non_parametric

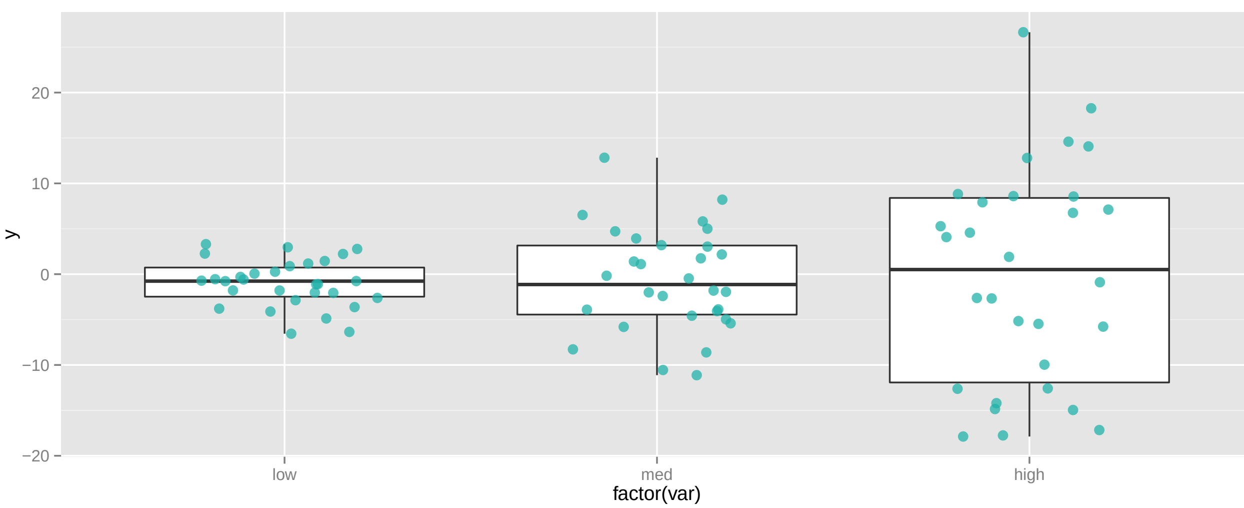
Joshua Burkhart

November 24, 2015

Class 16: Nonparametric methods for two sample problems

Parametric tests of homogeneity of variances

```
set.seed(10)
low <- rnorm(30, 0, 3)
med <- rnorm(30, 0, 6)
high <- rnorm(30, 0, 12)
var_df <- data.frame(low, med, high)
var_df_plot <- var_df %>%
  gather(var, y)
ggplot(var_df_plot, aes(x = factor(var), y = y)) +
  geom_boxplot() +
  geom_jitter(position = position_jitter(height = 0,
    width = 0.25), fill = "lightseagreen", colour = "lightseagreen", alpha = 0.75, size = 3)
```




```
# filter to just months 5 and 9
air5and9 <- airquality %>%
  filter(Month %in% c(5, 9))

air5and9 %>%
  group_by(Month) %>%
  dplyr::summarise(vars = var(Ozone, na.rm = TRUE),
            sds = sd(Ozone, na.rm = TRUE))
```

```
Source: local data frame [2 x 3]

  Month     vars      sds
```

```
  (int)    (dbl)    (dbl)
1     5 493.9262 22.22445
2     9 582.8276 24.14182
```

```r
# library(car)
# Levene's test
with(air5and9, leveneTest(Ozone ~ factor(Month), center = mean))
```

```
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value Pr(>F)
group  1  0.6934 0.4087
      53
```

```r
# library(car)
# Brown-Forsythe
with(air5and9, leveneTest(Ozone ~ factor(Month), center = median))
```

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  1  0.1954 0.6603
      53
```

```r
# filter to just months 8 and 9
air8and9 <- airquality %>%
  filter(Month %in% c(8, 9))

air8and9 %>%
  group_by(Month) %>%
  dplyr::summarise(vars = var(Ozone, na.rm = TRUE),
            sds = sd(Ozone, na.rm = TRUE))
```

```
Source: local data frame [2 x 3]

  Month      vars      sds
  (int)     (dbl)    (dbl)
1     8 1574.5985 39.68121
2     9  582.8276 24.14182
```

```r
# library(car)
# Levene's test
with(air8and9, leveneTest(Ozone ~ factor(Month), center = mean))
```

