HW3_SOLUTION

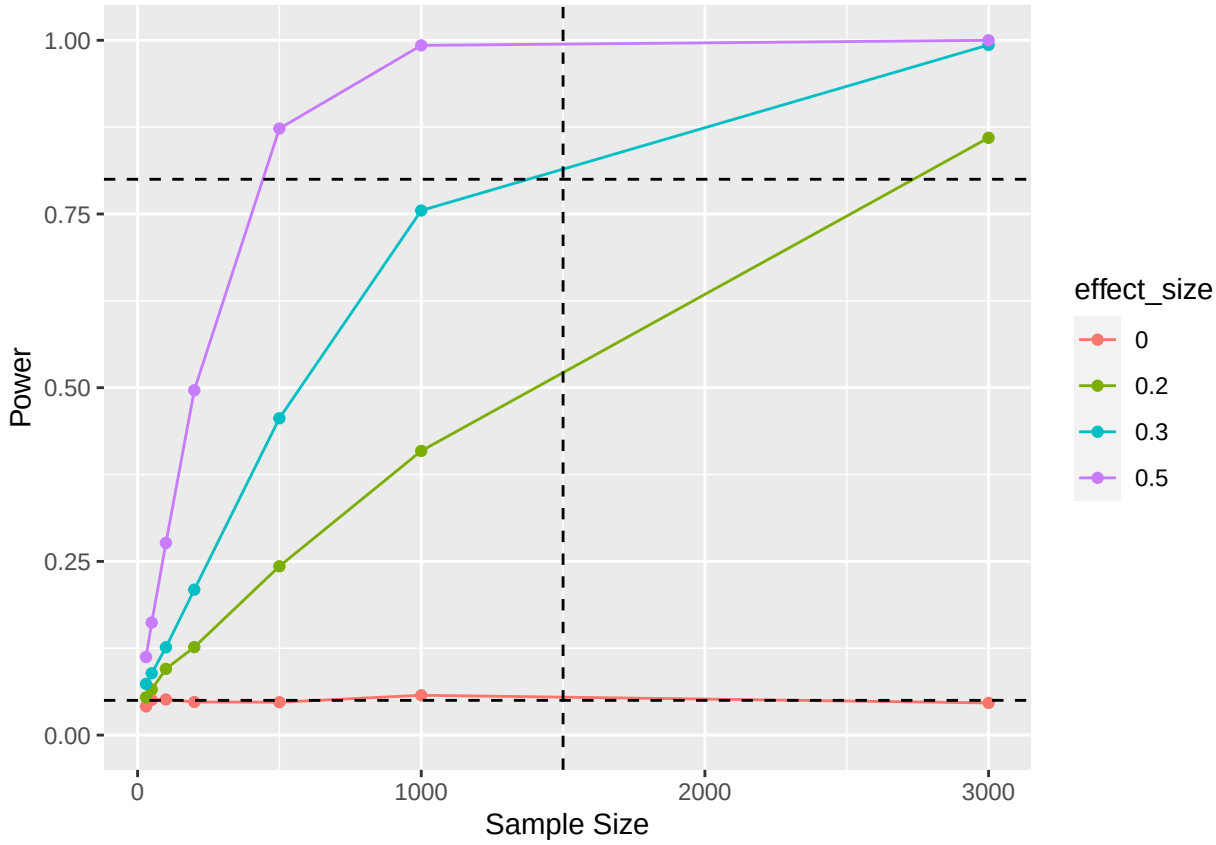
gustavoc@msu.edu

11/18/20201

Question 1

```
N=c(30,50,100,200,500,1000,3000)
EFFECTS=c(0,0.2,0.3,0.5)
nRep=3000

PAR=expand.grid(n=N,b=EFFECTS)
PAR$rej_rate=NA

for(i in 1:nrow(PAR)){

  n=PAR$n[i]
  b=PAR$b[i]

  countRejects=0
  for(j in 1:nRep){
      M=rep(0:1,each=n/2) # male dummy variable

      BMI.M=rnorm(mean=27.4,sd=sqrt(16.7),n=n/2)
      BMI.F=rnorm(mean=26.5,sd=sqrt(30),n=n/2)
      BMI=ifelse(M==1,BMI.M,BMI.F)
      Z=BMI-mean(BMI)
      signal=120-3*M+b*Z
      error=rnorm(n,sd=sqrt(300))
      SBP=signal+error
      pVal=summary(lm(SBP~M+BMI))$coef[3,4]
      countRejects=countRejects+(pVal<0.05)
    }
  PAR$rej_rate[i]=countRejects/nRep
}


library(ggplot2)
PAR$effect_size=factor(PAR$b)

p=ggplot(PAR,aes(x=n,y=rej_rate))
  p=p+geom_line(aes(colour=effect_size))
p=p+geom_point(aes(colour=effect_size))
  p=p+labs( x="Sample Size",y="Power")
p=p+ geom_vline(xintercept=1500,linetype='dashed')
  p=p+ geom_hline(yintercept=0.05,linetype='dashed')
  p=p+ geom_hline(yintercept=0.8,linetype='dashed',size=0.5)
```

```
p=p+ ylim(c(0,1))

plot(p)
```

