Στατιστική Στην Πληροφορική - 4η Σειρά Ασκήσεων
ΟΠΑ , Ακαδημαϊκό Έτος: 2020-2021

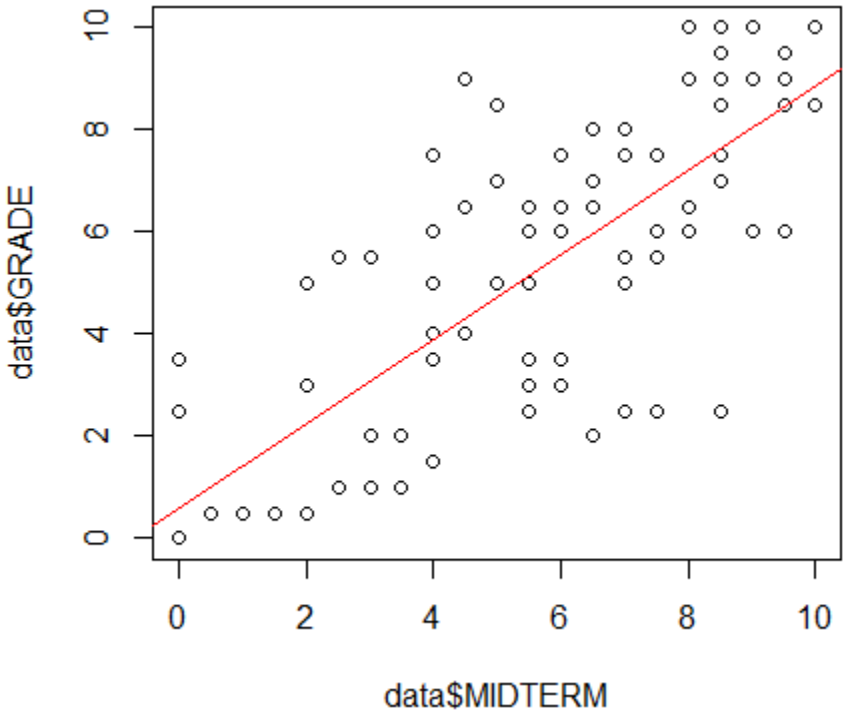
Ομάδα #0

- ΤΣΙΟΜΠΙΚΑΣ ΔΗΜΗΤΡΙΟΣ , 3180223
- ΠΑΝΑΓΙΩΤΟΥ ΠΑΝΑΓΙΩΤΗΣ , 3180139

1η Άσκηση

a. Τοποθετούμε τα δεδομένα σε data frame με όνομα data και
για να εξετάσουμε την **Γραμμικότητα** θα κάνουμε το scatterplot των 2 μεταβλητών :

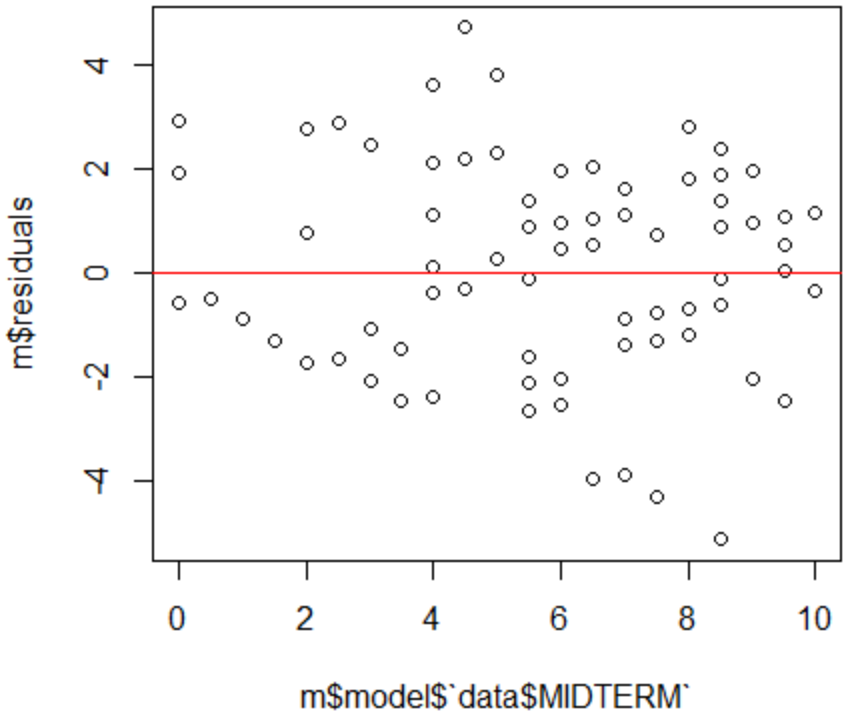
```
> plot(data$GRADE ~ data$MIDTERM)
> m <- lm(data$GRADE ~ data$MIDTERM)
> abline(m,col="red")
```



