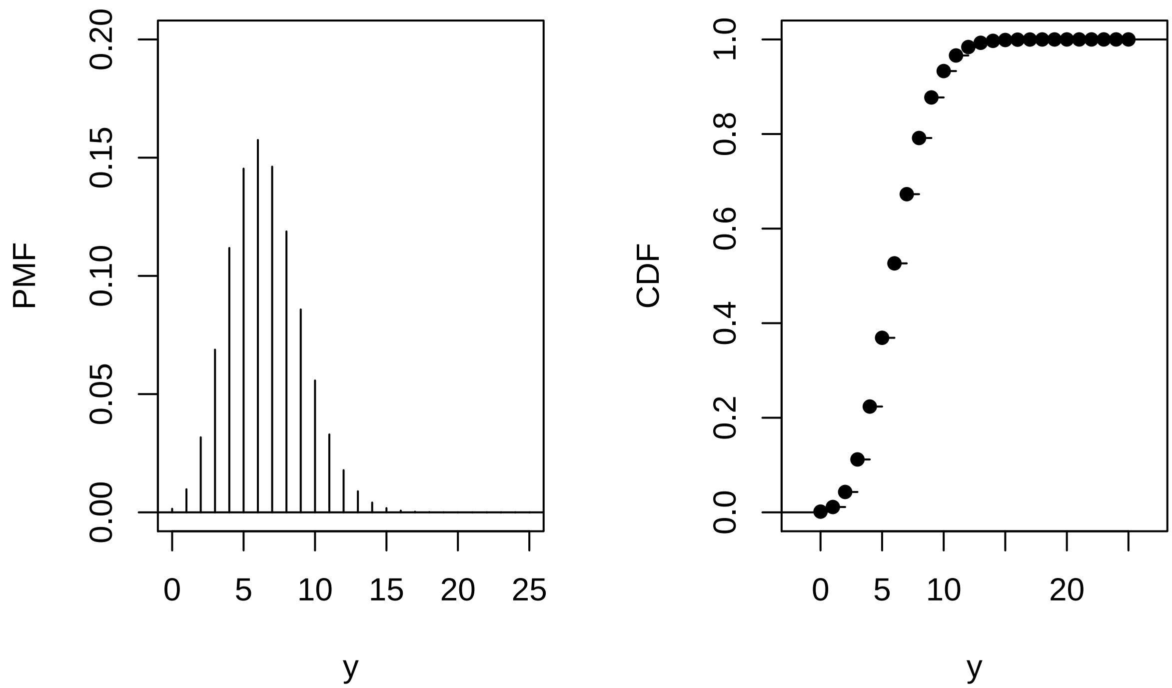
(c). Note that $Y \sim nib(r=2, p=0.2)$, the probability that taking more than 4 wells to find the 2nd successful well is

$$
\begin{aligned}
P(Y>4) &= 1-P(Y\leq 4) \\
&= 1-(P(Y=2)+P(Y=3)+P(Y=4)) \\
&= 1-((0.2)^2+\binom{2}{1}(0.2)^2(0.8)+\binom{3}{1}(0.2)^2(0.8)^2) \\
&= 0.8192
\end{aligned}
$$

Problem 5

(a). Graph the pmf and pdf of Y

```r
y = 0:25; prob = dpois(y,6.5)
cdf = c(0,cumsum(prob)); cdf.plot = stepfun(y,cdf,f=0)
op<-par(mfrow=c(1,2))
# Plot PMF
plot(y,prob,type="h",xlab="y",ylab="PMF",ylim=c(0,0.2))
abline(h=0)
# Plot CDF
plot.stepfun(cdf.plot,xlab="y",ylab="CDF",verticals=FALSE,do.points=TRUE,main="",pch=16)
```



(b).

- The probability that on a given day there are **exactly** 5 calls is

$$P(Y = 5) = \frac{6.5^5}{5!}e^{-6.5} = 0.1453689$$

- The probability that on a given day there are **at least** 5 calls is

$$\begin{aligned} P(Y \geq 5) &= 1 - P(Y < 5) = 1 - P(Y = 0) + P(Y = 1) + \cdots + P(Y = 4) \\ &= 1 - \sum_{y=0}^{4} \frac{y^{\lambda}}{y!}e^{-\lambda} = 1 - 0.2236718 = 0.7763282 \end{aligned}$$

- The probability that on a given day there are **at most** 5 calls is

$$\begin{aligned} P(Y \le 5) &= P(Y = 0) + P(Y = 1) + \cdots + P(Y = 5) \\ &= \sum_{y=0}^{5} \frac{y^{\lambda}}{y!} e^{-\lambda} = 0.3690407 \end{aligned}$$

(c). From the pmf plot of Y, we can esaily tell the mode of Y is 6.

