MA50259 - Coursework 2 Assignment

Candidate Number: 08226

Part 1: Randomized Complete Block Experiment

First, we load the required libraries for subsequent analysis.

```
library(tidyverse)
library(MASS)
```

We now set up the corresponding dataframe for the given randomised complete block design (RCBD) experiment. In such experimental designs, we are able to split up the total variation such that each source of variation is measured separately. Specifically, this is achieved by the incorporation of blocking factors or blocks, these of which contain observations from each of the treatment factors, thus forming a more homogeneous experimental unit on which to compare the treatments. Effectively, this design improves the accuracy of the statistical comparisons among treatments by systematically controlling/eliminating the variability among the blocks.

```
block <- c(rep("1 (1931)",5), rep("1 (1932)",5), rep("2 (1931)",5), rep("2 (1932)",5),
           rep("3 (1931)",5), rep("3 (1932)",5), rep("4 (1931)",5), rep("4 (1932)",5),
           rep("5 (1931)",5), rep("5 (1932)",5), rep("6 (1931)",5), rep("6 (1932)",5))
barley <- rep(c("Manchuria", "Svansota", "Velvet","Trebi","Peatland"), times = 12)
yield <- c(81, 105.4, 119.7, 109.7, 98.3,
           80.7, 82.3, 80.4, 87.2, 84.2,
           146.6, 142, 150.7, 191.5, 145.7,
           100.4, 115.5, 112.2, 147.7, 108.1,
           82.3, 77.3, 78.4, 131.3, 89.6,
           103.1, 105.1, 116.5, 139.9, 129.6,
           119.8, 121.4, 124, 140.8, 124.8,
           98.9, 61.9, 96.2, 125.5, 75.7,
           98.9, 89, 69.1, 89.3, 104.1,
           66.4, 49.9, 96.7, 61.9, 80.3,
           86.9, 77.1, 78.9, 101.8, 96,
           67.7, 66.7, 67.4, 91.8, 94.1)
df <- data.frame(block,barley,yield)

df$block <- as.factor(df$block)
df$barley <- as.factor(df$barley)
```

In the given experiment, location and years (time) serve as blocking factors, from which variation may arise from. They may also be referred to as nuisance factors, as their variability can affect the results/response, however the reality is that we are not interested in this variability. For example, different locations may have different soil composition or weather, therefore possibly contributing to the variability observed in the barley yield. Thus, as a result, the experimental error will reflect both random error and variability between locations (and time). The RCBD diminishes this problem by reducing the variance of the experimental error, by removing the variability between locations (and time) from the experimental error. By grouping

(blocking) the varieties of barley based on location (or proximity) and year (or time), variability within these groups are greatly reduced as locations in close proximity normally have similar soil characteristics and recording responses annually (defined temporal proximity) allows for yields which are more alike.

Ideally, both blocking effects would be investigated using a Latin Square design, but due to the specifications of the question given, we combine location and year into one blocking variable for the purpose of the following analysis. This ensures that each of the b = 12 blocks contains exactly t = 5 experimental units for the total 12 x 5 = 60 experimental units.

The RCBD can be represented by the following equation:

$$y_{ij} = \mu + \ b_i + \tau_j + \epsilon_{ij}$$

where b_i represents the block (combined location and year) effects, τ_j represents the treatment effects (due to different varieties of the barley) and ϵ_{ij} is the experimental error.The usual assumptions of normality of experimental error and homogeneity of variance of experimental error across levels of the treatment factors and blocks are made for this model.

