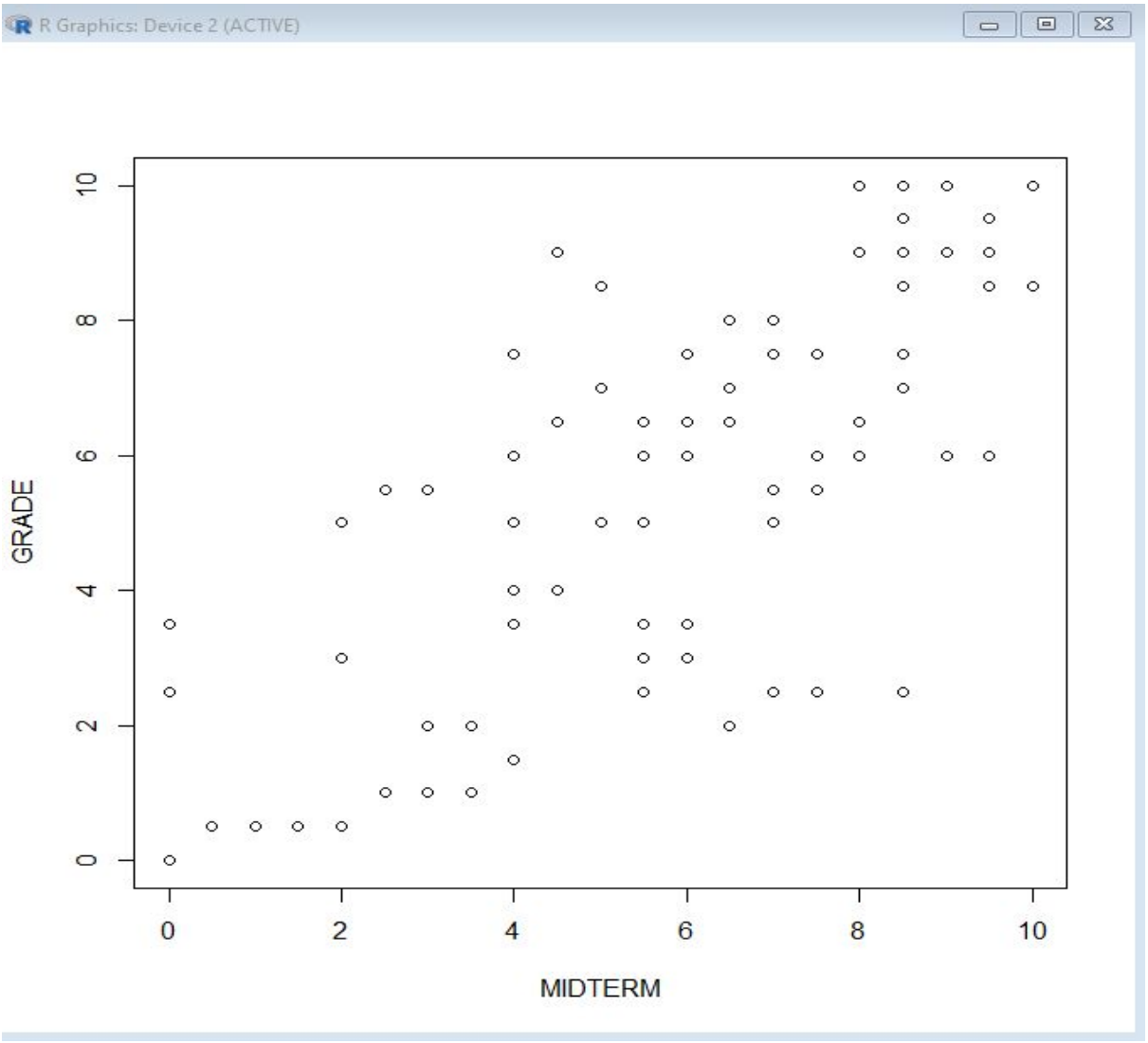
Στατιστική στην Πληροφορική
2020-2021

Φέκα Αγγελική A.M:3140290
Σταυρουλάκη Μαρία A.M:3160168

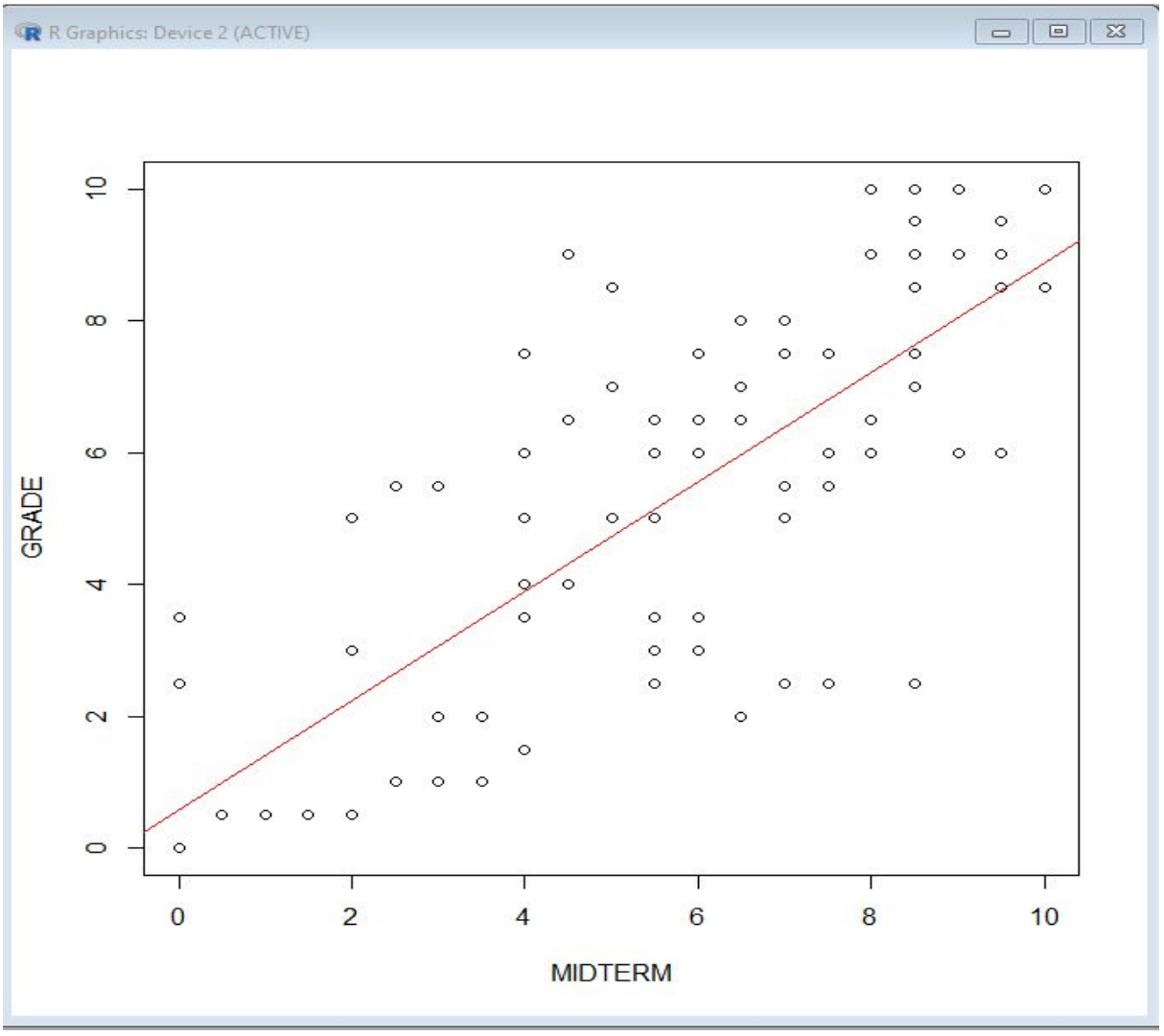
Άσκηση 1η :

a) Θα διερευνήσουμε τη σχέση μεταξύ του GRADE και του MIDTERM, όπου ως επεξηγηματική μεταβλητή θεωρούμε την MIDTERM. Παρακάτω εμφανίζεται το scatterplot των δύο μεταβλητών.