```
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value  Pr(>F)
group  1  7.0958 0.01021 *
      53
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# library(car)
# Brown-Forsythe
with(air8and9, leveneTest(Ozone ~ factor(Month), center = median))
```

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value  Pr(>F)
group  1  7.2717 0.00937 **
      53
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Wilcoxon Mann Whitney (WMW) test for stochastic ordering of alternatives

```r
# library(datasets)
data("esoph")
head(esoph)
```

```
  agegp     alcgp    tobgp ncases ncontrols
1 25-34 0-39g/day 0-9g/day      0        40
2 25-34 0-39g/day    10-19      0        10
3 25-34 0-39g/day    20-29      0         6
4 25-34 0-39g/day      30+      0         5
5 25-34     40-79 0-9g/day      0        27
6 25-34     40-79    10-19      0         7
```

```r
# unit of analysis is records for 88 age/alcohol/tobacco combinations

# collapse across age/tobacco combinations
tidy_esoph <- esoph %>%
  group_by(alcgp) %>%
  dplyr::summarise(cases = sum(ncases),
            controls = sum(ncontrols)) %>%
  gather(group, n, -alcgp) %>%
  mutate(alcgp = as.numeric(alcgp)) # required for wilcoxon test

# need to create data with unit of analysis = participant
esoph_data <- tidy_esoph %>%
  group_by(group) %>%
  do(data.frame(y = rep(.$alcgp, .$n)))

head(esoph_data)
```

```
Source: local data frame [6 x 2]
Groups: group [1]

   group     y
  (fctr) (dbl)
1  cases     1
2  cases     1
3  cases     1
```

```
4   cases     1
5   cases     1
6   cases     1
```

```r
esoph_data %>%
  group_by(group, y) %>%
  tally()
```

```
Source: local data frame [8 x 3]
Groups: group [?]

     group     y     n
    (fctr) (dbl) (int)
1    cases     1    29
2    cases     2    75
3    cases     3    51
4    cases     4    45
5 controls     1   415
6 controls     2   355
7 controls     3   138
8 controls     4    67
```

```r
# run the test!
wilcox.test(y ~ group, data = esoph_data)
```

```
    Wilcoxon rank sum test with continuity correction

data:  y by group
W = 135610, p-value < 0.00000000000000022
alternative hypothesis: true location shift is not equal to 0
```

Air quality data

```r
sum(1:55) #check with R
```

```
[1] 1540
```

```r
min_rank_sum <- cbind(min_R_A=sum(1:26), min_R_B=sum(1:29))
min_rank_sum #This is what R subtracts from W
```

```
     min_R_A min_R_B
[1,]     351     435
```

```r
pwilcox(377, 26, 29) #pi=P(WO) based on discrete distribution
```

```
[1] 0.5033348
```

```r
qwilcox(.5, 26, 29) #this would gotten us to W0 also
```

```
[1] 377
```

```r
qwilcox(.025, 26, 29)
```

```
[1] 261
```

```r
qwilcox(.975, 26, 29)
```

```
[1] 493
```

```r
air_ranks <- airquality %>%
  filter(Month %in% c(5, 9), !is.na(Ozone)) %>%
  mutate(oz_rank = rank(Ozone, ties.method = "average"))
# check that it worked
air_ranks %>%
  arrange(oz_rank) %>%
  head()
```

```
  Ozone Solar.R Wind Temp Month Day oz_rank
1     1       8  9.7   59     5  21     1.0
2     4      25  9.7   61     5  23     2.0
3     6      78 18.4   57     5  18     3.0
4     7      NA  6.9   74     5  11     4.5
5     7      49 10.3   69     9  24     4.5
6     8      19 20.1   61     5   9     6.0
```

```r
obs_rank_sum <- air_ranks %>%
  group_by(Month) %>%
  dplyr::summarise(rank_sum = sum(oz_rank))
obs_rank_sum
```

```
Source: local data frame [2 x 2]

  Month rank_sum
  (int)    (dbl)
1     5      635
2     9      905
```

```r
w_a <- obs_rank_sum$rank_sum[1] - min_rank_sum[1]
w_b <- obs_rank_sum$rank_sum[2] - min_rank_sum[2]
w_min <- min(w_a, w_b) #take the minimum as the test statistic
cbind(w_a, w_b, w_min)
```

