Rank Sum

Code ▾

Grace Davis

2023-02-15

Hide

```r
library(tidyverse)
library(ggalt)
library(latex2exp)
library(dplyr)
library(ggplot2)

Ucenter <- function(size){
  c <- prod(size) +size[1]*(size[2]+1)/2
  return(c)
}
```

Exercise 1

The data does not support the hypothesis of different location parameters because, as calculated by the wilcox test, the alternative hypothesis is true that the location shift is greater than 0.

Hide

```
mm.data <- read.table(header=TRUE, file="mm.txt")
mm.data <- data.frame(mm.data)
wilcox.test(SB2M~Treatment, alternative="greater", data=mm.data)
```

```
## Warning in wilcox.test.default(x = DATA[[1L]], y = DATA[[2L]], ...): cannot
## compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  SB2M by Treatment
## W = 62, p-value = 0.1917
## alternative hypothesis: true location shift is greater than 0
```

Exercise 2

As calculated here, the rank-sums for each of the treatment groups in the experiment are 117 and 93, respectively.

```
mm.data %>%
  mutate(r=rank(SB2M)) %>%
  group_by(Treatment) %>%
  summarize(R=sum(r), n=n()) %>%
  mutate(U=Ucenter(n)-R) %>%
  print.data.frame() ->
  U.stats
```

```
##   Treatment   R   n   U
## 1   control 117  10  38
## 2       new  93  10  62
```

Exercise 3

Below is the graphically representation of the exact null sampling distribution for an experiment of this size. The mean and variance are 105 and 175, respectively.

```
mm.data %>%
  mutate(r=rank(SB2M)) %>%
  group_by(Treatment) %>%
  summarize(R=sum(r), n=n()) %>%
  mutate(U=Ucenter(n)-R) %>%
  print.data.frame() ->
  U.stats
```

```
##   Treatment   R  n  U
## 1   control 117 10 38
## 2       new  93 10 62
```

```
Me  <- U.stats$n[1]*(1+sum(U.stats$n))/2
Va  <- prod(U.stats$n)*(1+sum(U.stats$n))/12
names(Me) <- "mean"
names(Va) <- "variance"
signif(c(Me, Va), digits=4)
```

```
##     mean variance
##      105      175
```

```r
RankSumNullDist <- function(n1=5, n2=5) {
  N <- n1+n2
  SUPPORT <- choose(N, n1)
  values <- utils::combn(N, n1, simplify = FALSE)
  rs <- unlist(lapply(values,sum))
  rs.tab <- table(rs)
  rs.dist <- as.data.frame(rs.tab/SUPPORT)
  rs.dist$rs<- as.integer(as.character(rs.dist$rs)) # convert factor level to integer
  return(rs.dist)
}

R0.dist <- RankSumNullDist(8, 8)

ggplot(data=R0.dist, aes(rs, Freq)) +
  geom_lollipop() +
  geom_vline(xintercept=U.stats$R[1], color="orange") +
  theme_classic() +
  labs(title="Rank-Sum Test Null Distribution",
    subtitle=TeX("$n_1 = n_2 = 8$", italic = TRUE) )
```
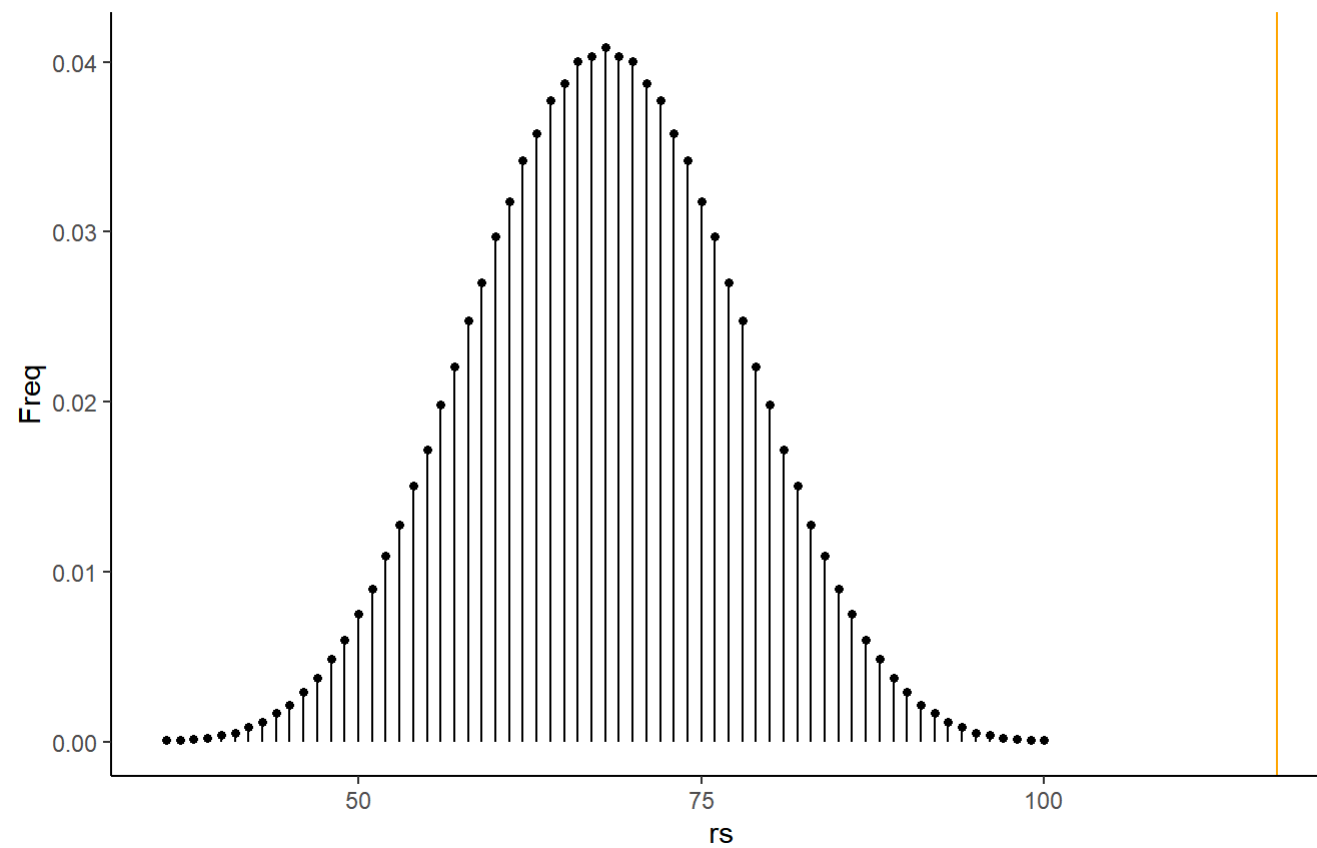
```
## Warning: Using the `size` aesthetic with geom_segment was deprecated in ggplot2 3.4.0.
## ℹ Please use the `linewidth` aesthetic instead.
```

