ΥΠΟΛΟΓΙΣΤΙΚΗ Στατιστική
ΕΡΓΑΣΙΑ 1

Δημήτριος Φούντας
A.M : 1112201600236

14 Δεκεμβρίου 2021

Εκφώνηση

Ο κύριος Χ θεωρεί πως είναι ειδικός στο να αναγνωρίζει διαφορετικές ποικιλίες κρασιών. Για αυτό τον λόγο προσκλήθηκε σε ένα συνέδριο οινολογίας, όπου σε μία δοκιμή έπρεπε να δοκιμάσει κ ποικιλίες κρασιών και να αναγνωρίσει ποιες είναι.

Ερώτηση 1

Στην αρχή σε 5 ποτήρια τοποθετήθηκαν 5 διαφορετικά κρασιά και ο κύριος Χ κλήθηκε να αναγνωρίσει τα κρασιά αυτά. Αναγνώρισε τρία από αυτά. Είναι τελικά ο κύριος Χ ειδικός;

Ερώτηση 1

Για να απαντήσουμε την παραπάνω ερώτηση πρέπει να κάνουμε ακριβή έλεγχο τυχαιοποίησης.
Η μηδενική μας υπόθεση H_0 θα είναι πως ο κύριος Χ επιλέγει τυχαία τα κρασιά κάθε φορά.
Η εναλλακτική μας υπόθεση H_1 είναι πως ο κύριος Χ είναι ειδικός, δηλαδή πως πετυχαίνει το είδος από περισσότερα κρασιά από αυτά που θα πετύχαινε αν απλά μάντευε στην τύχη.

Ερώτηση 1

Στην πραγματικότητα αυτό που μας ενδιαφέρει, αν θέλουμε να το πούμε πιο μαθηματικά, είναι το αν η συχνότητα με την οποία παρατηρούμε επιτυχίες (Έστω F) είναι σημαντικά μεγαλύτερη από τον αριθμό επιτυχιών που θα περιμέναμε να έχει κάποιος που επιλέγει τα κρασια στην τύχη (Έστω πως αυτή είναι ένας σταθερός αριθμός $F_{expected}$).

Ερώτηση 1

Άρα η ελεγχοσυνάρτηση μας θα είναι η συνάρτηση $T = F - F_{expected}$.
Συνεπώς οι υποθέσεις μας μπορούν πλέον να μεταφραστούν σε:
$H_0 : F = F_{expected}$
$H_1 : F > F_{expected}$
Όμως όπως είπαμε και πριν το $F_{expected}$ είναι σταθερός αριθμός, οπότε μπορούμε να τον αγνοήσουμε και να κάνουμε τον ισοδύναμο έλεγχο με ελεγχοσυνάρτηση το $T = F$.

Ερώτηση 1

Μπορούμε να υποθέσουμε χωρίς περιορισμό της γενικότητας ότι η σωστή σειρά που θα έπρεπε να μαντέψει ο κύριος Χ τα είδη κρασιών προκειμένου να τα πετύχει όλα είναι να μαντέψει πρώτα το α κρασί, μετά το β κρασί, μετά το γ κρασί, μετά το δ κρασί, μετά το ε κρασί (άρα η σειρά είναι η : (α,β,γ,δ,ε)). Έστω πάλι χωρίς περιορισμό της γενικότητας ότι ο κύριος Χ πέτυχε τα τρία πρώτα και απέτυχε να βρει τα δύο τελευταία.
Η τιμή της ελεγχοσυνάρτησης της παρατήρησης λοιπόν είναι η $tobs=3$.

Κώδικας

```
library('gtools')
cor_Wine=c("a","b","c","d","e")
comb=permutations(5,5,cor_Wine)
n=factorial(5)
tobs=3
t=numeric(n)
for (i in 1:n) {
  s=0
  for (j in 1:5) {
    if (comb[i,j]==cor_Wine[j]){
      s=s+1
    }
  }
  t[i]=s
  }
pvalue=length(t[tobs<=t])/n
```

Εξηγώντας τον κώδικα (1)