```
     w_a w_b w_min
[1,] 284 470   284
```

```r
p_min <- min(pwilcox(w_min, 26, 29), 1-pwilcox(w_min, 26, 29))
p_2tailed <- 2*p_min
c(w_min, p_2tailed)
```

```
[1] 284.0000000   0.1195006
```

```r
w1 <- wilcox.test(Ozone ~ Month, data = airquality,
            subset = Month %in% c(5, 9), correct = FALSE, exact = TRUE)
```

```r
w1
```

```

    Wilcoxon rank sum test

data:  Ozone by Month
W = 284, p-value = 0.1166
alternative hypothesis: true location shift is not equal to 0
```

```r
lowerz <- -1.559383
upperz <- -1.576241
pmin <- min(pnorm(lowerz), 1 - pnorm(upperz))
2*pmin
```

```
[1] 0.1189058
```

```r
# library(coin)
# coin is especially fussy about predictors as factors
air_ranks$Month <- as.factor(air_ranks$Month)
w2 <- wilcox_test(Ozone ~ Month, data = air_ranks, conf.int = TRUE, distribution = "exact")
w2
```

```

    Exact Wilcoxon-Mann-Whitney Test

data:  Ozone by Month (5, 9)
Z = -1.5691, p-value = 0.1182
alternative hypothesis: true mu is not equal to 0
95 percent confidence interval:
 -13   2
sample estimates:
difference in location
                    -6
```

```r
# and right back where we started from!
w2@statistic@linearstatistic # the uncorrected observed rank sum for group A
```

```
[1] 635
```

```r
w2@statistic@linearstatistic - min_rank_sum[1] # the corrected observed W for group A
```

```
[1] 284
```

```r
# we can also check what W was under the null
expectation(w2) # should give you 728
```

```
  5
728
```

```r
w_asymp <- wilcox_test(Ozone ~ Month, data = air_ranks, conf.int = TRUE)
w_asymp
```

```

	Asymptotic Wilcoxon-Mann-Whitney Test

data:  Ozone by Month (5, 9)
Z = -1.5691, p-value = 0.1166
alternative hypothesis: true mu is not equal to 0
95 percent confidence interval:
 -12.999940   1.999926
sample estimates:
difference in location
             -5.999917
```

```r
x <- air_ranks$Ozone[air_ranks$Month == 5]
y <- air_ranks$Ozone[air_ranks$Month == 9]
diffs <- sort(as.vector(outer(y, x, "-")))
median(diffs)
```

```
[1] 6
```

```r
# library(pairwiseCI)
# pairwiseCI can be especially fussy about predictors as factors
air_ranks$Month <- as.factor(air_ranks$Month)
pairwiseCI(Ozone ~ Month, data= air_ranks, method = "HL.diff") #Exact conditional nonparametric CI for
```

```

95 %-confidence intervals
 Method:  Difference in location (Hodges-Lehmann estimator)


    estimate lower upper
9-5        6    -2    13
```

```r
air_ranks %>%
  group_by(Month) %>%
  summarise(medians = median(Ozone))
```

```
Source: local data frame [2 x 2]

   Month medians
  (fctr)   (dbl)
1      5      18
2      9      23
```

```
n_a <- 26
n_b <- 29
u_a <- n_a*n_b + min_rank_sum[1] - obs_rank_sum$rank_sum[1]
u_b <- n_a*n_b + min_rank_sum[2] - obs_rank_sum$rank_sum[2]
cbind(u_a, u_b, min(u_a, u_b))
```

```
     u_a u_b
[1,] 470 284 284
```

Your turn

The hypothesis that babies born to mothers who smoked have different birthweights than babies whose mothers did not smoke (2-tailed test).

```
library(MASS)
data(birthwt)
```

```
select <- dplyr::select
```

1. The total rank sum

```
rank_sum <- sum(1:nrow(birthwt))
rank_sum
```

```
[1] 17955
```

2. Expected rank sums for groups A and B under the null hypothesis

```
na <- birthwt %>% filter(smoke==0) %>% nrow()
nb <- birthwt %>% filter(smoke==1) %>% nrow()

exp_R_A <- (na * (nrow(birthwt) + 1))/2
exp_R_B <- (nb * (nrow(birthwt) + 1))/2
exp_R_A
```

