Modern Inference (Statistical Learning) - University of Milano-Bicocca

Homework 1

To submit via e-mail by 06/04/2020 h 14:00.
You will have to submit your solution by the deadline. Feel free to choose the format for your solution (.txt, .tex, .pdf etc.), the nicer the better, and the number of files in attachment (one or more, but not too many). Answer with clarity and precision. All R code must be reproducible. For theoretical questions, try to provide a well-reasoned mathematical argument. Simulations can help form your intuition; but a purely empirical answer will only receive partial points. It is encouraged to discuss the problem sets with others, but every group needs to turn in a unique write-up. Use of sources (people, books, internet and so on) without citing them in homework sets results in failing grade.

1 Darwin's plants experiment

Reference:
A.C. Davison (2003)
Statistical Models
Cambridge University Press
Examples 1.1, 3.11, 3.16, 7.24, 7.28

Charles Darwin collected data over a period of years on the heights of Zea mays plants.

Population: Zea mays plants

Experimental Design: The plants were descended from the same parents and planted at the same time. Half of the plants were *self-fertilized*, and half were *cross-fertilized*, and the purpose of the experiment was to compare their *heights*. To this end Darwin planted them in pairs in different pots.

Data:

	Pot	Cross	Self
1	I	23.500	17.375
2	I	12.000	20.375
3	I	21.000	20.000
4	II	22.000	20.000
5	II	19.125	18.375
6	II	21.500	18.625
7	III	22.125	18.625
8	III	20.375	15.250
9	III	18.250	16.500
10	III	21.625	18.000
11	III	23.250	16.250
12	IV	21.000	18.000
13	IV	22.125	12.750
14	IV	23.000	15.500
15	IV	12.000	18.000

The focus of interest is the relation between the height of a plant and something that can be controlled by the experimenter, namely whether it is self or cross-fertilized. The essence of the model is to regard the height as random with a distribution that depends on the type of fertilization, which is fixed for each plant.

Note that in order to minimize differences in humidity, growing conditions, and lighting, Darwin had taken the trouble to plant the seeds in pairs in the same pots. The height of a plant would therefore also depend on these factors, which are not of interest, not only on the type of fertilization.

Questions:

1. Does the height depend on the type of fertilization?
2. Can we estimate the height difference, and assess the uncertainty of our estimate?

Hypothesis/Conjecture: Height of a plant depends on the type of fertilization.

Analysis Plan: Darwin asked his cousin, Francis Galton, whether the difference in heights between the types of plants was too large to have occurred by chance, and was in fact due to the effect of fertilization. See Galton model.
In discussing this experiment many years later, R. A. Fisher pointed out that the Galton model is inappropriate: comparison of different pairs would involve differences in humidity, growing conditions, and lighting, which are not of interest, whereas comparisons within pairs would depend only on the type of fertilization. See Fisher model.

1.1 Galton's model

Galton assumed a model where the height of a self-fertilized plant is

$$Y = \mu + \sigma\varepsilon$$

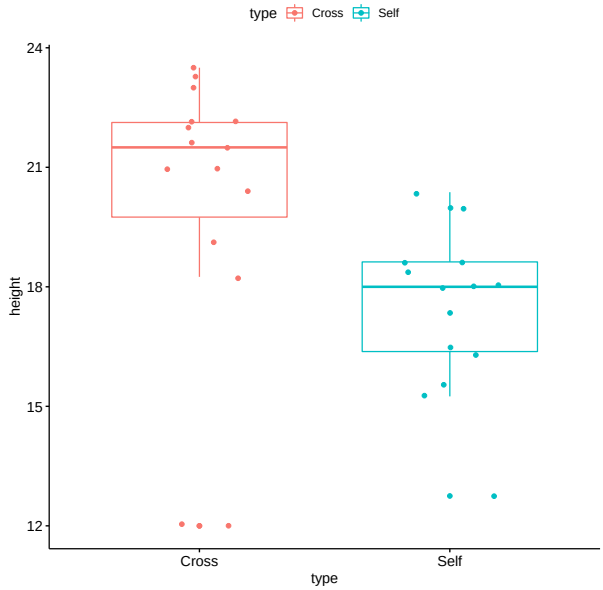
and of a cross-fertilized plant is

$$X = \mu + \theta + \sigma\epsilon$$

where μ, θ and σ are unknown parameters, and ε and ϵ are independent random variables with mean zero and unit variance.

Observations from self-fertilized plants $Y_1, \ldots, Y_{15}$ are i.i.d. as Y, and observations from cross-fertilized plants $X_1, \ldots, X_{15}$ are i.i.d. as X.

The comparison between two independent samples are usually visualized by a boxplot:



Under this model Darwin's questions translate to:

G1 Is the average height increase $\theta \neq 0$?

G2 Can we estimate θ, and assess the uncertainty of our estimate?

TASK 1: Answer G1 and G2 by specifying:

1. **Analysis Plan:** Choose the appropriate statistical methodology to address G1 and G2. The methods should satisfy the assumptions of Galton's model. You can add further assumptions to the model. Specify *all* the assumptions that you need.
2. **Code:** Write the R code to get the results. You can add plots, statistical summaries, etc. Your code must be replicable.
3. **Claim:** Comment your results and write the answers to G1 and G2.

1.2 Fisher's model

In order to minimize differences in humidity, growing conditions, and lighting, Darwin had taken the trouble to plant the seeds in pairs in the same pots.

Comparison of different pairs would therefore involve these differences, which are not of interest, whereas *comparison within pairs* would depend only on the type of fertilization.