In order to determine whether a significant difference exists between the barley varieties, we run an ANOVA to test the following hypotheses:

$$H_0 : \tau_{Man} = \tau_{Svan} = \tau_{Vel} = \tau_{Tre} = \tau_{Peat} = 0$$

$$H_1 : \tau_i \neq 0 \ for \ some \ i \in \{Man, Svan, Vel, Tre, Peat\}$$

where Man represents Manchuria barley, $Svan$ represents Svansota barley, Vel represents Velvet barley, Tre represents Trebi barley and $Peat$ represents Peatland barley. τ_{Man}, τ_{Svan}, τ_{Vel}, τ_{Tre} and τ_{Peat} correspond to the mean yield (measured in bushels per acre) according to each respective variety of barley.

```
#b
mod_0 <- aov(yield ~ block + barley, data = df)
summary(mod_0)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## block       11  31913  2901.2  17.000 2.06e-12 ***
## barley       4   5310  1327.5   7.779 7.85e-05 ***
## Residuals   44   7509   170.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

By inspecting the ANOVA table, it is clear that the calculated p-value of 7.85e-05 is virtually zero and is thus, significantly less than 0.05. Therefore, we can reject the null hypothesis and infer that there is a significant difference between at least one of the values corresponding to the mean yield, based on the different treatment effects. In conclusion, there exists an effect on the yield due to the variety of barley.

In order to determine which variety/varieties account for this difference, we can compare the overall means corresponding to each variety, as well as run Tukey's HSD post hoc procedure for further analysis.

```
model.tables(mod_0,type="means")$tables$barley
```

```
## barley
## Manchuria  Peatland  Svansota     Trebi    Velvet
##  94.39167 102.54167  91.13333 118.20000  99.18333
```

```
TukeyHSD(mod_0,"barley")
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = yield ~ block + barley, data = df)
##
## $barley
##                           diff        lwr       upr     p adj
## Peatland-Manchuria   8.150000  -7.018286 23.318286 0.5502442
## Svansota-Manchuria  -3.258333 -18.426620 11.909953 0.9726113
## Trebi-Manchuria     23.808333   8.640047 38.976620 0.0005085
## Velvet-Manchuria     4.791667 -10.376620 19.959953 0.8959315
## Svansota-Peatland  -11.408333 -26.576620  3.759953 0.2222818
## Trebi-Peatland      15.658333   0.490047 30.826620 0.0400003
## Velvet-Peatland     -3.358333 -18.526620 11.809953 0.9694406
## Trebi-Svansota      27.066667  11.898380 42.234953 0.0000712
## Velvet-Svansota      8.050000  -7.118286 23.218286 0.5620467
## Velvet-Trebi       -19.016667 -34.184953 -3.848380 0.0075363
```

The means and Tukey comparisons show the average yield for Trebi barley to be approximately 118.2 bushels per acre. It is significantly more than the average yield produced by all of the other barely varieties. For example, looking at the two extremes, the greatest significance is seen between Trebi and Svansota as this comparison returns the smallest p-value of 0.0000712. Likewise, the least significance is seen between Trebi and Peatland as this comparison returns the largest p-value of 0.0400003 (when looking at the significant p-values). P-values for the two other comparisons between Trebi and the remaining barley varieties fall in-between these two values. All of these four values are considerably less than 0.05, thus implying the existence of a significant difference. In fact, average yield for Trebi barley is the only one that shows any significance in these comparisons between the other barely varieties, and hence, accounts for the result observed in the previous ANOVA output.

The estimated value of the response variance σ^2 can also be derived from the ANOVA output above. More specifically, we examine the residual mean square value to get the respective value as follows:

```
#c
sigma2.est<-anova(mod_0)$`Mean Sq`[3]
sigma2.est
```

```
## [1] 170.6605
```

Now, with an estimated σ^2 of 170.6605, we can determine how many more experimental runs would have been needed in order to achieve this same variance level, had the experiment not being blocked. From this, we can then determine whether carrying out blocking in the experiment was beneficial or not.

We first fit a new model, this time without the blocking factor included, and then carry out the same analysis using ANOVA.

