Biostatistics I: Hypothesis testing

Continuous data: M-sample tests

Eleni-Rosalina Andrinopoulou

Department of Biostatistics, Erasmus Medical Center

e.andrinopoulou@erasmusmc.nl

@erandrinopoulou

Erasmus MC
University Medical Center Rotterdam

In this Section

- Analysis of variance (ANOVA) / F-test
- M-sample Kruskal-Wallis test
- Examples

Analysis of variance / F-test: Theory

Assumptions

- The variables are continuous
- The samples are independent
- Homogeneity of variance
- The data are normally distributed

Analysis of variance / F-test: Theory

Scenario

Is the mean BMI of the students different in groups 1, 2 and 3?

Connection with linear regression

$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$, where x_{1i} and x_{2i} indicate whether the subject is in group 2 or 3

$H_0 : y_i = \beta_0 \Rightarrow H_0 : \beta_1 = \beta_2 = 0$

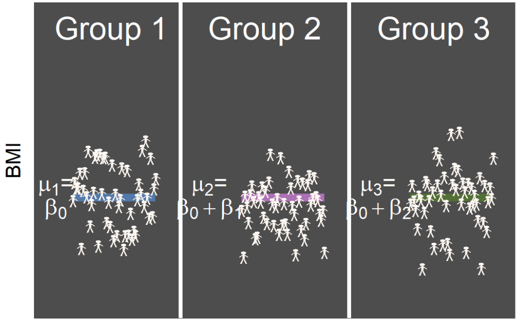
- ANOVA: each category's mean is compared to a grand mean
- Regression: dummy coded variables

Alternatively

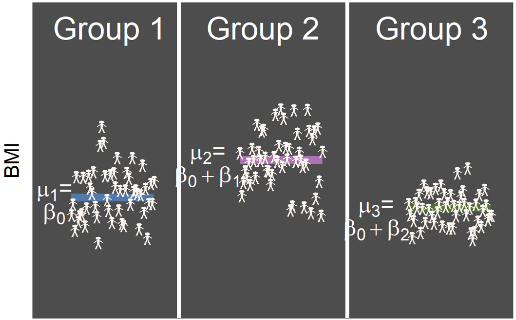
$H_0 : \mu_1 = \mu_2 = \mu_3$
$H_1 : \mu_1 \neq \mu_2$ or $\mu_2 \neq \mu_3$ or $\mu_1 \neq \mu_3$

Analysis of variance / F-test: Theory

Null hypothesis



Alternative hypothesis



Analysis of variance / F-test: Theory

Notes...

- It generalizes the t-test to more than two groups
  - Multiple two-sample t-tests $\Rightarrow$ increased Type I error
- ANOVA tests if at least two groups were different from each other (it won't test which groups were different)
- A two way ANOVA will allow for 2 independent variables - multivariable regression models

