

# MA677 HW1

Chenghao Meng

2021/2/9

## 1 Reproduce the Experiment

As the book mentioned, they have done a few experiments to reach the conclusion that  $m_{min} = 69$  and  $m_{min} = 73$ . So, we would like to reproduce the experiments by visualizing the distribution of the function:

$$\alpha(p) = \sum_{m \leq k \leq n} b(n, p, k).$$

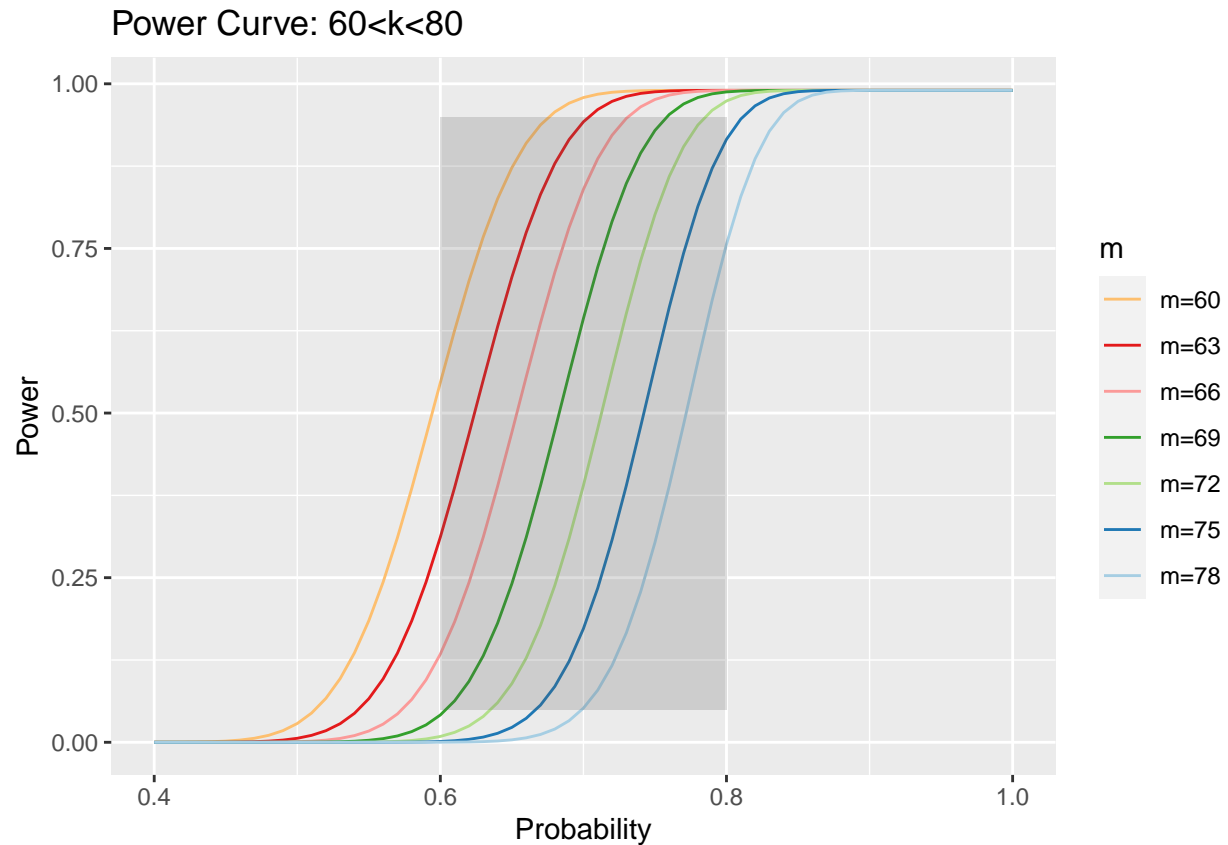
```
# Load the package
pacman::p_load(tidyverse)
options(scipen=200)

set.seed(1)
k <- seq(60,80,by = 3)
p <- seq(0.4,1,by = 1/100)

# Create empty dataframe
df <- data.frame(k = NA, prob = NA, power = NA)[-1,]

# Use for loop to create the dataframe
for (i in k){
  df <- rbind(df,data.frame(k = paste("m=",i,sep=""),
                           prob = p,
                           power = cumsum(dbinom(x = i,
                                                  size = 100,
                                                  prob = p))
  )
)
}

# Plot the graph
ggplot() +
  geom_line(data = df,aes(x = prob,y = power, color = k)) +
  geom_rect(aes(xmin = 0.6,xmax = 0.8,ymin = 0.05,ymax = 0.95),alpha = 0.2) +
  scale_color_brewer(palette = "Paired",direction = -1) +
  labs(col="m",title = "Power Curve: 60<k<80") +
  xlab("Probability") + ylab("Power")
```



```
df %>%
  group_by(k) %>%
  summarise(max_power = max(power),
            min_power = min(power))
```

```
## # A tibble: 7 x 3
##   k      max_power min_power
##   <chr>      <dbl>    <dbl>
## 1 m=60      0.990  2.44e- 5
## 2 m=63      0.990  1.80e- 6
## 3 m=66      0.990  9.06e- 8
## 4 m=69      0.990  3.07e- 9
## 5 m=72      0.990  6.85e-11
## 6 m=75      0.990  9.84e-13
## 7 m=78      0.990  8.82e-15
```

From the results above, we can see that when  $69 \leq k \leq 75$ , the power curves move accross the bottom and the top limits of the grey box, which are the thresholds of the type I and type II error. So, the smallest and the biggest values for  $m$  to thwarts the type I and type II errors are within this range.

