Lsn 10

Clark

Admin

Awhile back I was asked to consult on a project with DPE examining the surface that the sled pull was conducted on. They wanted to demonstrate that there was an effect due to the surface. They took 25 volunteers and had them do the sled pull on grass, turf, and sand.

Their sources of variation diagram was:

Consider the following output

```
ACFT=read.csv("https://raw.githubusercontent.com/nick3703/MA376/master/ACFT.csv")
ACFT<-ACFT %>% mutate(Participantf=as.factor(Participant))
levels(ACFT$Surface)<-c("G","S","S","T")
contrasts(ACFT$Surface)=contr.sum
ACFT %>% group_by(Surface)%>%summarise(n=n(),mean=mean(Event),sd=sd(Event))
```

```
## # A tibble: 3 x 4
##   Surface     n  mean    sd
##   <fct>   <int> <dbl> <dbl>
## 1 G          25  108.  13.4
## 2 S          25  109.  13.8
## 3 T          25  100.  11.2
```

Calculate the observed effects for each method

Consider the ANOVA table

```
single.lm<-lm(Event~Surface,data=ACFT)
anova(single.lm)
```

```
## Analysis of Variance Table
##
## Response: Event
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## Surface    2  1026.7  513.37  3.1123 0.05054 .
## Residuals 72 11876.2  164.95
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

How much variation is explained by Surface? What does the p-value tell us?

What source of variation are we not considering?

Based on the output below, does participant seem to matter?

```
ACFT %>% group_by(Participantf)%>%summarise(n=n(),mean=mean(Event),sd=sd(Event))
```

```
## # A tibble: 25 x 4
##    Participantf     n  mean    sd
##    <fct>        <int> <dbl> <dbl>
##  1 1                3  90.3 1.53 
##  2 2                3  86.3 2.08 
##  3 3                3 103   4.58 
##  4 4                3 107.  5.86 
##  5 5                3 114   7.21 
##  6 6                3  92.7 0.577
##  7 7                3  93.3 4.16 
##  8 8                3  97.3 7.02 
##  9 9                3 105.  4.04 
## 10 10               3 109.  5.86 
## # ... with 15 more rows
```

Why is this a blocked study design?

The statistical model we should be analyzing is:

```
M<-5000
stats.df<-data.frame(trial=seq(1,M),stat=NA)
ACFT.mod<-ACFT
for(j in 1:M){
  ACFT.mod$shuffled.cat<-sample(ACFT$Surface)
  shuff.lm<-lm(Event~shuffled.cat,data=ACFT.mod)
  stats.df[j,]$stat<-anova(shuff.lm)$"F value"[1]
}
```

What is the mean/sd of our shuffled null distribution?

Our P value can be found by:

```
stats.df %>% filter(stat>3.1123)%>% summarise(pval=n()/M)
```

```
##    pval
## 1 0.047
```

But we still want to know whether, after adjusting for participant, the variability is statistically significant:

```
M<-5000
stats.df<-data.frame(trial=seq(1,M),stat=NA)
ACFT.mod<-ACFT
for(j in 1:M){
  ACFT.mod <- ACFT.mod %>% group_by(Participantf)%>%sample_n(3)
  ACFT.mod$shuffled.cat <- rep(c("S","T","G"),25)
  shuff.lm<-lm(Event~shuffled.cat,data=ACFT.mod)
  stats.df[j,]$stat<-anova(shuff.lm)$"F value"[1]
}

stats.df %>% filter(stat>3.1123)%>% summarise(pval=n()/M)
```

```
##   pval
## 1    0
```

Let's talk about what I'm doing here.

What happens to the mean/sd of our F distribution?

How rare is our F value now?

What is our conclusion?

Let's find Participant (block) effect for a few of our participants

```
effects=ACFT %>% group_by(Participantf)%>%summarise(effect=mean(Event)-mean(ACFT$Event))
```

Note that we have imposed a sum to zero constraint

```
sum(effects$effect)
```

```
## [1] -9.947598e-14
```

We can come up with an adjusted surface effect by:

```
adj.effect=ACFT %>% mutate(person.means=rep(effects$effect,3)) %>%
  mutate(adj.effect=Event-person.means)
```

Let's compare the block-adjusted values to the original values, how does the variation compare?

```
var(adj.effect$adj.effect)
```

```
## [1] 24.63964
```

```
var(ACFT$Event)
```

```
## [1] 174.3647
```

Does the group means change?

```
adj.effect %>% group_by(Surface) %>% summarise(means=mean(adj.effect),sd=sd(adj.effect))
```

```
## # A tibble: 3 x 3
##   Surface means    sd
##   <fct>   <dbl> <dbl>
## 1 G        108.  3.43
## 2 S        109.  3.32
## 3 T        100.  3.22
```

Our block-adjusted F-statistic is:

```
adj.mod<-lm(adj.effect ~ Surface+Participantf,data=adj.effect)
anova(adj.mod)
```

```
## Analysis of Variance Table
##
## Response: adj.effect
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## Surface       2 1026.75  513.37  30.934 2.338e-09 ***
## Participantf 24    0.00    0.00   0.000         1
## Residuals    48  796.59   16.60
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To simulate the F statistic we can do:

```
M<-5000
stats.df<-data.frame(trial=seq(1,M),stat=NA)
ACFT.mod<-adj.effect
for(j in 1:M){
  ACFT.mod$shuffled.cat<-sample(ACFT$Surface)
  shuff.lm<-lm(adj.effect~shuffled.cat+Participantf,data=ACFT.mod)
  stats.df[j,]$stat<-anova(shuff.lm)$"F value"[1]
}
```

Are validity conditions met for using F-distribution for our F statistic in this study?

Why/why not?

ANOVA Table without blocking, SS calculations:

```r
anova(single.lm)
```

```
## Analysis of Variance Table
##
## Response: Event
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## Surface    2  1026.7  513.37  3.1123 0.05054 .
## Residuals 72 11876.2  164.95
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA Table with blocking, SS calculations:

```r
full.lm<-lm(Event~Surface+Participantf,data=ACFT)
anova(full.lm)
```

```
## Analysis of Variance Table
##
## Response: Event
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## Surface       2  1026.7  513.37  30.934 2.338e-09 ***
## Participantf 24 11079.7  461.65  27.818 < 2.2e-16 ***
## Residuals    48   796.6   16.60
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So, should we always block? Let's consider again the ANOVA table:

If we create blocks and there are no true block effects, what happens to our F statistic?

In the best case scenario we have decreased the numerator of $MSE = \frac{SSE}{df}$ by enough to offset the loss of df in the denominator. As a rule of thumb before blocking you should, a-priori, think about the variation in our response and consider whether it makes logical sense that variation in our blocks would impact the variability in our response. You cannot be wrong though if your statistical model reflects your experiment. If you randomized within block you should build a model that contains blocks.