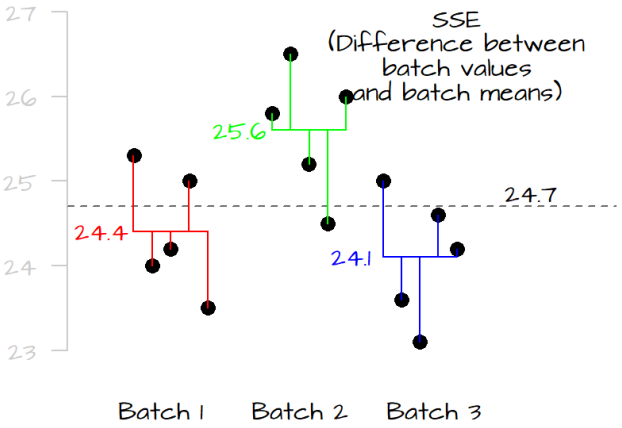


ΤΜΗΜΑ ΜΟΡΙΑΚΗΣ ΒΙΟΛΟΓΙΑΣ & ΓΕΝΕΤΙΚΗΣ ΔΠΘ

ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ | DEMOCRITUS UNIVERSITY OF THRACE

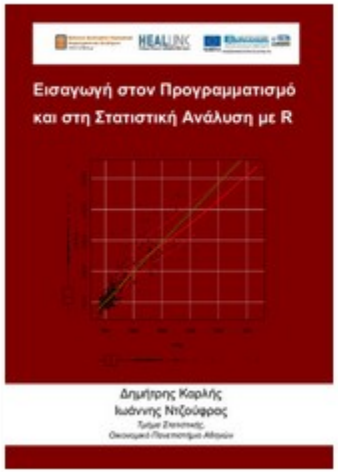
Βιοστατιστική

Εφαρμογές και Παραδείγματα Ανάλυσης Διακύμανσης

Αριστοτέλης Παπαγεωργίου, Τμ. ΜΒΓ ΔΠΘ, apapage@mbg.duth.gr

Πηγές υλικού

- https://mgimond.github.io/Stats-in-R/ANOVA.html
- https://statsandr.com/blog/anova-in-r/
- https://www.edanz.com/blog/anova-explained
- Βιβλίο Χ. Νικολάου



Εισαγωγή στον προγραμματισμό και στη στατιστική ανάλυση με R

Author(s) :*Ntzoufras, Ioannis; Karlis, Dimitrios*
Type : Undergraduate textbook

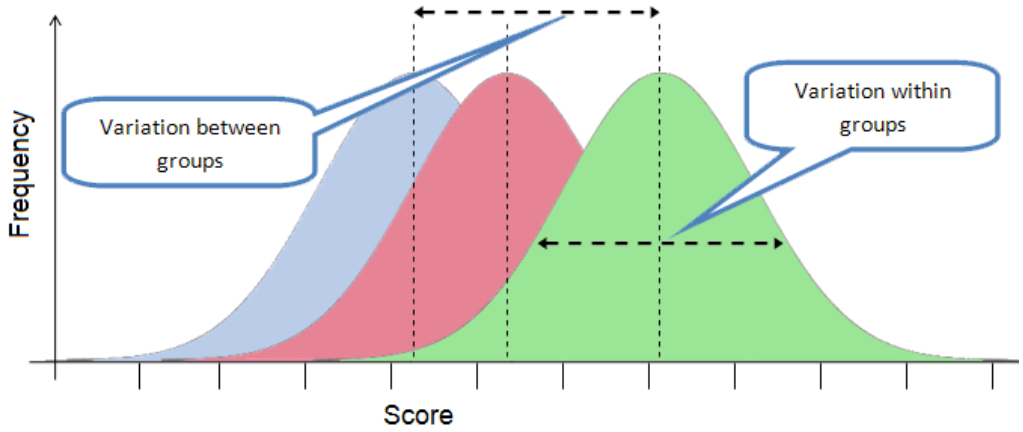
Μεθοδολογία της έρευνας στις επιστήμες υγείας

Author(s) :*Lagoumintzis, Georgios; Vlachopoulos, Georgios; Koutsogiannis, Konstantinos*
Type : Undergraduate textbook

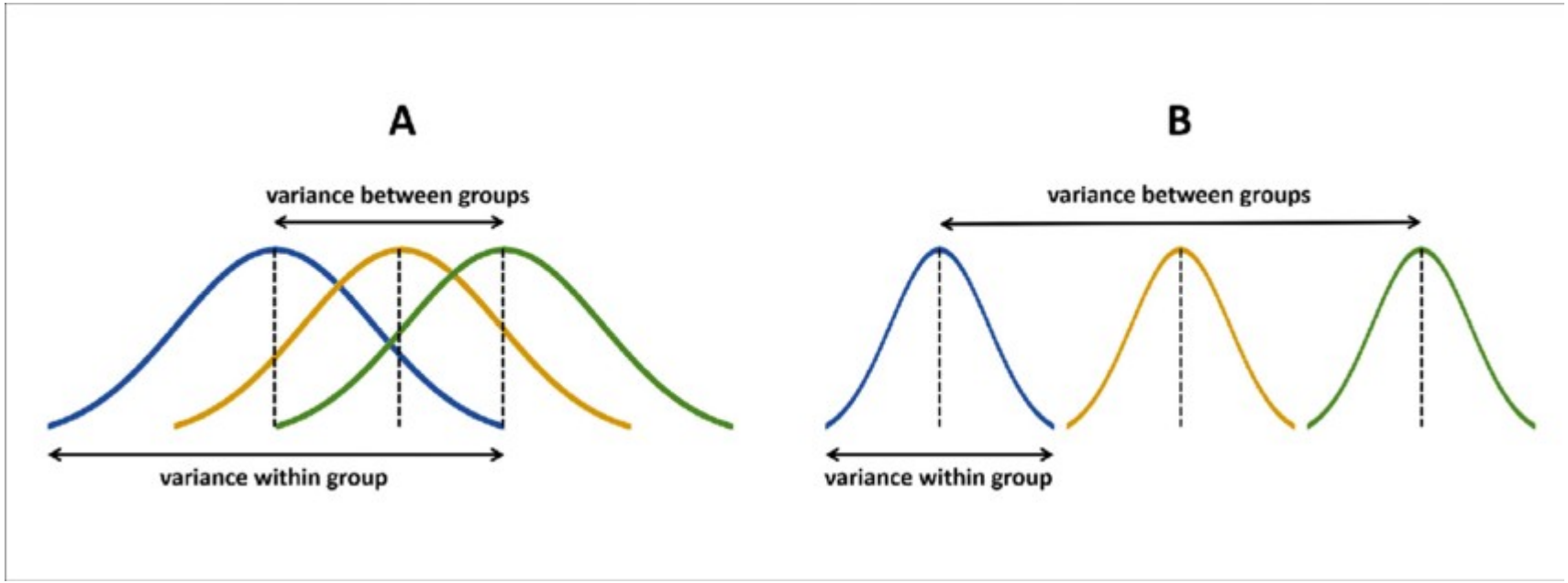
- https://repository.kallipos.gr/

Ανάλυση διακύμανσης

- Η ANOVA χρησιμοποιείται για τη σύγκριση διαφορών μέσων μεταξύ περισσότερων από δύο ομάδων
  - Αυτό το κάνει εξετάζοντας τη διακύμανση στα δεδομένα και πού βρίσκεται αυτή η διακύμανση
  - Η ANOVA συγκρίνει την ποσότητα της διακύμανσης μεταξύ των ομάδων με την ποσότητα της διακύμανσης εντός των ομάδων



Σε ποια περίπτωση διαφέρουν οι μέσοι;



Το ερώτημα της ANOVA είναι αν η διακύμανση που παρατηρούμε ανάμεσα στις τιμές ενός δείγματος, οφείλονται στην ύπαρξη διαφορετικών ομάδων, ή στην τυχαιότητα.

Μαθηματική απεικόνιση της ANOVA

- $x_{ij} = \mu_i + \varepsilon_{ij}$
- Για **i** ομάδες και **j** παρατηρήσεις, **x** είναι τα επιμέρους σημεία δεδομένων, **ε** είναι η απόκλιση από τον μέσο της ομάδας για κάθε σημείο δεδομένων και **μ** είναι οι μέσοι όροι του πληθυσμού της κάθε ομάδας
- Έτσι, κάθε σημείο δεδομένων ($\mathbf{x}_{ij}$) είναι το άθροισμα του μέσου όρου της ομάδας του συν την απόκλιση από αυτόν