Rank-Sum Test Null Distribution

$n_1 = n_2 = 8$

Exercise 4

The p-value of the two-sided test based on your exact sampling distribution is 0.1821731.

Hide

```
Ucenter <- function(size){
  c <- prod(size) +size[1]*(size[2]+1)/2
      return(c)
}

mm.data %>%
  mutate(r=rank(SB2M)) %>%
  group_by(Treatment) %>%
  summarize(R=sum(r), n=n()) %>%
  mutate(U=Ucenter(n)-R) %>%
  print.data.frame() ->
  U.stats
```

```
##   Treatment   R  n  U
## 1   control 117 10 38
## 2       new  93 10 62
```

```
Me  <- U.stats$n[1]*(1+sum(U.stats$n))/2
Va  <- prod(U.stats$n)*(1+sum(U.stats$n))/12
names(Me) <- "mean"
names(Va) <- "variance"
signif(c(Me, Va), digits=4)
```

```
##     mean variance
##      105      175
```

```
SDRT <- sqrt(Va)
Ze <- (max(U.stats$R)-Me)/SDRT
pval <- pnorm(Ze, lower.tail = FALSE)
names(pval) <- "p-value"
  pval
```

```
##    p-value
## 0.1821731
```

Exercise 5

The p-value generated by the wilcox test is 0.1917 which is pretty close to the one generated by the two-side test, 0.1822.

<div style="text-align: right">Hide</div>

```
wilcox.test(SB2M~Treatment, alternative="greater", data=mm.data)
```

```
## Warning in wilcox.test.default(x = DATA[[1L]], y = DATA[[2L]], ...): cannot
## compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  SB2M by Treatment
## W = 62, p-value = 0.1917
## alternative hypothesis: true location shift is greater than 0
```