Háskóli Íslands

STÆ203G Líkindareikningur og tölfræði

HAG206G Líkindareikningur og tölfræði

MAS201F Líkindareikningur og tölfræði

Vor 2016

Tölvuverkefni 3 - fyrir R notendur

Sett fyrir: Miðvikudaginn 2. mars 2016.

Skiladagur: Þriðjudagurinn 15. mars 2016.

Verkefnið

Í þessari æfingu lærið þið að nota lýsandi tölfræði með aðstoð R fyrir mælingar á samfelldri breytu í einni vídd. Þið eigið að beita tölulegum aðferðum við að reikna helstu reiknistærðir og myndrænum aðferðum til að teikna myndir sem sýna gögnin á lýsandi hátt.

1. Sólarhringsúrkoma hvers dags er mæld klukkan 9:00 á morgnana. Mælingar á mestu sólarhringsúrkomu (mm) á Fagurhólsmýri innan hvers árs fyrir árin 1924 til 2007 ($n = 84$) er að finna í gagnaskránni `precipitation_fagurhm.txt`. Ein leið til að lesa gögnin inn er að nota

```
max_urkoma <- read.table("precipitation_fagurhm.txt")
```

Notið þessi gögn til að reikna stærðirnar sem eru gefna hér fyrir neðan. Látum x standa fyrir árlega hámarkssólarhringsúrkomu og x_i er þá i-ta mælingin í úrtakinu.

   (a) Reiknið eftirfarandi stærðir og setjið þær saman í eina töflu;
   meðaltal úrtaks, $\bar{x}$. R: `mean`
   miðgildi úrtaks, $\tilde{x}$. R: `median`
   dreifni úrtaks, s^2. R: `var`
   staðalfrávik úrtaks, s. R: `sd`
   fyrsta fjórðungsmark úrtaks, Q_1. R: `quantile`
   þriðja fjórðungsmark úrtaks, Q_3. R: `quantile`
   fjórðungsbil úrtaks, IQR.

   (b) Hvort er meðaltal úrtaksins stærra eða minna en miðgildi úrtaksins? Hvers vegna er meðaltal úrtaksins stærra (minna) en miðgildi úrtaksins fyrir þessi gögn?

   (c) Á hvaða bili liggja 50% af mælingunum sem eru sitthvoru megin við miðgildið?

   (d) Reiknið logarann af hverju staki í gögnunum og kallið nýju stökin y ($y = \ln(x)$). Reiknið sömu stærðir og hér fyrir ofan fyrir y, þ.e., reiknið meðalgildi úrtaks, $\bar{y}$, miðgildi úrtaks, $\tilde{y}$, dreifni úrtaks, s_y^2, staðalfrávik úrtaks, s_y, fyrsta fjórðungsmark úrtaks, $Q_{1,y}$, þriðja fjórðungsmark úrtaks, $Q_{3,y}$, fjórðungsbil úrtaks, IQR_y, og setjið þær saman í aðra töflu.

2. Hér á að teikna myndir af gögnunum. Byggt á myndunum á að meta hvort að normaldreifingin lýsi gögnunum nægjanlega vel. Einnig á að meta hvort að normaldreifingin lýsi logranum af gögnunum nægjanlega vel. Teiknið eftirfarandi myndir.

   (a) Teiknið árlega hámarkssólarhringsúrkomu á móti tíma (nota ártalið). Á hvaða ári rigndi mest?

(b) Kassarit. Sýnir kassaritið einhverja útlaga? (R: `boxplot`).

