Eloszlások jellemzése

(1) Az alábbi állítások közül mennyi igaz? A szórás a várható értektől való átlagos négyzetes eltérés. A medián az elsõ (alsó) kvartilis. A relatív gyakorisági hisztogram a sűrűségfüggvény becslése. Boxdiagram esetén a dobozban van az adatok pontosan 25%-a. A vércsoport és az elmúlt hónapban Csongrád megyében lehullott csapadék mennyisége diszkrét változók.

##Eredmény: 1 "A relatív gyakorisági hisztogram a sűrűségfüggvény becslése."

---

(2) Olvassuk be a „mice” csomagból a „nhanes” adatbázist. Adjuk meg az adatbázisban lévő hiányzó értékek számát.

##Módszer:

> library(mice)

> input = data.frame(nhanes)

> attach(input)

> a = nhanes[is.na(nhanes)]

> length(a)

##Eredmény: 27

---

(3) Olvassuk be a „MASS” csomagból a „survey” adatbázist, mely egyetemi hallgatóktól származó adatokat tartalmaz. Két tizedes jegyre kerekítve adjunk pontbecslést a fiúk („Sex” = „Male”) életkorának („Age”) szórására.

##Módszer:

```
> library(MASS)
> input = data.frame(survey)
> attach(input)
> sd(Age[Sex == "Male"], na.rm = TRUE)
```

Eredmény: 6.07

---

(4) Olvassuk be a „MASS” csomagból a „survey” adatbázist, mely egyetemi hallgatóktól származó adatokat tartalmaz. A „Pulse” változó a hallgatók pulzusát adja meg. A pulzusszámok között van kiugró érték (outlier)? Ha igen, adjuk meg a második legkisebb outlier értékét, ha nincs, válaszként adjunk 0-t.

##Módszer:

```
> library(MASS)
> input = data.frame(survey)
> attach(input)
> fiuk = survey[survey[,1] == "Male", 6]
> boxplot(fiuk)
```

Eredmény: 40

---

(5) Olvassuk be a „datasets” csomagból a „ToothGrowth” adatbázist, mely tengerimalacok foghosszúságával kapcsolatos kísérlet eredményét tartalmazza. Két tizedes jegyre kerekítve adjunk pontbecslést a foghosszúság („len” változó) várható értékére.

##Módszer:

> library(datasets)

> input = data.frame(ToothGrowth)

> attach(input)

> mean(len)

Eredmény: 18.81

Várható értékek

(1) Az alábbi állítások közül melyik igaz a szignifikancia szintre (? -ra)?

##Válasz: A hibás döntés valószínűsége, ha a nullhipotézis igaz, vagyis annak a valószínűsége, hogy tévesen elvetjük az igaz nullhipotézist.

---

(2) Olvassuk be a „car” csomagból a „Blackmore” adatbázist. Vizsgáljuk az „exercise” változót a „group” változó csoportjai szerint. Az alábbiak közül melyik próbával tesztelhetjük a várható értékek azonosságát?

##Válasz: Egyik sem

---

(3) Egy vizsgálatban arra voltak kíváncsiak, hogy változik-e az átlagos reakcióidő hipnózis hatására. 50 páciens reakcióidejét mérték meg éberen és hipnózisban. Az átlagos reakcióidő éberen 360, hipnózisban 320, a reakcióidők különbségeinek szórása pedig 30. Normális eloszlást feltételezve készítsünk 99% megbízhatósági szintû konfidencia intervallumot a várható értékek különbségére. Két tizedesjegyre kerekítve adjuk meg a kapott konfidencia intervallum felső végpontját.

##Módszer:

```
> konf.int_felso = function(n, xvonas, sd, alpha) { xvonas + qt(1-alpha/2, df = n-1)*sd/sqrt(n) }
```

```
> konf.int_felso(50, 360-320, 30, 0.01)
```

##Eredmény: 51.37

Olvassuk be a „datasets” csomagból a „ToothGrowth” adatbázist, mely tengerimalacok foghosszúságával kapcsolatos kísérlet eredményét tartalmazza. Normális eloszlást feltételezve, 5%-os szignifikancia szinten teszteljük azt a nullhipotézist, hogy a foghosszúság („len” változó) várható értéke azonos a különbözõ kapott dózisok esetén („dose”) változó. Két tizedesjegyre kerekítve adjuk meg a kapott F-statisztika értékét.

