Multivariate analysis of variance (MANOVA)

Multivariate analysis of variance

- Standard ANOVA has just one response variable.

- What if you have more than one response?

- Try an ANOVA on each response separately.

- But might miss some kinds of interesting dependence between the responses that distinguish the groups.

Packages

```
library(car)
library(tidyverse)
```

Small example

- Measure yield and seed weight of plants grown under 2 conditions: low and high amounts of fertilizer.

- Data (fertilizer, yield, seed weight):

```
url <- "http://www.utsc.utoronto.ca/~butler/d29/manova1.txt"
hilo <- read_delim(url, " ")
```

```
##
## -- Column specification --------------------------
## cols(
##   fertilizer = col_character(),
##   yield = col_double(),
##   weight = col_double()
## )
```
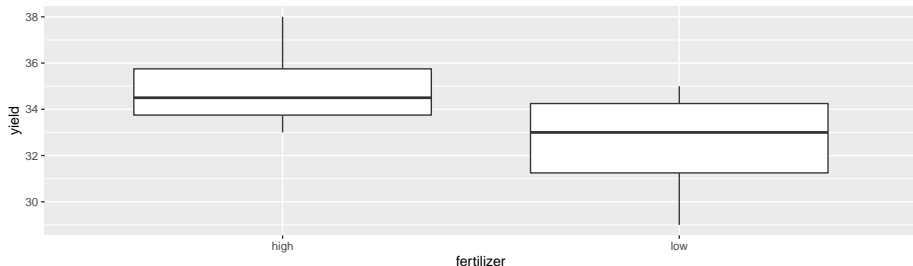
- 2 responses, yield and seed weight.

The data

`hilo`

fertilizer	yield	weight
low	34	10
low	29	14
low	35	11
low	32	13
high	33	14
high	38	12
high	34	13
high	35	14

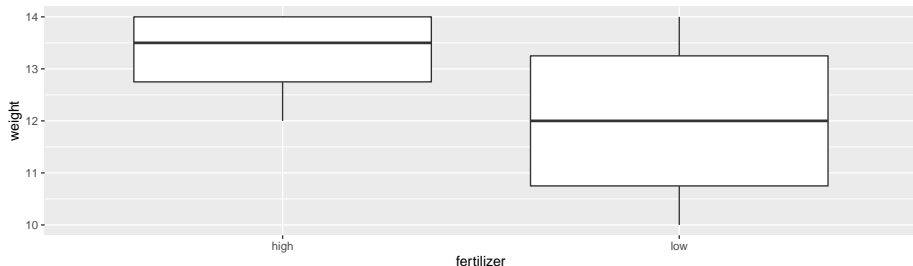# Boxplot for yield for each fertilizer group

```
ggplot(hilo, aes(x = fertilizer, y = yield)) + geom_boxplot()
```


Yields overlap for fertilizer groups.

Boxplot for weight for each fertilizer group

```
ggplot(hilo, aes(x = fertilizer, y = weight)) + geom_boxplot()
```


Weights overlap for fertilizer groups.

ANOVAs for yield and weight

```
hilo.y <- aov(yield ~ fertilizer, data = hilo)
summary(hilo.y)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## fertilizer   1   12.5  12.500   2.143  0.194
## Residuals    6   35.0   5.833
```

```
hilo.w <- aov(weight ~ fertilizer, data = hilo)
summary(hilo.w)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## fertilizer   1  3.125   3.125   1.471  0.271
## Residuals    6 12.750   2.125
```

Neither response depends significantly on fertilizer. But…

Plotting both responses at once

- Have two response variables (not more), so can plot the response variables against *each other*, labelling points by which fertilizer group they're from.

- First, create data frame with points $(31, 14)$ and $(38, 10)$ (why? Later):

```r
d <- tribble(
  ~line_x, ~line_y,
  31, 14,
  38, 10
)
```

- Then plot data as points, and add line through points in d:

```r
g <- ggplot(hilo, aes(x = yield, y = weight,
                      colour = fertilizer)) + geom_point() +
  geom_line(data = d,
            aes(x = line_x, y = line_y, colour = NULL))
```

