LSN 9 - ACFT Sled Pull (2.2)

CDT I. M. Smrt

Admin

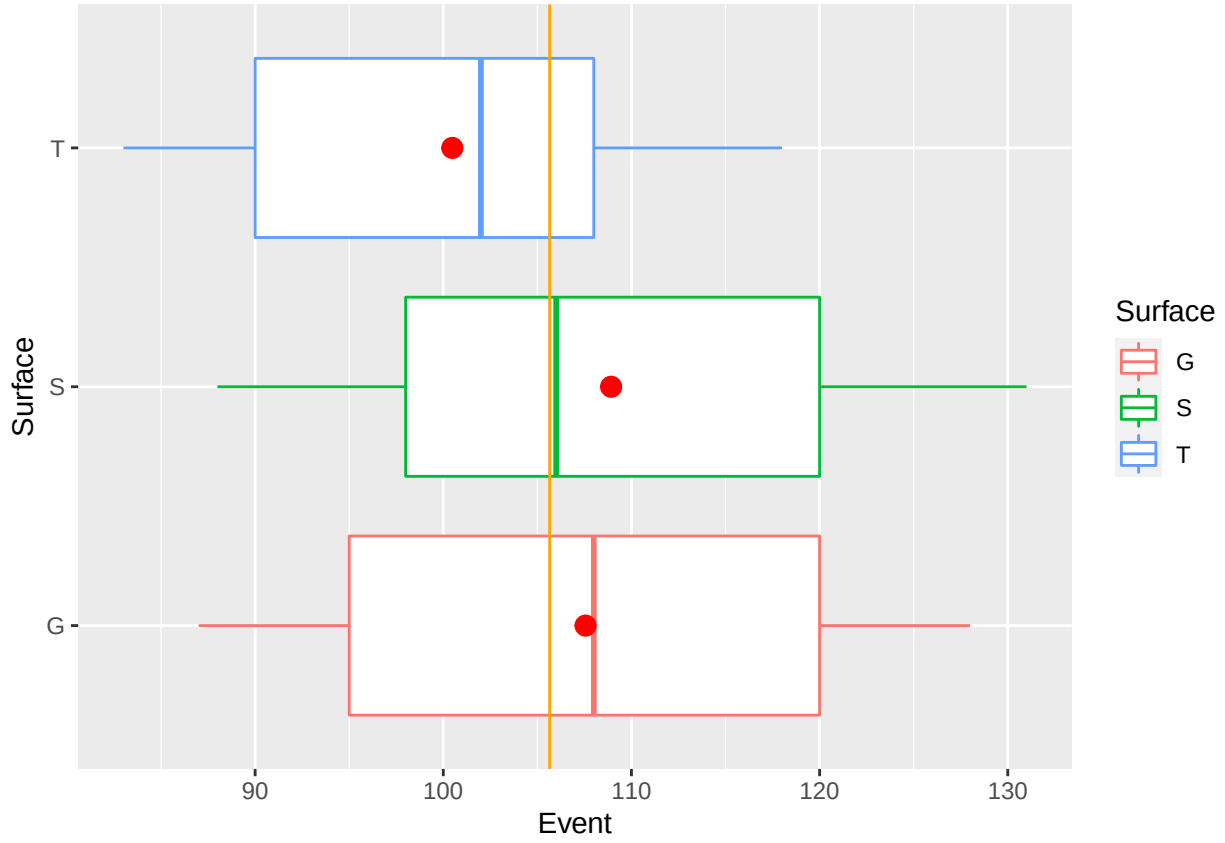
The Center for Data Analysis and Statistics (CDAS) at D/Math USMA was asked to consult on a project with DPE examining the surface that the sled pull was conducted on. They wanted to demonstrate that there was an effect due to the surface. They took 25 volunteers and had them do the sled pull on grass, turf, and sand.