1)Στην πρώτη γραμή καλούμε το πακέτο *gtools* που περιέχει διάφορες χρήσιμες συναρτήσεις, στον κώδικα μας χρησιμοποιούμε μόνο την συνάρτηση *permutations* μέσα από αυτό το πακέτο η οποία μας επιστρέφει πίνακα που περιέχει όλες τις δυνατές διατάξεις που του ζητάμε.
2)Στη δεύτερη γραμμή του κώδικα τοποθετούμε σε ένα διάνυσμα την σειρά των ειδών των κρασιών που θεωρήσαμε πως είναι η σωστή χ.π.γ.
3)Στην τρίτη γραμμή βρίσκω όλες τις δυνατές μεταθέσεις των 5 στοιχείων μας.
4)Στην τέταρτη γραμμή έχω την παρατηρούμενη τιμή της ελεγχοσυνάρτησης μου η οποία είναι το πλήθος ειδών κρασιών που μάντεψε σωστά.
5)Στην πέμπτη γραμμή αρχικοποιώ ένα διάνυσμα μεγέθους $n = 120$ στο οποίο θα τοποθετήσω τις τιμές της ελεγχοσυνάρτησης μου.

Εξηγώντας τον κώδικα (2)

Αμέσως μετά για κάθε πιθανή μετάθεση που βρίσκετε μέσα στον πίνακα *comb* μετράμε το πόσα είδη κρασιών βρίσκονται στην σωστή σειρά και εισάγω το πλήθος τους στην ελεγχοσυνάρτηση μου, ως ένα στοιχείο του διανύσματος της ελεγχοσυνάρτησης t.
Τέλος βρίσκω το *pvalue* μετρώντας το πόσα στοιχεία της ελεγχοσυνάρτησης έχουν τιμή μεγαλύτερη ή ίση από αυτή της παρατηρώμενης τιμής, προς το πλήθος όλων των διαφορετικών μεταθέσεων n, το οποίο είναι ίσο με το πλήθος όλων των διαφορετικών τιμών που μπορει να πάρει η ελεγχοσυνάρτηση.

Συμπέρασμα

Τρέχοντας το πρόβλημα βρίσκουμε ως *pvalue* τον αριθμό 0.09166667. Αυτό σημαίνει πως δεν μπορώ να απορρίψω την υπόθεση πως ο κύριος Χ επέλεξε τυχαία την σειρά με τα είδη κρασιών.

Ερώτηση 2

Στην συνέχεια σε 10 ποτήρια τοποθετήθηκαν 10 διαφορετικά είδη κρασιών και ο κύριος Χ κλήθηκε να αναγνωρίσει τα είδη αυτά. Αναγνώρισε τέσσερα από αυτά. Είναι τελικά ο κύριος Χ ειδικός; Περιγράψτε την κατανομή της ελεγχοσυνάρτησης.Επαναλάβεται την τυχαιοποίηση 100 φορές και περιγράψτε όλα τα *pvalue* που βρήκατε.

Ερώτηση 2

Για να απαντήσουμε την παραπάνω ερώτηση πρέπει να κάνουμε ακριβή έλεγχο τυχαιοποίησης.
Η μηδενική μας υπόθεση H_0 θα είναι πως ο κύριος Χ επιλέγει τυχαία τα κρασιά κάθε φορά.
Η εναλλακτική μας υπόθεση H_1 είναι πως ο κύριος Χ είναι ειδικός, δηλαδή πως πετυχαίνει το είδος από περισσότερα κρασιά από αυτά που θα πετύχαινε αν απλά μάντευε στην τύχη.

Ερώτηση 2

Στην πραγματικότητα αυτό που μας ενδιαφέρει, αν θέλουμε να το πούμε πιο μαθηματικά, είναι το αν η συχνότητα με την οποία παρατηρούμε επιτυχίες (Έστω F) είναι σημαντικά μεγαλύτερη από τον αριθμό επιτυχιών που θα περιμέναμε να έχει κάποιος που επιλέγει τα κρασια στην τύχη (Έστω πως αυτή είναι ένας σταθερός αριθμός $F_{expected}$).