```
[1] 10925
```

```
exp_R_B
```

```
[1] 7030
```

3. Minimum rank sums for groups A and B

```r
min_rank_sum <- cbind(min_R_A=sum(1:na),min_R_B=sum(1:nb))
min_rank_sum
```

```
     min_R_A min_R_B
[1,]    6670    2775
```

4. The expected (corrected) W statistic under the null hypothesis for groups A and B

```r
w_a_null <- exp_R_A - min_rank_sum[1] #qwilcox(.5,na,nb)
w_b_null <- exp_R_B - min_rank_sum[2] #qwilcox(.5,nb,na)
w_a_null
```

```
[1] 4255
```

```r
w_b_null
```

```
[1] 4255
```

```r
pwilcox(w_a,na,nb)
```

```
[1] 0.0000000000000000000000000000000000007554078
```

```r
qwilcox(.5,na,nb)
```

```
[1] 4255
```

5. The regions of rejection (=.05, two tailed) for the expected (corrected) W statistic for groups A and B

```r
w_lo <- qwilcox(.025,na,nb)
w_hi <- qwilcox(.975,na,nb)
w_lo
```

```
[1] 3536
```

```r
w_hi
```

```
[1] 4974
```

6. Observed (uncorrected) rank sums for groups A and B

```r
birthwt_ranks <- birthwt %>%
  filter(smoke %in% c(0,1), !is.na(bwt)) %>%
  mutate(bwt_rank = rank(bwt, ties.method = "average"))

birthwt_ranks %>%
  arrange(bwt_rank) %>%
  head()
```

```
  low age lwt race smoke ptl ht ui ftv  bwt bwt_rank
1   1  28 120    3     1   1  0  1   0  709      1.0
2   1  29 130    1     0   0  0  1   2 1021      2.0
3   1  34 187    2     1   0  1  0   0 1135      3.0
4   1  25 105    3     0   1  1  0   0 1330      4.0
5   1  25  85    3     0   0  0  1   0 1474      5.0
6   1  27 150    3     0   0  0  0   0 1588      6.5
```

```
obs_rank_sum <- birthwt_ranks %>%
  group_by(smoke) %>%
  dplyr::summarise(rank_sum = sum(bwt_rank))

obs_rank_sum
```

```
Source: local data frame [2 x 2]

  smoke rank_sum
  (int)    (dbl)
1     0  11919.5
2     1   6035.5
```

7. Observed (corrected) rank sums for groups A and B

```
w_a_obs <- obs_rank_sum$rank_sum[1] - min_rank_sum[1]
w_b_obs <- obs_rank_sum$rank_sum[2] - min_rank_sum[2]
w_a_obs
```

```
[1] 5249.5
```

```
w_b_obs
```

```
[1] 3260.5
```

8. The observed (minimum corrected) W statistic and its p-value

```
w_min <- min(w_a_obs, w_b_obs) #take the minimum as the test statistic
p_min <- min(pwilcox(w_min, na, nb), 1-pwilcox(w_min, na, nb))
p_2tailed <- 2*p_min
c(w_min, p_2tailed)
```

```
[1] 3260.500000000    0.006538324
```

9. The z-statistic and its p-value (either exact or asymptotic depending on your sample size)

```
mu_w_a <- (na * (nrow(birthwt) + 1))/2
mu_w_a
```

```
[1] 10925
```

```r
sd_w_a <- (na * nb * (nrow(birthwt) + 1))/2
sd_w_a
```

```
[1] 808450
```

```r
z_a <- (w_a_obs + .5 - mu_w_a)/sqrt(sd_w_a)
z_a
```

```
[1] -6.311597
```

```r
lowerz <- z_a + .5
upperz <- z_a - .5
pmin <- min(pnorm(lowerz), 1 - pnorm(upperz))
2*pmin
```

```
[1] 0.000000006187954
```

10. The Hodges-Lehmann estimate and its 95% confidence interval

```r
birthwt_ranks$smoke <- as.factor(birthwt_ranks$smoke)
pairwiseCI(bwt ~ smoke, data= birthwt_ranks, method = "HL.diff")
```

```


95 %-confidence intervals
 Method:  Difference in location (Hodges-Lehmann estimator)


    estimate lower upper
1-0     -307  -512   -85
```

11. What do the results suggest about your null hypothesis?

    The results suggest mothers who smoke during pregnancy give birth to children with lower median weights than mothers who don't smoke during pregnancy.