(d). Note that $E(Y) = Var(Y) = \lambda$, and $Var(Y) = E(Y^2) - [E(Y)]^2$, therefore $E(Y^2) = Var(Y) + [E(Y)]^2 = \lambda + \lambda^2$. Hence,

$$\begin{aligned} E(g(Y)) &= 150 + 100E(Y) + 5E(Y^2) \\ &= 150 + 100(6.5) + 5(6.5 + 6.5^2) \\ &= 1043.75 \end{aligned}$$

(e). In this context, of total $N = 30$ calls, they are split into two classes, class 1={calles involving illegal consumption of alcohol} and class 2={calls not involving illegal consumption of alcohol}. Randomly pick 5 cases(calls), define

$$Y = \# \text{ of cases involving illegal consumption of alcohol}$$

Hence, $Y \sim hyper(N = 30, r = 12, N - r = 18, n = 5)$. The probability that **at least** 4 of these cases will involve illegal consumption of alcohol is

$$\begin{aligned} P(Y \ge 4) &= P(Y = 4) + P(Y = 5) \\ &= \frac{\binom{12}{4}\binom{18}{1}}{\binom{30}{5}} + \frac{\binom{12}{5}\binom{18}{0}}{\binom{30}{5}} \\ &= 0.0680813 \end{aligned}$$

Problem 6

(a).

- chemical compound = `"trials"`
- compound is active = `"success"`
- $p = P(\text{“compound is active”})$

(b). The probability that a single pool of k compounds will test positive

$$\begin{aligned} P(\text{a single pool of k compounds test positive}) &= P(\text{at least one compound tests positive}) \\ &= 1 - P(\text{None of k compounds tests positive}) \\ &= 1 - (1 - p)^k \end{aligned}$$

(c). Consider m pools as a series of Bernoulli trials, each pool has either `"active"` or `"inactive"` two statuses, and $P(\text{a single pool tests active}) = 1 - (1 - p)^k$(denoted as p_l) is the same as all pools. Note

$$Y = \# \text{ of tests needed to screen the entire library under (ii)}$$

Interval	preferred
$0 < p < 0.001$	$m = 2000, k = 50$
$0.001 < p < 0.006$	$m = 5000, k = 20$
$0.006 < p < 0.024$	$m = 10000, k = 10$
$0.024 < p < 0.2$	$m = 20000, k = 5$

and further Y can be written as

$$Y = Y_1 + Y_2 + \cdots + Y_m$$

Where $Y_j = \#$ of tests needed to screen the jth pool, $j = 1, 2, \ldots, m$ and

$$Y_j = \begin{cases} 1, & \text{with probability } 1 - p_l \\ k+1, & \text{with probability } p_l \end{cases}$$

Note that $E(Y_j) = (1)(1 - p_l) + (k+1)(p_l) = 1 + kp_l$, therefore

$$\begin{aligned} E(Y) &= E(Y_1 + Y_2 + \cdots + Y_m) = mE(Y_1) \\ &= m(1 + kp_l) = mk(1 - (1-p)^k) + m (= N(1 - (1-p)^k) + m) \end{aligned}$$

(d).