```
#d
mod_0a <- aov(yield ~ barley, data = df)
summary(mod_0a)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## barley       4   5310  1327.5   1.852  0.132
## Residuals   55  39422   716.8
```

```
anova(mod_0a)$`Mean Sq`[2]/sigma2.est
```

```
## [1] 4.199978
```

Thus, it can be inferred that blocking was highly beneficial and efficient in this case, as we would have needed approximately 320% or 4.2 times more experimental runs if blocking had not been used.

Part 2: Observational Studies

1. Apgar Scores of Babies

In this cohort study, we investigate the developed long-term effects of asymmetric delayed foetal growth. This was done by comparing the proportions of low birth weight babies with corresponding Apgar scores determined at birth, in two groups. More specifically, after consulting earlier ultrasound scans, the exposed group consisted of babies with asymmetric delayed foetal growth and the control group consisted of babies with symmetric foetal growth. Here, Apgar scores are the outcomes of interest, such that a score of 7 or higher is an indication of a healthy baby.

We first set up the corresponding dataframe for the analysis. Note that the encoding is completed as follows:
• for the factor `apgar`, 1 represents babies with low Apgar scores (< 7) and 0 represents babies with high Apgar scores (≥ 7)
• for the factor `symmetry`, 1 represents babies with asymmetrical foetal growth and 0 represents babies with symmetric foetal growth

```
apgar <- factor(rep(c(1, 0), times = c(35, 72)))
#....<7 = 1, >/= 7 = 0

symmetry <- factor(rep(c(1, 0, 1, 0), times = c(33, 2,58,14)))
#.... asy = 1, sym = 0

growth <- data.frame(apgar,symmetry)
```

We can better visualise the data by the summary table below:

```
levels(growth$apgar) <- c(">/= 7", "<7")
levels(growth$symmetry) <- c("Symmetric", "Asymmetric")

cont.table <-
  growth %>%
  group_by(symmetry, apgar) %>%
  summarise(n=n()) %>% spread(key = apgar, value = n) %>%
  mutate(Total =`<7`+`>/= 7`, Risk = `<7`/Total, Odds = `<7`/`>/= 7`)
print(cont.table)
```

```
## # A tibble: 2 x 6
## # Groups:   symmetry [2]
##   symmetry   `>/= 7`  `<7` Total  Risk  Odds
##   <fct>        <int> <int> <int> <dbl> <dbl>
## 1 Symmetric       14     2    16 0.125 0.143
## 2 Asymmetric      58    33    91 0.363 0.569
```

```
summaries.rr<-summary(glm(apgar~symmetry,data=growth,
                          family = binomial(link = "log"))) %>% coefficients()

print(exp(summaries.rr[2,"Estimate"]))  #....>1 so asymmetric growth negatively affects apgar score
```

```
## [1] 2.901099
```

```
exp(c(summaries.rr[2,"Estimate"]-1.96*summaries.rr[2,"Std. Error"],
      summaries.rr[2,"Estimate"]+1.96*summaries.rr[2,"Std. Error"]))
```

```
## [1]  0.7713345 10.9114461
```

From the summary table, we can see that babies in the exposed group, that is those with asymmetric foetal growth, have a higher risk of a low Apgar score when compared to babies with symmetric foetal growth. Likewise, the derived relative risk of approximately 2.9 implies that the risk of a low Apgar score is 2.9 greater in the asymmetric group than the symmetric group, meaning that asymmetric foetal growth could be a risk factor for low Apgar scores. However, the presence of a positive relative risk value does not necessarily indicate that this association is statistically significant. One must consider the respective confidence interval and p-value(s) to determine significance.

When considering the derived confidence interval, we are 95% confident that the population risk of low Apgar scores in the group with asymmetric foetal growth is between 0.8 and 10.9 times the risk of low Apgar scores in the group with symmetric foetal growth. However, this interval contains one so we fail to reject the null hypothesis that the population relative risk is one with a significance level of 0.05. In other words, the risk of low Apgar scores in babies with asymmetric foetal growth versus babies with symmetric foetal growth is not significantly different.

We can further solidify this outcome by carrying out a Chi-squared test of the following hypotheses:

$$H_0 : \ p_1 = \ p_0$$

$$H_1 : \ p_1 \neq \ p_0$$

where p_1 and p_0 represent the risk of low Apgar scores in babies with asymmetric foetal growth and symmetric foetal growth respectively.