Ερώτηση 2

Άρα η ελεγχοσυνάρτηση μας θα είναι η συνάρτηση $T = F_{expected} - F$.
Συνεπώς οι υποθέσεις μας μπορούν πλέον να μεταφραστούν σε:
$H_0 : F = F_{expected}$
$H_1 : F > F_{expected}$
Όμως όπως είπαμε και πριν το $F_{expected}$ είναι σταθερός αριθμός, οπότε μπορούμε να τον αγνοήσουμε και να κάνουμε τον ισοδύναμο έλεγχο με ελεγχοσυνάρτηση το $T = -F$. Δεν διαλέγουμε την ελεγχοσυνάρτηση $T = F$ για καθαρά τεχνικούς λόγους. Προκειμένου να απορρίπτω την μηδενική υπόθεση για μικρές τιμές της T.

Ερώτηση 2

Μπορούμε να υποθέσουμε χωρίς περιορισμό της γενικότητας ότι η σωστή σειρά που θα έπρεπε να μαντέψει ο κύριος Χ τα είδη κρασιών προκειμένου να τα πετύχει όλα είναι να μαντέψει πρώτα το α κρασί, μετά το β κρασί, μετά το γ κρασί, μετά το δ κρασί, μετά το ε κρασί,μετά το ζ, μετά το η, μετά το θ, μετά το ι, μετά το κ (άρα η σειρά είναι η : (α,β,γ,δ,ε,ζ,η,θ,ι,κ)).
Έστω πάλι χωρίς περιορισμό της γενικότητας ότι ο κύριος Χ πέτυχε τα τέσσερα πρώτα και απέτυχε να βρει τα έξι τελευταία.
Η τιμή της ελεγχοσυνάρτησης της παρατήρησης λοιπόν είναι η $tobs$=-4 .

Κώδικας

```
p=numeric(100)
for (k in 1:100) {
  Cor_wine=c(letters[1:10])
  T=-4
  B=999
  Tobs=numeric(B)
  for (i in 1:B) {
    Wine=sample(Cor_wine)
    s=0
    for (j in 1:10) {
      if (Wine[j]==Cor_wine[j]){
        s=s+1
      }
    }
    Tobs[i]=-s
  }
  Pvalue=(length(Tobs[Tobs<=T])+1)/B
  p[k]=Pvalue
}
```

Κώδικας Συνέχεια

```
p
length(p[p>0.05])
hist(Tobs)
shapiro.test(p)
```

Εξηγώντας τον κώδικα (1)

Αρχικά αφού μου ζητείται να παρατηρήσω τα *pvalues* και να εξηγήσω για ποιο λόγο παρατηρώ τα συγκεκριμένα *pvalues*, αρχικοποιώ ένα διάνυσμα p το οποίο θα γεμίσω με όλα τα διαφορετικά *pvalues* που θα βρω σε 100 επαναλήψεις του προγράμματος.
Στην συνέχεια όπως και στο προηγούμενο ερώτημα δημιουργώ ένα διάνυσμα το οποίο θεωρώ πως περιέχει τα είδη των κρασιών στην σωστή σειρά.
Στην επόμενη γραμμή του κώδικα βάζω στην μεταβλητή T τον αντίθετο αριθμό της συχνότητας επιτυχίας του κυρίου X στην παρατήρηση που μας δίνεται από την εκφώνηση.
Αμέσως μετά αρχικοποιώ ένα διάνυσμα B διάστασης 999 στο οποίο θα εισάγω την τιμή της ελεγχοσυνάρτησης μου για 999 τυχαίες αναμεταθέσεις. Οι αναμεταθέσεις αυτές για να διευκρινήσω είναι σαν τυχαίες μαντεψιές που θα έκανε ο κύριος X αν δεν ήταν ειδικός στα κρασιά.
Μπορώ να πάρω οποιοδήποτε B αρκεί να είναι μικρότερο από το 10!

Εξηγώντας τον κώδικα (2)

Για κάθε μετάθεση που παίρνω τυχαία χρησιμοποιώντας την συνάρτηση *sample* ελέγχω πόσα είδη κρασιών βρίσκονται στην σωστή σειρά. Ο αντίθετος αριθμός του πλήθους αυτών των κρασιών αποτελεί την τιμή της ελεγχοσυνάρτησης για την συγκεκριμένη μετάθεση. Το *pvalue* μου τελικά είναι το πλήθος όλων των τιμών που δίνει η ελεγχοσυνάρτηση για τις μεταθέσεις οι οποίες είναι μικρότερες από την τιμή που δίνει η ελεγχοσυνάρτηση για την παρατήρηση που μας δινεται από την υπόθεση $+1$ και όλο αυτο διαιρεμένο με το B$+1$ που είναι ο αριθμός που μας ζητείτε να επαναλάβουμε την τυχαιοποίηση $+$ την παρατήρηση της υπόθεσης.

