tp2

me

2023-01-31

Statistics for Biology 2: TP 2

Exercise 2

1. Upload tauber.csv. Assign column height to variable H. How many children in the sample are taller than 110 cm? How many have height 110 or less?

```
tauber <- read.table("tauber.csv", header = TRUE, sep = ";")
summary(tauber)
```

```
##     gender               age            height          weight
##  Length:2891        Min.   :49.00   Min.   : 96.0   Min.   :13.00
##  Class :character   1st Qu.:66.00   1st Qu.:110.0   1st Qu.:19.00
##  Mode  :character   Median :69.00   Median :114.0   Median :20.00
##                     Mean   :68.57   Mean   :113.7   Mean   :20.67
##                     3rd Qu.:72.00   3rd Qu.:117.0   3rd Qu.:22.00
##                     Max.   :86.00   Max.   :139.0   Max.   :40.00
```

```
H <- tauber$height
length(H[H>110])
```

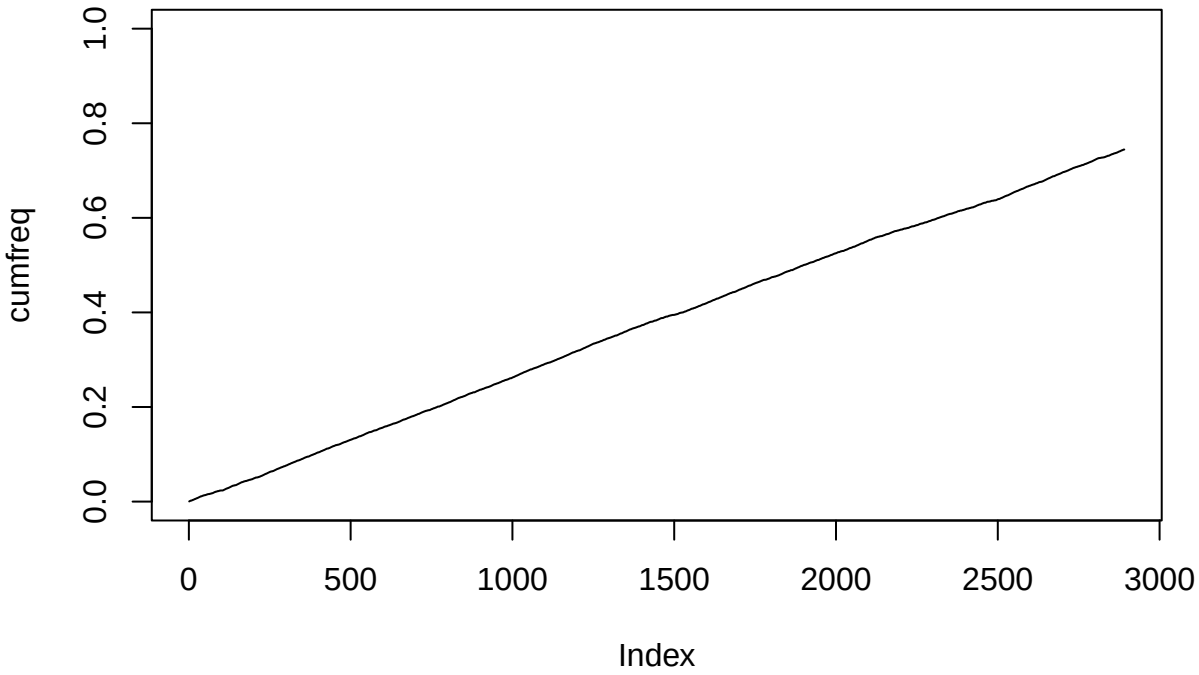
```
## [1] 2152
```

```
length(H[H<=110])
```

```
## [1] 739
```