Fisher considered the model

$$Y_i = \mu_i + \sigma\varepsilon_i, \quad X_i = \mu_i + \theta + \sigma\epsilon_i, \quad i = 1, \ldots, n$$

The parameter μ_i represents the effects of the planting conditions for the ith pair, and ε_i and ϵ_i are independent random variables with mean zero and unit variance.

The μ_i could be eliminated by using the differences

$$D_i = X_i - Y_i$$

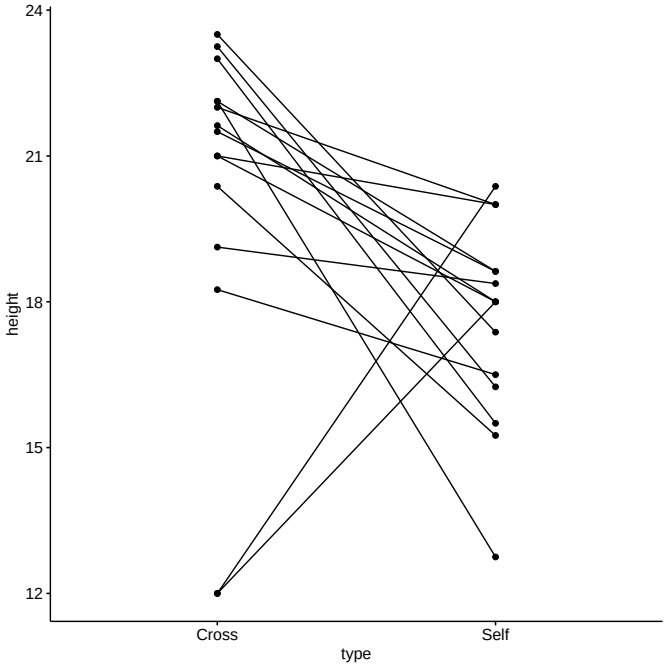
which have mean θ and variance $2\sigma^2$.

Under this model Darwin's questions translate to:

F1 Is the average height increase $\theta \neq 0$?

F2 Can we estimate θ, and assess the uncertainty of our estimate?

The comparison within pairs can be visualized by the following plot:



TASK 1: Answer F1 and F2 by specifying:

1. **Analysis Plan:** Choose the appropriate statistical methodology to address F1 and F2. The methods should satisfy the assumptions of Fisher's model. You can add further assumptions to the model. Specify *all* the assumptions that you need.
2. **Code:** Write the R code to get the results. You can add plots, statistical summaries, etc. Your code must be replicable.
3. **Claim:** Comment your results and write the answers to F1 and F2.

1.3 A more general model

Consider a more general matched pair model for the Darwin data:

$$Y_i = \mu_i + \sigma_i \varepsilon_i, \quad X_i = \mu_i + \theta + \tau_i \epsilon_i, \quad i = 1, \ldots, n$$

with ε_i and ϵ_i are independent random variables with continuous distribution and mean zero and unit variance.

The height differences may be written as

$$D_i = \theta + (\tau_i \epsilon_i - \sigma_i \varepsilon_i)$$

Assume that ε_i and ϵ_i are independent and symmetrically distributed around 0 (e.g. Uniform$(-i, i)$), then D_i is symmetrically distributed around θ, i.e.

$$D_i - \theta \stackrel{d}{=} \theta - D_i \qquad i = 1, \ldots, n$$

Under this model, Darwin's first question translate to:

S1 Is $\theta \neq 0$?

Note that if $\theta = 0$, then the probability that D_i falls on either side of 0 is 1/2. This suggests that we can base a test on

$$T = \sum_{i=1}^{n} \mathbb{1}\{D_i > 0\}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function. Note that using $>$ or $\geq$ in the indicator function is equivalent because the probability that $D_i = 0$ is 0 since we have assumed that the distribution of D_i is continuous.

TASK 3: Answer S1 by specifying

1. **Analysis Plan:** Specify a statistical test to answer S1 satisfying the assumptions of the more general model.
2. **Code:** Write the R code to get the results. Your code must be replicable.
3. **Claim:** Comment your results and write the answer to S1.
4. Let $X \sim f(x)$ and $Y \sim g(y)$ be continuously distributed random variables with density functions f and g, and assume that f is symmetric around 0, g is symmetric around 0, and X and Y are independent. Prove that $Z = X - Y$ is symmetric around 0

2 Galieo's inclined plane experiment (1604)

How Things Fall?

Reference: Galileo's Great Discovery: How Things Fall

`https://www.springer.com/cda/content/document/cda_downloaddocument/9781461454434-c1.pdf`

- Researcher: Galileo Galilei (1565-1642)
- Question of interest: If a ball rolls down a ramp, what is the relationship between time and distance?

- Aristotle theory/hypothesis/model: Constant velocity (zero acceleration): distance $\propto$ time
- Galileo theory/hypothesis/model: Increasing velocity (constant acceleration): distance $\propto$ time ^{2}
- Galileo's goal: reject Aristotle theory/hypothesis/model in favour of his conjecture. Here Galileo's conjecture is fixed a priori, i.e. before seeing the data
- Experimental data:

time	distance
1	33.00
2	130.00
3	298.00
4	526.00
5	824.00
6	1192.00
7	1620.00
8	2104.00

TASK 4: Help Galileo with his goal:

1. **Analysis Plan** Choose the statistical methodology and specify all required assumptions
2. **Code** Write the R code to get the results. Your code must be replicable.
3. **Claim** Comment on the results.