```
chisq.test(x=growth$symmetry,y=growth$apgar,correct=F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  growth$symmetry and growth$apgar
## X-squared = 3.4912, df = 1, p-value = 0.0617
```

```
summaries.rr[2,"Pr(>|z|)"]
```

```
## [1] 0.1150584
```

From the above output, we can see that the calculated p-value of 0.0617 is more than 0.05 and so we fail to reject the null hypothesis that the risk of low Apgar scores in babies with asymmetric foetal growth versus those with symmetric foetal growth are not significantly different. Extending this to the normal approximation to the sampling distribution of the estimated risk ratio, we can also see that the p-value of 0.1150584 is more than 0.05. We can now conclude that there is insufficient evidence of the symmetry of the growth delay (asymmetric foetal growth) affecting the risk of low Apgar scores.

In theory, the odds ratio can be utilised to estimate the relative risk of low Apgar scores. However, in cohort studies, such as this one, the problem arises when the outcome occurs in more than 10% of the population of interest [3,7]. Therefore, in the context of this study, if low Apgar scores are rare, then the odds ratio and relative risk may be comparable, but the odds ratio will overestimate the risk if low Apgar scores are more common [9]. In such cases, the odds ratio should be avoided, and the relative risk will be a more accurate estimation of risk [5].

In this study, low Apgar scores are not particularly rare as they arise in approximately 13% of babies with symmetric foetal growth and 36% of babies with asymmetric foetal growth, and so, the odds ratio would exaggerate the estimated strength of association. More specifically, we expect the odds ratio to be further from 1 compared to the previously derived relative risk of 2.901099. We can verify this by deriving the corresponding odds ratio; this value should be significantly larger than the previously derived relative risk.

```
summaries.or<-glm(apgar~symmetry,data=growth,
                  family = binomial(link = "logit")) %>%
  summary %>% coefficients()

print(exp(summaries.or[2,"Estimate"]))
```

```
## [1] 3.982759
```

```
exp(c(summaries.or[2,"Estimate"]-1.96*summaries.or[2,"Std. Error"],
      summaries.or[2,"Estimate"]+1.96*summaries.or[2,"Std. Error"]))
```

```
## [1]  0.8521042 18.6155249
```

With a derived value of 3.982759, we observe the expected result of a substantial increase in the odds ratio when compared to the relative risk. Thus, in this case, it would not be preferable to report the odds ratio as it largely overestimates the relative risk due to the commonality of low Apgar scores in the study population.

There are several potential variables which may have confounded the association between low Apgar scores and asymmetric foetal growth. For example, if maternal and/or neonatal infection occurred during any part of the gestational period, this would predispose the baby to risk factors including low Apgar scores, among other things such as premature births, low birth weight and birth asphyxia. Another confounding variable to consider is the occurrence of pre-eclampsia and/or eclampsia as this condition is known to result in foetal growth restriction, therefore predisposing affected babies to low Apgar scores. Lastly, whether the mother smoked during any part of the gestational period should be considered as a confounding variable as this action is known to increase problems during foetal development and result in birth defects.

2. Asthma Deaths in New Zealand

In this case-control study, we investigate the possible association between prescribed Fenoterol and asthma deaths in New Zealand during the late 1970s.

We first set up the corresponding dataframe for the analysis. Note that the encoding is completed as follows:
• for the factor `fenoterol`, Yes represents those who used prescribed Fenoterol and No represents those who did not use prescribed Fenoterol
• for the factor `cases`, 1 represents those with asthma who died from asthma and 0 represents those with asthma who did not die from asthma

