Computerzitting 1:
Kansverdelingen en transformaties in R
2022-2023

1 Kansen en kwantielen

1.1 Kansen van de normale verdeling, t-verdeling χ^2-verdeling en F-verdeling

We willen de kans berekenen dat een normaal verdeelde toevalsvariabele $X \sim N(\mu, \sigma^2)$ waarden aanneemt in een bepaald interval. Hiervoor maken we gebruik van de normale verdelingsfunctie, aangezien $P(X \leq a) = F(a)$ en $P(a < X < b) = F(b) - F(a)$. We zullen hiervoor het commando `pnorm` gebruiken.

1. Normale verdeling

   - Om de kans te berekenen dat $X < x$ voor $X \sim N(\mu, \sigma^2)$ geven we het volgende in:

     ```
     pnorm(x,mean=mu,sd=sigma,lower.tail=TRUE)
     ```

   - Willen we echter $P(X > x)$ berekenen, dan typen we

     ```
     pnorm(x,mean=mu,sd=sigma,lower.tail=FALSE)
     ```

2. t-verdeling

   - Om de kans te berekenen dat $X < x$ voor $X \sim t_n$ geven we het volgende in:

     ```
     pt(x,n)
     ```

   - Willen we echter $P(X > x)$ berekenen, dan typen we

     ```
     pt(x,n,lower.tail=FALSE)
     ```

3. χ^2 verdeling

   - Om de kans te berekenen dat $X < x$ voor $X \sim \chi^2_n$ geven we het volgende in:

     ```
     pchisq(x,n)
     ```

   - Willen we echter $P(X > x)$ berekenen, dan typen we

```
pchisq(x,n,lower.tail=FALSE)
```

4. F-verdeling

   - Om de kans te berekenen dat $X < x$ voor $X \sim F_{m,n}$ geven we het volgende in:

     ```
     pf(x,n,m)
     ```

   - Willen we echter $P(X > x)$ berekenen, dan typen we

     ```
     pf(x,n,m,lower.tail=FALSE)
     ```

5. Voor elke andere kans, bv. $P(a < X < b)$ moet je gebruik maken van de regels die zijn gezien tijdens de oefenzitting.

Oefening 1 Bereken de volgende kansen in R:

- $P(X < 8)$ als $X \sim N(5, 4)$

  .....

- $P(X \leq 3.45)$ als $X \sim t_6$. Doe hetzelfde voor $X \sim \chi^2_5$.

  .....

  Opmerking: Een vierkantswortel wordt door het commando `sqrt` ingegeven. Zo staat `sqrt(3)` voor $\sqrt{3}$.

- $P(X > 2.5)$ als $X \sim F_{25,2}$

  .....

- $P(2 < X < 5)$ als $X \sim N(3, 4)$

  .....

1.2 Kwantielen

1. Normale verdeling

   - Om de kwantielen van de normale verdeling te berekenen, passen we eenzelfde werkwijze toe. We gebruiken nu enkel het commando `qnorm` dat $\Phi^{-1}(p)$ berekent voor $0 < p < 1$.
   - Gegeven een $a \in (0, 1)$ zoeken we het bijbehorend $100a\%$ kwantiel, m.a.w. zoek $x \in \mathbb{R}$ zodat $P(X < x)a$. Dit verkrijg je in R door `qnorm(a,mean=mu,sd=sigma,lower.tail=TRUE)` in te geven. We kunnen ook x vinden zodat $P(X > x) = a$. Dit gaat met de volgende code `qnorm(a,mean=mu,sd=sigma,lower.tail=FALSE)`.

2. t-verdeling

   - Gegeven een $a \in (0,1)$ zoeken we het bijbehorend $100a\%$ kwantiel, m.a.w. zoek $x \in \mathbb{R}$ zodat $P(X < x) = a$ met $X \sim t_n$. Dit verkrijg je in R door `qt(a,n)`. We kunnen ook x vinden zodat $P(X > x) = a$. Dit gaat met de volgende code `qt(a,n,lower.tail=FALSE)`.

3. χ^2 verdeling

   - Gegeven een $a \in (0,1)$ zoeken we het bijbehorend $100a\%$ kwantiel, m.a.w. zoek $x \in \mathbb{R}$ zodat $P(X < x) = a$ met $X \sim \chi^2_n$. Dit verkrijg je in R door `qchisq(a,n)`. We kunnen ook x vinden zodat $P(X > x) = a$. Dit gaat met de volgende code `qchisq(a,n,lower.tail=FALSE)`.