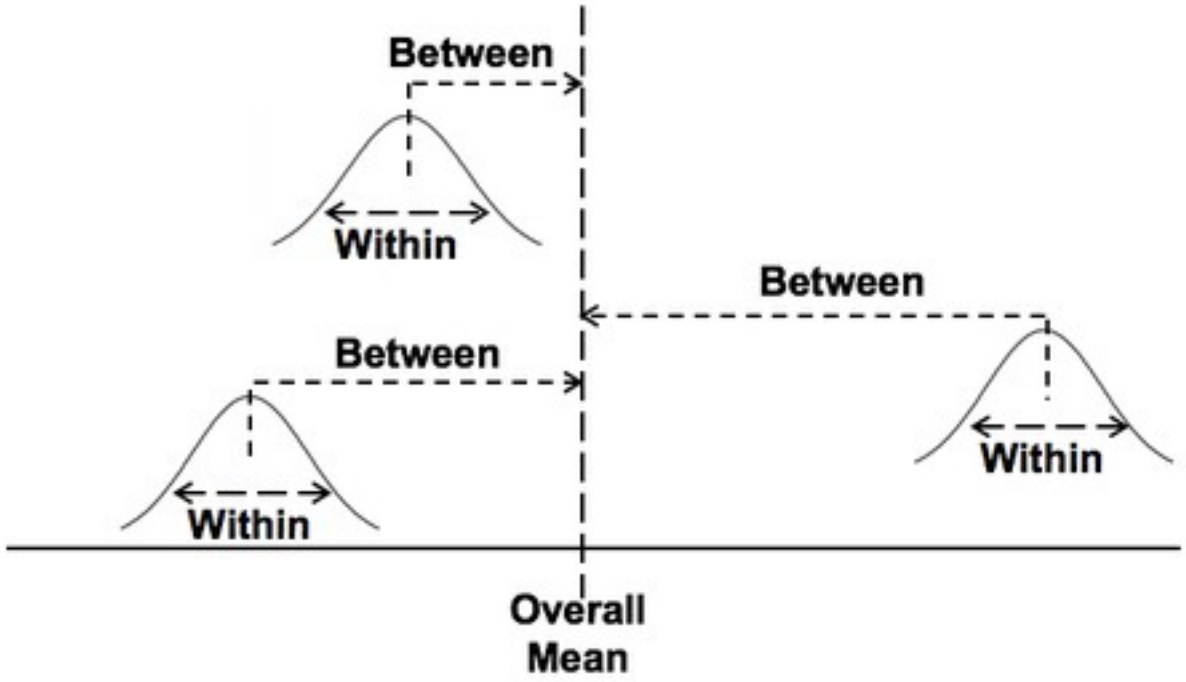
Έλεγχος υπόθεσης ANOVA

- Σε ένα σύνολο δεδομένων που χωρίζεται σε ομάδες, κάνουμε τον έλεγχο της υπόθεσης ως εξής:
  - Ho: οι μέσοι των ομάδων στον πληθυσμό είναι ίσοι
  - H1: τουλάχιστον για μία ομάδα, ο μέσος όρος διαφέρει
- Όπως και σε άλλους στατιστικούς ελέγχους, υπολογίζουμε σε ένα δείγμα ένα στατιστικό (εδώ ο **λόγος F**) και την πιθανότητα να λάβουμε το στατιστικό αυτό τυχαία, όταν ισχύει η H0 (**p-value**)
  - Όταν p-value <= 0,05 θεωρούμε ότι απορρίπτεται η H0 και ισχύει η H1

Υπολογίζουμε το F

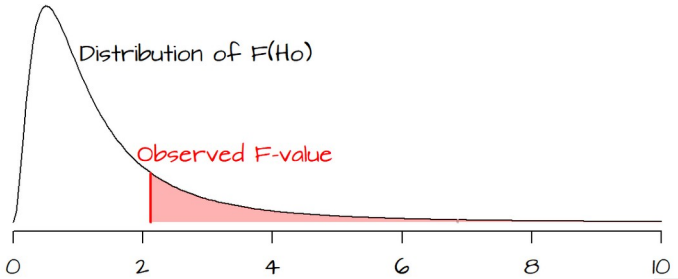
- Διακύμανση ανάμεσα στις ομάδες
  - Άθροισμα τετραγώνων των αποκλίσεων των μέσων των ομάδων από τον συνολικό μέσο του δείγματος
  - $\text{Between SS} = n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + n_3(\bar{x}_3 - \bar{x})^2$
  - Διαιρούμε το BSS με τους βαθμούς ελευθερίας (n-1, όπου n ο αριθμός των ομάδων) και παίρνουμε τη μέση διακύμανση ανάμεσα στις ομάδες
- Διακύμανση μέσα στις ομάδες
  - Άθροισμα τετραγώνων των αποκλίσεων των παρατηρήσεων από τον μέσο της ομάδας στην οποία ανήκουν / βαθμούς ελευθερίας για κάθε ομάδα
  - Προσθέτοντας τις διακυμάνσεις των ομάδων βρίσκουμε τη μέση διακύμανση μέσα στις ομάδες
  - $SS_R = s^2_{group1}(n_{group1} - 1) + s^2_{group2}(n_{group2} - 1) + s^2_{group3}(n_{group3} - 1)$

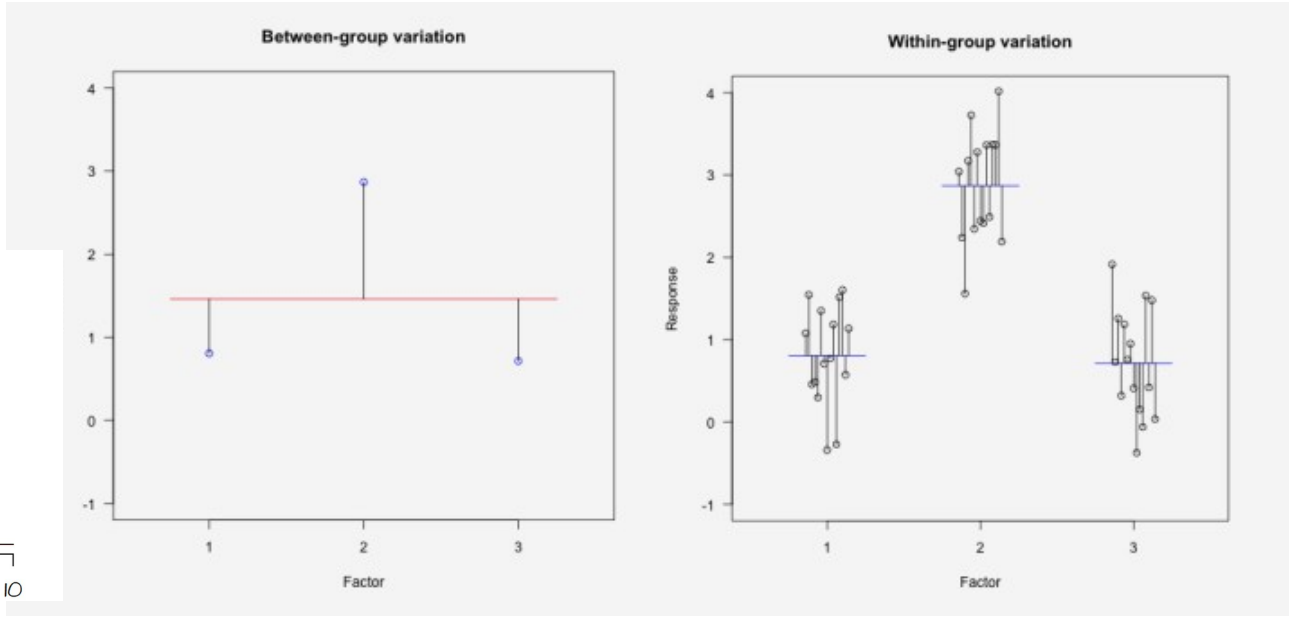
Υπολογίζουμε το F



Υπολογίζουμε το F

- Ο λόγος F υπολογίζεται από τη διαίρεση του μέσου αθροίσματος τετραγώνων ανάμεσα στις ομάδες προς το μέσο άθροισμα τετραγώνων μέσα στις ομάδες
- $$\frac{\text{Mean Between-group SS}}{\text{Mean Within-group SS}}$$





Αποτέλεσμα ANOVA

	Df	SS	Mean SS	F ratio	P(>F)
Between-group	2	44.475	22.2373	52.022	4.306e-12***
Within-group	42	17.953	0.4275		

Δεχόμαστε ή απορρίπτουμε την Ho;

Παράδειγμα ANOVA με R

Θα χρειαστούμε το πακέτο **palmerpenguins**

```
> library(palmerpenguins)
> library(tidyverse)
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✔ dplyr     1.1.4     ✔ readr     2.1.5
✔ forcats   1.0.0     ✔ stringr   1.5.1
✔ ggplot2   3.5.0     ✔ tibble    3.2.1
✔ lubridate 1.9.3     ✔ tidyr     1.3.1
✔ purrr     1.0.2
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
> dat <- penguins %>%
  select(species, flipper_length_mm)
> summary(dat)
      species    flipper_length_mm
 Adelie   :152   Min.   :172.0
 Chinstrap: 68   1st Qu.:190.0
 Gentoo   :124   Median :197.0
                 Mean   :200.9
                 3rd Qu.:213.0
                 Max.   :231.0
                 NA's   :2
>
```

344 άτομα από 3 είδη πιγκουίνων
8 μεταβλητές
Κρατάμε το μήκος του πτερυγίου για το παράδειγμα (*flipper_length_mm*)

```
ggplot(dat) +
  aes(x = species, y = flipper_length_mm, color = species
) +
  geom_jitter() +
  theme(legend.position = "none")
Warning message:
Removed 2 rows containing missing values or values
outside the scale range (`geom_point()`).
>
```

Ερώτηση:

Διαφέρουν τα τρία είδη πιγκουίνων ως προς το μήκος του πτερυγίου (*flipper_length_mm*);

