

Spelregels:

- **Deadline: VRIJDAG 22 SEPTEMBER!**
- **U mag alleen werken of in tweetallen.**
- **Stuur de antwoorden per email naar de werkcollegedocent.**
- **Zet namen en studentnummers duidelijk bovenaan het document.**

1. Laat X_1, \dots, X_n onafhankelijke, continue stochastische variabelen zijn met verdelingsfunctie F . Toon aan dat voor de bijbehorende ordestatistieken $X_{(1)}, \dots, X_{(n)}$ geldt dat

$$\mathbb{E}F(X_{(i)}) = \frac{i}{n+1}$$

voor $i = 1, 2, \dots, n$.

2. Download het tekstbestand
<http://www.epsilon-uitgaven.nl/bijlagen/E76/E76-tweelingdata.txt>.
Bekijk de beschrijving van de dataset in de eerste regels van de file. Schrijf een kort R script dat de data uit dit bestand met behulp van bijvoorbeeld het commando `scan` of het commando `read.table` inlaadt en dat een vector `man` creëert die de lengtes van de mannen bevat en een vector `vrouw` die de lengtes van de vrouwen bevat.
3. Van lengtedata wordt vaak aangenomen dat ze normaal verdeeld zijn. Maak histogrammen van de vectoren `man` en `vrouw` om een eerste idee te krijgen van de redelijkheid van deze aanname in het onderhavige geval. Wat is uw conclusie op basis van deze figuren?
4. Maak om de hypothese van normaliteit verder te onderzoeken QQ-plots van de vectoren `man` en `vrouw`. Leg uit hoe plausibel u de normaliteit van de data acht op basis van deze figuren.
5. Mannen zijn “over het algemeen” vaak langer dan vrouwen. Laat in één enkele figuur boxplots zien van de vectoren `man` en `vrouw` om in te schatten of hetzelfde geldt voor deze data. Wat is uw conclusie?
6. We weten niet hoe de data verzameld zijn en of de aanname van onafhankelijkheid daarom redelijk is. Gebruik auto-correlatie plots om te onderzoeken in welke mate de vectoren `man` en `vrouw` op te vatten zijn als (realisaties van) vectoren onafhankelijke stochastische variabelen.
7. Formuleer op basis van de gedane observaties een plausibel statistisch model voor de twee sets lengtedata.