Observational Assignment 1

Nick Graetz

February 1, 2020

1.(a)

In a randomized experiment, we want to make sure the only thing varying between the treatment and control groups is the treatment variable under study (seeding). Flying the plane through the clouds during non-seeding days ensures that there is no additional difference introduced between the two groups simply related to the plane moving through the clouds. If the pilots knew the result of the randomization (whether or not seeding was to occur), it may have also introduced an additional difference by them altering their behavior based on that knowledge (e.g. flying in a different pattern while seeding vs. not seeding).

1.(b)(c)

```
## Calculate observed test stat (sample mean difference)
obs_test <- cloud[treatment==1, mean(value)] - cloud[treatment==0, mean(value)]
print(obs_test)
```

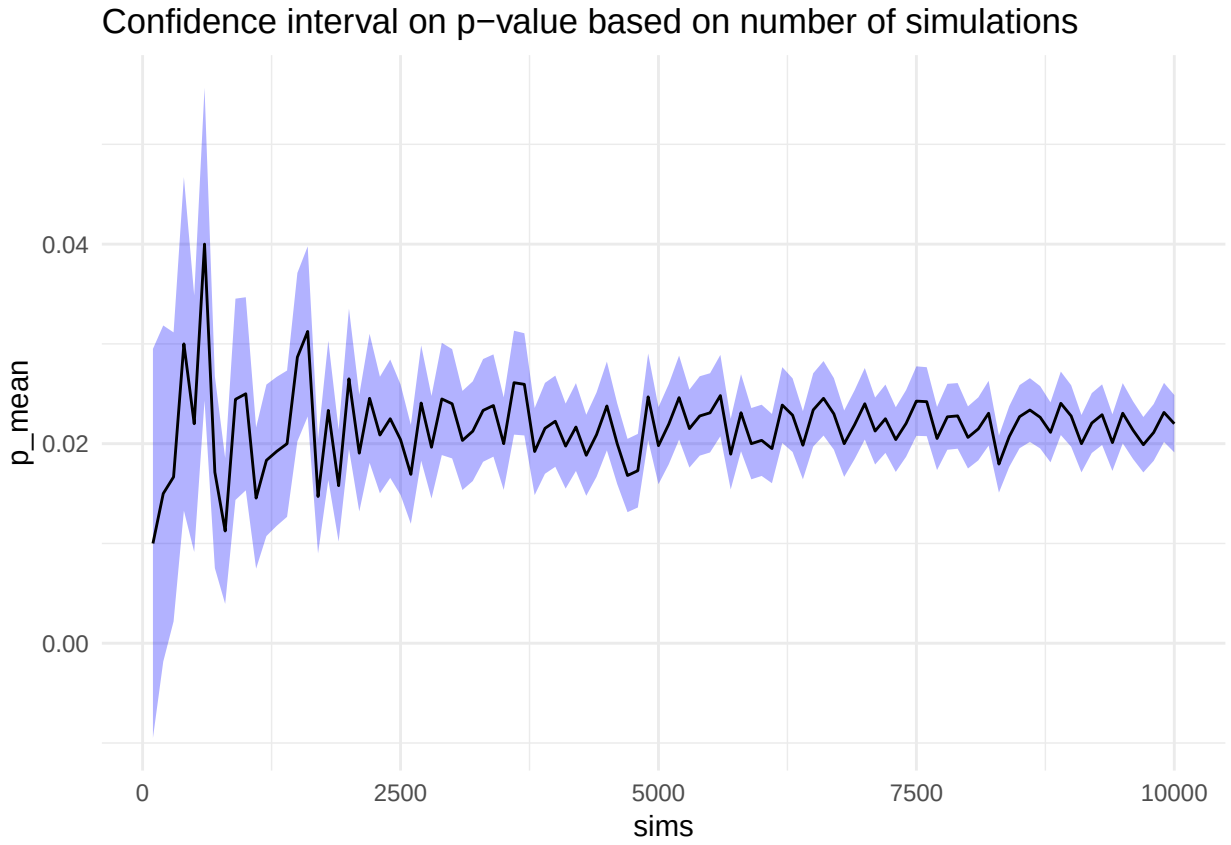
```
## [1] 277.4
```

```
## MC p-values
mc_results <- mc_pvalue(10000, cloud, obs_test, mc_test='mean')
print(mc_results)
```

```
##    p_mean    p_lower    p_upper
## 1: 0.0216 0.01875068 0.02444932
```