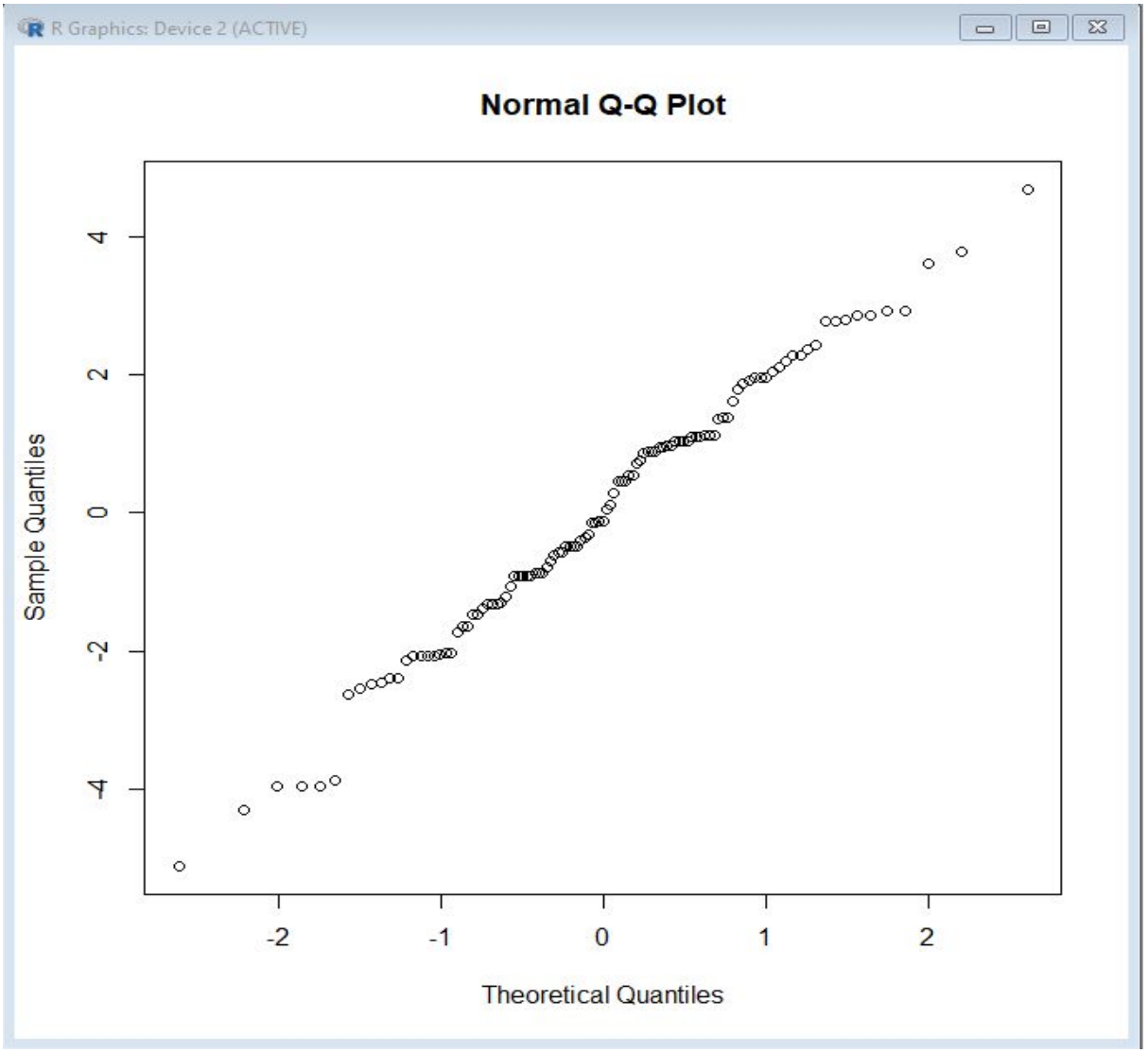
Όπως φαίνεται στο σχήμα παρατηρούμε μια αύξουσα σχέση μεταξύ των μεταβλητών και επομένως θα μπορούσαμε να υποθέσουμε ότι υπάρχει μια γραμμική σχέση μεταξύ την μεταβλητής MIDTERM και της μεταβλητής GRADE. Για να εξάγουμε ένα ασφαλές αποτέλεσμα θα κάνουμε γραμμική παλινδρόμηση.