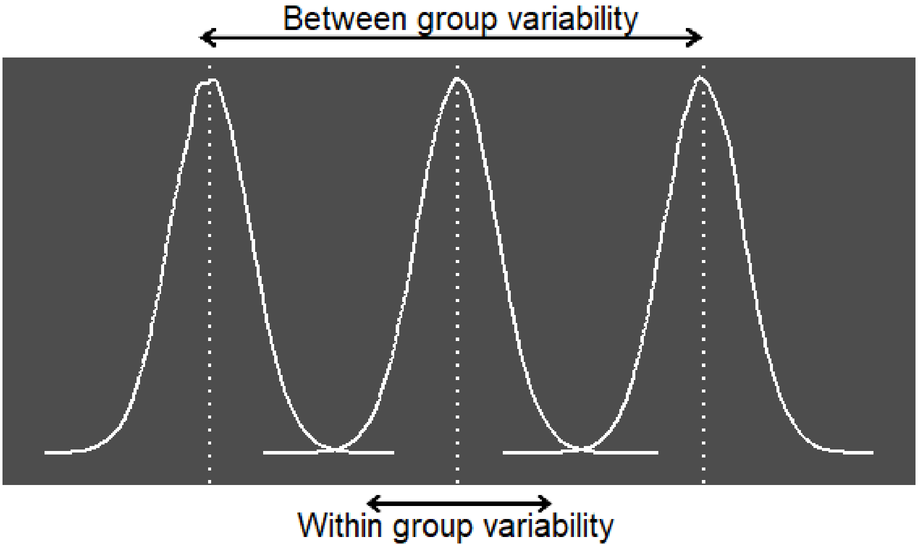
Analysis of variance / F-test: Theory

Test statistic

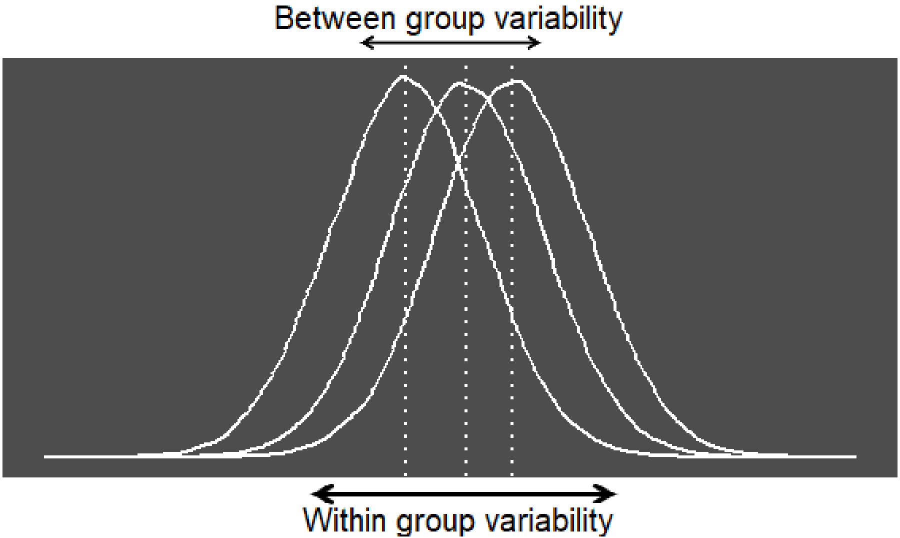
- Within group variation
- Between group variation
- F statistic: the ratio of between group variation to within group variation

Under the null hypothesis that the group means are the same $\Rightarrow$ the between-group variability will be similar to the within-group variability

Analysis of variance / F-test: Theory



Analysis of variance / F-test: Theory



Analysis of variance / F-test: Theory

Sum of squares due to differences between groups:

$SS_{between} = \sum_j n_j(\bar{x}_j - \bar{x})^2$

Sum of squares due to variability within groups:

$SS_{within} = \sum_j \sum_i (x_{ij} - \bar{x}_j)^2$

Total sum of squares:

$SS_{total} = SS_{between} + SS_{within} = \sum_j \sum_i (x_{ij} - \bar{x})^2 = \sum_j n_j(\bar{x}_j - \bar{x})^2 + \sum_j \sum_i (x_{ij} - \bar{x}_j)^2$

Mean squares:

$MS_{between} = \frac{SS_{between}}{m-1}$

$MS_{within} = \frac{SS_{within}}{n-m}$

Test statistic: $F = \frac{MS_{between}}{MS_{within}}$

Analysis of variance / F-test: Theory

Sampling distribution

- F-distribution with $df_1 = m - 1$ and $df_2 = n - m$
- Critical value and p-value