Παρατηρούμε ότι υπάρχει γραμμικότητα στα δεδομένα

Ομοσκεδαστικότητα :

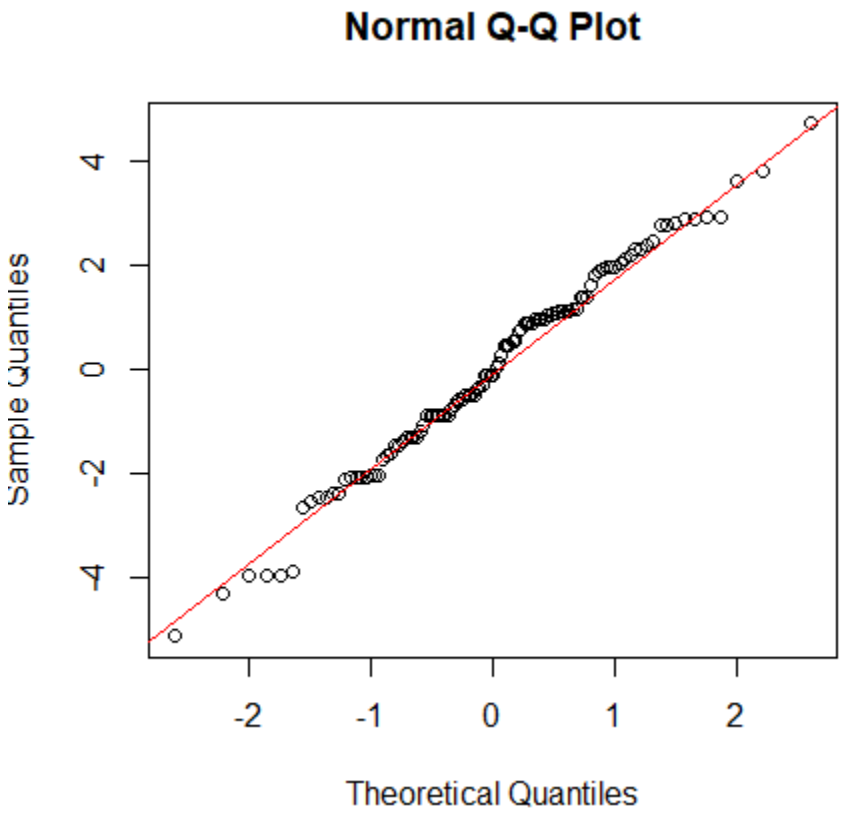
```
> plot(m$residuals ~ m$model$`data$MIDTERM`)
> newM <- lm(m$residuals ~ m$model$`data$MIDTERM`)
> abline(newM,col="red")
```



Παρόλο που φαίνεται αρκετές τιμές να είναι μακριά από την κόκκινη γραμμή , έχουμε ομοσκεδαστικότητα εφόσον οι περισσότερες είναι κοντά σε αυτή.

<u>Κανονικότητα</u> :

```
> qqnorm(m$residuals)
> qqline(m$residuals, col="red")
```



Παρατηρούμε στο Normal Quantile plot ότι τα δεδομένα μας δεν αποκλίνουν πολύ από την ευθεία άρα τα δεδομένα φαίνεται να ακολουθούν την κανονική κατανομή.

b. Ο Εκτιμητής b_1 θα είναι :

```
> m$coefficients
 (Intercept) data$MIDTERM
   0.5756001    0.8290192
> m$coefficients[2] -> b1
> b1
data$MIDTERM
   0.8290192
```

Άρα έχουμε :

- **b_1 = 0.8290192**
- **SEb_1 = 0.06328825**
- **df = 109**
- **t = 1.981967**

Άρα το διάστημα εμπιστοσύνης 95% είναι :

```
> SEb1
[1] 0.06328825
> t <- -qt(0.025,df=109)
> b1 + c(-1,1) * t * SEb1
[1] 0.7035840 0.9544545
```

[0.7035840,0.9544545].

c.