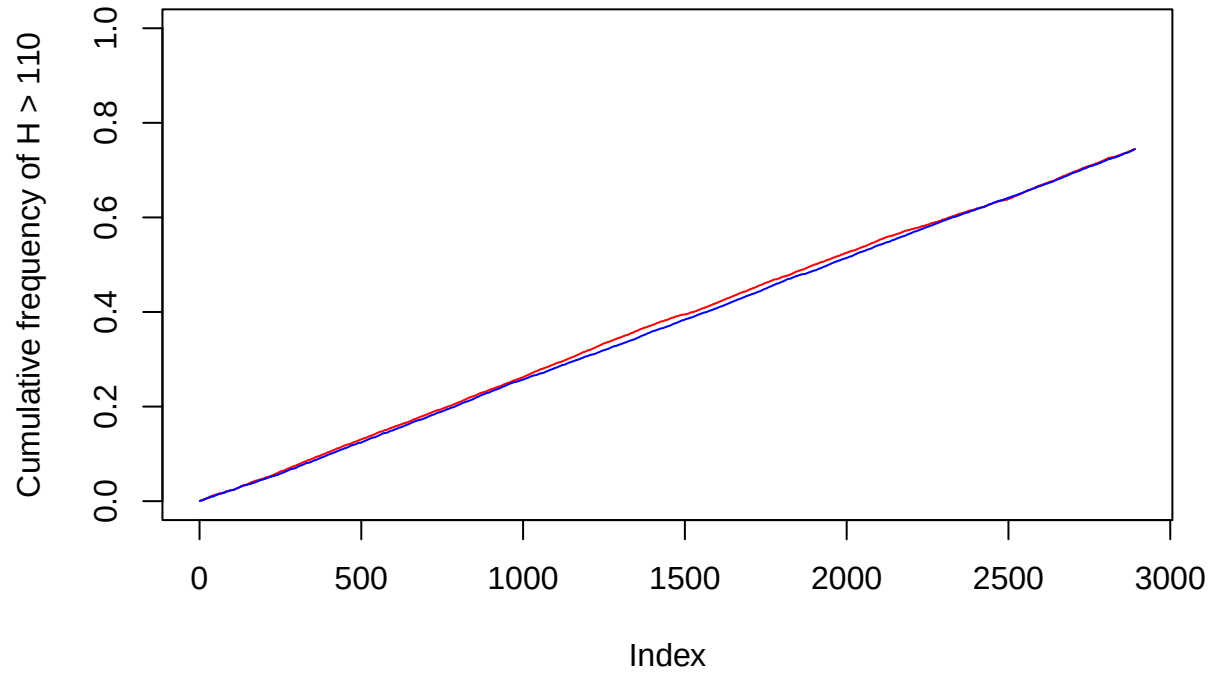
2. Plot cumulated frequencies for the event H>110, with ordinates in the interval (0,1): plot option ylim=c(0,1).

```
cumfreq <- cumsum(H>110)/length(H)
plot(cumfreq, ylim=c(0,1), type="l")
```



3. Assign to vector rH a random permutation of H. Superpose on the same graphics cumulated frequencies for rH>110, in blue.