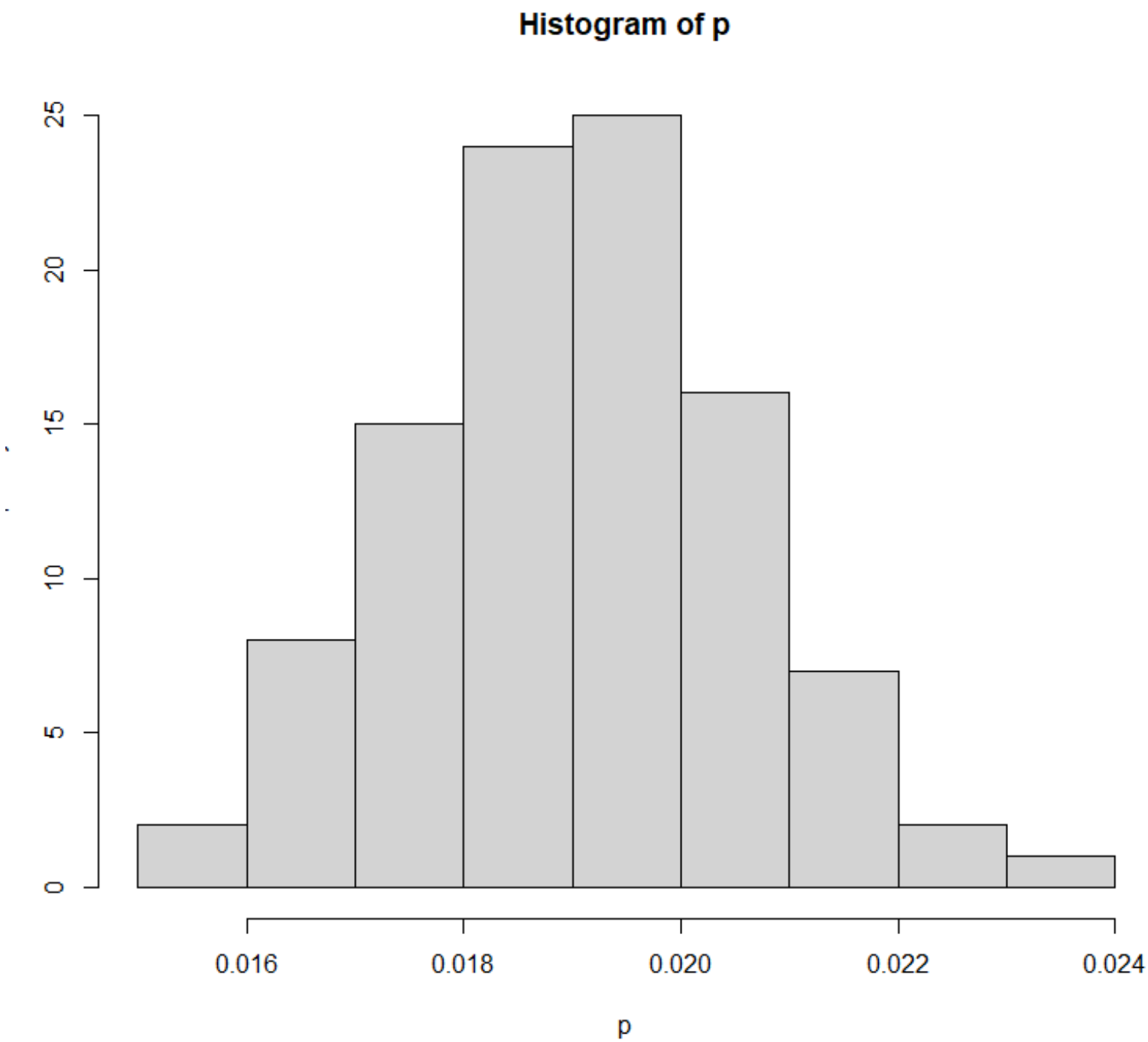
```
> p
  [1] 0.01801802 0.01901902 0.03103103 0.01601602 0.01701702 0.01801802
 0.01701702 0.01601602 0.02602603 0.02502503 0.02502503
 [12] 0.01601602 0.01501502 0.02702703 0.01701702 0.02202202 0.01601602
 0.02502503 0.01701702 0.02202202 0.02002002 0.01501502
 [23] 0.01701702 0.01601602 0.02502503 0.02202202 0.01601602 0.01201201
 0.02102102 0.01701702 0.02502503 0.02502503 0.01801802
 [34] 0.02302302 0.02502503 0.02202202 0.01901902 0.01501502 0.02202202
 0.02002002 0.01801802 0.01401401 0.02102102 0.02102102
 [45] 0.01401401 0.02302302 0.02202202 0.01801802 0.02302302 0.01501502
 0.02202202 0.02602603 0.02102102 0.01601602 0.01501502
 [56] 0.01901902 0.02202202 0.02702703 0.01801802 0.02302302 0.02302302
 0.02602603 0.02602603 0.02302302 0.01901902 0.01701702
 [67] 0.01801802 0.02002002 0.01501502 0.02502503 0.02402402 0.01701702
 0.02002002 0.03003003 0.02002002 0.01501502 0.02202202
 [78] 0.01801802 0.01601602 0.02202202 0.01701702 0.01501502 0.01801802
 0.02302302 0.01501502 0.01601602 0.02102102 0.01401401
 [89] 0.01401401 0.02002002 0.02102102 0.01901902 0.02202202 0.01701702
 0.02102102 0.01701702 0.01701702 0.02402402 0.02102102
[100] 0.02602603
> length(p[p>0.05])
[1] 0
```