(c) Tíðnirit með flatarmál sem er jafnt einum. Teiknið ofan á tíðniritið þéttifall normaldreifingar með meðalgildi jafnt og $\bar{x}$ og staðalfráviki jafnt og s. Hér má nota eftirfarandi skipanir í R:

```
x_dat <- max_urkoma[,2]
h <- hist(x_dat, plot=F)
ylim <- range(0, h$density, dnorm(0)/sd(x_dat))
hist(x_dat, freq=F, ylim=ylim)
curve(dnorm(x, mean=mean(x_dat), sd=sd(x_dat)), add=T)
```

(d) Dreififall úrtaks. Teiknið ofan á dreififall úrtaksins dreififall normaldreifingar með meðalgildi jafnt og $\bar{x}$ og staðalfráviki jafnt og s. Hér má nota eftirfarandi skipanir í R:

```
n <- length(x_dat)
x_sort <- sort(x_dat)
plot(x_sort,(1:n)/n,type="s",ylim=c(0,1))
```

og bæta svo við skipunum byggðum á `pnorm`.

(e) Sætisfall úrtaks. Teiknið ofan á sætisfall úrtaksins sætisfall normaldreifingar með meðalgildi jafnt og $\bar{x}$ og staðalfráviki jafnt og s. Hér má nota eftirfarandi skipanir í R:

```
n <- length(x_dat)
x_sort <- sort(x_dat)
p <- (1:n)/(n + 1)
plot(p,x_sort,type="s",xlim=c(0,1))
```

og bæta svo við skipunum byggðum á `qnorm` sem gefur gildi á sætisfalli normaldreifingar sem fall af `p`.

(f) Teiknið sömu myndir og lýst er í liðum (a), (b), (c), (d) og (e) fyrir y.

(g) Lýsir normaldreifing gögnunum nægjanlega vel? Lýsir lognormaldreifing gögnunum nægjanlega vel? Notið myndirnar úr liðunum hér fyrir ofan til að rökstyðja svör ykkar. Þegar gert er ráð fyrir að logrinn af slembistærð fylgi normaldreifingu þá fylgir slembistærðin lognormaldreifingu.

3. Hér á að teikna tvær myndir, þá fyrri byggða á gögnunum og hina byggða á logranum af gögnunum. Látum $x_{(1)} < x_{(2)} < ... < x_{(n)}$ tákna röðuðu gögnin. (R: `x_sort <- sort(x_dat)`)

   (a) Teiknið $x_{(i)}$ á móti $\mu + \sigma\Phi^{-1}(i/(n+1))$, $i = 1, ..., n$. Teiknið á sömu mynd línu með hallatölu 1 og skurðpunkt 0. (R: Φ^{-1}: `qnorm`)

   Ábending: Metið μ með $\bar{x}$ og σ með s.

(b) Teiknið $x_{(i)}$ á móti $\exp(\mu_y + \sigma_y \Phi^{-1}(i/(n+1)))$, $i = 1, ..., n$. Teiknið á sömu mynd línu með hallatölu 1 og skurðpunkt 0.

Ábending: Metið μ_y með $\bar{y}$ og σ_y með s_y.

(c) Því nær sem að punktarnir eru beinu línunni því betur lýsir dreifingin sem er lögð til gögnunum. Notið þessar tvær myndir til að svara eftirfarandi spurningum. Lýsir normaldreifing gögnunum nægjanlega vel? Lýsir lognormaldreifing gögnunum nægjanlega vel?

4. Metið líkurnar á því að sólarhringsúrkoma á Fagurhólsmýri sé meiri en 115 mm, það er, metið $P(X > 115)$

(a) með því að finna hlutfall mælinga stærri en 115 mm.

(b) með því að gera ráð fyrir að normaldreifingin lýsi mælingunum nægjanlega vel.

Ábending:
$$P(X > 115) = 1 - \Phi\left(\frac{115 - \mu}{\sigma}\right)$$
Metið μ með $\bar{x}$ og σ með s.

(c) með því að gera ráð fyrir að normaldreifingin lýsi logranum af mælingunum nægjanlega vel.

Ábending:
$$P(X > 115) = 1 - \Phi\left(\frac{\ln(115) - \mu_y}{\sigma_y}\right)$$
Metið μ_y með $\bar{y}$ og σ_y með s_y.

(d) Ef þið ættuð að meta $P(X > 130)$, hvaða aðferð munduð þið nota? Aðferðina í (a), (b), (c) eða einhverja aðra aðferð? Útskýrið val ykkar.