```
fenoterol <- factor(rep(c("Yes", "No"), times = c(249, 336)))

cases <- factor(c(rep(1,60), rep(0,189), rep(1,57), rep(0,279)))

study <- data.frame(fenoterol,cases)
```

Both cases and controls were chosen among persons who were admitted to hospital for asthma. The cases comprised 117 persons with asthma who died of asthma; the controls were 468 persons with asthma who did not die of asthma. We can better visualise the data by the summary table below:

```
levels(study$fenoterol) <- c("No", "Yes")
levels(study$cases) <- c("Controls", "Cases")

cont.table <-
  study %>%
  group_by(cases, fenoterol) %>%
  summarise(n=n()) %>% spread(key = cases, value = n) %>%
  mutate(Total =`Controls`+`Cases`, Odds = `Cases`/`Controls`)
print(cont.table)
```

```
## # A tibble: 2 x 5
##   fenoterol Controls Cases Total  Odds
##   <fct>        <int> <int> <int> <dbl>
## 1 No             279    57   336 0.204
## 2 Yes            189    60   249 0.317
```

In order to measure the association between prescribed Fenoterol and asthma deaths, we consider the odds ratio as derived below:

```
#a
summaries.or<-glm(cases~fenoterol,data=study, family = binomial(link = "logit")) %>%
  summary %>% coefficients
print(exp(summaries.or[2,"Estimate"]))
```

```
## [1] 1.553885
```

```
exp(c(summaries.or[2,"Estimate"]-1.96*summaries.or[2,"Std. Error"],
      summaries.or[2,"Estimate"]+1.96*summaries.or[2,"Std. Error"]))
```

```
## [1] 1.034501 2.334032
```

The crude odds ratio for the data is 1.553885 with 95% confidence interval (1.034501, 2.334032). Therefore, we are 95% confident that the population odds of death due to asthma in the group of those who used prescribed Fenoterol are between approximately 1.03 and 2.33 times the odds of death due to asthma in the group of those who did not use prescribed Fenoterol. This interval does not contain one so we reject the null hypothesis that the population odds ratio is one with a significance level of 0.05. In other words, the odds of death due to asthma in those who used prescribed Fenoterol versus those who did not use prescribed Fenoterol is significantly different.

Considering the fact that disease severity may be associated with Fenoterol prescription, making it a potential confounder, we stratify the data by whether oral steroids were prescribed or not, this of which is associated with severity.

We now set up the stratified dataframe for further analysis. Note that the additional variable is encoded as follows:
• for the factor `steroids`, 1 represents those who used prescribed oral steroids and 0 represents those who did not use prescribed oral steroids

```
#b

fenoterol <- factor(rep(c(1, 0, 1, 0), times = c(64, 73, 185, 263)))



cases <- factor(c(rep(1,26), rep(0,38), rep(1,7), rep(0,66),
                  rep(1,34), rep(0,151), rep(1,50), rep(0,213)))
  #rep(c("0", "1","0","1"), times = c(60, 57,189,279)))


steroids <- factor(rep(c(1, 0), times = c(137, 448)))

strat_study <- data.frame(steroids, fenoterol, cases)
```

We can better visualise the stratified data by the summary tables below:

```
levels(strat_study$fenoterol) <- c("No", "Yes")
levels(strat_study$cases) <- c("Controls", "Cases")
levels(strat_study$steroids) <- c("No steroids", "Steroids")

cont.table.ster <-
  strat_study %>% filter(steroids=="Steroids") %>%
  group_by(cases, fenoterol) %>%
  summarise(n=n()) %>% spread(key = cases, value = n) %>%
  mutate(Total =`Controls`+`Cases`, Odds = `Cases`/`Controls`)

cont.table.ster
```

```
## # A tibble: 2 x 5
##   fenoterol Controls Cases Total  Odds
##   <fct>        <int> <int> <int> <dbl>
## 1 No              66     7    73 0.106
## 2 Yes             38    26    64 0.684
```

```
cont.table.no_ster <-
  strat_study %>% filter(steroids=="No steroids") %>%
  group_by(cases, fenoterol) %>%
  summarise(n=n()) %>% spread(key = cases, value = n) %>%
  mutate(Total =`Controls`+`Cases`, Odds = `Cases`/`Controls`)
cont.table.no_ster
```

```
## # A tibble: 2 x 5
##   fenoterol Controls Cases Total  Odds
##   <fct>        <int> <int> <int> <dbl>
## 1 No             213    50   263 0.235
## 2 Yes            151    34   185 0.225
```

We now calculate the odds ratio and corresponding confidence interval for each stratum.

