Data Analyse en Statistiek

Hella Snoek & Marthe Schut

Inhoud

0.1 M2.2 Meesjes ****

Je vindt helaas een dood meesje in de tuin. Het lijkt op een koolmeesje maar het zou ook een pimpelmeesje kunnen zijn. Deze twee vogeltjes lijken erg veel op elkaar. Er zijn manierenom pimpelmeesjes van koolmeesjes te onderscheiden met behulp van uiterlijke kenmerken. Maar je bent een Natuurkundige en geen Bioloog. Online vind je een dataset met informatie over het massa en de spanwijdte van beide soorten meesjes.

Voordat we aan deze opdracht beginnen moeten we eerst een nieuwe versie downloaden van de `DAS_DatasetGenerator.py`. Zonder de nieuwe versie werkt deze opgave niet. Download ook het bestand `M2.2_Meesjes.py` en zorg dat deze in dezelfde folder staat als het `DAS_DatasetGenerator.py` bestand.

We genereren eerst een twee datasets met behulp van de volgende regel code:

```
m_km, span_km, m_pm, span_pm = ds.datasetVogeltjes()
```

De variabelen hebben de volgende betekenis:

```
m_km    : de massa van een koolmeesje in gram
span_km : de spanwijdte van een koolmeesje in cm
```

De laatste twee variabelen zijn de datapunten voor pimpelmeesjes. De twee variabelen van de koolmeesjes horen bij elkaar. Van elk meesje in de dataset zijn zowel de massa als de spanwijdte gemeten. De dataset is zo geordend dat als je het n-de punt uit de `m_km`-lijst bij het n-de punt uit de `span_km`-lijst hoort. Dit zijn de gegevens van het n-de meesje. Pas dus op dat je de lijsten in de juiste volgorde houdt! Voor de twee variabelen van de pimpelmeesjes geldt precies hetzelfde.

We gaan eerst naar de twee massaverdelingen van de meesjes kijken.

- **M2.2a) Plot de massaverdelingen van beide meesjes in een histogram. Laat in een legenda zien welke meesje bij welke kleur hoort. Maak ook een apart histogram waarin je spanwijdtes van de twee soorten meesjes plot. Maak de twee histogrammen netjes af en zorg dat duidelijk is welke distributie bij welk soort meesje hoort.**

  TIP: Gebruik de plot optie `alpha=0.8` zodat je histogrammen wat doorzichtig worden. Zo kan je het achterste histogram ook nog altijd goed zien.

- **M2.2b) Maak een tabel waarin je voor beide soorten meesjes de gemiddeldes, de standaarddeviaties en de varianties noteert. Let goed op de notatie en denk ook even aan de eenheden.**

We meten nu de massa op van het meesje dat je gevonden hebt. Gebruik de volgende regel

code om dat te doen:

```
mees_m_laag, mees_m_hoog = ds.meetMassaMeesje()
```

Je krijgt nu een onderwaarde `mees_m_laag` en een bovenwaarde `mees_m_hoog` terug. Deze geven de onzekerheid op de meting aan. Het gemiddelde van deze twee is de gemeten massa, de centrale waarde. De waarde van de massa van de mees ligt **zeker** tussen de boven- en onderwaarde in. NB. Als je een foutmelding krijgt dat `meetMassaMeesje()` niet bestaat controleer dan of je wel een nieuwe `DAS_DatasetGenerator.py` hebt downgeload voor Module 2.

Met deze informatie kunnen we nu met de Frequentist Methode de kans uitrekenen dat onze mees een Koolmeesje is.

- **M2.2c) Gebruik de dataset `m_km` om de kans uit te rekenen dat je een koolmeesje vindt die een massa heeft die in het gebied `mees_m_laag` en `mees_m_hoog` in ligt. Dit noem je ook wel de voorwaardelijke kans $P(\texttt{est}_m \text{flest}_m) pimpelmeesje, berekendusook \mathbf{P(m_{obs} \mid pimpelmees)}$.**
- **M2.2d) Als je kijkt naar de uitkomst van M2.2c), wat vogeltje denk je dan dat het is?**

De frequentist methode, zoals we die hierboven gebruiken, is uiteindelijk een ratio tussen twee getallen. Deze twee getallen hebben een onzekerheid volgens de Poisson verdeling.