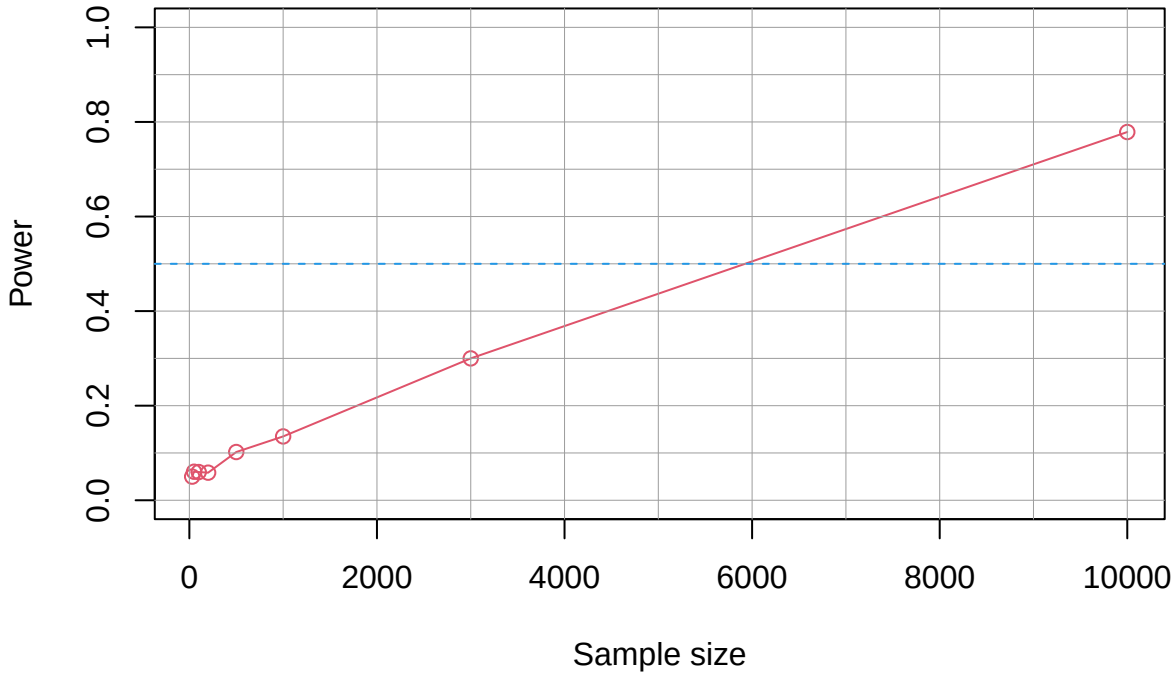


1.2): We need a sample size of at least 1,500 to reach a power of 0.8 with an effect size.

Question 2

```
N=c(30,50,100,200,500,1000,3000,10000)

REJ_RATE2=rep(0,length(N))
bM=0.4
bF=0.2
for(i in 1:length(N)){
  n=N[i]
  for(k in 1:nRep){
    M=rep(0:1,each=n/2) # male dummy variable

    BMI.M=rnorm(mean=27.4,sd=sqrt(16.7),n=n/2)
    BMI.F=rnorm(mean=26.5,sd=sqrt(30),n=n/2)
    BMI=ifelse(M==1,BMI.M,BMI.F)
    Z=BMI-mean(BMI)

    signal=120-3*M+bM*M*Z+bF*(1-M)*Z
    error=rnorm(n,sd=sqrt(300))
    SBP=signal+error
```

```
    pVal=summary(lm(SBP~M+BMI+BMI*M))$coef[4,4]
    REJ_RATE2[i]=REJ_RATE2[i]+(pVal<0.05)/nRep
  }
}
plot(REJ_RATE2~N,ylab='Power', xlab='Sample size',ylim=c(0,1),type='o',col=2)
abline(h=seq(from=0,to=1,by=.1),col=8,lwd=.5)
abline(v=seq(from=0,to=max(N),by=1000),lwd=.5,col=8)
abline(h=.5,col=4,lty=2)
```



Answer: Yes, the power analysis suggests that a power of 50% can be achieved with N~6,000.