Type I error

- Normally $\alpha = 0.05$

Draw conclusions

- Compare test statistic (F) with the critical value or the p-value with α

Analysis of variance / F-test: Application

Scenario

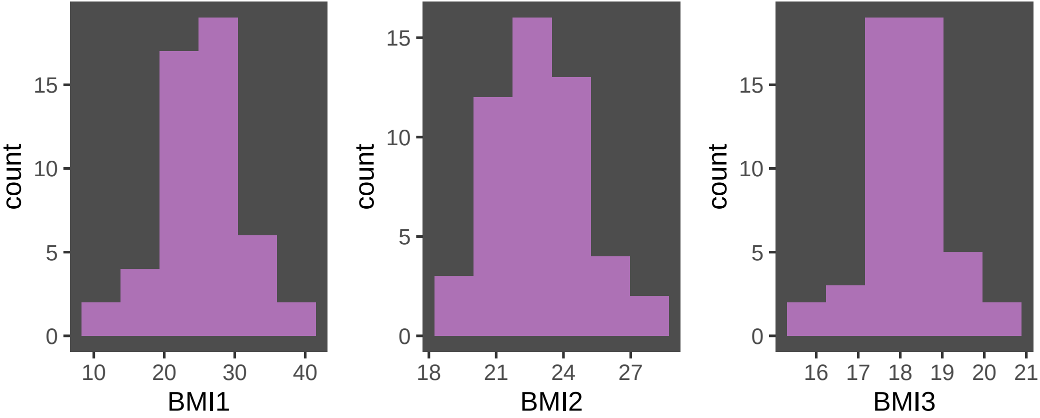
Is the mean BMI of the students different in groups 1, 2 and 3?

Hypothesis

$H_0 : \mu_1 = \mu_2 = \mu_3$
$H_1 : \mu_1 \neq \mu_2$ or $\mu_2 \neq \mu_3$ or $\mu_1 \neq \mu_3$

Analysis of variance / F-test: Application

Collect and visualize data



Analysis of variance / F-test: Application

Test statistic

Let's assume that:

- Sample mean of group 1: $\bar{x}_1 = 25.1633$
- Sample mean of group 2: $\bar{x}_2 = 22.8823$
- Sample mean of group 2: $\bar{x}_3 = 18.1985$
- Number of subjects in the groups: $n_1 = n_2 = n_3 = 50$

$SS_{between} = \sum_j n_j(\bar{x}_j - \bar{x})^2 = 1260.8$ and $SS_{within} = \sum_j \sum_i (x_{ij} - \bar{x}_j)^2 = 1721.1$

Total sum of squares:
$SS_{total} = SS_{between} + SS_{within} = 2981.9$

Mean squares:
$MS_{between} = \frac{SS_{between}}{m-1} = \frac{1260.8}{3-1} = 630.4$ and $MS_{within} = \frac{SS_{within}}{n-m} = \frac{1721.1}{150-3} = 11.7$

$F = \frac{MS_{between}}{MS_{within}} = \frac{630.4}{11.7} = 53.8$

Analysis of variance / F-test: Application

Degrees of freedom

$df_1 = m - 1 = 3 - 1 = 2$ and $df_2 = n - m = 150 - 3 = 147$

Type I error

$\alpha = 0.05$

Critical values

Using R we get the critical value from the F-distribution:
critical value$_\alpha$ = critical value$_{0.05}$

```r
qf(p = 0.05, df1 = 2, df2 = 147, lower.tail = FALSE)
```

```
[1] 3.057621
```

Analysis of variance / F-test: Application

Draw conclusions

We reject the H_0 if:

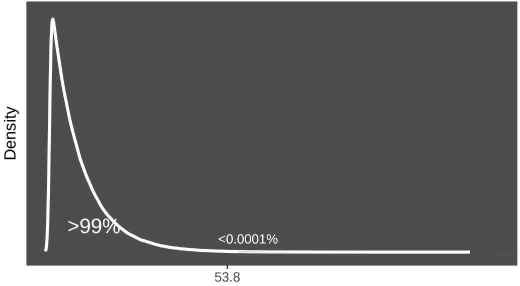
- $F >$ critical value$_\alpha$

We have $53.8 > 3.06 \Rightarrow$ we reject the H_0

Using R we obtain the p-value from the F-distribution:

```
pf(q = 53.8, df1 = 2, df2 = 147,
   lower.tail = FALSE)
```

```
[1] 2.932458e-18
```



M-sample Kruskal-Wallis test: Theory

Assumptions

- Within and between groups observations are independent of one another

M-sample Kruskal-Wallis test: Theory

Scenario

Is the distribution of the score values of the students different in groups 1, 2 and 3?