Αποτελέσματα

Μπορούμε να δούμε πως τα *pvalues* είναι αρκετά μικρά ώστε να ισχυριστούμε πως ο κύριος Χ είναι ειδικός στο να αναγνωρίζει κρασιά, δηλαδή να απορρίψουμε την μηδενική υπόθεση. Η εικόνα εμφανίστηκε όταν έτρεξα το πρόγραμμα με B=999, δηλαδή 1000 επαναλήψεις. Αρχικά επέλεξα να το τρέξω για B=99, δηλαδή 100 επαναλήψεις αλλά εμφανίζονταν συχνά *pvalues* τα οποία ξεπερνούσαν την τιμή 0,05. Προφανώς όσο περισσότερο ανεβάζω την τιμή του B οι τιμές των *pvalues* θα γίνονται όλο και μικρότερες, αυτό είναι όμως αχρείαστο καθώς έχω ήδη τα αποτελέσματα που χρειάζομαι για να απορριψω την μηδενική υπόθεση με σιγουριά.

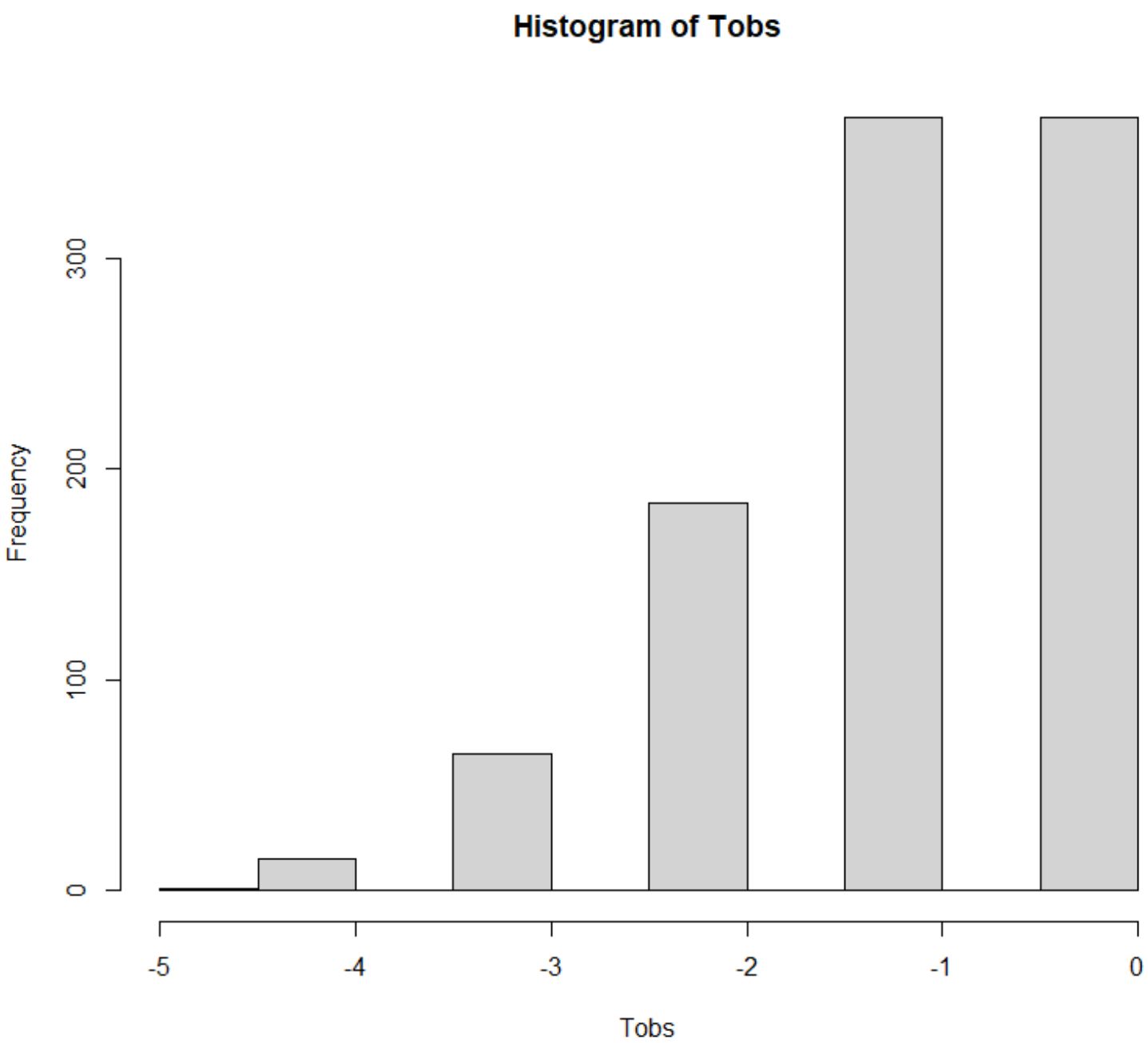
Αποτελέσματα

Ακόμα αν ανεβάσουμε το πλήθος των επαναλήψεων της τυχαιοποίησης στις 10000, μπορούμε να παρατηρήσουμε πως το ιστόγραμμα των 