Θα χρησιμοποιήσουμε t-έλεγχο σημαντικότητας με τις εξής υποθέσεις :

- H0 : $b_1 = 0$ (Δεν έχουν σχέση)
- Ha : b_1 != 0 (Έχουν σχέση)

Έχουμε :

- **b_1 = 0.8290192**
- **SEb_1 = 0.06328825**

```
> t <- b1/SEb1
> t
[1] 13.0991
```

- **t = 13.0991**

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.57560    0.38438   1.497    0.137
data$MIDTERM  0.82902    0.06329  13.099   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 1.926 on 109 degrees of freedom
  (16 observations deleted due to missingness)
Multiple R-squared:  0.6115,    Adjusted R-squared:  0.608
F-statistic: 171.6 on 1 and 109 DF,  p-value: < 2.2e-16
```

Βρέθηκε μέσω του summary(m) ότι το pvalue είναι <2.2e-16 άρα είναι πολύ μικρότερο του α = 0.05 . Επομένως, μπορούμε να απορρίψουμε τη μηδενική υπόθεση και να συμπεράνουμε ότι ο βαθμός της προόδου με τον τελικό βαθμό έχουν σχέση.

d. Η εκτίμηση του τελικού βαθμού για όσους φοιτητές έγραψαν 7 στην πρόοδο είναι :
$\mu 7 = b_1 * 7 + b_0 = 6.378735$
Και έχουμε :

- **sMid = 2.902179**
- **SEm7 = 0.3021594**
- **n = 127**
- **xbar = 5.342342**
- **sum = 926.491 (άθροισμα Σ(χι – χbar) ^ 2)**

Καταλήγουμε λοιπόν στο διάστημα εμπιστοσύνης 95%:

```
> m7 + c(-1,1) * t * SEm7
[1]  2.420718 10.336751
```

[2.420718,10.336751].

e. Έχουμε :
$y = b_1 * 7 + b_0 = 6.378735$
SEy = 2.913588
Άρα το Διάστημα πρόβλεψης 95% είναι :

```
> y + c(-1,1) * SEy
[1] 3.465147 9.292322
```

[3.465147 , 9.292322].

2η Άσκηση

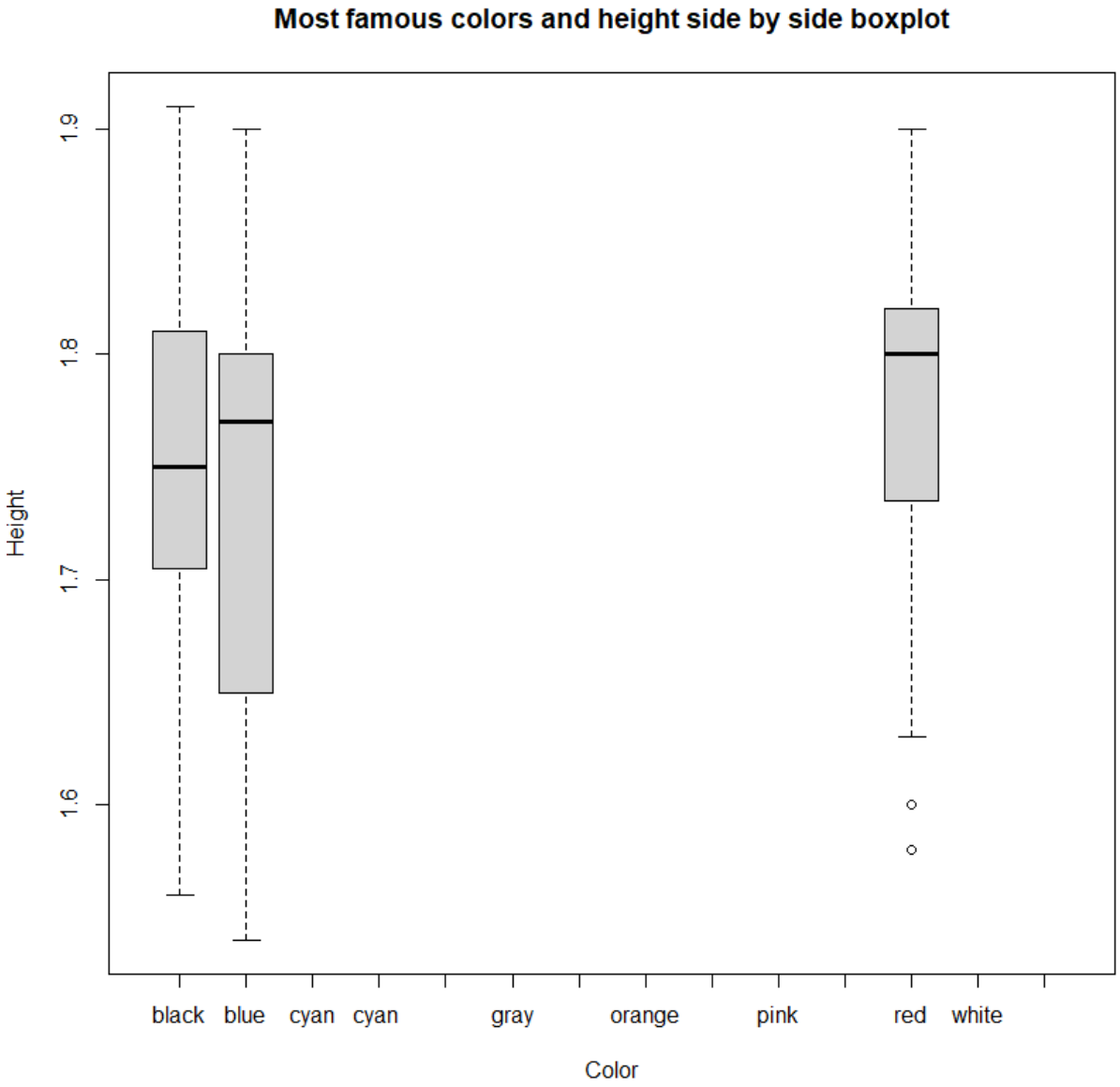
a. Παρατηρούμε πως τα 3 πιο δημοφιλή χρώματα είναι :