```
summaries.ster<-summary(glm(cases~fenoterol,data=strat_study,
                            family=binomial(link = "logit"),
                            subset=steroids=="Steroids")) %>% coefficients()
print(exp(summaries.ster[2,"Estimate"]))
```

```
## [1] 6.451128
```

```
summaries.ster
```

```
##               Estimate Std. Error   z value     Pr(>|z|)
## (Intercept)  -2.243745  0.3974532 -5.645305 1.648882e-08
## fenoterolYes  1.864255  0.4719601  3.950026 7.814256e-05
```

```
 exp(c(summaries.ster[2,"Estimate"]-1.96*summaries.ster[2,"Std. Error"]
,summaries.ster[2,"Estimate"]+1.96*summaries.ster[2,"Std. Error"]))
```

```
## [1]  2.557968 16.269574
```

The crude odds ratio for the steroids prescribed stratum is 6.451128 with 95% confidence interval (2.557968, 16.269574). Therefore, we are 95% confident that when considering persons who used oral steroids, the population odds of death due to asthma in the group of those who used prescribed Fenoterol are between approximately 2.56 and 16.27 times the odds of death due to asthma in the group of those who did not use prescribed Fenoterol. This interval does not contain one so we reject the null hypothesis that the population odds ratio is one with a significance level of 0.05. In other words, when considering persons who used oral steroids, the odds of death due to asthma in those who also used prescribed Fenoterol versus those who did not use prescribed Fenoterol is significantly different.

```
summaries.no_ster<-summary(glm(cases~fenoterol,data=strat_study,
                            family=binomial(link = "logit"),
                            subset=steroids=="No steroids")) %>% coefficients()
print(exp(summaries.no_ster[2,"Estimate"]))
```

```
## [1] 0.9592053
```

```
summaries.no_ster
```

```
##                 Estimate Std. Error    z value     Pr(>|z|)
## (Intercept)  -1.44926916  0.1571459 -9.2224440 2.904041e-20
## fenoterolYes -0.04165015  0.2464328 -0.1690122 8.657870e-01
```

```
 exp(c(summaries.no_ster[2,"Estimate"]-1.96*summaries.no_ster[2,"Std. Error"]
,summaries.no_ster[2,"Estimate"]+1.96*summaries.no_ster[2,"Std. Error"]))
```

```
## [1] 0.5917575 1.5548173
```

The crude odds ratio for the non-steroids prescribed stratum is 0.9592053 with 95% confidence interval (0.5917575, 1.5548173). Therefore, we are 95% confident that when considering persons who did not use oral steroids, the population odds of death due to asthma in the group of those who used prescribed Fenoterol are between approximately 0.59 and 1.55 times the odds of death due to asthma in the group of those who did not use prescribed Fenoterol. This interval contains one so we fail to reject the null hypothesis that the population odds ratio is one with a significance level of 0.05. In other words, when considering persons who did not use oral steroids, the odds of death due to asthma in those who also used prescribed Fenoterol versus those who did not use prescribed Fenoterol is not significantly different.

Overall, the effect of Fenoterol prescription on asthma deaths seems to be different for persons prescribed with oral steroids compared to those not prescribed (odds risk of 6.451128 compared to odds risk of 0.9592053 for those prescribed with Fenoterol). Simply put, the use of oral steroids along with Fenoterol may have an increased association with asthma deaths compared to just the use of Fenoterol. However, in this analysis, we are trying to determine whether any association exists between only Fenoterol and asthma deaths. As discussed before, steroid prescription is a potential confounder and from the stratified analysis, we can observe that there is some difference between the two strata. Now, to determine if any confounding of this nature is in fact present, we can calculate the regression adjusted odds ratio as seen below:

```
#c
summaries.or<-glm(cases~fenoterol+steroids,data=strat_study,
                  family = binomial(link = "logit")) %>%
  summary %>% coefficients
print(exp(summaries.or[2,"Estimate"])) # similar to or from part a....no evidence of confounding
```

```
## [1] 1.536691
```

```
exp(c(summaries.or[2,"Estimate"]-1.96*summaries.or[2,"Std. Error"],
      summaries.or[2,"Estimate"]+1.96*summaries.or[2,"Std. Error"]))
```

```
## [1] 1.022143 2.310265
```

The regression adjusted odds ratio is 1.536691 with 95% confidence interval (1.022143, 2.310265). The interval does not contain one so we reject the null hypothesis that the corresponding population odds ratio is one with a significance level of 0.05. In other words, after controlling (adjusting) for the effect of steroid prescription, the odds of death due to asthma in those who used prescribed Fenoterol versus those who did not use prescribed Fenoterol is still significantly different. Now, we are 95% confident that the population odds of death due to asthma in those who used prescribed Fenoterol are between approximately 1.02 and 2.31 times the population odds of death due to asthma in those who did not use prescribed Fenoterol.

When comparing the crude odds ratio of 1.553885 and the adjusted odds ratio of 1.536691, we can see that there is very little difference between the two values. Likewise, the crude and adjusted confidence intervals tend to overlap, and so, we can conclude that steroid prescription is not a serious confounder of the association between Fenoterol prescription and asthma deaths.

There are several potential variables which may have confounded the association between Fenoterol prescription and asthma deaths. For example, the smoker status of a person, whether it be never, current or former, may have an effect on asthma development/exacerbation and thus, death due to asthma. Closely linked to this is the exposure to secondhand smoke, such that the frequency of this exposure is also a potential confounder. Likewise, when considering environmental factors, areas with increased air pollution may cause asthma development/exacerbation and thus, death due to asthma. Consequently, the area(s) in which a person lives/frequents, may be a confounding variable. These are considered along with the usual confounders such as age and gender.

References

[1] Bondell, H. and Reich, B., 2008. Simultaneous Factor Selection and Collapsing Levels in ANOVA. Biometrics, 65(1), pp.169-177. doi: 10.1111/j.1541-0420.2008.01061.x.

[2] Cleveland, W., 1993. Visualizing data. Summit, N.J.: Hobart Press.

[3] Cummings, P., 2009. The Relative Merits of Risk Ratios and Odds Ratios. Archives of Pediatrics & Adolescent Medicine, 163(5), pp.438-445. doi: 10.1001/archpediatrics.2009.31.

[4] Fisher, R.A., 1937. The design of experiments. The design of experiments., (2nd Ed).

[5] Knol, M., Algra, A. and Groenwold, R., 2012. How to Deal with Measures of Association: A Short Guide for the Clinician. Cerebrovascular Diseases, 33(2), pp.98-103. doi: 10.1159/000334180.

[6] Lawson, J., 2014. Design and Analysis of Experiments with R (Vol. 115). CRC press.

[7] Ranganathan, P., Aggarwal, R. and Pramesh, C., 2015. Common pitfalls in statistical analysis: Odds versus risk. Perspectives in Clinical Research, 6(4), pp.222-224. doi: 10.4103/2229-3485.167092.

[8] Venables, W. and Ripley, B., 2011. Modern applied statistics with S. New York: Springer.

[9] Wang, Z., 2013. Converting Odds Ratio to Relative Risk in Cohort Studies with Partial Data Information. Journal of Statistical Software, 55(5). doi: 10.18637/jss.v055.i05.