Connection with linear regression

$rank(y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$, where x_{1i} and x_{2i} indicate whether the subject is in group 2 or 3

$H_0 : y_i = \beta_0 \Rightarrow H_0 : \beta_1 = \beta_2 = 0$

Alternatively

H_0 : the samples (groups) are from identical populations
H_1 : at least one of the samples (groups) comes from a different population than the others

M-sample Kruskal-Wallis test: Theory

Test statistic

- Rank all data from all groups together (assign any tied values the average of the ranks they would have received had they not been tied)
- Calculate:
  - if we do not have ties: $H = \frac{12}{n(n+1)} \sum_j n_j \bar{r}_j^2 - 3(n+1)$, where
    - n is the total number of observations in all groups
    - n_j is the number of observations in group j
    - $\bar{r}_j = \frac{\sum_{i=1}^{n_j} r_{ij}}{n_j}$ where r_{ij} is the rank of observation i from group j
  - if we have ties: $H = \frac{H}{1 - \frac{\sum_j (T_j^3 - T_j)}{(n^3 - n)}}$ where T_j is the number of tied values in group j

M-sample Kruskal-Wallis test: Theory

Sampling distribution

- χ^2-distribution with $df = m - 1$, where m is the total number of groups
- Critical value and p-value

For small sample size, we can use the exact distribution

Type I error

- Normally $\alpha = 0.05$

Draw conclusions

- Compare test statistic (H) with the critical values or the p-value with α

M-sample Kruskal-Wallis test: Application

Scenario

Is the distribution of the score values of the students different in groups 1, 2 and 3?

Hypothesis

H_0 : the samples (groups) are from identical populations
H_1 : at least one of the samples (groups) comes from a different population than the others

M-sample Kruskal-Wallis test: Application

Collect and visualize data

Groups	values	rank	mean rank per group
1	8.88	5	9.17
1	9.54	6	9.17
1	13.12	13	9.17
1	10.14	8	9.17
1	10.26	9	9.17
1	13.43	14	9.17
2	9.92	7	3.25
2	6.47	1	3.25
2	7.63	2	3.25
2	8.11	3	3.25
3	15.90	15	10.40
3	12.44	11	10.40
3	12.60	12	10.40
3	11.44	10	10.40
3	8.78	4	10.40

M-sample Kruskal-Wallis test: Application

Hypothesis

H_0 : the samples (groups) are from identical populations
H_1 : at least one of the samples (groups) comes from a different population than the others

Test statistic

No ties:

$$H = \frac{12}{n(n+1)} \sum_j n_j \bar{r}_j^2 - 3(n+1) = \frac{12}{15(15+1)}(9.17^2 * 6 + 3.25^2 * 4 + 10.40^2 * 5) - 3(15+1) = 6.38$$

Degrees of freedom

$$df = number\ of\ groups - 1 = 3 - 1 = 2$$

Type I error

$$\alpha = 0.05$$

M-sample Kruskal-Wallis test: Application

Critical values

Using `R` we get the critical values from the exact distribution:
critical value$_{\alpha}$ = critical value$_{0.05}$

```
qchisq(p = 0.05, df = 2, lower.tail = FALSE)
```

```
[1] 5.991465
```

M-sample Kruskal-Wallis test: Application

Draw conclusions

We reject the H_0 if:

- $H >$ critical value$_\alpha$

We have $6.38 > 5.99 \Rightarrow$ we reject the H_0

Using R we obtain the p-value from the exact distribution:

```
pchisq(q = 6.38, df = 2,
       lower.tail = FALSE)
```

```
[1] 0.04117187
```