Υποθέσεις της ANOVA

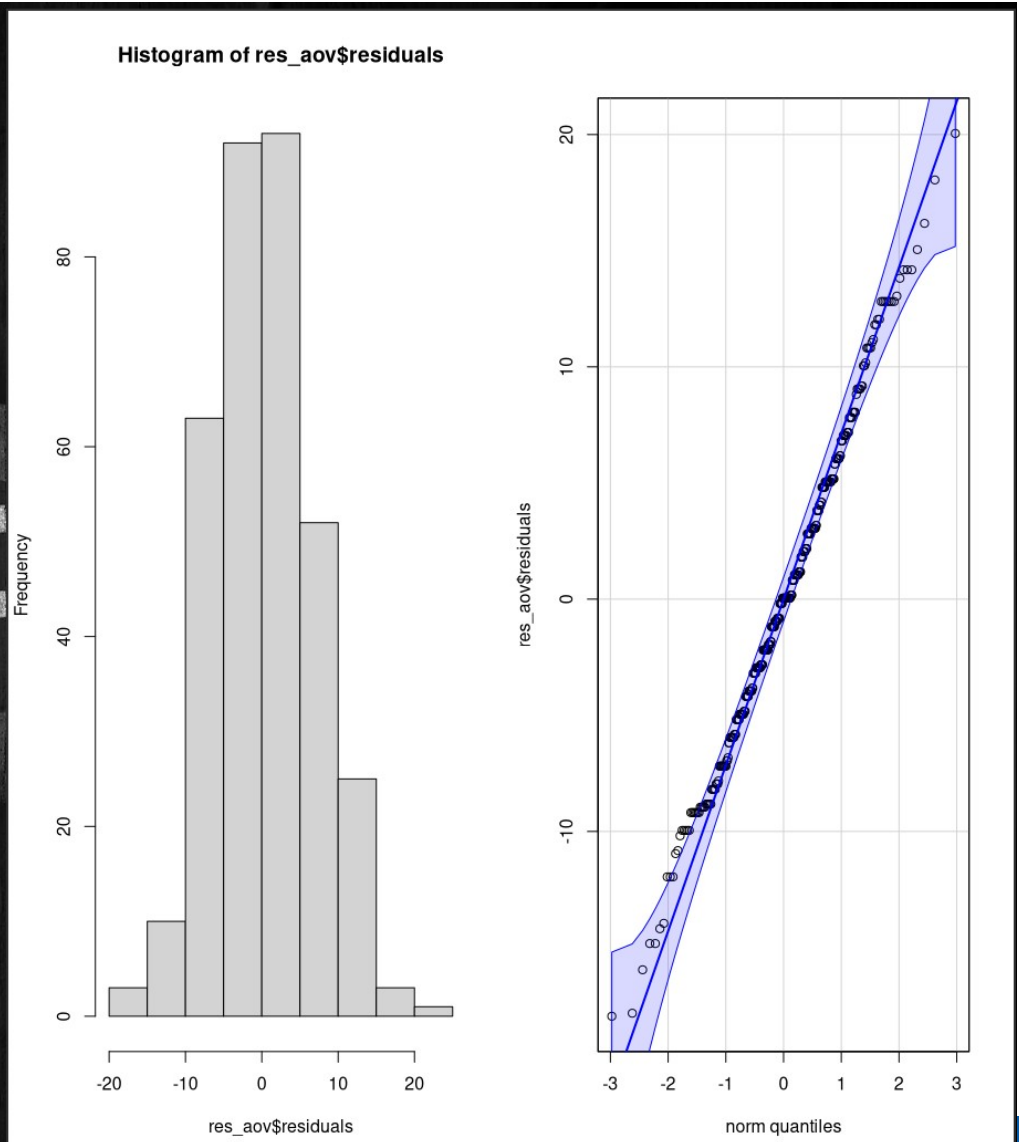
- H_0: $\mu_{Adelie}=\mu_{Chinstrap}=\mu_{Gentoo}$
  - τα τρία είδη είναι ίσα για το μήκος πτερυγίου
- H_1: τουλάχιστον ένας μέσος είναι διαφορετικός
  - Τουλάχιστον ένα είδος διαφέρει από τα άλλα δύο για το μήκος πτερυγίου

- Η εντολή της ANOVA στην R είναι:
  - *res_aov <- aov(flipper_length_mm ~ species, data = dat)*
- Αποθηκεύουμε το αποτέλεσμα της ANOVA στο αντικείμενο **res_aov**

Έλεγχος κανονικότητας

- Μία από τις προϋποθέσεις για τη χρήση της ANOVA είναι οι τιμές του δείγματος να ακολουθούν την κανονική κατανομή
  - Όταν το δείγμα είναι μεγάλο (n≥30) τότε δεν υπάρχει λόγος να κάνουμε τον έλεγχο αυτό.
    - Στο παράδειγμα αυτό,το δείγμα είναι μεγάλο, αλλά θα κάνουμε τον έλεγχο για εξάσκηση
- Για να ελέγξουμε την κανονικότητα στις τιμές του δείγματος, αυτό πρέπει να γίνει χωριστά για κάθε ομάδα του δείγματος (εδώ τα είδη των πιγκουίνων)
  - Μπορούμε να κάνουμε έλεγχο κανονικότητας στις υπολειμματικές τιμές της ANOVA για όλο το δείγμα

```
> par(mfrow = c(1, 2)) # combine plots

# histogram
hist(res_aov$residuals)

# QQ-plot
qqPlot(res_aov$residuals, id = FALSE)
> 
```

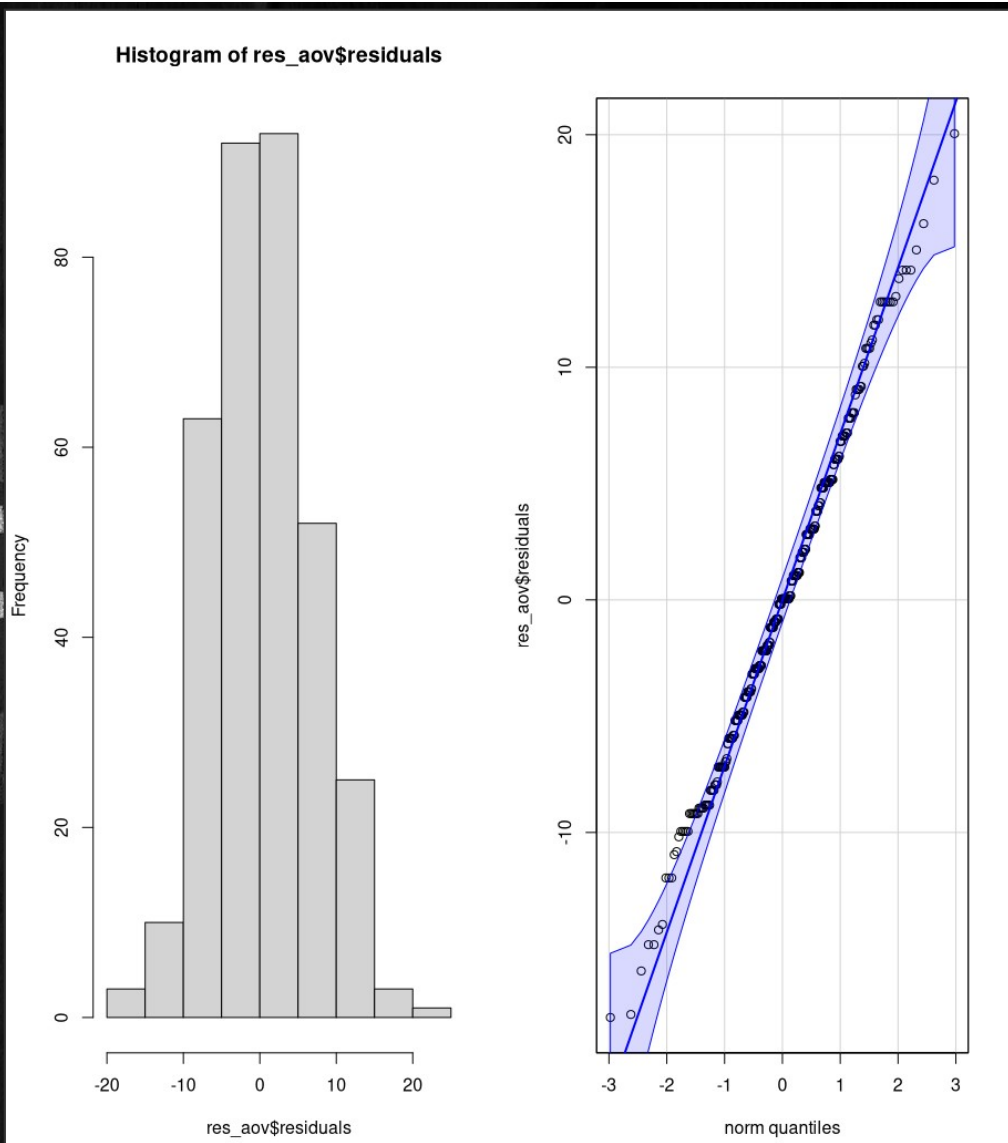


Οπτικός έλεγχος κανονικότητας στις υπολειμματικές τιμές της ANOVA

```
> par(mfrow = c(1, 2)) # combine plots

# histogram
hist(res_aov$residuals)

# QQ-plot
qqPlot(res_aov$residuals, id = FALSE)
> shapiro.test(res_aov$residuals)

        Shapiro-Wilk normality test

data:  res_aov$residuals
W = 0.99452, p-value = 0.2609

>
```

Στατιστικός έλεγχος κανονικότητας στις υπολειμματικές τιμές της ANOVA με τον έλεγχο Shapiro-Wilk