##Módszer:

```
> library(datasets)
```

```
> input = data.frame(ToothGrowth)
```

```
> attach(input)
```

```
> oneway.test(len ~ dose, var.equal = TRUE)
```

##Eredmény: 67.42

Vegyes feladatok

(1) Egy felmérés szerint Amerikában a mobiltelefonok 50%-án fut Android, 45%-án Apple iOS, 3%-án Windows és 2%-án másfajta operációs rendszer. Kíváncsiak vagyunk, hogy Magyarországon ugyanezek az arányok tapasztalhatóak-e, ezért megkérdezzük 70 ismerõsünket. Azt kapjuk, hogy közülük 52 Android, 9 Apple iOS, 5 Windows, 4 pedig másfajta operációs rendszert használ. Teszteljük 5 százalékos szignifikancia szinten azt a nullhipotézist, hogy hazánkban ugyanaz az operációs rendszerek megoszlása, mint Amerikában.

##Válasz: A várt gyakoriságok némelyike túl alacsony, ezért nem tudunk megbízható döntést hozni.

---

(2) Az admission.csv adatbázis üzleti fõiskolákra jelentkezõk kétféle pontszámát tartalmazza (GPA és GMAT), valamint azt, hogy a jelentkezõ felvételt nyert-e (De = admit, notadmit, borderline), letölthetõ a következõ webhelyrõl:

http://www.biz.uiowa.edu/faculty/jledolter/DataMining/admission.csv

5%-os szignifikancia szinten teszteljük azt a nullhipotézist, hogy a GPA pontszámok az 'admit' csoportban normális eloszlásból származnak, majd adjuk meg a kapott p-értéket.

##Módszer:

```
> url = "http://www.biz.uiowa.edu/faculty/jledolter/DataMining/admission.csv"

> admission = read.csv(url)

> attach(admission)

> v = GPA[De=="admit"]
> ks.test(v,"pnorm",mean(v),sd(v))
```

##Eredmény: 0.9699

---

(3) James Bond szerint a martini jobb rázva, mint keverve. Azt szeretnénk tesztelni, hogy tényleg tud-e különbséget tenni a kétféle elkészítési mód között. 16 pohár martini mindegyikérõl el kell döntenie, hogy rázva vagy keverve készült. A 16 esetbõl 13-szor helyesen dönt. Teszteljük azt a nullhipotézist, hogy James Bond csak véletlenszerûen találgat. Az alábbi állítások közül melyik igaz?

##Válasz: p < 0,05, elutasítjuk a nullhipotézist, James Bond nem csak találgat.

---

(4) Egy új gyógyszer hatásának tesztelése érdekében megmérték 15 beteg szisztolés vérnyomását kezelés elõtt és után. Normális eloszlást feltételezve, milyen statisztikai próbát használ annak eldöntésére, hogy a kezelés hatásos volt vagy nem?

##Válasz: páros t-próba

---

(5) Az admission.csv adatbázis üzleti fõiskolákra jelentkezõk kétféle pontszámát tartalmazza (GPA és GMAT), valamint azt, hogy a jelentkezõ felvételt nyert-e (De = admit, notadmit, borderline), letölthetõ a következõ webhelyrõl:

http://www.biz.uiowa.edu/faculty/jledolter/DataMining/admission.csv

Normális eloszlást feltételezve hasonlítsuk össze az 'admit' és a 'notadmit' csoport átlagos GPA pontszámát, majd három tizedes jegyre kerekítve adjuk meg a kapott próbastatisztika értékét.

##Módszer:

```
> rm(list=ls())
> url = "http://www.biz.uiowa.edu/faculty/jledolter/DataMining/admission.csv"
> database = read.csv(url)
> egyesek = database[database[,3]=="admit", 1]
> kettesek = database[database[,3]=="notadmit", 1]
> t.test(egyesek, kettesek, conf.level = 0.9, var.equal = TRUE)
```

(t = megoldás)

##Eredmény: 17.926

---

(6) Az alábbi állítások közül melyik igaz a QQ-ábrára?

##Válasz: n-elemû minta esetén a QQ-ábra i-edik pontjának y-koordinátája a rendezett minta i-edik eleme.