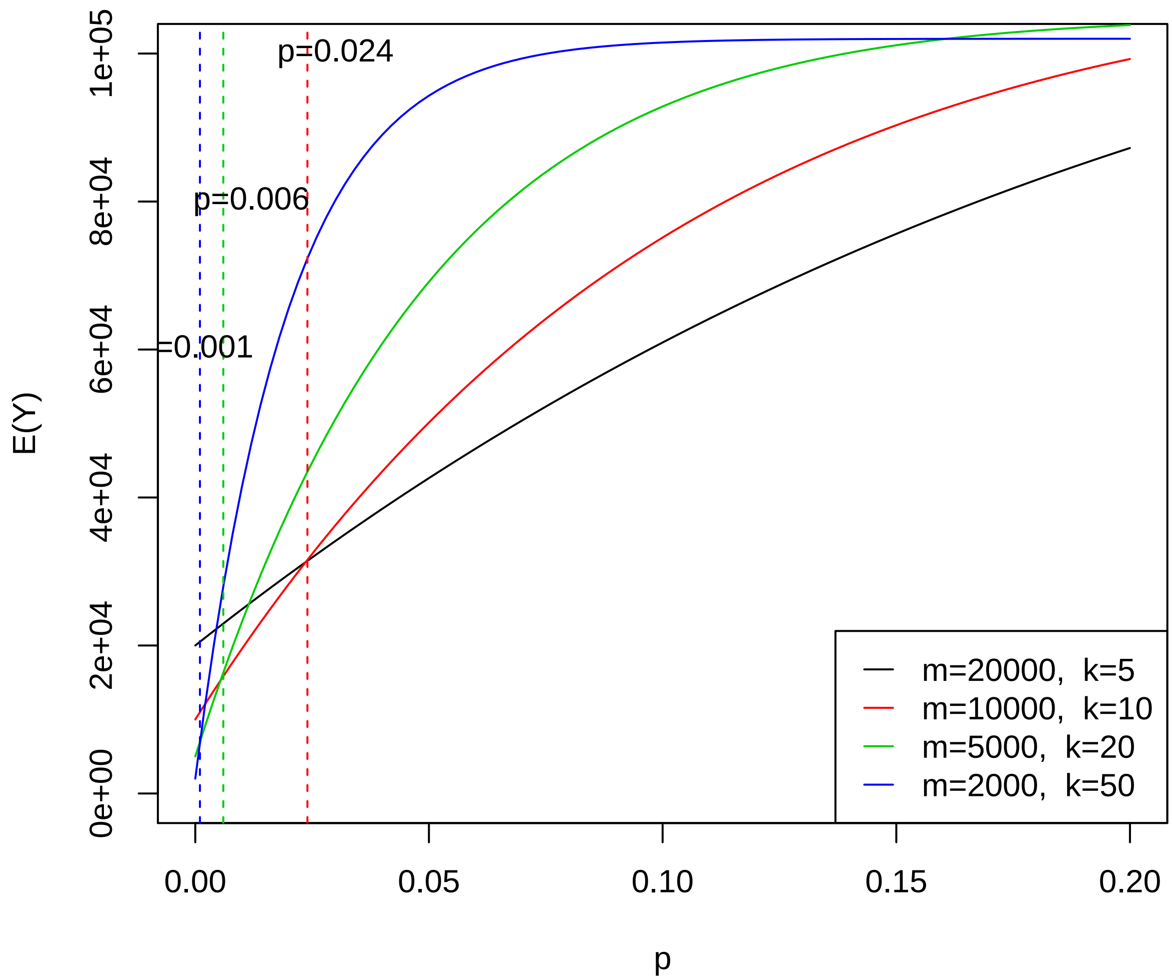
- plan (i): $E(Y) = N$
- plan (ii): $E(Y) = N(1 - (1-p)^k) + m \approx m$ (since $(1-p)^k \approx 1$ as $p \to 0$).

(e).

```
myfun<-function(p,m,k) m*k*(1-(1-p)^k)+m

##find the local minimum value for E(Y)
p=seq(0,0.1,by=0.001)
y=cbind(myfun(p,20000,5),myfun(p,10000,10),myfun(p,5000,20),myfun(p,2000,50))
ymin=apply(y,1,which.min)
localmin=p[c(which(ymin==4)[2],which(ymin==2)[1],which(ymin==1)[1])]

## graphs of E(Y) as a function of p
par(mfrow=c(1,1))
curve(myfun(x,20000,5),0,0.2,col=1,ylim=c(0,100000),xlab="p",ylab="E(Y)")
curve(myfun(x,10000,10),0,0.2,col=2,xlab="p",ylab="E(Y)",add=T)
curve(myfun(x,5000,20),0,0.2,col=3,xlab="p",ylab="E(Y)",add=T)
curve(myfun(x,2000,50),0,0.2,col=4,xlab="p",ylab="E(Y)",add=T)
abline(v=localmin,lty=2,col=c(4,3,2))
text(0,6e4,"p=0.001")
text(0.012,8e4,"p=0.006")
text(0.03,1e5,"p=0.024")
legend("bottomright", legend=c(
  paste("m=20000, ","k=5"),paste("m=10000, ","k=10"),
  paste("m=5000, ","k=20"),paste("m=2000, ","k=50")
),col=1:4,lty=rep(1,4),seg.len = 1)
```



(f). It is possible to happen that the test is positive but the actual is negative and the test is negative but the actual is positive.