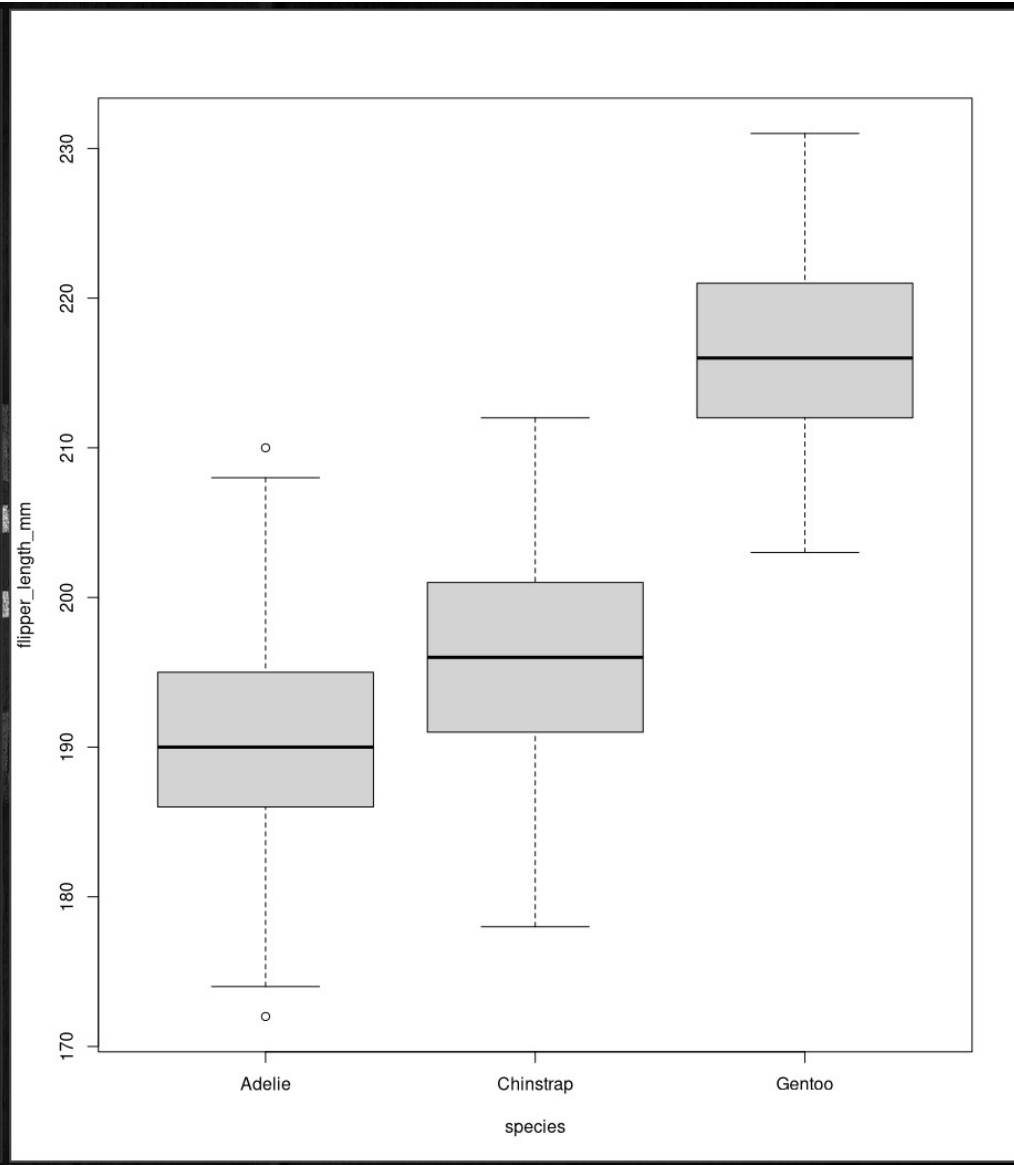
p-value = 0.2609
Τι συμπέρασμα βγάζουμε;



Έλεγχος ομοιογένειας - ισότητας διακύμανσης

- Μια ακόμη προϋπόθεση για τη χρήση της ANOVA είναι οι διακυμάνσεις των ομάδων στο δείγμα να είναι ίσες
  - Οπτικός έλεγχος με boxplot
  - Στατιστικός έλεγχος με test Levene

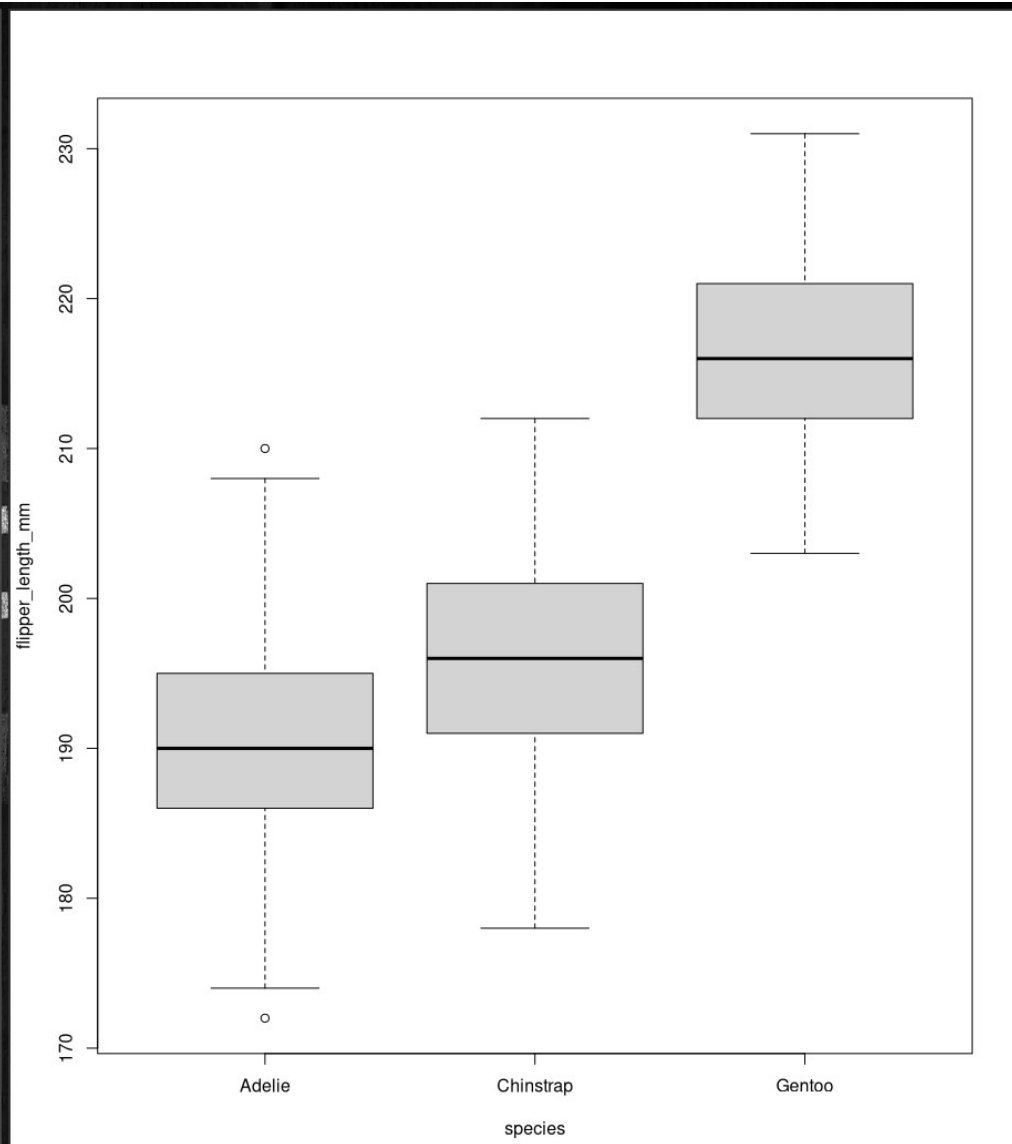
```
> # Boxplot
boxplot(flipper_length_mm ~ species, data = dat)
>
```



Οπτικός έλεγχος ομοιογένειας: οι διακυμάνσεις για κάθε είδος φαίνονται παρόμοιες

```
> # Boxplot
boxplot(flipper_length_mm ~ species, data = dat)
> leveneTest(flipper_length_mm ~ species,  data = dat)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group   2  0.3306 0.7188
      339
>
```

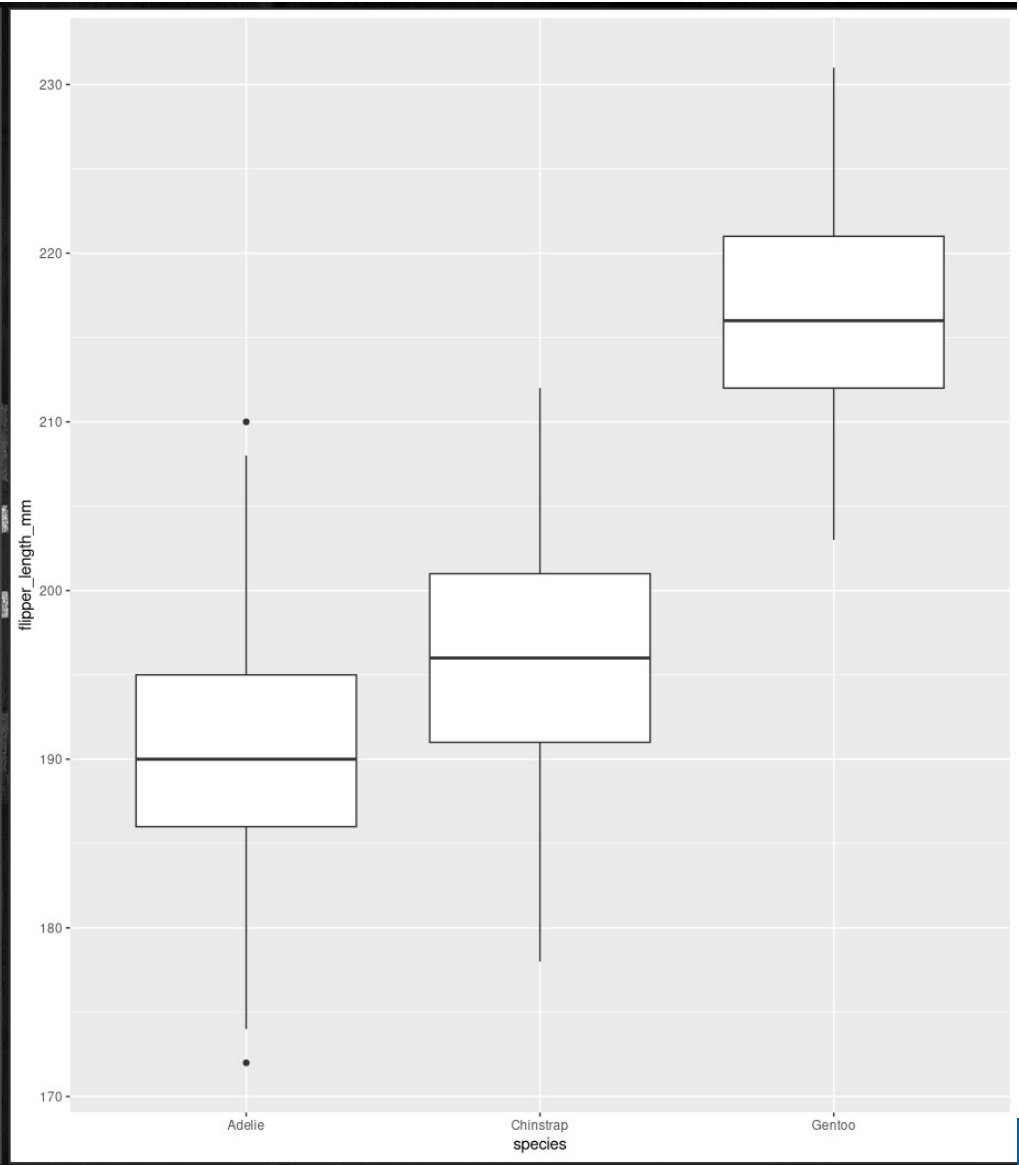
Στατιστικός έλεγχος ομοιογένειας με τον έλεγχο Levene