Το καινούργιο διάγραμμα scatterplot με την γραμμική παλινδρόμηση ελαχίστων τετραγώνων(κόκκινη γραμμή) φαίνεται παρακάτω:



Όπως βλέπουμε από το σχήμα η ομοσκεδαστικότητα, κοιτάμε τις διαφορές του GRADE από την γραμμική παλινδρόμηση (residuals). Παρατηρούμε ότι η διασπορά των τιμών πάνω και κάτω από την γραμμική παλινδρόμηση δεν είναι ομοιογενής, όμως δεν φαίνεται να είναι και πολύ κακή προσέγγιση. Επομένων μπορούμε να υποθέσουμε ότι ισχύει η ομοσκεδαστικότητα.

Για τον έλεγχο την κανονικότητας δίνουμε το normal quantile plot των υπόλοιπων:



Η κατανομή των υπόλοιπων φαίνεται να είναι κοντά στην κανονική.

b) Από τη γραμμική παλινδρόμηση, η τιμή του εκτιμητή του β_1 είναι b_1 = 0.82902 και η τιμή του εκτιμητή του β_0 είναι b_0 = 0.57560.

Το 95% διάστημα εμπιστοσύνης είναι $b_1 \pm t_* SE_{b1}$ = [0.7037832, 0.9542552], όπου SE_{b1} = 0.06328825 και χρησιμοποιήσαμε για την κατανομή t με df = n − 2 = 125 βαθμούς ελευθερίας, οπότε t_* = 1.97882.

c) Παρατηρώντας το scatterplot του προηγούμενου υποερωτήματος φαίνεται ότι οι μεταβλητές ακολουθούν γραμμική, αύξουσα, θετική κατεύθυνση. Θα το διαπιστώσουμε με τον έλεγχο σημαντικότητας H_0:β_1= 0, H_α:$\beta_1 \neq 0$. Το στατιστικό ελέγχου είναι t = 13.099, το οποίο δίνει p-value = 0.137, όπου χρησιμοποιήσαμε την κατανομή t με df = 125 βαθμούς ελευθερίας. Άρα απορρίπτουμε τη μηδενική υπόθεση, δηλαδή ότι δεν υπάρχει σχέση μεταξύ των μεταβλητών.

d) Χρησιμοποιώντας το πρόγραμμα της R εκτιμούμε ότι ο τελικός βαθμός που θα επιτύγχαναν φοιτητές, οι οποίοι στην εξέταση προόδου έλαβαν 7 είναι: 6.378735

```
> m$coefficients
(Intercept)     MIDTERM
  0.5756001   0.8290192
> b1 * 7 + m$coefficients[1]
 MIDTERM
6.378735
```

Και το 95% διάστημα εμπιστοσύνης είναι: [5.960928, 6.796541]