```
> tail(names(sort(table(data$color))), 3)
[1] "red"   "blue"  "black"
```

Μαύρο , μπλε και κόκκινο με αυτή τη σειρά.

```
> subData <- data[which(data$color == "black" | data$color == "bl
ue" | data$color == "red"),]
```

Διαχωρίζουμε τα δεδομένα μας και ξεκινάμε την ανάλυση :



Παρατηρούμε ότι υπάρχουν 2 outliers στο boxplot του κόκκινου χρώματος , παρόλα αυτά δεν φαίνεται να υπάρχουν άλλα outliers στα boxplots το οποίο συνιστά ότι τα δεδομένα ακολουθούν την κανονική κατανομή. Επίσης παρατηρούμε ότι το κόκκινο και το μαύρο προτιμούνται από κυρίως άτομα με μεγάλο ύψος ενώ το μπλε προτιμάται ανεξαιρέτως ύψους.

b. Θα χρησιμοποιήσουμε ANOVA έλεγχο σημαντικότητας όπου :

- μ1 = Ύψος
- μ2 = Χρώματα
- H0 : μ1 = μ2 (δεν έχει σχέση η επιλογή χρώματος με το ύψος)
- Ha : μ1 != μ2 (έχει σχέση η επιλογή χρώματος με το ύψος)

Κάνοντας ANOVA στην R παίρνουμε το εξής αποτέλεσμα :