p-value = 0.7188
Τι συμπέρασμα βγάζουμε;

```
> ggplot(dat) +
  aes(x = species, y = flipper_length_mm) +
  geom_boxplot()
Warning message:
Removed 2 rows containing non-finite outside the scale range
(`stat_boxplot()`).
> aggregate(flipper_length_mm ~ species, data = dat, function(x) round
(c(mean = mean(x), sd = sd(x)), 2))
    species flipper_length_mm.mean flipper_length_mm.sd
1    Adelie                189.95                6.54
2 Chinstrap                195.82                7.13
3    Gentoo                217.19                6.48
>
```

Υπάρχουν στατιστικά σημαντικές διαφορές ανάμεσα στους μέσους των τριών ειδών;



ANOVA με oneway.test

```
> oneway.test(flipper_length_mm ~ species, data = dat, var.equal = TRUE)

        One-way analysis of means

data:  flipper_length_mm and species
F = 594.8, num df = 2, denom df = 339, p-value < 2.2e-16

>
```

Τι συμπέρασμα βγάζουμε;

ANOVA με aov()

```
> res_aov <- aov(flipper_length_mm ~ species, data = dat)
summary(res_aov)
             Df Sum Sq Mean Sq F value Pr(>F)
species       2  52473   26237   594.8 <2e-16 ***
Residuals   339  14953      44
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
2 observations deleted due to missingness
>
```

Τι συμπέρασμα βγάζουμε;

ANOVA με aov()

```
> library(report)
> report(res_aov)
The ANOVA (formula: flipper_length_mm ~ species) suggests that:

  - The main effect of species is statistically significant and large (F(2, 339)
= 594.80, p < .001; Eta2 = 0.78, 95% CI [0.75, 1.00])

Effect sizes were labelled following Field's (2013) recommendations.
>
```

Welch ANOVA

- Τι θα συνέβαινε αν οι διακυμάνσεις δεν ήταν όμοιες;
  - Στην περίπτωση αυτή θα μπορούσαμε να εκτελούσαμε την ANOVA με oneway.test και var.equal = FALSE

```
> oneway.test(flipper_length_mm ~ species, data = dat, var.equal = FALSE)

        One-way analysis of means (not assuming equal variances)

data:  flipper_length_mm and species
F = 614.01, num df = 2.00, denom df = 172.76, p-value < 2.2e-16
```

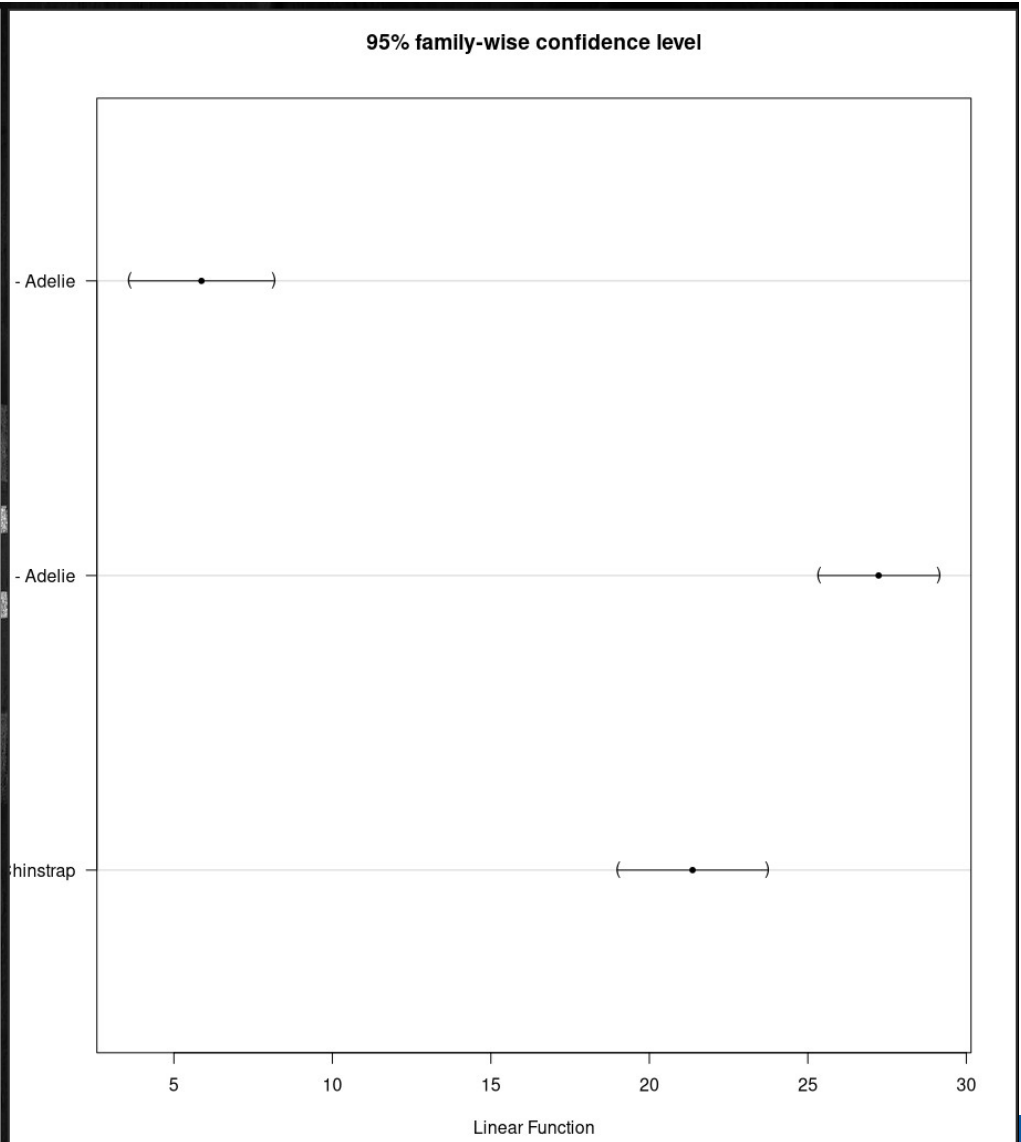
Μετά την ANOVA - Post-hoc tests

- Η ANOVA δεν κάνει συγκρίσεις ανά δύο ομάδες, ούτε εξηγεί ποια ομάδα διαφέρει από ποια
  - Το μόνο που γνωρίζουμε είναι ότι τουλάχιστον ένα είδους πιγκουίνου έχει διαφορετικό μέσο όρο μήκους πτερυγίου από τα άλλα
- Για τον λόγο αυτό μπορούμε να κάνουμε μια σειρά συγκρίσεων που λέγονται Post-hoc tests
  - Στο παράδειγμά μας θα κάνουμε το τεστ Tuckey HSD με το αντικείμενο res_aov από την ANOVA

```
> post_test <- glht(res_aov, linfct = mcp(species = "Tukey"))
> summary(post_test)

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = flipper_length_mm ~ species, data = dat)

Linear Hypotheses:
                       Estimate Std. Error t value Pr(>|t|)
Chinstrap - Adelie == 0   5.8699     0.9699   6.052   <1e-08 ***
Gentoo - Adelie == 0     27.2333     0.8067  33.760   <1e-08 ***
Gentoo - Chinstrap == 0  21.3635     1.0036  21.286   <1e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)

>
```

Τι συμπέρασμα βγάζουμε;

```
The following object is masked from ‘package:dplyr’:

    select


Attaching package: ‘TH.data’

The following object is masked from ‘package:MASS’:

    geyser

> post_test <- glht(res_aov, linfct = mcp(species = "Tukey"))
> summary(post_test)

        Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = flipper_length_mm ~ species, data = dat)

Linear Hypotheses:
                     Estimate Std. Error t value Pr(>|t|)
Chinstrap - Adelie == 0   5.8699     0.9699   6.052   <1e-08 ***
Gentoo - Adelie == 0     27.2333     0.8067  33.760   <1e-08 ***
Gentoo - Chinstrap == 0  21.3635     1.0036  21.286   <1e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)

> plot(post_test)
>
```