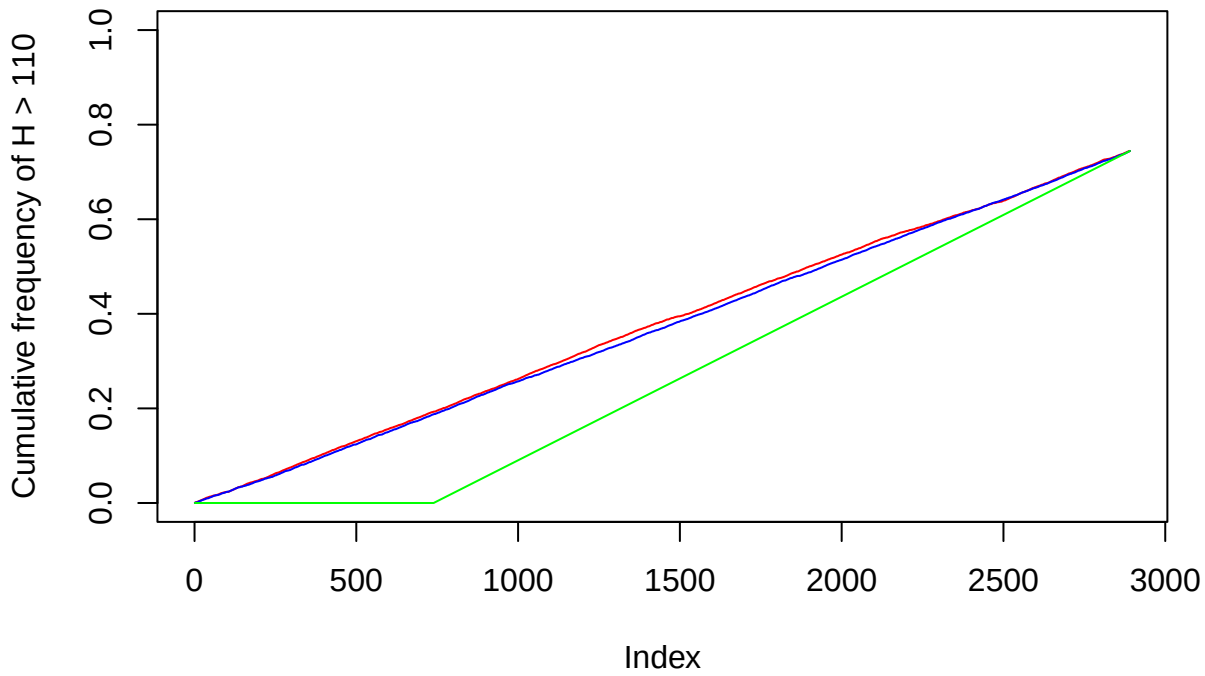
```
set.seed(420)
rH <- sample(H)
plot(cumfreq, ylim=c(0,1), type="l", col="red", ylab = "Cumulative frequency of H > 110")
lines(cumsum(rH>110)/length(rH), col="blue")
```



4. Assign to vector iH the values of H, sorted in increasing order. Superpose on the same graphics cumulated frequencies for iH>110, in green. In which interval is the green curve constant?

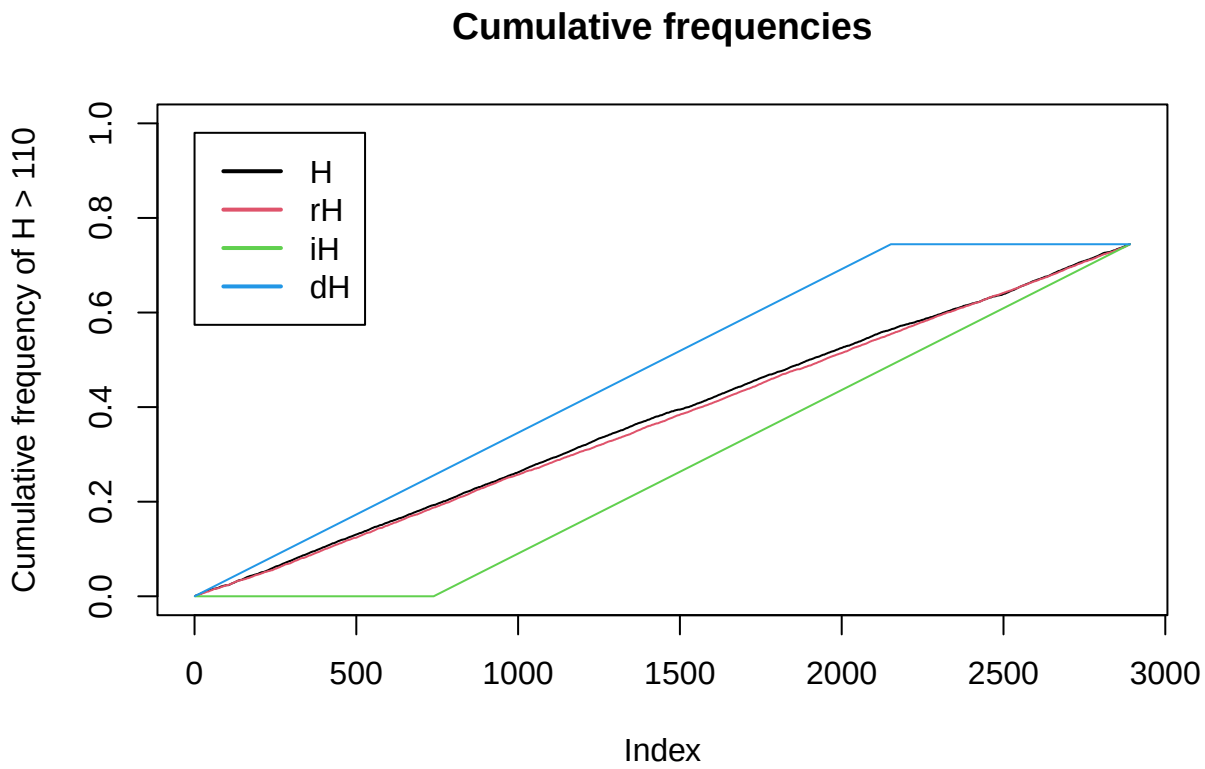
```
iH <- sort(H)
plot(cumfreq, ylim=c(0,1), type="l", col="red", ylab = "Cumulative frequency of H > 110")
```

```
lines(cumsum(rH>110)/length(rH), col="blue")
lines(cumsum(iH>110)/length(iH), col="green")
```