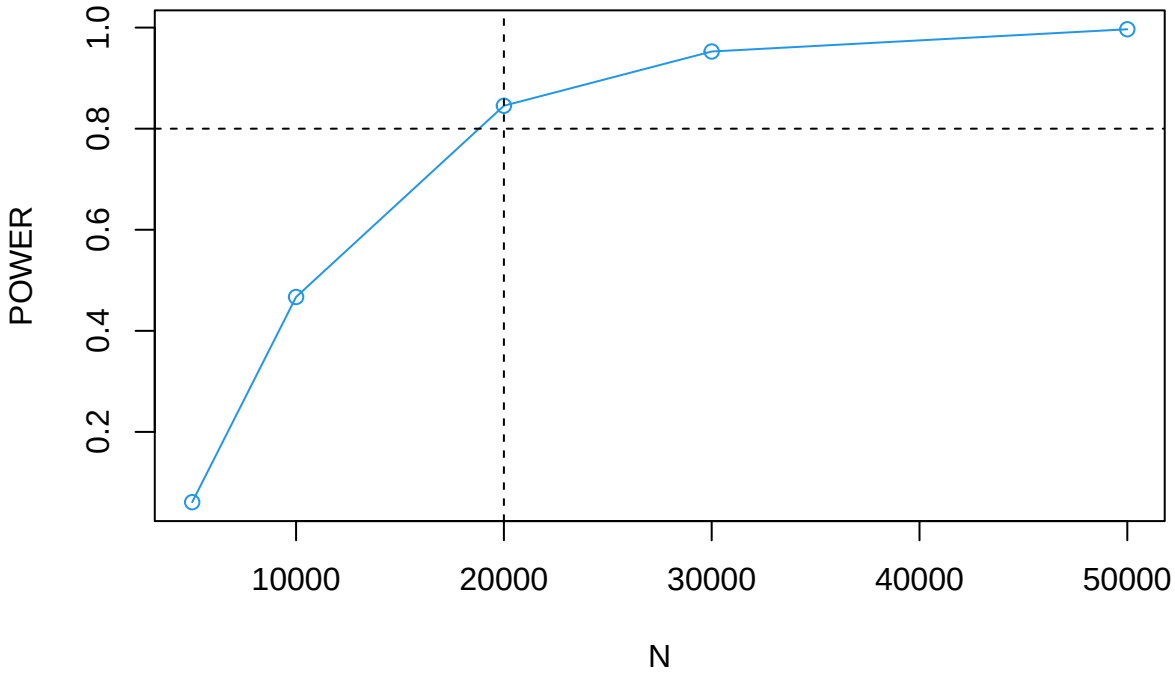
Question 3

```
risk0=1/500
risk1=risk0*0.3

N=c(5000,10000,20000,30000,50000)


POWER=rep(NA,length(N))

for(i in 1:length(N)){

  n=N[i]
  countRejects=0
  for(j in 1:nRep){
    Y0=rbinom(size=1,prob=risk0,n=round(n*.6))
    Y1=rbinom(size=1,prob=risk1,n=round(n*.4))

    Y=c(Y0,Y1)
    Z=c(rep(0,length(Y0)),rep(1,length(Y1)))
    fm=glm(Y~Z,family='binomial')
```

```
      pval=pnorm(summary(fm)$coef[2,3],lower.tail=TRUE)
      countRejects=countRejects+(pval<0.05)
    }
    POWER[i]=countRejects/nRep
  }
```

```
## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge
```

```
  plot(POWER~N,type='o',col=4);abline(h=.8,lty=2);abline(v=20000,lty=2)
```



For a vaccine with 70% efficacy, one would need about 20,000 participants to achieve 80% power to detect a reduction in risk induced by the vaccine. As a reference, Pfizer's BNT162b2 mRNA Covid-19 Vaccine safety and efficacy trial enrolled 43,548 participants (Polack et al., NEJM, 2020). They may have targeted higher power, risk may have been lower, they may have designed it to have results earlier than 6 months, or they may have designed the trial to detect efficacy even for a vaccine with efficacy lower than 70%. A main factor explaining why a very large sample size is required is the low risk for unvaccinated people (1/500 in our example).