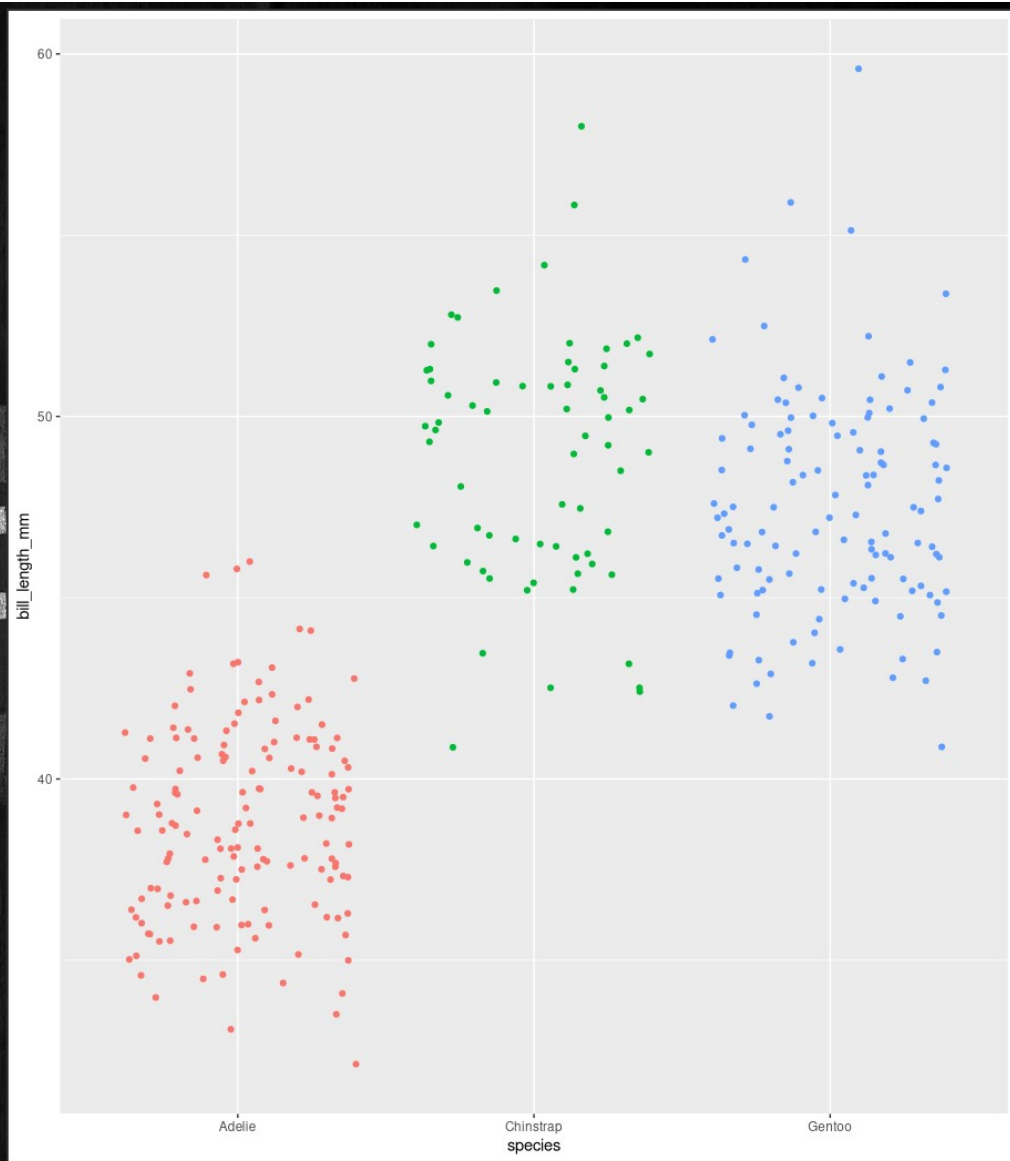


```
> str(penguins)
tibble [344 × 8] (S3: tbl_df/tbl/data.frame)
 $ species          : Factor w/ 3 levels "Adelie","Chinstrap",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ island           : Factor w/ 3 levels "Biscoe","Dream",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ bill_length_mm   : num [1:344] 39.1 39.5 40.3 NA 36.7 39.3 38.9 39.2 34.1 42 ...
 $ bill_depth_mm    : num [1:344] 18.7 17.4 18 NA 19.3 20.6 17.8 19.6 18.1 20.2 ...
 $ flipper_length_mm: int [1:344] 181 186 195 NA 193 190 181 195 193 190 ...
 $ body_mass_g      : int [1:344] 3750 3800 3250 NA 3450 3650 3625 4675 3475 4250 ...
 $ sex              : Factor w/ 2 levels "female","male": 2 1 1 NA 1 2 1 2 NA NA ...
 $ year             : int [1:344] 2007 2007 2007 2007 2007 2007 2007 2007 2007 2007 ...
>
```

Επαναλαμβάνουμε όλη τη διαδικασία (ANOVA) για το μήκος του ράμφους *bill_length_mm*

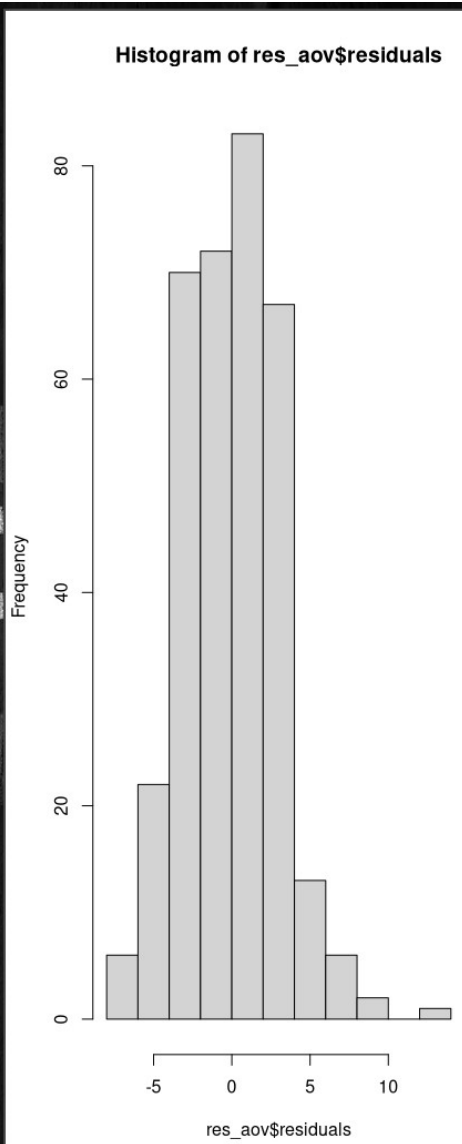
```
> dat <- penguins %>%
  dplyr::select(species, bill_length_mm)

> ggplot(dat) +
  aes(x = species, y = bill_length_mm, color = species) +
  geom_jitter() +
  theme(legend.position = "none")
Warning message:
Removed 2 rows containing missing values or values outside the scale
range (`geom_point()`).
> res_aov <- aov(bill_length_mm ~ species, data = dat)
> shapiro.test(res_aov$residuals)

        Shapiro-Wilk normality test

data:  res_aov$residuals
W = 0.98903, p-value = 0.01131

>
```

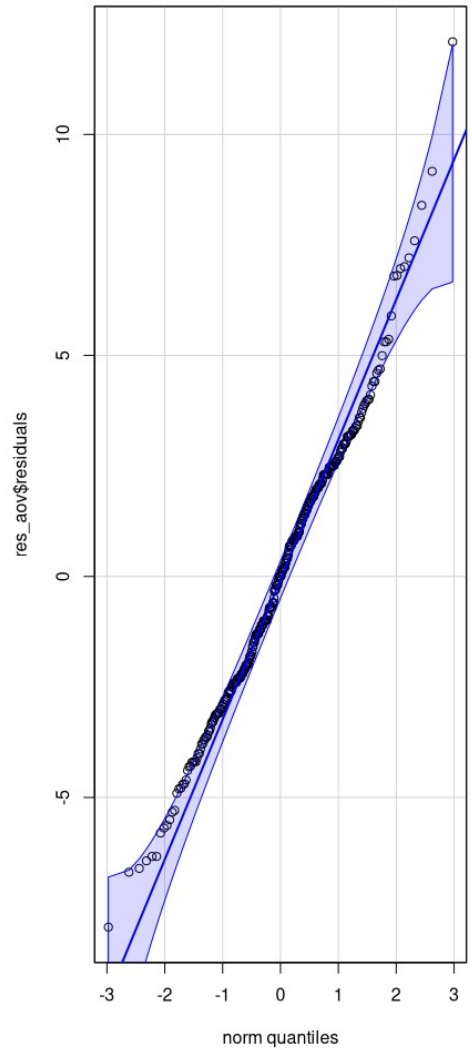
To test Shapiro-Wilk έδωσε
p-value = 0.01131
Τι σημαίνει αυτό;



```
> par(mfrow = c(1, 2)) # combine plots

# histogram
hist(res_aov$residuals)

# QQ-plot
qqPlot(res_aov$residuals, id = FALSE)
> 
```

Οι υπολειμματικές τιμές για το μήκος του ράμφους δεν ακολουθούν την κανονική κατανομή