100 *pvalues* που αποθηκεύουμε μοιάζει όλο και περισσότερο με το ιστόγραμμα μίας ομοιόμορφης κατανομής. Δοκίμασα να απορρίψω αυτή την υπόθεση χρησιμοποιόντας την συνάρτηση $shapiro.test()$ όμως το *pvalue* που μου επιστράφηκε δεν επέτρεψε κάτι τέτοιο.
Στην επόμενη διαφάνεια παραθέτω την εικόνα του ιστογράμματος για 10000 επαναλήψεις.


Histogram of p
0
5
10
15
20
25
0.016
0.018
0.020
0.022
0.024
p

Κατανομή Ελεγχοσυνάρτησης

Κάνοντας ένα ιστόγραμμα μπορούμε να παρατηρήσουμε πως η κατανομή της ελεγχοσυνάρτησης παίρνει πολύ πιο συχνά μεγάλες τιμές από ότι μικρές. Όμως η ερώτηση που προκύπτει είναι πώς θα υπολογίσουμε ακριβώς την συνάρτηση πυκνότητας πιθανότητας της κατανομής της ελεγχοσυνάρτησης. Για να απαντήσουμε στο παραπάνο ερώτημα θα χρειαστούμε πρώτα να εξηγήσουμε την έννοια των διαταραχών.