The plot
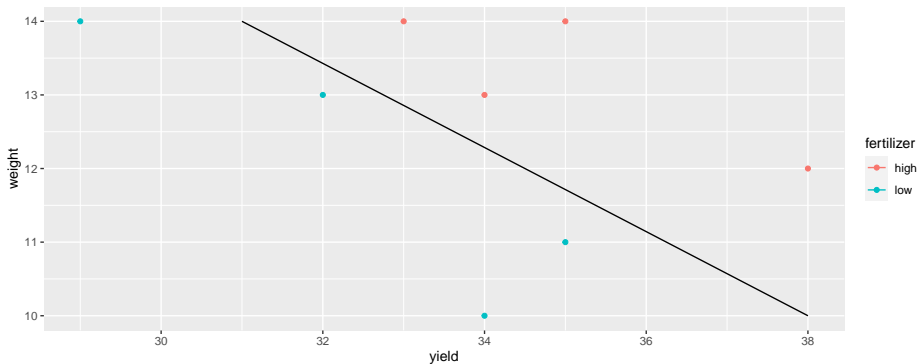
Comments

- Graph construction:
  - Joining points in `d` by line.
  - `geom_line` inherits `colour` from `aes` in `ggplot`.
  - Data frame `d` has no `fertilizer` (previous `colour`), so have to unset.
- Results:
  - High-fertilizer plants have both yield and weight high.
  - True even though no sig difference in yield or weight individually.
  - Drew line separating highs from lows on plot.

MANOVA finds multivariate differences

- Is difference found by diagonal line significant? MANOVA finds out.

```
response <- with(hilo, cbind(yield, weight))
hilo.1 <- manova(response ~ fertilizer, data = hilo)
summary(hilo.1)
```

```
##             Df  Pillai approx F num Df den Df  Pr(>F)
## fertilizer  1 0.80154   10.097      2      5 0.01755 *
## Residuals   6
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Yes! Difference between groups is *diagonally*, not just up/down (weight) or left-right (yield). The *yield-weight combination* matters.

Strategy

- Create new response variable by gluing together columns of responses, using `cbind`.
- Use `manova` with new response, looks like `lm` otherwise.
- With more than 2 responses, cannot draw graph. What then?
- If MANOVA test significant, cannot use Tukey. What then?
- Use *discriminant analysis* (of which more later).

Another way to do MANOVA

Install (once) and load package `car`:

```
library(car)
```

Another way…

```
hilo.2.lm <- lm(response ~ fertilizer, data = hilo)
hilo.2 <- Manova(hilo.2.lm)
hilo.2
```

```
##
## Type II MANOVA Tests: Pillai test statistic
##            Df test stat approx F num Df den Df  Pr(>F)
## fertilizer  1   0.80154   10.097      2      5 0.01755 *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Same result as small-m manova.

- Manova will also do *repeated measures*, coming up later.

Another example: peanuts

- Three different varieties of peanuts (mysteriously, 5, 6 and 8) planted in two different locations.

- Three response variables: `y`, `smk` and `w`.

```
u <- "http://www.utsc.utoronto.ca/~butler/d29/peanuts.txt"
peanuts.orig <- read_delim(u, " ")
```

```
##
## -- Column specification ------------------------------
## cols(
##   obs = col_double(),
##   location = col_double(),
##   variety = col_double(),
##   y = col_double(),
##   smk = col_double(),
##   w = col_double()
## )
```

The data

`peanuts.orig`

obs	location	variety	y	smk	w
1	1	5	195.3	153.1	51.4
2	1	5	194.3	167.7	53.7
3	2	5	189.7	139.5	55.5
4	2	5	180.4	121.1	44.4
5	1	6	203.0	156.8	49.8
6	1	6	195.9	166.0	45.8
7	2	6	202.7	166.1	60.4
8	2	6	197.6	161.8	54.1
9	1	8	193.5	164.5	57.8
10	1	8	187.0	165.1	58.6
11	2	8	201.5	166.8	65.0
12	2	8	200.0	173.8	67.2

Setup for analysis

```
peanuts <- peanuts.orig %>%
  mutate(
    location = factor(location),
    variety = factor(variety)
  )
response <- with(peanuts, cbind(y, smk, w))
head(response)
```

```
##          y    smk    w
## [1,] 195.3 153.1 51.4
## [2,] 194.3 167.7 53.7
## [3,] 189.7 139.5 55.5
## [4,] 180.4 121.1 44.4
## [5,] 203.0 156.8 49.8
## [6,] 195.9 166.0 45.8
```

Analysis (using `Manova`)

```
peanuts.1 <- lm(response ~ location * variety, data = peanuts)
peanuts.2 <- Manova(peanuts.1)
peanuts.2
```

```
##
## Type II MANOVA Tests: Pillai test statistic
##                  Df test stat approx F num Df den Df
## location          1   0.89348  11.1843      3      4
## variety           2   1.70911   9.7924      6     10
## location:variety  2   1.29086   3.0339      6     10
##                     Pr(>F)
## location          0.020502 *
## variety           0.001056 **
## location:variety  0.058708 .
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comments

- Interaction not quite significant, but main effects are.

- Combined response variable (y,smk,w) definitely depends on location and on variety

- Weak dependence of (y,smk,w) on the location-variety *combination.*

- Understanding that dependence beyond our scope right now.