1) Create a sources of variation diagram:

2) Consider the following boxplot. Explain what you see.

```
ACFT <- read.csv("https://raw.githubusercontent.com/jkstarling/MA376/main/ACFT.csv", stringsAsFactors =
ACFT <- ACFT %>% mutate(Participantf=as.factor(Participant)) %>% arrange(Surface, Participant)

ACFT %>% ggplot(aes(x=Event, y=Surface, col=Surface)) +
  geom_boxplot() + stat_summary(fun=mean, geom="point", shape=20, size=5, color="red", fill="red") + ge
```



3) Consider the following output. Calculate the observed effects for each surface. What is the sum of all effects?

```
contrasts(ACFT$Surface)=contr.sum
ACFT.mn.sd <- ACFT %>% group_by(Surface) %>%
  summarise(n=n(),mean=mean(Event),sd=sd(Event)); ACFT.mn.sd
```

```
## # 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
```

```
event.mn <- mean(ACFT$Event); event.mn
```

```
## [1] 105.6533
```

```
#
```

4) Consider the ANOVA table. How much variation is explained by Surface? What does the p-value tell us?

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

```
#
```

5) What other source of variation can we consider?

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

```
ACFT %>% group_by(Participantf) %>%
  summarise(n=n(),mean=mean(Event),sd=sd(Event)) %>%
  head(10)
```

```
## # A tibble: 10 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
```

7) Can we consider this a blocked study design? What is our updated statistical model?

8) Consider the simulation below. Calculate the p-value, mean, and standard deviation from this simulation. Does this model take into account blocking?

```
set.seed(376)
M<-1000
stat <- rep(NA, M)  # create empty vector
for(j in 1:M){
  ACFT$shuffled.cat <- sample(ACFT$Surface)
  shuff.lm <- ACFT %>% lm(Event~shuffled.cat,data=.)
  stat[j] <- anova(shuff.lm)$"F value"[1]
}
```

```
#
```

9) Consider another simulation which takes into account the blocking by participant. What is going on in this code? What happens to the mean/sd of our F distribution? How rare is our F value now? What is our conclusion?

```
set.seed(376)
M<-1000
stat <- rep(NA, M)  # create empty vector
for(j in 1:M){
  ACFT.mod <- ACFT %>% group_by(Participantf) %>% sample_n(3)
  ACFT.mod$shuffled.cat <- rep(c("S","T","G"),25)
  shuff.lm <- ACFT.mod %>% lm(Event~shuffled.cat,data=.)
  stat[j] <- anova(shuff.lm)$"F value"[1]
}
```

```
#
```

10) Calculate the Participant (block) effect for our participants. Did we impose a 'sum to zero constraint'?

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

```
##  [1] -15.320000 -19.320000  -2.653333   1.680000   8.346667 -12.986667
##  [7] -12.320000  -8.320000  -0.320000   3.680000   3.680000  10.680000
## [13]  15.013333  15.346667   8.680000   9.013333 -12.986667 -17.653333
## [19] -15.653333  16.346667  19.346667   4.680000  -8.653333  18.013333
## [25]  -8.320000
```

```
#
```

11) Consider the adjusted surface effects below. How does the variance of the block-adjusted values compare to the original values? Do the group means change?

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

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

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

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

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

```
## # 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
```

12) Our block-adjusted F-statistic and simulated F-statistics are given below. Calculate the p-value for the simulation. What do you notice about the values in the row associated with the `Participantf` variable?

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

```
## Analysis of Variance Table
##
## Response: adjeffect
##              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
```

```
set.seed(376)
M<-1000
stat <- rep(NA, M)  # create empty vector
for(j in 1:M){
  adj.effect$shuffled.cat <- sample(adj.effect$Surface)
  shuff.lm <- adj.effect %>% lm(adjeffect ~ shuffled.cat + Participantf, data=.)
  stat[j] <- anova(shuff.lm)$"F value"[1]
}
```

```
#
```

13) Are validity conditions met for using F-distribution for our F statistic in this study? Why/why not?

14) Compare the ANOVA Table for the model with blocking with the ANOVA table from the adjusted model and the original model. What do you notice about the SST for each model? What about the MSE? ($MSE = \frac{SSE}{df}$)

```
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
```

```
anova(adj.mod)
```

```
## Analysis of Variance Table
##
## Response: adjeffect
```

```
##              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
```

```
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
```

15) So, should we always block? If we create blocks and there are no true block effects, what happens to our F statistic?