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

```r
library(car) # may need to install first
library(tidyverse)
library(MVTests) # also may need to install
```

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://ritsokiguess.site/datafiles/manova1.txt"
hilo <- read_delim(url, " ")
```

```
## Rows: 8 Columns: 3
## -- Column specification --------------------------------
## Delimiter: " "
## chr (1): fertilizer
## dbl (2): yield, weight
##
## i Use `spec()` to retrieve the full column specification fo
## i Specify the column types or set `show_col_types = FALSE`
```

- 2 responses, yield and seed weight.

The data
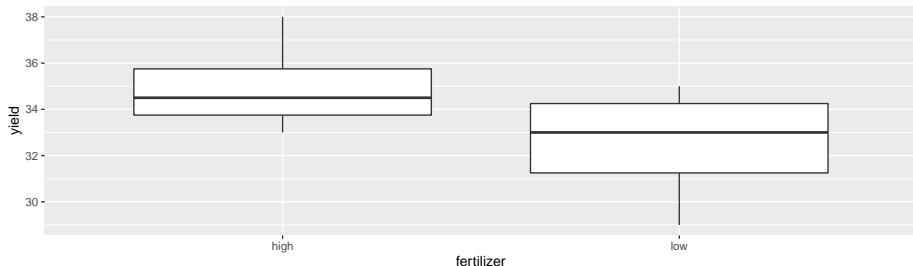
`hilo`

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

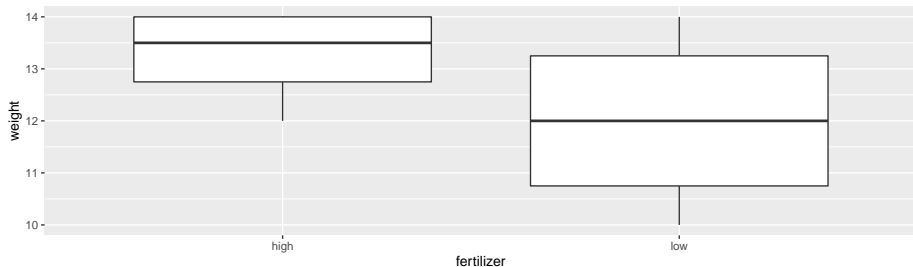# Boxplot for yield for each fertilizer group

```
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):

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

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

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

The plot
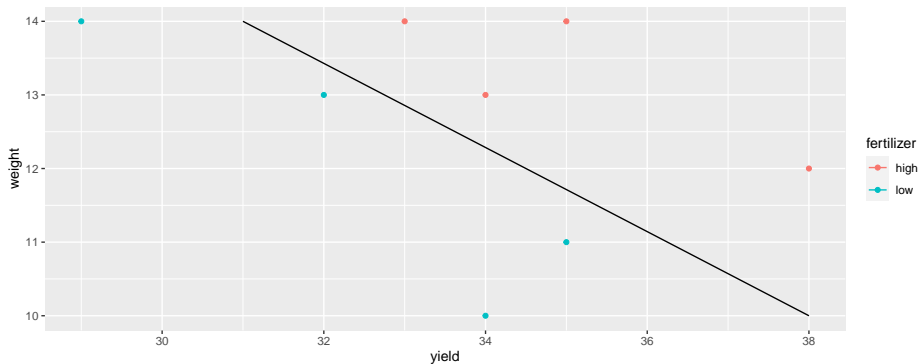
Comments

- 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

using `Manova` from package `car`:

```
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.

Assumptions

- normal distribution of *each* response within each treatment group (actually *multivariate* normal with correlations between response variables)

  - assess with boxplots as usual (or normal quantile plots)

- each response has same spread across all treatment groups, and constant correlation between response variables across treatment groups

  - assess with Box M test. This is very sensitive to unequal spread, so there is only a problem if P-value less than say 0.001. (BoxM in package MVTests.)

Assumptions for yield-weight data

For normal quantile plots, need "extra-long" with all the data values in one column:

```
hilo %>%
  pivot_longer(-fertilizer, names_to = "xname",
               values_to = "xvalue") %>%
  ggplot(aes(sample = xvalue)) + stat_qq() +
    stat_qq_line() +
    facet_grid(xname ~ fertilizer, scales = "free") -> g
```

There are only four observations per response variable - treatment group combination, so graphs are not very informative (over):

The plots

g
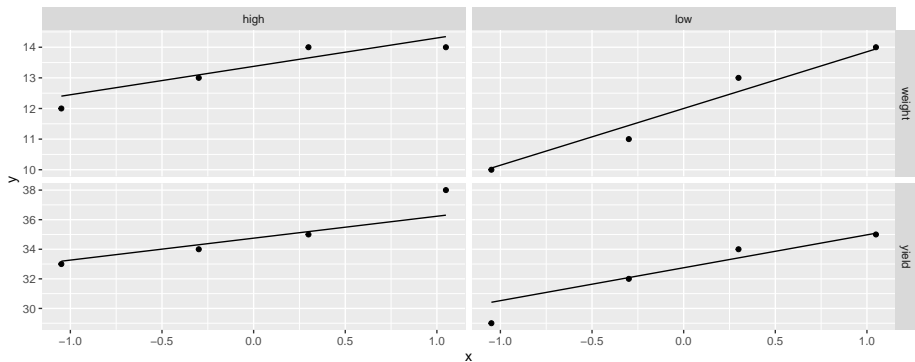
Box M test

- Make sure package MVTests loaded first.
- Then pull out the numeric columns from dataframe, and (separately) the grouping column:

```
hilo %>% select(yield, weight) -> numeric_values
summary(BoxM(numeric_values, hilo$fertilizer))
```

```
##          Box's M Test
##
## Chi-Squared Value = 1.002964 , df = 3  and p-value: 0.801
```

No problem at all with unequal spreads.

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://ritsokiguess.site/datafiles/peanuts.txt"
peanuts.orig <- read_delim(u, " ")
```

```
## Rows: 12 Columns: 6
## -- Column specification ------------------------------------
## Delimiter: " "
## dbl (6): obs, location, variety, y, smk, w
##
## i Use `spec()` to retrieve the full column specification fc
## i Specify the column types or set `show_col_types = FALSE`
```

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.orig %>%
  mutate(
    location = factor(location),
    variety = factor(variety)
  ) -> peanuts
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   Pr(>F)
## location          1   0.89348  11.1843      3      4 0.020502 *
## variety           2   1.70911   9.7924      6     10 0.001056 **
## location:variety  2   1.29086   3.0339      6     10 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.

Normal quantile plots

Setup for normal quantile plots

```
peanuts %>%
  pivot_longer(y:w, names_to = "xname",
               values_to = "xvals") -> d
```
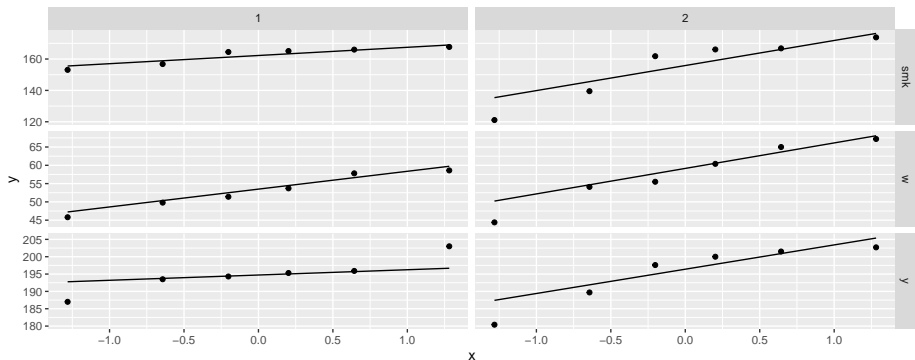
Make the plots (save to show on next page)

```
ggplot(d, aes(sample = xvals)) + stat_qq() +
  stat_qq_line() +
  facet_grid(xname ~ location, scales = "free") -> g1

ggplot(d, aes(sample = xvals)) + stat_qq() +
  stat_qq_line() +
  facet_grid(xname ~ variety, scales = "free") -> g2
```
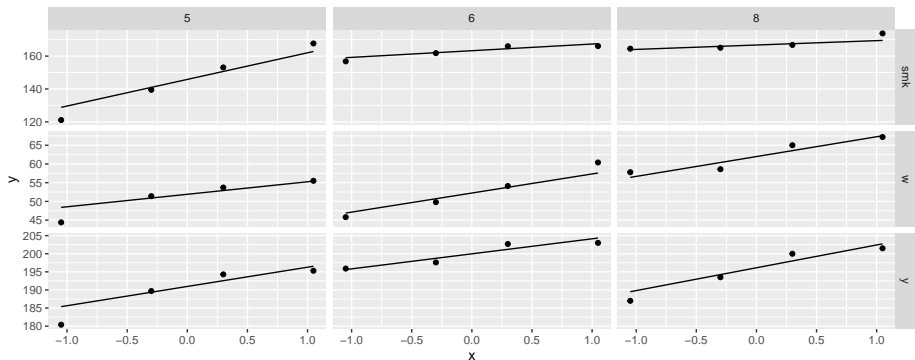
The plot for location

g1

The plot for variety

g2

Comments

- this time there are only six observations per location and four per variety, so normality is still difficult to be confident about

- y at location 1 seems to be the worst for normality (long tails / outliers), and maybe y at location 2 is skewed left, but the others are not bad

- there is some evidence of unequal spread (slopes of lines), but is it bad enough to worry about? (Box M-test, over).

Box's M tests

- One for location, one for variety:

```
peanuts %>% select(y:w) -> numbers
summary(BoxM(numbers, peanuts$location))
```

```
##          Box's M Test
##
## Chi-Squared Value = 12.47797 , df = 6  and p-value: 0.0521
```

```
summary(BoxM(numbers, peanuts$variety))
```

```
##          Box's M Test
##
## Chi-Squared Value = 10.56304 , df = 12  and p-value: 0.567
```

- Neither of these P-values is low enough to worry about. (Remember, the P-value here has to be *really* small to indicate a problem.)