Based on the plot below running our MC function with different numbers of simulations, the p-value seems to converge fairly well by the time we get up to running with 10,000 simulations. We would want to iterate further out to make sure.

```
p_cis <- rbindlist(lapply(seq(100,10000,100),mc_pvalue,
                          cloud, obs_test, mc_test='mean'))
p_cis[, sims := seq(100,10000,100)]
ggplot(data=p_cis) +
  geom_ribbon(aes(x=sims,
                  ymin=p_lower,
                  ymax=p_upper),
              alpha=0.3,fill='blue',color=NA) +
  geom_line(aes(x=sims,
                y=p_mean),
            size=0.5,color='black') +
  theme_minimal() +
  labs(title = 'Confidence interval on p-value based on number of simulations')
```



1.(d)

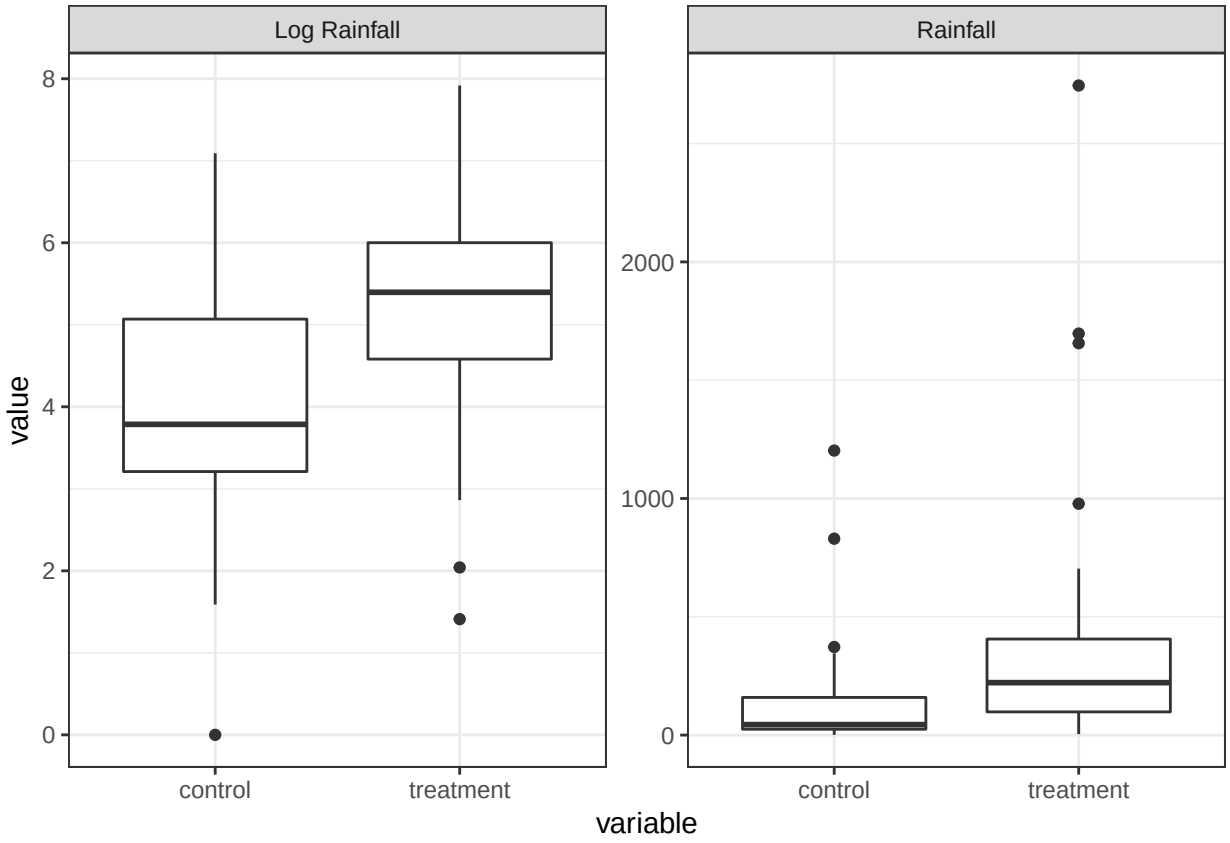
```
## Calculate observed test stat (sample variance difference)
obs_test <- cloud[treatment==1, var(value)] - cloud[treatment==0, var(value)]
## MC p-values
mc_results <- mc_pvalue(10000, cloud, obs_test, mc_test='variance')
print(mc_results)
```

```
##    p_mean    p_lower    p_upper
## 1: 0.0891 0.08351619 0.09468381
```