Μη παραμετρικό τεστ Kruskal-Wallis

- Κάνει ακριβώς την ίδια σύγκριση με την ANOVA χωρίς να είναι απαραίτητη η προϋπόθεση της κανονικής κατανομής για τα δεδομένα
  - Πρόκειται για επέκταση του μη παραμετρικού τεστ Mann-Whitney
- Ισχύουν οι ίδιες υποθέσεις, όπως στην ANOVA
  - H_0: $\mu_{Adelie}=\mu_{Chinstrap}=\mu_{Gentoo}$
    - τα τρία είδη είναι ίσα για το μήκος ράμφους
  - H_1: τουλάχιστον ένας μέσος είναι διαφορετικός
    - Τουλάχιστον ένα είδος διαφέρει από τα άλλα δύο για το μήκος του ράμφους
  -

```
> par(mfrow = c(1, 2)) # combine plots

# histogram
hist(res_aov$residuals)

# QQ-plot
qqPlot(res_aov$residuals, id = FALSE)
> kruskal.test(bill_length_mm ~ species, data = dat)

        Kruskal-Wallis rank sum test

data:  bill_length_mm by species
Kruskal-Wallis chi-squared = 244.14, df = 2, p-value < 2.2e-16

>
```

Τι συμπέρασμα βγάζουμε;

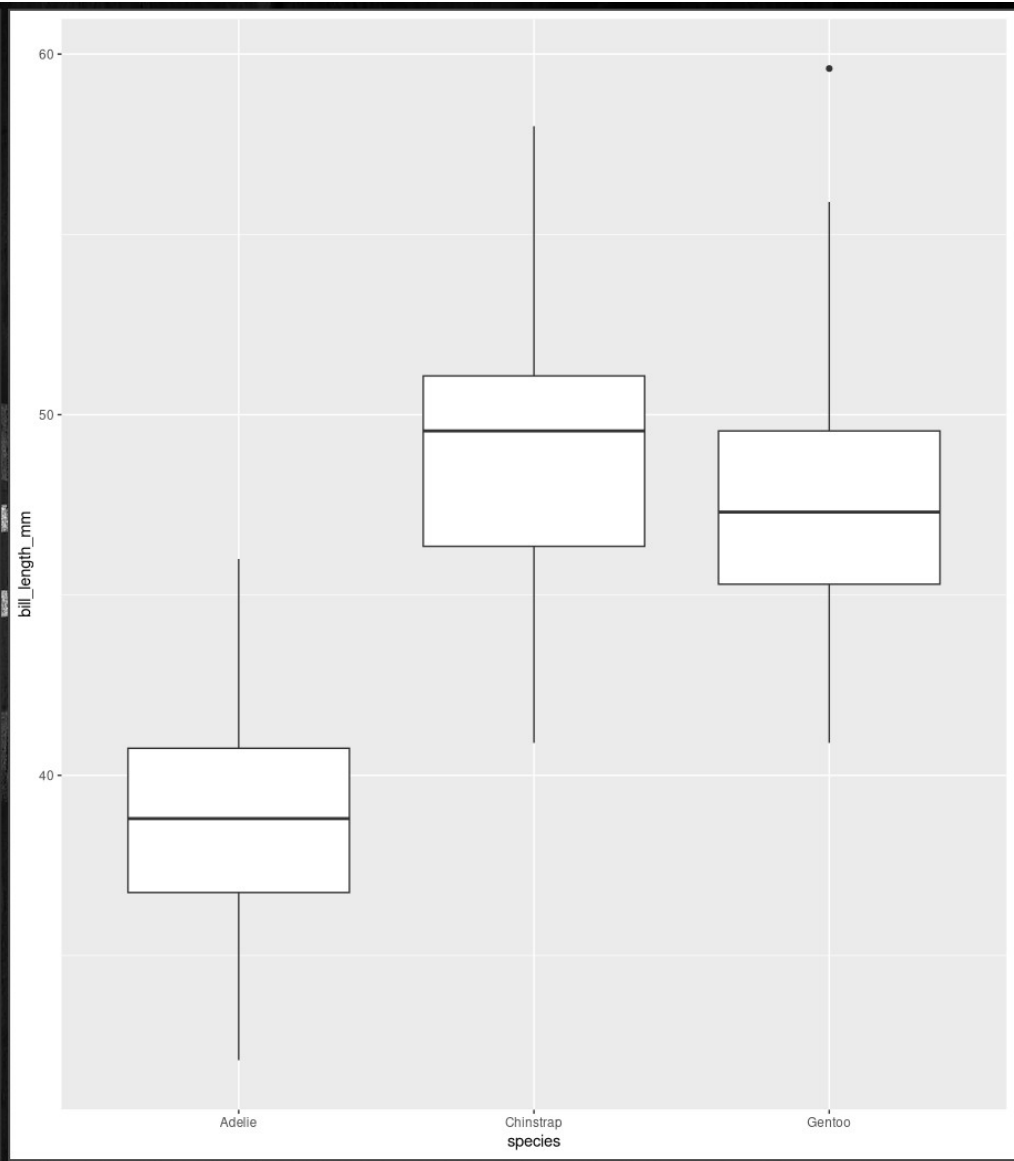
Post-hoc μετά τον μη παραμετρικό έλεγχο

- Όπως και στην ANOVA, έτσι και στον μη παραμετρικό έλεγχο Kkruskal-Wallis μπορούμε να κάνουμε μια σειρά Post-hoc tests
  - Στο παράδειγμά μας θα κάνουμε το τεστ Dunn από το πακέτο της R *FSA*

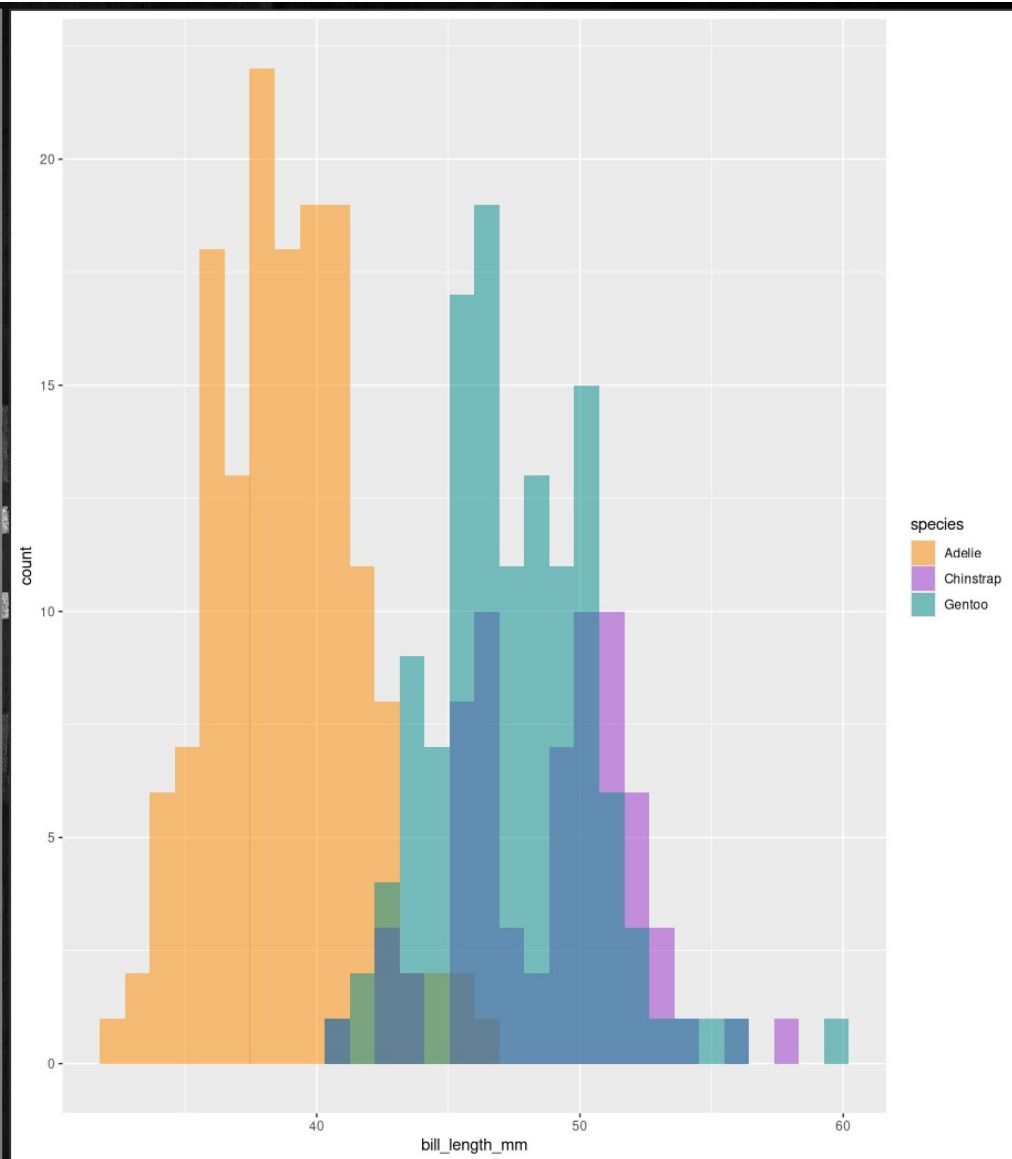
```
> library(FSA)
Registered S3 methods overwritten by 'FSA':
  method       from
  confint.boot car
  hist.boot    car
## FSA v0.9.5. See citation('FSA') if used in publication.
## Run fishR() for related website and fishR('IFAR') for related book.

Attaching package: ‘FSA’

The following object is masked from ‘package:car’:

    bootCase

> dunnTest(bill_length_mm ~ species, data = dat, method = "holm")
Dunn (1964) Kruskal-Wallis multiple comparison
  p-values adjusted with the Holm method.

          Comparison          Z      P.unadj        P.adj
1 Adelie - Chinstrap -12.753511 2.980163e-37 5.960326e-37
2    Adelie - Gentoo -13.135630 2.057716e-39 6.173147e-39
3 Chinstrap - Gentoo   1.767498 7.714481e-02 7.714481e-02
Warning message:
Some rows deleted from 'x' and 'g' because missing data.
>
```

