LegumeLegacy Doctoral Network
SS1 Foundations of Statistics Training Event
5th – 6th September 2023

SS1.1 – Recap of probability distributions

The class will be divided into groups of three or four. In your group, answer the following questions. In each case, identify the random variable (using words) and say what possible values it can take, then give the random variable appropriate mathematical notation and define its distribution, and state any assumptions made. Write out the answers using including appropriate mathematical formatting and notation.

1) Suppose that items in a vending machine get stuck on the way out at random with probability 0.03. You arrive at the vending machine, put your money in and select your item. What is the probability you get your item?

2) A recent study showed that 30% of animals in a certain large population suffer from a disease. A random sample of eight of the animals is taken from the population. What is the probability that three or more of the animals have the disease?

3) Anne is playing a computer game and has a probability of winning of 0.27. She keeps playing until she has won a game (and then stops). What is the probability that she plays seven games in total?

4) The height of a particular tree species follows a normal distribution with mean 3.1 metres and standard deviation 0.4 metres.
    a. If a tree is selected at random from this population, what is the probability that its height will be greater than 3.5 metres?
    b. If 40 trees are selected at random from the population, what is the probability that their average is lower than 3?

5) A grassland species of interest is present in a field with a proportion of 8%. After a harvest, an experimenter took grab samples (each weighing approx. the same) at random from the field and continued taking samples until five samples contained the species of interest. What is the probability that ten samples were needed in total?

6) Suppose the number of typos on a page of a book follows a Poisson distribution with parameter $\lambda = 1$. What is the probability that a page is error free?

Reading material suggestion:

- *Diez, Çetinkaya-Rundel and Barr (2019; updated 2022) OpenIntro Statistics, 4th Edition. Ch. 1 to 4.*

The LegumeLegacy project has received funding from the European Union’s Horizon 2021 Doctoral Network programme under the Marie Skłodowska-Curie grant agreement No. 101072579

SS1.2 – Hypothesis testing

Example
An experiment was carried out involving 19 people. The people were randomly assigned to group 1 (nine people) or group 2 (ten people).

- The people in group 1 were each given a set of 40 coins, in each bundle was ten each of: 1c, 2c, 5c and 10c coins. Each participant was asked to separate the coins and create four neat bundles of the same coins. The participants were timed doing their activity.
- Group 2 were then asked to do the same activity, but group 2 were told to keep one hand behind their backs and only use their other hand to perform the activity and the time was recorded.
- When group 2 had completed their task, they were asked to repeat the task, but this time they were allowed to use two hands. Once again, each participant was timed in how long it took them to separate out their bundle.

The data are recorded in the file: SS1-2_data_coins.csv. The variables in the dataset are:
group: which group the participant belonged to (1 or 2)
person: a unique identifier for each participant in the experiment (1 – 19)
numhands: the number of hands permitted for the separation of coins
time: the time in seconds that it took to complete the task at hand.

Question: Is there evidence that it takes longer to separate the coins using one compared to two hands? State the hypotheses, test statistic value and conclusion, and discuss assumptions, of any test that you carry out. Feel free to use software to help with the analysis.

Reading material suggestion:

- *Diez, Çetinkaya-Rundel and Barr (2019; updated 2022) OpenIntro Statistics, 4th Edition. Ch. 5 to 7.*

The LegumeLegacy project has received funding from the European Union's Horizon 2021 Doctoral Network programme under the Marie Skłodowska-Curie grant agreement No. 101072579

SS1.3 – Simple linear regression

Example
In grassland ecology, it is known that increasing the number of species, and, in particular, combining grasses with legumes, can improve yield outputs. A grassland biodiversity experiment that examined mixtures of grasses and legumes was carried out at 30 sites across Europe where the number and proportion of species was manipulated across multiple plots at each site. The experiment was established and measured for two years at all sites and continued into a third year at 24 of the 30 sites.

In year one of the experiment, the average percentage of legumes of all plots at each site was recorded and stored in the variable called legY1. In year two, a standardised measurement of the benefits of mixing was quantified across all plots at each site and stored in the variable sdeY2. The data is recorded in SS1-3_grassland.csv.

Using a simple linear regression approach, does the percentage of legume in year 1 affect the standardised benefits of mixing in year 2?

Interpret the model and discuss any assumptions / limitations with the analysis.

Reading material suggestion:

- *Diez, Çetinkaya-Rundel and Barr (2019; updated 2022) OpenIntro Statistics, 4th Edition. Section 7.5.*
- *Chatterjee and Hadi (2006) Regression analysis by example. Ch. 2 – Simple Linear Regression.*
- *Brophy et al (2017) Major shifts in species' relative abundance in grassland mixtures alongside positive effects of species diversity in yield: A continental-scale experiment. Journal of Ecology; 105, 1210-1222.*

The LegumeLegacy project has received funding from the European Union's Horizon 2021 Doctoral Network programme under the Marie Skłodowska-Curie grant agreement No. 101072579

LegumeLegacy Doctoral Network
SS1 Foundations of Statistics Training Event
5th – 6th September 2023

SS1.4 – ANOVA methods

Example
An experiment was carried out in controlled plant growth chambers to identify the effect of temperature on the growth rate of a particular plant species. The plants were grown in individual pots at varying temperatures: 2, 4, 6, 8 and 10^{O}C. There were four replicates for each temperature.

The data are recorded in the file: SS1-4_data_plant_growth.csv. The variables in the dataset are:

treat: indicator variable 1 to 5 to indicate the five different treatments
temp: the temperature at which the plant was grown
resp: the growth rate response variable.

Question 1: Analyse this data to identify the effect of temperature on plant growth rate. What do you recommend with regards to the optimal temperature for the growth rate for this plant species? (Answer this question completely before moving on to question 2.)

Question 2: Suppose new information comes to light: plants that were grown at 2, 4, 8 and 10^{O}C were grown under normal light conditions, while plants grown at 6^{O}C were exposed to only half the normal light. With this new information re-analyse the data. Explain the revised analysis, conclusions and recommendations. Are there any limitations with your analysis given the design of the experiment?

Reading material suggestions:

- *Diez, Çetinkaya-Rundel and Barr (2019; updated 2022) OpenIntro Statistics, 4th Edition. Section 7.5.*
- *Hector et al (2010) Analysis of variance with unbalanced data: an update for ecology and evolution. Journal of Animal Ecology, 79, 308–316.*

The LegumeLegacy project has received funding from the European Union's Horizon 2021 Doctoral Network programme under the Marie Skłodowska-Curie grant agreement No. 101072579