4. F-verdeling

   - Gegeven een $a \in (0,1)$ zoeken we het bijbehorend $100a\%$ kwantiel, m.a.w. zoek $x \in \mathbb{R}$ zodat $P(X < x) = a$ met $X \sim F_{m,n}$. Dit verkrijg je in R door `qf(a,n,m)`. We kunnen ook x vinden zodat $P(X > x) = a$. Dit gaat met de volgende code `qf(a,n,m,lower.tail=FALSE)`.

Oefening 2 Bereken volgende kwantielen (percentielen):

- Bereken x zodanig dat $P(X < x) = 0.1 = a$, waarbij $X \sim N(7,1)$.
  Oplossing: $x =$ .....
- Indien X normaal verdeeld is met $\mu = 25$ en $\sigma = 5$, wat is dan
  1. het 91ste percentiel van de verdeling?
     .....
     .....
  2. het zesde percentiel van de verdeling?
     .....
     .....
- Vraag het 25ste percentiel van t_{67} en χ^2_{16} op.

1.3 Genereren van een steekproef van grootte N uit een verdeling

We willen een steekproef $X_1, \ldots, X_N$ genereren uit een veranderlijke X.

1. Normale verdeling

   - Genereer een steekproef van grootte N uit $X \sim N(\mu, \sigma^2)$. Gebruik het commando `rnorm(N,mean=mu,sd=sigma)`

2. t-verdeling

   - Genereer een steekproef van grootte N uit $X \sim t(n)$. Gebruik het commando `rt(N,n)`

3. χ^2-verdeling

   - Genereer een steekproef van grootte N uit $X \sim \chi^2_n$. Gebruik het commando `rchisq(N,n)`

4. F-verdeling

   - Genereer een steekproef van grootte N uit $X \sim F_{m,n}$. Gebruik het commando `rf(N,n,m)`

Oefening 3

- Genereer een steekproef van grootte 25 uit de standaard normale verdeling. Bereken gemiddelde, mediaan, variantie en IQR.
- Genereer een steekproef van grootte 83 uit een F-verdeling $F_{27,15}$.

2 Transformaties voor een normale verdeling

Om na te gaan of een bepaalde variabele normaal verdeeld is, kunnen we een normale kwantielplot of normale QQ-plot opstellen (andere mogelijkheden zijn bv. een boxplot of een histogram maken).

Strategie:

- Er bestaat een lineair verband tussen de kwantielen van $N(0,1)$ en $N(\mu,\sigma^2)$:

$$\underbrace{Q_{N(\mu,\sigma^2)}(p)}_{\substack{\downarrow \\ \text{kwantiel} \\ \text{van } N(\mu,\sigma^2)}} = \mu + \sigma \underbrace{Q_{N(0,1)}(p)}_{\substack{\downarrow \\ \text{kwantiel} \\ \text{van } N(0,1)}}.$$

  Dus, wanneer we de koppels $(Q_{N(0,1)}(p), Q_{N(\mu,\sigma^2)}(p))$ uitzetten in een grafiek, bekomen we een rechte met intercept μ en richtingscoëfficiënt σ.

- In praktijk zijn de kwantielen $Q_{N(\mu,\sigma^2)}(p)$ niet gekend, dus gebruiken we de empirische kwantielen $\hat{Q}_n(p)$ met $p = \frac{i-0.5}{n}$ voor $i = 1, \ldots, n$. Dan is

$$\hat{Q}_n(\frac{i-0.5}{n}) = x_{(i)}.$$

Besluit: De QQ-plot is een grafiek waarbij de kwantielen van de standaard normale verdeling op de X-as worden uitgezet en de geordende gegevens op de Y-as. Als de grafiek volgens een rechte verloopt, kunnen we besluiten dat de onderliggende verdeling van de gegevens de normale verdeling is.

Nu gaan we deze normale QQ-plot gebruiken om de normaliteit van gegevens na te gaan en indien nodig een transformatie te zoeken zodanig dat de getransformeerde gegevens uit een normale verdeling komen.

Op Blackboard staan drie datasets (*statdata1.xls*, *statdata2.xls* en *statdata3.xls*) die moeten gebruikt worden voor de volgende oefeningen. Sla deze datasets op op je computer en importeer ze daarna in *R*. Voor je ze importeert, check je eerst of de namen van de variabelen geen spaties bevatten (bv. niet data 1, wel data1).