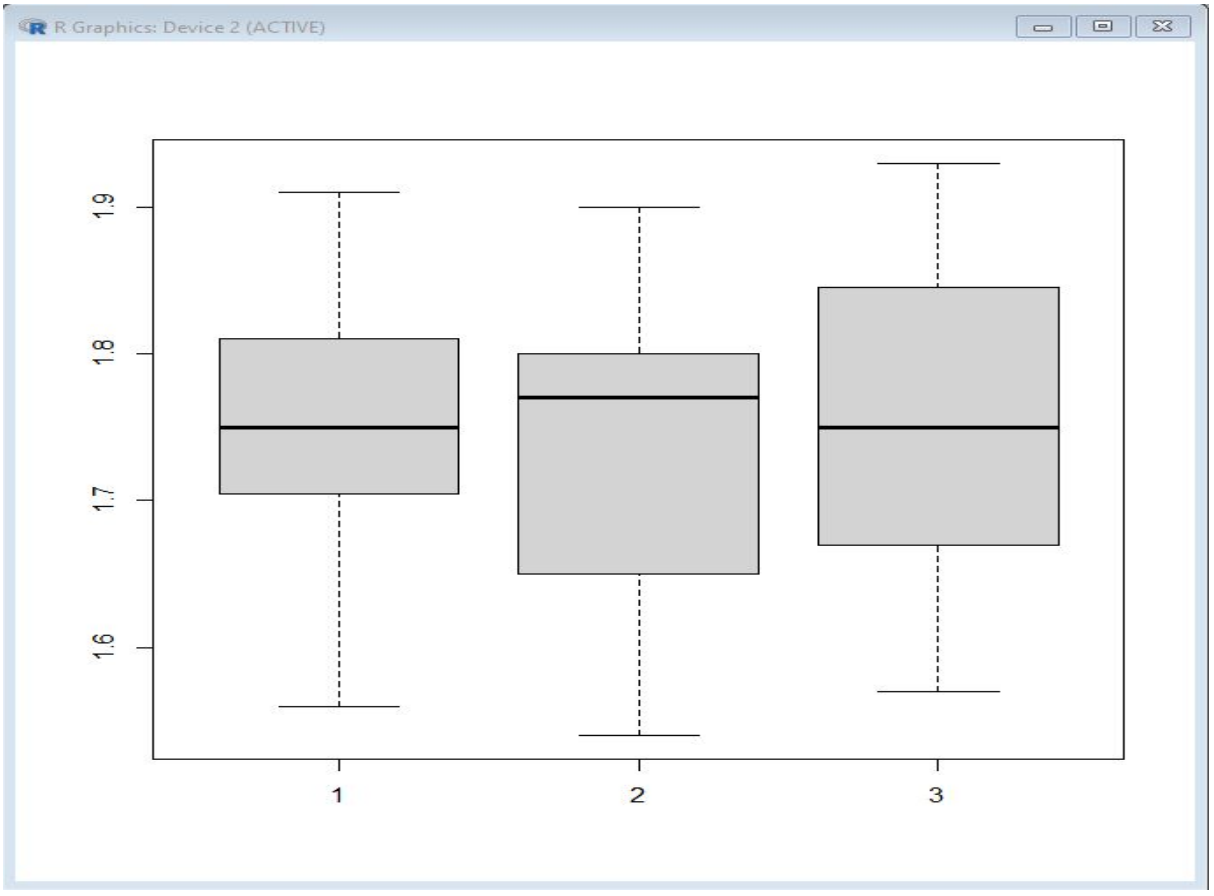
e) Χρησιμοποιώντας το πρόγραμμα της R προβλέπουμε ότι ο τελικός βαθμός που θα επιτύχει ένας τυχαία επιλεγμένος φοιτητής που έχει γράψει 7 στην πρόοδο είναι: 6.378735

```
> predict(m, newdata = data.frame(MIDTERM = 7.0), interval = "prediction")
       fit      lwr      upr
1 6.378735 2.537905 10.21956
```

Και το 95% διάστημα εμπιστοσύνης είναι: [2.537905, 10.21956]

Άσκηση 2η :

a) Τα 3 δημοφιλέστερα χρώματα ήταν τα εξής: Μαύρο(28), μπλε(24) και μωβ(19). Στην τρίτη θέση βρισκόταν και το κόκκινο με επισης 19 εμφανίσεις αλλά επιλέξαμε τυχαία το μωβ.



Οι δειγματικές μέσες τιμές είναι οι παρακάτω:

Χρώμα	Μαύρο	Μπλε	Μωβ
Μέση τιμή ύψους	1.757143	1.735652	1.745789

Από τον παραπάνω πίνακα φαίνεται ότι δεν υπάρχει αισθητή διαφορά στα μέσα ύψη.

b) Έστω H_0: οι τιμές δεν διαφέρουν και H_α: οι τιμές διαφέρουν

```
> tapply(number[si>0], as.factor(colors[si[si>0]]), mean)
       b       bl        p
55.08333 48.96429 79.84211

> tapply(number[si>0], as.factor(colors[si[si>0]]), sd)
       b       bl        p
 30.57267  32.01907 161.44220

>  aov(number[si>0]~as.factor(colors[si[si>0]])) -> ares
> anova(ares)
Analysis of Variance Table

Response: number[si > 0]
                            Df Sum Sq Mean Sq F value Pr(>F)
as.factor(colors[si[si > 0]])  2  11435  5717.7  0.7501 0.4762
Residuals                     68 518323  7622.4
```