5. Assign to vector dH the values of H, sorted in decreasing order. Superpose on the same graphics cumulated frequencies for dH>110, in red. In which interval is the red curve constant ?

```
dH <- sort(H, decreasing=TRUE)
plot(cumfreq, ylim=c(0,1), type="l", col=1, ylab = "Cumulative frequency of H > 110", main="Cumulative
lines(cumsum(rH>110)/length(rH), col=2)
lines(cumsum(iH>110)/length(iH), col=3)
lines(cumsum(dH>110)/length(dH), col=4)
legend(x=0, y=0.98, legend=c("H", "rH", "iH", "dH"), col=1:4, lwd=2)
```



Exercise 3 . Import and export of external data

Import the imcenfant.txt file using the read.table() function. Pay attention to the header, sep, dec, row.names parameters.

```
D <- read.table("imcenfant.txt", header = TRUE, dec=",")
D$SEXE <- factor(D$SEXE)
D$zep <- factor(D$zep)
```

1. Display the contents of this object, which will be named D.

```
# View(D)
summary(D)
```

```
##  SEXE   zep         poids             an             mois
##  F:71   N: 41   Min.   :10.50   Min.   :3.000   Min.   : 0.000
##  G:81   O:111   1st Qu.:15.00   1st Qu.:3.000   1st Qu.: 3.000
##                 Median :16.00   Median :3.000   Median : 6.000
##                 Mean   :16.28   Mean   :3.303   Mean   : 5.618
##                 3rd Qu.:17.50   3rd Qu.:4.000   3rd Qu.: 9.000
##                 Max.   :22.80   Max.   :4.000   Max.   :11.000
##      taille
##  Min.   : 88.5
##  1st Qu.: 98.0
##  Median :101.0
##  Mean   :100.7
##  3rd Qu.:103.6
##  Max.   :111.5
```

2. Verify the names of the columns of D. Name the rows.

```
colnames(D)
```

```
## [1] "SEXE"   "zep"    "poids"  "an"     "mois"   "taille"
```

```
rownames(D) <- 1:length(D[,1])
```

3. Add a IMC column to D. Extract from D the lines for children with IMC<15 and age<=3.5. Give the number of childrens checking the above conditions.

```
D$IMC <- D$poids/(D$taille*0.01)^2
D$age <- D$an + D$mois*0.1
D[D$IMC>15 & D$age<=3.5,]
```

```
##     SEXE zep poids an mois taille      IMC age
## 1      F   0  16.0  3    5  100.0 16.00000 3.5
## 10     G   0  16.7  3    3  100.0 16.70000 3.3
## 22     G   0  16.8  3    5  101.5 16.30712 3.5
## 58     G   0  15.0  3    3   98.0 15.61849 3.3
## 60     F   0  17.0  3    2  103.0 16.02413 3.2
## 61     G   0  14.5  3    4   98.0 15.09788 3.4
## 70     F   0  16.7  3    4  100.0 16.70000 3.4
## 72     F   0  16.6  3    4   98.0 17.28446 3.4
## 73     F   0  17.0  3    4  100.0 17.00000 3.4
## 75     F   0  16.0  3    3   98.0 16.65973 3.3
## 77     F   0  17.0  3    4  100.5 16.83127 3.4
## 81     G   0  15.0  3    3   98.0 15.61849 3.3
## 87     G   0  14.5  3    2   92.0 17.13138 3.2
## 88     G   0  17.0  3    3   99.0 17.34517 3.3
## 89     G   0  19.0  3    4  107.0 16.59534 3.4
## 90     F   0  18.0  3    3  100.0 18.00000 3.3
## 97     G   0  14.5  3    2   95.5 15.89869 3.2
## 131    G   0  18.5  3    5  104.0 17.10429 3.5
## 134    G   N  13.0  3    3   92.0 15.35917 3.3
## 138    G   N  16.5  3    1  101.0 16.17488 3.1
```