Τι συμπέρασμα βγάζουμε;

```
> library(FSA)
> dunnTest(bill_length_mm ~ species, data = dat, method = "holm")
Dunn (1964) Kruskal-Wallis multiple comparison
  p-values adjusted with the Holm method.

         Comparison          Z      P.unadj        P.adj
1 Adelie - Chinstrap -12.753511 2.980163e-37 5.960326e-37
2    Adelie - Gentoo -13.135630 2.057716e-39 6.173147e-39
3 Chinstrap - Gentoo   1.767498 7.714481e-02 7.714481e-02
Warning message:
Some rows deleted from 'x' and 'g' because missing data.
> ggplot(dat) +
  aes(x = species, y = bill_length_mm) +
  geom_boxplot()
Warning message:
Removed 2 rows containing non-finite outside the scale range
(`stat_boxplot()`).
>
```



```
> library(FSA)
> dunnTest(bill_length_mm ~ species, data = dat, method = "holm")
Dunn (1964) Kruskal-Wallis multiple comparison
  p-values adjusted with the Holm method.

         Comparison          Z      P.unadj        P.adj
1 Adelie - Chinstrap -12.753511 2.980163e-37 5.960326e-37
2    Adelie - Gentoo -13.135630 2.057716e-39 6.173147e-39
3 Chinstrap - Gentoo   1.767498 7.714481e-02 7.714481e-02
Warning message:
Some rows deleted from 'x' and 'g' because missing data.
> ggplot(dat) +
  aes(x = species, y = bill_length_mm) +
  geom_boxplot()
Warning message:
Removed 2 rows containing non-finite outside the scale range
(`stat_boxplot()`).
> # Histogram example: bill length by species
ggplot(data = penguins, aes(x = bill_length_mm)) +
  geom_histogram(aes(fill = species), alpha = 0.5, position = "identit
y") +
  scale_fill_manual(values = c("darkorange","darkorchid","cyan4"))
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
Warning message:
Removed 2 rows containing non-finite outside the scale range
(`stat_bin()`).
>
```



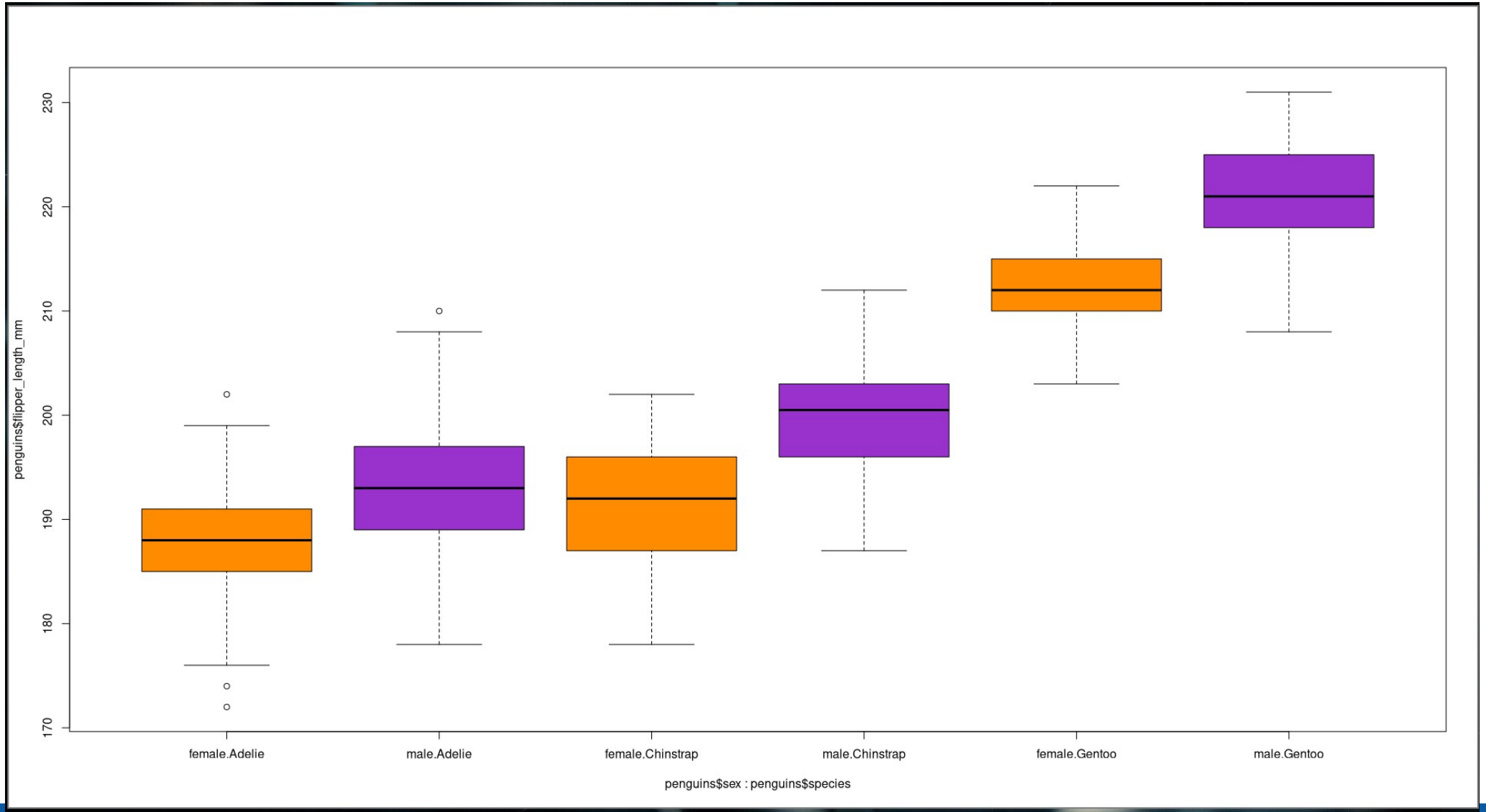
Two way ANOVA in R

- Όταν έχουμε περισσότερες από μία κατηγορικές μεταβλητές
- Παράδειγμα από το σύνολο δεδομένων *penguins*
- Τιμές μήκους πτερυγίου σε πιγκουίνους κάτω από την επίδραση:
  - Τριών διαφορετικών ειδών
  - Δύο διαφορετικών φύλων

```
> str(penguins)
tibble [344 × 8] (S3: tbl_df/tbl/data.frame)
 $ species          : Factor w/ 3 levels "Adelie","Chinstrap",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ island           : Factor w/ 3 levels "Biscoe","Dream",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ bill_length_mm   : num [1:344] 39.1 39.5 40.3 NA 36.7 39.3 38.9 39.2 34.1 42 ...
 $ bill_depth_mm    : num [1:344] 18.7 17.4 18 NA 19.3 20.6 17.8 19.6 18.1 20.2 ...
 $ flipper_length_mm: int [1:344] 181 186 195 NA 193 190 181 195 193 190 ...
 $ body_mass_g      : int [1:344] 3750 3800 3250 NA 3450 3650 3625 4675 3475 4250 ...
 $ sex              : Factor w/ 2 levels "female","male": 2 1 1 NA 1 2 1 2 NA NA ...
 $ year             : int [1:344] 2007 2007 2007 2007 2007 2007 2007 2007 2007 2007 ...
> 
```

Αλληλεπίδραση μεταξύ κατηγορικών μεταβλητών

- Η two way ANOVA μπορεί να γίνει όταν οι δύο μεταβλητές είναι ανεξάρτητες αλλά και όταν είναι εξαρτημένες
  - Όταν είναι ανεξάρτητες ακολουθούμε τον “αθροιστικό” (additive) τρόπο με τελεστή **+**
  - Όταν είναι εξαρτημένες ακολουθούμε τον “παραγοντικό” (factorial) τρόπο με τελεστή *****
- Απεικονίζουμε τα δεδομένα για μια πρώτη εκτίμηση
  - *boxplot(penguins$flipper_length_mm ~ penguins$sex * penguins$species, col = c("darkorange","darkorchid"))*


penguins$flipper_length_mm
170
180
190
200
210
220
230
female.Adelie
male.Adelie
female.Chinstrap
male.Chinstrap
female.Gentoo
male.Gentoo
penguins$sex : penguins$species

```
> factorial <- aov(penguins$flipper_length_mm ~ penguins$sex * penguins$species)
> summary(factorial)
                          Df Sum Sq Mean Sq F value  Pr(>F)
penguins$sex               1   4246    4246 132.777 < 2e-16 ***
penguins$species           2  50185   25093 784.583 < 2e-16 ***
penguins$sex:penguins$species   2    329     165   5.144 0.00631 **
Residuals                327  10458      32
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
11 observations deleted due to missingness
> additive <- aov(penguins$flipper_length_mm ~ penguins$sex + penguins$species)
> summary(additive)
                Df Sum Sq Mean Sq F value Pr(>F)
penguins$sex     1   4246    4246   129.5 <2e-16 ***
penguins$species   2  50185   25093   765.3 <2e-16 ***
Residuals        329  10787      33
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
11 observations deleted due to missingness
>
```

Τι συμπέρασμα βγάζουμε;

Αρχικά κάνουμε την 2-way ANOVA θεωρώντας τους παράγοντες **sex** και **species** εξαρτημένους και βλέπουμε ότι πράγματι υπάρχει ισχυρή αλληλεπίδραση μεταξύ τους

Thank you