Το p-value είναι 0.4762, το οποίο είναι αρκετά μεγάλο για να απορριφθεί η μηδενική υπόθεση. Άρα η διαφορά ανάμεσα στις τρεις τιμές δεν είναι στατιστικά σημαντική.

Άσκηση 3η :

a) Το υπόδειγμα για τη σχέση μεταξύ βαθμού προόδου και επιτυχίας, χρησιμοποιώντας λογιστική παλινδρόμηση στα δεδομένα εντάσσεται ως εξής: Υπόθεση για πληθυσμό: p(x) = ποσοστό '1' στον υποπληθυσμό όπου $X = x$ Log(p(x)/1-p(x))=β1x+β0 Όπου β1=0.3610592, SEb1 = 0.1073019.Η επεξηγηματική μεταβλητή είναι ο βαθμός προόδου και η μεταβλητή απόκρισης η επιτυχία.

b) Βάσει του υποδείγματος από το ερώτημα a, εκτιμάται ότι το ποσοστό επιτυχίας των φοιτητών όταν παίρνουν βαθμό 5 στην πρόοδο είναι: $\widehat{p}(5) = 0.6038$

c) Κάνουμε τον έλεγχο σημαντικότητας H_0:$\beta_1 = 0$ και H_α:$\beta_1 \neq 0$

Γνωρίζουμε ότι το στατιστικό ελέγχου είναι $z = \frac{b_1}{SE_{b1}}$, όπου από το ερώτημα a έχουμε: b_1 = 0.6397 και με το λογισμικό βρίσκουμε ότι SE_{b1} = 0.1166. Επομένως απο το στατιστικό έλεγχο το z = 5.488.

Επιπλέον το p-value βάση της κανονικής κατανομής είναι 4.06×10^{-8}

Άρα απορρίπτουμε την μηδενική υπόθεση β_1 = 0, δηλαδή τα δεδομένα υποδεικνύουν ότι η επιτυχία στο μάθημα σχετίζεται με τον βαθμό προόδου.

d) Χρησιμοποιώντας το παραπάνω υπόδειγμα μπορούμε να προβλέψουμε αν θα περάσει το μάθημα ένας φοιτητής που πήρε 5 στην πρόοδο. Συγκεκριμένα από το ερώτημα b έχουμε: $\widehat{p}(5) = 0.6038 > \frac{1}{2}$, άρα προβλέπουμε ότι ο φοιτητής αυτός θα περάσει το μάθημα.

Άσκηση 4η :

Για να εκτιμήσουμε την πιθανότητα εμφάνισης της κορώνας σύμφωνα με την Αρχή της Μέγιστης Πιθανότητας, πρέπει να γνωρίζουμε ότι η συνάρτηση πιθανοφάνειας που δίνεται είναι η συχνότητα εμφάνισης x εάν επαναλαμβάναμε τη δειγματοληψία πάρα πολλές φορές, υπό διαφορετικές υποθέσεις.

Υπολογίζουμε τον εκτιμητή μέγιστης πιθανοφάνειας (*Maximum Likelihood Estimator (MLE)*): $\widehat{\theta}_{MLE}$ της εξίσωσης λ'(θ) = 0, όπου εφόσον έχουμε 44 κορώνες σε 100 ρίψεις ισχύει ότι: λ(θ) = xlogθ + (n - x)log(1- θ) και λ'(θ) = $\frac{x}{\theta} - \frac{(n-x)}{(1-\theta)} = 0$. Επομένως έχουμε ότι $\widehat{\theta}_{MLE}$ = $\frac{44}{100} = \frac{11}{25}$.

Άρα η Αρχή της Μέγιστης Πιθανότητας δίνει εκτίμηση 0.44 για την πιθανότητα εμφάνισης κορώνας.