4. Export the new data table to your directory as a text file.

```
write.table(D, "new_imcenfant.txt")
```

Exercise 4 Data analysis

1. Save in your working directory the Poids_naissance.txt file available in AmeTICE
2. Import this data set into R, creating an object that will be of data.frame type, named E. To do this, use the read.table() function by correctly filling the parameters header, sep, dec, row.names

```
E <- read.table("Poids_naissance.txt", header=TRUE)
row.names(E) <- E$ID
```

3. Add a PTL1 (number of preterm antecedents) variable with three modalities (where the third will be rated 2 and will correspond to 2 or more preterm antecedents).

```
E$PTL1 <- E$PTL
E$PTL1[E$PTL1 > 2] <- 2
```

4. Same question for FVT (number of visits to a doctor) to add FVT1.

```
E$FVT1 <- E$FVT
E$FVT1[E$FVT1 > 2] <- 2
```

5. Order the data frame according to increasing birth weights (BWT)

```
res1 <- E[order(E$BWT),]
```

6. Extract individuals with black or white mothers who smoke.

```
res2 <- E[E$RACE>1 & E$SMOKE==1,]
```

Exercise 1

From past experience, it is known that a certain surgery has a 90% chance to succeed. This surgery is going to be performed on 5 patients. Let X be the random variable equal to the number of successes out of the 5 attempts.

1. Which probability distribution do you propose as a model for X? What are the values? What are the probabilities of the different values? What is their sum?

```
set.seed(42) # for reproducibility
pbinom(0:5, 5, 0.9)
```

```
## [1] 0.00001 0.00046 0.00856 0.08146 0.40951 1.00000
```

```
sum(pbinom(0:5, 5, 0.9))
```

```
## [1] 1.5
```

2. Compute the theoretical mean, variance, standard-deviation, median, first and third quartiles of that distribution. Why are the median and third quartile both equal to 5?

```
summary(rbinom(100, p=0.9, size=5))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00    4.00    5.00    4.42    5.00    5.00
```

3. What is the probability that each of the 5 surgeries will be successful?

```
dbinom(x=(0:5), p=0.9, size=5)
```

```
## [1] 0.00001 0.00045 0.00810 0.07290 0.32805 0.59049
```

3 of the 5 will be successful ?

```
dbinom(3, p=0.9, size=5)
```

```
## [1] 0.0729
```

at most 3 surgeries will be successful?

```
pbinom(q=3, prob=0.9, size=5)
```

```
## [1] 0.08146
```

at least 3 will be successful?

```
pbinom(q=3, prob=0.9, size=5, lower.tail = FALSE)
```

```
## [1] 0.91854
```

from 2 to 4 will be successful?

```
sum(dbinom(x=(2:4), p=0.9, size=5))
```

```
## [1] 0.40905
```