1.(e)

```
plot <- copy(cloud)
plot[, Type := 'Rainfall']
plot_log <- copy(plot)
plot_log[, value := log(value)]
plot_log[, Type := 'Log Rainfall']
plot <- rbind(plot, plot_log)
plot[, variable := factor(variable, levels=c('control','treatment'))]
boxes <- ggplot() +
  geom_boxplot(data=plot,
```

```
                aes(x=variable,
                    y=value)) +
  facet_wrap(~Type, scales='free_y') +
  theme_bw()
print(boxes)
```



The additive treatment effect model implies that the distribution of observed outcomes among treated is the same as among control. Our boxplot of rainfall in normal space suggests this is not the case (much more dispersion and extreme outliers among treated). In log space, the disperions are roughly equal between treated and control groups.

1.(f)

```
## Calculate observed test stat for multiplicative model
## (sample mean difference in log space)
cloud_log <- copy(cloud)
cloud_log[, value := log(value)]
obs_test <- cloud_log[treatment==1, mean(value)] - cloud_log[treatment==0, mean(value)]
print(obs_test)
```

```
## [1] 1.144458
```

```
## MC p-values
mc_results_mult <- mc_pvalue(10000, cloud_log, obs_test, mc_test='mean')
print(mc_results_mult)
```

```
##    p_mean     p_lower     p_upper
## 1: 0.0072 0.005542883 0.008857117
```

1.(g)

```
## Calculate Wilcoxon rank sum test stat with confidence intervals
wilcox.test(cloud_log[treatment==1, value],
            cloud_log[treatment==0, value],
            conf.int=TRUE)$conf.int
```

```
## [1] 0.2816254 2.0967816
## attr(,"conf.level")
## [1] 0.95
```

```
# Compare point estimate of treatment effect under Fisher and Wilcoxon.
# Mean difference
obs_test
```

```
## [1] 1.144458
```

```
# Median of the difference between a sample from x and a sample from y.
wilcox.test(cloud_log[treatment==1, value],
            cloud_log[treatment==0, value],
            conf.int=TRUE)$estimate
```

```
## difference in location
##                1.26038
```

1.(h)

```
multiplicative_treatment_effect <- exp(obs_test)
```

We can conclude with a high degree of confidence ($p < 0.05$) that the seeding treatment resulted in 3.14 times higher rainfall ($p = 0.0072$).