```
Analysis of Variance Table

Response: subData$height
              Df  Sum Sq   Mean Sq F value Pr(>F)
subData$color  2 0.01104 0.0055186  0.7328 0.4844
Residuals     67 0.50455 0.0075306
```

- Df των χρωμάτων = 2
- Df των υψών = 70
- Pvalue = 0.4844

Άρα αφού pvalue > α δεν μπορούμε να απορρίψουμε τη μηδενική υπόθεση το οποίο σημαίνει ότι η επιλογή χρώματος δεν έχει σχέση με το ύψος.

3η Άσκηση

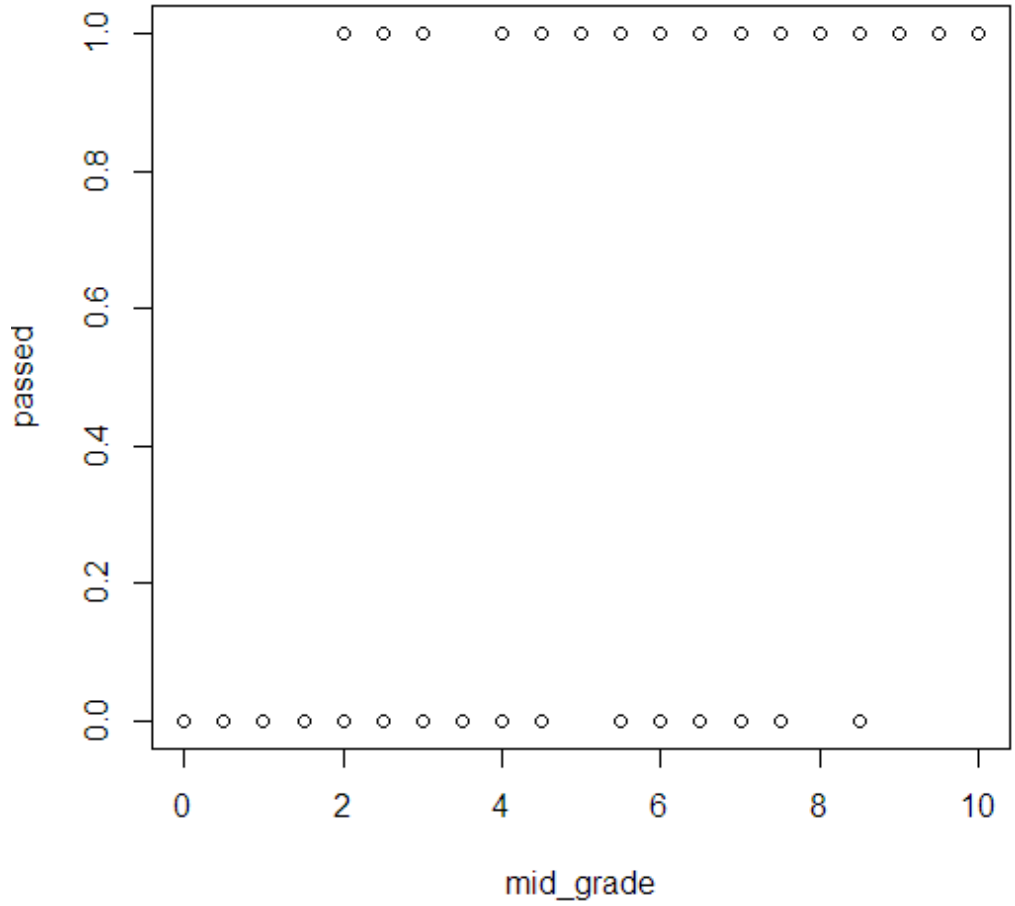
a. Αρχικά θα αναπαραστήσουμε την σχέση μεταξύ βαθμού προόδου (επεξηγηματική μεταβλητή) και επιτυχίας (μεταβλητή απόκριση) με ένα scatterplot (1 = επιτυχία , 0 = αποτυχία).

Πρώτα, θα βάλουμε τα δεδομένα μας στην R και θα αφαιρέσουμε τις πλειάδες όπου midterm= NA, καθώς δε μας αφορούν.

Στη συνέχεια, γράφουμε:

```
> data$GRADE[data$GRADE<5] <- 0
> data$GRADE[data$GRADE>=5] <- 1
> passed = data$GRADE
> mid_grade = data$MIDTERM
> fit = glm(passed ~ mid_grade, data = data, family = binomial)
> data2 <- data.frame(mid_grade=seq(min(data$MIDTERM), max(data$MIDTERM),len=11$
> data2$passed = predict(fit, newdata=data2, type="response")
> plot(passed~mid_grade, data=data)
```

Με αποτέλεσμα το παρακάτω plot:



Στο plot μπορούμε να δούμε με το μάτι πως όσοι γράψανε καλά στην πρόοδο , είχαν περισσότερη επιτυχία στο πέρασμα του μαθήματος. Απαντάμε λοιπόν στην εκφώνηση, λέγοντας πως η λογιστική παλινδρόμηση είναι κατάλληλη ως υπόδειγμα.
Άρα έχουμε τη συνάρτηση : $p(x) = \frac{1}{1+e^{-(\beta_1 x+\beta_0)}}$

όπου p η πιθανότητα επιτυχίας του x, δηλαδή του βαθμού της προόδου.

Αρκεί να βρούμε τώρα τα β_1 και β_0, κάνοντας λογιστική παλινδρόμηση.