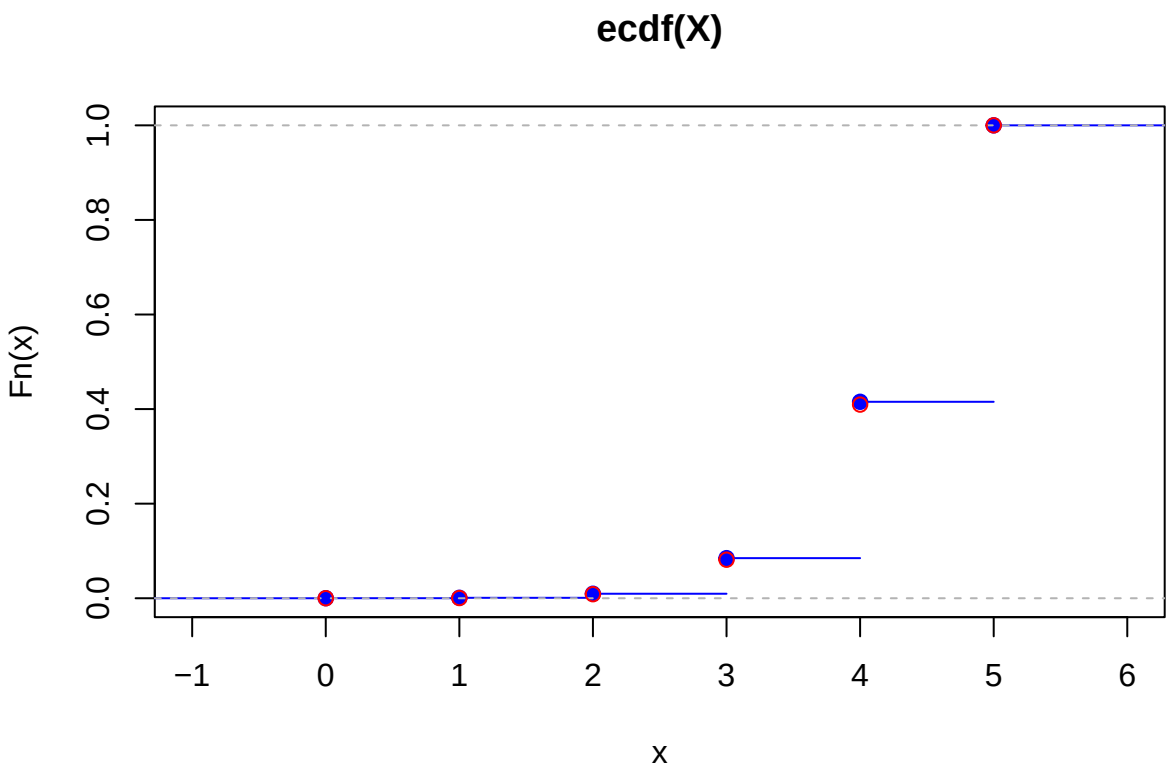
4. Assign to X a simulated sample of size N=100 of the binomial distribution with parameters 5 and 0.9. (rbinom())Compute the relative frequencies of the different values. Compare with the theoretical probabilities. Repeat (several times) for N=1e4, N=1e6.

```
for (n in rep(c(100, 1e4, 1e6), times = 3)) {
  X <- rbinom(n, 5, 0.9)
  print(table(X)/sum(X))
}
```

```
## X
##          3          4          5
## 0.01569507 0.08968610 0.11883408
## X
##            1            2            3            4            5
## 0.0001111976 0.0017791616 0.0168575559 0.0724785945 0.1311686868
## X
##            0            1            2            3            4            5
## 2.888713e-06 9.399428e-05 1.763893e-03 1.622323e-02 7.291489e-02 1.312098e-01
## X
##           3           4           5
## 0.008714597 0.071895425 0.137254902
## X
##            1            2            3            4            5
## 6.692694e-05 1.873954e-03 1.753486e-02 7.448968e-02 1.291244e-01
## X
##            0            1            2            3            4            5
## 3.111527e-06 1.017914e-04 1.816465e-03 1.624684e-02 7.289374e-02 1.311900e-01
## X
##           3           4           5
## 0.008695652 0.069565217 0.139130435
## X
##            1            2            3            4            5
## 0.0001114852 0.0021628130 0.0176369596 0.0726437602 0.1304153939
## X
##            0            1            2            3            4            5
## 2.666325e-06 1.066530e-04 1.801992e-03 1.620082e-02 7.272136e-02 1.313603e-01
```

5. Assign to X a simulated sample of size N=1e4. Plot the ecdf of X in blue. Superpose the theoretical cumulative probabilities of the binomial distribution with parameters 5 and 0.9 as red points.

```
X <- rbinom(1e4, 5, 0.9)
plot(ecdf(X), col="blue")
points(0:5, pbinom(0:5, 5, 0.9), col="red")
```


ecdf(X)
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
−1
0
1
2
3
4
5
6
x