Histogram of Tobs
Frequency
0
100
200
300
-5
-4
-3
-2
-1
0
Tobs

Χρήσιμοι Ορισμοί

Σταθερό σημείο σε μιά συνάρτηση s λέμε το σημείο για το οποίο ισχύει $s(x) = x$. Το x αυτό πρέπει να ανήκει στο πεδίο ορισμού και στο πεδίο τιμών την συνάρτησης.
Διαταραχές καλούνται όλες οι μεταθέσεις s, $n \in N$ στοιχείων οι οποίες δεν έχουν κανένα σταθερό σημείο. Δηλαδή: $s(x)! = x ... \forall x \in 1, 2, 3, ..., n$.

Υπολογισμός διαταραχών

Έστω $D(n)$ το πλήθος των διαταραχών n αντικειμένων τότε:

$$D(n) = n! \sum_{k=0}^{n} \frac{(-1)^k}{k!}$$

.

Τώρα που γνωρίζουμε όλα τα παραπάνω μπορούμε να ασχοληθούμε με τον υπολογισμό της κατανομής που μας ενδιαφέρει.

Υπολογισμός διαταραχών

Ο σκοπός μας είναι να μετρήσουμε την πιθανότητα σε μια τυχαία μετάθεση n στοιχείων να έχουμε κ αριθμό επιτυχιών. Δηλαδή κ στοιχεία να βρίσκονται στην σωστή θέση και τα υπόλοιπα στοιχεία να βρίσκονται σε λάθος θέση.
Ο δειγματικός χώρος του προβλήματος μας είναι όλες οι δυνατές μεταθέσεις n στοιχείων.

Υπολογισμός διαταραχών

Για να μετρήσουμε όλους τους τρόπους με τους οποίους σε μια τυχαία μετάθεση n στοιχείων έχουμε k αριθμό επιτυχιών, αρκεί να βρούμε με πόσους τρόπους μπορούμε να καλύψουμε τις σωστές θέσεις, οι οποίες καλύπτονται με $\binom{n}{k}$ τρόπους. Τα υπόλοιπα στοιχεία τα οποία έχουν πλήθος $(n-k)$ πρέπει να βρίσκονται σε λάθος θέσεις, δηλαδή να μην αποτελούν σταθερά σημεία σε μία μετάθεση πλήθους $(n-k)$ στοιχείων. Άρα χρειαζόμαστε $(n-k)$ διαταραχές, οι οποίες υπολογίζονται από τον τύπο:

$$D(n-k) = (n-k)! \sum_{r=0}^{n-k} \frac{(-1)^r}{r!}$$

.

Υπολογισμός Κατανομής

Άρα όλοι οι τρόποι με τους οποίους σε μια τυχαία μετάθεση n στοιχείων έχουμε k αριθμό επιτυχιών έχουν πλήθος:

$$\binom{n}{k} * (n-k)! \sum_{r=0}^{n-k} \frac{(-1)^r}{r!} = \frac{n!}{k!} * \sum_{r=0}^{n-k} \frac{(-1)^r}{r!}$$

.

Άρα για να βρούμε την πιθανότητα εμφάνισης k επιτυχιών αρκεί να διαιρέσουμε με το πλήθος των δυνατών μεταθέσεων. Δηλαδή το $n!$.

Υπολογισμός Κατανομής

Συνεπώς πλέον έχουμε υπολογίσει με ακρίβεια την συνάρτηση πυκνότητας πιθανότητας της κατανομής μας για μία μετάθεση n στοιχείων.

$$f[x = k] = \frac{1}{k!} * \sum_{r=0}^{n-k} \frac{(-1)^{r+1}}{r!}$$

Όπου $k \leq n$
Μπορούμε λοιπόν να παρατηρήσουμε πως όσο μεγαλώνει ο αριθμός επιτυχιών τόσο μεγαλώνει και η τιμή του $k!$ άρα τόσο μικραίνει η τιμή της πιθανότητας. Συνεπώς δεν αποτελεί έκπληξη που για μεγάλες τιμές του k οι πιθανότητες εμφάνισης είναι εξαιρετικά μικρές.