- **M2.2e) Schrijf de formule uit hoe de onzekerheden van de noemen en deler zich propageren naar de onzekerheid op de uitgerekende kans. Noteer deze formule en bereken met behulp van deze formule de onzekerheden uit op de kansen die je in M2.2c) hebt berekend.**

Je besluit ook de spanwijdte van de mees op te meten. Misschien geeft dat wel meer uitsluitsel.

```
mees_span_laag, mees_span_hoog = ds.meetLengteMeesje()
```

De output volgt dezelfde logica als hiervoor.

- **M2.2f) Gebruik dezelfde methode als hiervoor om beide kansen $P(w_{\text{obs}} \mid \text{koolmees})$ en $P(w_{\text{obs}} \mid \text{pimpelmees})$ uit te rekenen maar nu door (alleen) gebruik te maken van de informatie van de spanwijdtes. Noteer ook de onzekerheden op de uitgerekende kansen.**

- **M2.2g) Op basis van deze informatie, wat denk je nu dat het voor vogeltje is?**

We kunnen nu natuurlijk ook de gecombineerde informatie gebruiken. Hiervoor gaan we eerst de data visualiseren.

- **M2.2h) Maak een tweedimensionale scatterplot die de tweedimensionale dataset van de massa versus de spanwijdte voor zowel de pimpelmezen als de koolmezen.**
  TIP gebruik de opties `'o',markersize=3, alpha=0.4` in de plot functie. Zorg dat beide datasets weer hun eigen kleur hebben en vergeet de legenda niet.

Het valt misschien op dat er een verband lijkt te zijn tussen beide variabelen. We gaan daar eerst naar kijken naar de covariantie en de correlatie tussen de massa en de spanwijdte voor beide vogelsoorten.

- **M2.2i) Bereken de covariantie en de correlatie tussen de massa en de spanwijdte voor zowel de koolmeesje als de pimpelmeesjes meetgegevens.**

- **M2.2j) Als je naar de berekende correlaties kijkt wat valt dan op, wat voor verband zit er tussen de twee variabelen? Als je toch even als een Bioloog nadenkt, is dit dan wat je verwacht?**

We gaan terug naar de kansberekeningen.

- **M2.2k) Combineer nu de gegevens en bereken de kansen** $P(m_{\textbf{obs}}$ **en** $w_{\textbf{obs}} \mid$ **koolmees)** **en** $P(m_{\textbf{obs}}$ **en** $w_{\textbf{obs}} \mid$ **pimpelmees).**

- **M2.2l) Welk vogeltje denk je nu dat het is? Beredeneer je antwoord.**

Na al deze berekeningen lopen we een eindje in de tuin. Op de plek waar we eerder het meesje aantroffen zit nu een ander meesje hartstochtelijk te zingen. Aan de zang hoor je direct dat dit een pimpelmeesje is. Je schat in dat er een kans is van 90% dat dit pimpelmeesje bij het andere meesje hoorde, en dat dat dus ook een pimpelmees is.

- **M2.2m) Bereken nu de kans dat het inderdaad een pimpelmeesje is geweest:** $P(\textbf{pimpelmees} \mid m_{\textbf{obs}} \textbf{ en } w_{\textbf{obs}})$. **Bereken hier alleen de centrale waarde.**
  TIP: Maak hierbij gebruik van de vergelijking van Bayes. Om $P(m_{\text{obs}} \text{ en } w_{\text{obs}})$ te berekenen kun je gebruiken maken van de volgende formule: $P(C) = P(C \mid D) \cdot P(D) + P(C \mid \text{niet } D) \cdot P(\text{niet } D)$.