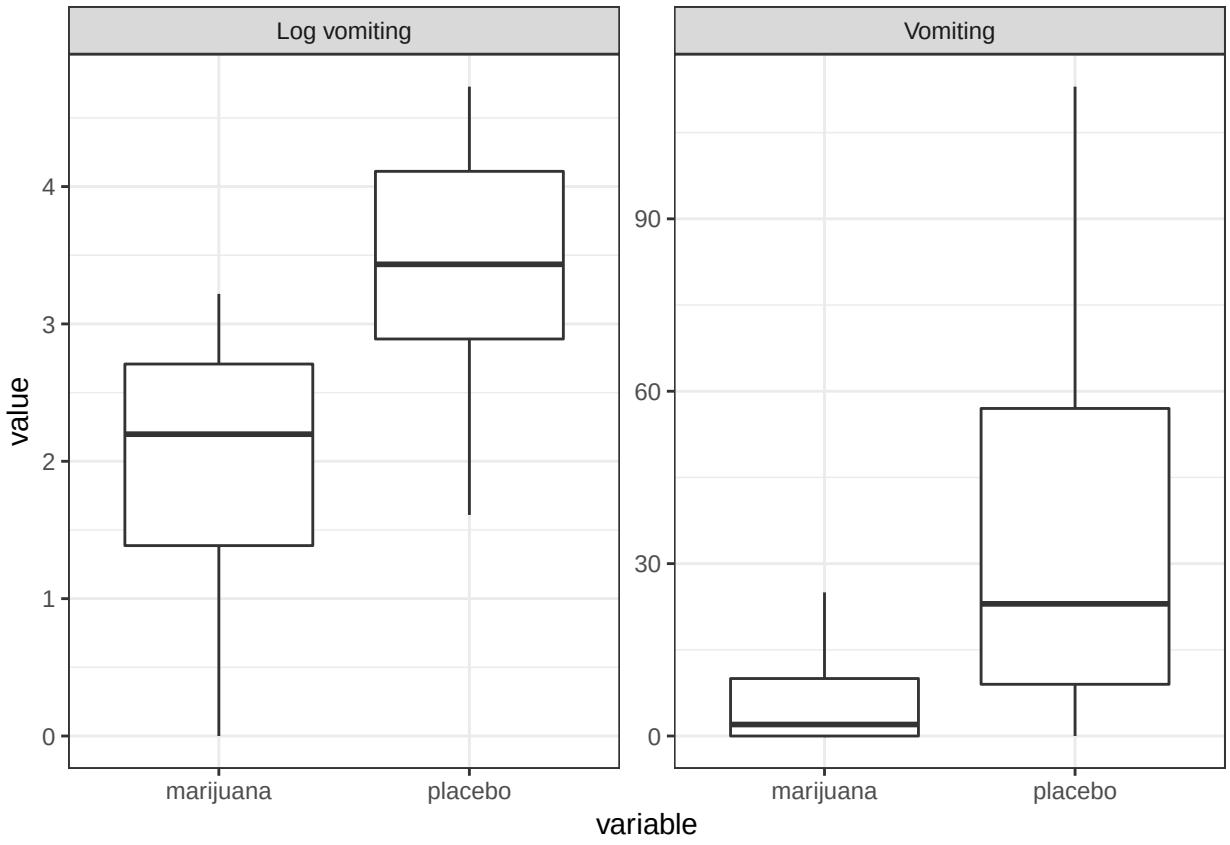
2.(a)

If we knew the order of treatment and placebo, we could examine the average treatment effect for those receiving the treatment first against those receiving the treatment second.

2.(b)

```
m <- fread('marijuana.csv')
m_plot <- melt(m, id.vars = 'V1')
m_plot[, Type := 'Vomiting']
```

```
m_plot_log <- copy(m_plot)
m_plot_log[, value := log(value)]
m_plot_log[, Type := 'Log vomiting']
m_plot <- rbind(m_plot,m_plot_log)
boxes <- ggplot() +
  geom_boxplot(data=m_plot,
               aes(x=variable,
                   y=value)) +
  facet_wrap(~Type, scales='free_y') +
  theme_bw()
print(boxes)
```