Oefening 4 Maak een normale QQ-plot van *statdata1*. In *R* gebeurt dit met de volgende code: `qqnorm(statdata1$data1)`. Merk op dat je de namen van de variabelen in de dataset kan opvragen via `names(statdata1)`.

Je kan in *R* ook eerst een variabele *data1* maken door `data1=statdata1$data1` en daarna een QQ-plot maken via `qqnorm(data1)`.

Wat besluit je omtrent de verdeling van de gegevens?

.....

Als de QQ-plot geen linear verband toont, zijn de gegevens niet normaal verdeeld. Men kan met behulp van de QQ-plot wel een idee krijgen over een mogelijke transformatie, zodat de getransformeerde gegevens wel normaal verdeeld zijn.
Voor de eerste dataset is het verloop van de grafiek ongeveer gelijk aan de exponentiële functie. Omdat de logaritmische functie de inverse functie is van de exponentiële functie, gaan we *data1* transformeren naar $\log(data1)$.

Dit gebeurt a.d.h.v. de volgende code `logdata1=log(data1)`.

Oefening 5 Maak opnieuw een normale QQ-plot van de getransformeerde data. Wat besluit je omtrent de verdeling?

.....

Oefening 6 Komen de gegevens van *statdata3* uit een normale verdeling? Zo nee, kan je een transformatie vinden zodanig dat de getransformeerde gegevens wel uit een normale verdeling komen?

.....

.....

3 Verdeling van gegevens nakijken

We zullen deze techniek toepassen op de dataset uit Computerzitting 0. Importeer de dataset *resultaten.xls* eerst. We proberen de verdeling van de variabelen te achterhalen. Technieken die hierbij kunnen helpen zijn:

- Boxplot: Naast het detecteren van uitschieters, kunnen symmetrie-eigenschappen van de verdeling hiermee besproken worden.

- Histogram: Dit geeft een goede weergave van de vorm van de verdeling. Symmetrie-eigenschappen kunnen hier ook worden opgemerkt.
- QQ-plot: Hiermee kan nagegaan worden of de gegevens aan een bepaalde verdeling voldoen.

We bespreken de verdeling van de variabele *lengte*.

- Maak een boxplot van de variabele *lengte* (zie ook Computerzitting 0): `boxplot(resultaten$lengte)`

  Wat besluit je?

  .....

- Maak een histogram van de variabele *lengte*: `hist(resultaten$lengte)`

  Wat besluit je?

  .....

- Maak een normale QQ-plot van de variabele *lengte*? Wat besluit je?

  .....

Herhaal de vorige procedure voor de lengte van mannen en vrouwen apart. We herhalen nog eens hoe je van de variable lengte enkel de gegevens van de vrouwen neemt R:

```
lengteV=resultaten$lengte[resultaten$geslacht=="V"]
```

- Wat besluit je voor de lengte van de mannen?

  .....

- Wat besluit je voor de lengte van de vrouwen?

  .....

4 Extra Oefeningen

Oefening 7 De tijd die nodig is om een toelatingsexamen voor een bepaalde cursus te beëindigen, is normaal verdeeld met een gemiddelde van 110 minuten en een standaarddeviatie van 20 minuten.

1. Welke proportie van studenten heeft in 2 uur het examen beëindigd?

   .....

   .....

2. Hoelang moet het examen dan duren opdat 90% van de studenten het examen beëindigd heeft?

.....

.....

Oefening 8 Veronderstel dat de pH van bodemstalen uit een bepaalde geografische streek normaal verdeeld is met gemiddelde 6 pH en standaarddeviatie 0.1 pH. Stel dat de pH van een willekeurig geselecteerde bodemstaal uit die streek bepaald is.

1. Wat is dan de kans dat de resulterende pH-waarde tussen 5.90 pH en 6.15 pH ligt?

   .....

   .....

2. Wat is dan de kans dat de resulterende pH-waarde groter is dan 6.10 pH?

   .....

   .....

3. Wat is dan de kans dat de resulterende pH-waarde ten hoogste 5.95 pH is?

   .....

   .....

4. Welke waarde zal overtroffen worden door slechts 5% van al de mogelijke pH-waarden?

   .....

   .....