```
> summary(fit)

Call:
glm(formula = passed ~ mid_grade, family = binomial, data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3358  -0.5486   0.3148   0.6696   1.8437

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.7771     0.6171  -4.500 6.80e-06 ***
mid_grade     0.6397     0.1166   5.488 4.06e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 149.078  on 110  degrees of freedom
Residual deviance:  96.598  on 109  degrees of freedom
AIC: 100.6

Number of Fisher Scoring iterations: 5
```

Επομένως, έχουμε **β_1 = 0.6397** και **β_0 = -2.7771**

```
> p <- function(x)
+ {prob <- 1/(1+exp(-(0.6397*x-2.7771)))
+ return(prob)}
```

Έχουμε πλέον έτοιμη τη συνάρτηση.

b. Το μόνο που έχουμε να κάνουμε είναι να βάλουμε στη συνάρτηση που βρήκαμε στο προηγούμενο ερώτημα, όπου x=5.

```
> p(5)
[1] 0.6038182
```

Άρα, **0.6038182** είναι το ποσοστό επιτυχίας των φοιτητών όταν παίρνουν βαθμό 5 στην πρόοδο.

c. Θα κάνουμε τον έλεγχο σημαντικότητας:

- H0: $b_1=0$
- Hα: $b_1!=0$

Θυμόμαστε από τις διαλέξεις πως στατιστικός έλεγχος: $z = \frac{b_1}{SE_{b_1}}$

```
> summary(fit)

Call:
glm(formula = passed ~ mid_grade, family = binomial, data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3358  -0.5486   0.3148   0.6696   1.8437

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.7771     0.6171  -4.500 6.80e-06 ***
mid_grade     0.6397     0.1166   5.488 4.06e-08 ***
---
```

Στο σχήμα βλέπουμε πως b_1=0.6397. Δεξιά του το SEb_1=0.116. Εκτελώντας την πράξη ή κοιτάζοντας στο σχήμα βλέπουμε πως: ***z*=5.488**.
Το **p_value = 4.06*10^{-8}**, είμαστε πλέον στη θέση να απορρίψουμε την μηδενική υπόθεση. Συμπεραίνουμε λοιπόν, πως όντως σχετίζεται ο βαθμός προόδου με την επιτυχία.

d. Είδαμε πως **p(5)= 0.6038182 > 0.5**

Μπορούμε να προβλέψουμε πως ο φοιτητής θα περάσει τις εξετάσεις του μαθήματος.

4η Άσκηση

Ρίχνουμε το νόμισμα: **N=100 φορές**.
Εμφανίσεις κορώνας: **κ=44**, πιθανότητα εμφάνισης θ ∈ (0,1)
Συνάρτηση πιθανοφάνειας: $L(\theta) = \binom{100}{44}\theta^{44}(1-\theta)^{56}$.

Θα εκτιμήσουμε την πιθανότητα εμφάνισης κορώνας σύμφωνα με την αρχή της Μέγιστης Πιθανοφάνειας.
Από τη συνάρτηση λογαριθμικής Πιθανοφάνειας (log-likehood) έχουμε τον τύπο:

$$\hat{\theta}_{MLE} = \arg\max_{\theta\in\Theta} \lambda(\theta)\,, \text{όπου } \lambda(\theta) = \log L(\theta)$$

Θα βρούμε για ποιο θ μεγιστοποιείται το λ(θ), και αυτό θα μας δώσει τη πιθανότητα που ψάχνουμε.
(Θέτω $\binom{100}{44}$ = λ , καθώς δε θα μας χρειαστεί ο υπολογισμός του)

Το λ(θ) μεγιστοποιείται όταν λ'(θ)=0 ⇔
⇔ $(\log L(\theta))' = 0$ ⇔
⇔ $(\log(\lambda\theta^{44}(1-\theta)^{56})' = 0$ ⇔
⇔ $(\log\lambda + 44\log\theta + 56\log(1-\theta))' = 0$ ⇔

⇔$(\log\lambda)' + 44(\log\theta)' + 56(\log(1-\theta))' = 0$ ⇔

⇔ $0 + 44/\theta - 56/(1-\theta) = 0$ ⇔

⇔ $44/\theta = 56/1-\theta$ ⇔
⇔ $44-44\theta = 56\theta$ ⇔

⇔$44 = 100\theta$⇔

⇔$\boldsymbol{\theta = 44/100}$

Επομένως, $\hat{\theta}_{MLE} = 44/100$