2.(c)

```
# Test of dilated treatment effect for matched
dilated.treateffect.matchedpair.test.func=function(Delta0,treated,
                                                   control,k,alternative="higher",
                                                   returntype="pval"){
  # Create vectors for Ri and Zi, and find total
  # number in experiment and  # number of treated subjects
  Ri=c(treated,control)
  Zi=c(rep(1,length(treated)),rep(0,length(control)))
```

```
  N=length(Ri)
  m=length(treated)
  # Calculate adjusted responses and rho=r_{C(k)}
  A=Ri-Zi*Delta0
  sorted.A=sort(A)
  rho=sorted.A[k]
  # q=1 if adjusted response>=rho, 0 otherwise
  q=(A>=rho)
  qpaired=cbind(q[1:length(treated)],q[(length(treated)+1):(2*length(treated))])
  qsplus=apply(qpaired,1,sum)
  no.discordant.pairs=sum(qsplus==1)
  # Test statistic = # of discordant pairs in which treated has qs=1
  teststat.obs= sum(qpaired[qsplus==1,1])
  # For returning the p-value,
  # p-value computed using hypergeometric distribution, see Notes 5
  if(returntype=="pval"& alternative=="lower"){
    returnval=pbinom(teststat.obs,no.discordant.pairs,.5)
  }
  if(alternative=="higher"){
    returnval=1-pbinom(teststat.obs-1,no.discordant.pairs,.5)
  }
  # For returning the test statistic minus its expected value
  if(returntype=="teststat.minusev"){
    returnval=teststat.obs-no.discordant.pairs*.5
  }
  returnval
}

calculate_dilated <- function(k,treated,control) {
  # Search for endpoints of lower and upper .025 confidence interval
  pval.Delta0.func=function(Delta0,treated,control,k,alternative){
    dilated.treateffect.matchedpair.test.func(Delta0,treated,control,k,alternative)-.02
  }
  upper.ci.limit=uniroot(pval.Delta0.func,c(-10000,10000),
                         treated=treated,control=control,
                         k=k,alternative="lower")$root
  lower.ci.limit=uniroot(pval.Delta0.func,c(-10000,10000),
                         treated=treated,control=control,
                         k=k,alternative="higher")$root
  # Hodges Lehmann estimate
  # Consider grid of values, find smallest value such
  # that test statistic is less than its expectation,
  # and largest value such that test statistic is
  # greater than its expectation, and average these values
  grid=seq(-1000,2000,1)
  teststat.minus.ev.grid=rep(0,length(grid))
```

```
  for(i in 1:length(teststat.minus.ev.grid)){
    teststat.minus.ev.grid[i]=
      dilated.treateffect.matchedpair.test.func(grid[i],treated,control,
                                                k,returntype="teststat.minusev")
  }
  sup=max(grid[(teststat.minus.ev.grid>.0001)==TRUE])
  inf=min(grid[(teststat.minus.ev.grid<(-.0001))==TRUE])
  hl.est=(sup+inf)/2
  return(data.table(k=as.character(k), HL=as.character(round(hl.est,1)),
                    ci95=paste0('(',round(lower.ci.limit,2),', ',
                                round(upper.ci.limit,2), ')')))
}
```

2.(d)

```
hl_results <- rbindlist(lapply(c(8,15,23), calculate_dilated, m[,marijuana], m[,placebo]
hl_results
```

```
##     k    HL      ci95
## 1:  8     0  (-18, 0)
## 2: 15 -20.5 (-31, -8)
## 3: 23   -47 (-95, -9)
```

3.

There are many problematic assumptions in Mr. X's argument. First, their conceptualization of potential outcomes is flawed because money is probably being spent on public health research in all years. As we don't observe what would have unfolded in those years if vaccinations, peniciliin, etc. had not been introduced, it is a hard comparison to make. There are also many other variables changing over time that may affect the crude death rate. Additionally, many of these public health innovations may primarily affect infants while the crude mortality rate captures changes at all ages. For example, an aging population drives up the crude mortality rate because older age is associated with higher mortality rates, which may be offsetting huge improvements to infant and child mortality.