## Reproduce Figure 3.7

```
m = c(69,73)
df2 = data.frame(k = NA, prob = NA, power = NA)[-1,]
for (i in m){
  df2 <- rbind(df2,data.frame(k = paste("m=",i,sep=""),
```

```

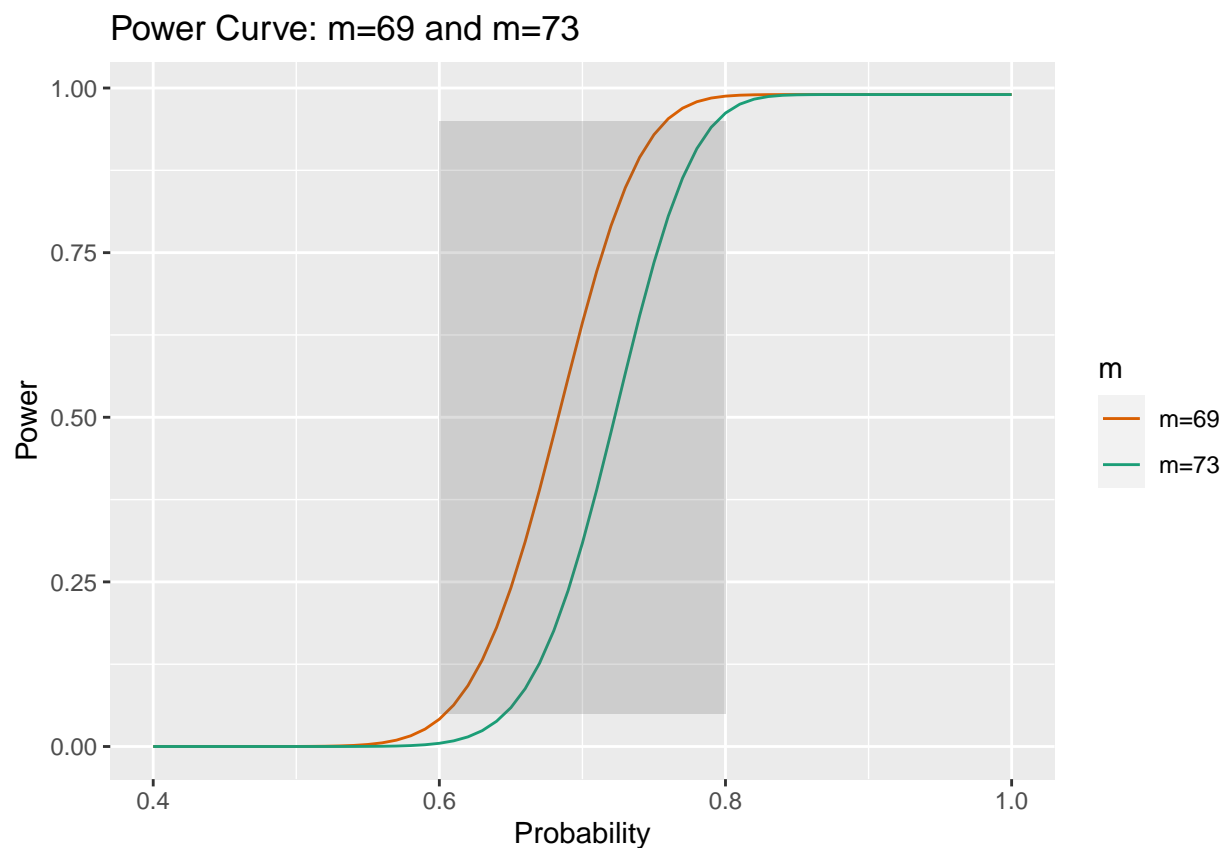
        prob = p,
        power = cumsum(dbinom(x = i,
                               size = 100,
                               prob = p))
      )
    }
  }
}

```

```

ggplot() +
  geom_line(data = df2,aes(x = prob,y = power, color = k)) +
  geom_rect(aes(xmin = 0.6,xmax = 0.8,ymin = 0.05,ymax = 0.95),alpha = 0.2) +
  scale_color_brewer(palette = "Dark2",direction = -1) +
  labs(col="m",title = "Power Curve: m=69 and m=73") +
  xlab("Probability") + ylab("Power")

```



For the grey box, the bottom line refers to the critical value of type I error, which is 5%, and the top line refers to the critical value of the type II error, which is 95% ( $1 - 5\%$ ).

For the lines, the orange line refers to the smallest  $m$  that allows the power curve move across the bottom and the top lines of the grey box, which are the thresholds of type I and type II error. The same, the green line refers to the biggest  $m$  that allows the power curve move across the bottom and top of the grey box.