HW2, Theory of Inference 2015/6

Jonathan Rougier
School of Mathematics
University of Bristol UK

This homework is about some of the inferential issues which arise when doing experiments to assess causal effects. In this case, the question is whether drinking a glass of milk in the morning causes children to grow faster.

Note that where there is no mark given, this type of question is unlikely to come up in an exam.

1. The experimental protocol is as follows. The experiment will run for one school year, for all of the children in one class. The heights of the children will be measured at the beginning and end of the year; the outcome will be the change in height for each of the children. The 'treatment' group will be given a 100 ml glass of milk each morning, at the start of class. The 'control' group will be given a 100 ml glass of water. Half of the boys will be allocated to the treatment group, and half of the girls will be allocated to the treatment group. All of the other children will be in the control group. Selection of the subset of boys and girls in the treatment group will be random.

   (a) Why does the control group get a glass of water rather than nothing at all?

   Answer. The treatment involves an additional drink, plus that the drink is milk. To isolate the effect of the milk from the effect of the additional drink, we need to put an additional drink into the control.

Also, being 'treated' has a placebo effect, so we want the control group to have a 'treatment' as well.

(b) Why is allocation to treatment or control 'blocked' by boy or girl?

Answer. We would like to derive results separately for boys and girls, because they might be different. Therefore we want to divide our resources evenly between them. Following any other scheme, we might end up with mostly boys, or mostly girls, in which case the other half of the experiment would be less conclusive.

(c) Why is allocation to treatment or control within blocks done at random?

Answer. Any other scheme risks some form of 'confounding', in which the allocation scheme and the treatment cannot be separated, and so the treatment effect cannot be isolated. For example, if we allocated on the basis of first letter of the surname, then the fact that some surnames are more common in some ethnic groups means that treatment and ethnic group are confounded.

(d) Provide an R snippet that identifies the children to be in the treatment group.

Answer. This is how an experienced R user (e.g. me!) might do it:

```
## sex of each child, ordered alphabetically, is provided

male <- c(FALSE, TRUE, TRUE, FALSE, FALSE, TRUE, FALSE, TRUE,
          TRUE, FALSE, TRUE, FALSE, FALSE) # etc etc

## here is the caclulation

ismale <- which(male)
isfemale <- which(!male)
```

```
istreat <- lapply(list(ismale, isfemale), function(x) {
  sample(x, size = floor(length(x) / 2))
})

treat <- rep(FALSE, length(male))
treat[unlist(istreat)] <- TRUE
```

If you are interested in R, make sure you understand the functions used here: `lapply`, `sample`, `unlist`.

2. The standard notation for a statistical model is

$$\boldsymbol{Y} \sim f(\theta) \quad \theta \in \Omega$$

where f is the name of the model, θ are the parameters, and Ω is the parameter space, which is usually a convex subset of Euclidean space. This is synomymous with the statement

$$\Pr(\boldsymbol{Y} \doteq \boldsymbol{y}; \theta) = f(\boldsymbol{y}; \theta)$$

(a) The *Maximum Likelihood Estimator (MLE)* of θ is defined as

$$\hat{\theta}(\boldsymbol{y}) := \underset{\theta \in \Omega}{\operatorname{argmax}} f(\boldsymbol{y}; \theta).$$

Show that if $g : \theta \mapsto \psi$ is bijective, then the MLE of ψ satisfies $\hat{\psi}(\boldsymbol{y}) = g(\hat{\theta}(\boldsymbol{y}))$. That is, the MLE is invariant to bijective transformations. This is one of its most attractive features. [10 marks]

Answer. Under the mapping, we have $f_\psi(y; \psi) = f(y; g^{-1}(\psi))$, with

$\psi \in \Psi := g(\Omega)$. Then, letting $\psi \in \Psi$ be arbitrary,

$$\begin{aligned}
f_\psi(y;\psi) &= f(y; g^{-1}(\psi)) \\
&\le f(y; \hat\theta(y)) && \text{because } \hat\theta \text{ is the MLE} \\
&= f(y; g^{-1} g\hat\theta(y)) \\
&= f(y; g^{-1}\hat\psi(y)) && \text{setting } \hat\psi := g\hat\theta \\
&= f_\psi(y; \hat\psi(y))
\end{aligned}$$

and hence $\hat\psi = g\hat\theta$ is the MLE for ψ, because its probability for y is at least as large as an arbitrary point selected from Ψ.

(b) If $\theta = (\theta_1, \theta_2)$, then the MLE of θ_1 is defined to be the first component of the MLE of θ, and similarly for generalisations. Show that if $g : \theta \mapsto \psi$ is *not* bijective, but can be embedded within a bijective mapping by adding additional elements, then the MLE of ψ is $g(\hat\theta(\boldsymbol{y}))$. [5 marks]

Answer. This is mostly about notation. Let

$$\theta \mapsto (g(\theta), h(\theta)) =: (\psi, \phi)$$

be bijective, for careful choice of h. Then by the result in Q2a, $(\hat\psi, \hat\phi) = (g\hat\theta, h\hat\theta)$, and hence $\hat\psi(\boldsymbol{y}) = g(\hat\theta(\boldsymbol{y}))$.

3. Let $Y_1, \dots, Y_m$ be the change in heights of the control group, and $Z_1, \dots, Z_n$ be the change heights of the treatment group. A standard model would be

$$f(\boldsymbol{y}, \boldsymbol{z}; \mu_c, \mu_t, \sigma^2) = \prod_{i=1}^m \phi(y_i; \mu_c, \sigma^2) \cdot \prod_{j=1}^n \phi(z_j; \mu_t, \sigma^2)$$

where ϕ is the Normal density function with specified expectation and variance. The treatment effect is defined as $\psi := \mu_t - \mu_c$. Show that the MLE for the treatment effect is

$$\hat\psi(\boldsymbol{y}, \boldsymbol{z}) = \bar z - \bar y \quad \text{(corrected typo.)}$$

where $\bar{y}$ and $\bar{z}$ are the sample means of $\boldsymbol{Y}$ and $\boldsymbol{Z}$. [10 marks]

Obviously, this is a very attractive model, since it has such an intuitive measure of the treatment effect. Note that the sample sizes have no effect in the estimator of the treatment effect, although, as we will see later, they have an important effect in the uncertainty about the treatment effect.

Answer. Write the parameters as $\theta := (\mu_c, \mu_t, \sigma^2)$. We would like to know the MLE of $\psi := \mu_t - \mu_c$. This can be embedded in a bijective function of θ, for example, $\theta \mapsto (\psi, \mu_t, \sigma^2)$. Therefore we know from the result in Q2b that we have to find the MLE of θ, and the MLE for ψ will follow directly.

To find the MLE of θ, we take logarithms (since log is an increasing function on the positive reals, and probabilities are non-negative), and then maximise. I.e.

$$\begin{aligned}\hat{\theta}(\boldsymbol{y}, \boldsymbol{z}) &= \operatorname*{argmax}_{\mu_c, \mu_t, \sigma^2} \left\{ \sum\nolimits_i \log \phi(y_i; \mu_c, \sigma^2) + \sum\nolimits_j \log \phi(z_j; \mu_t, \sigma^2) \right\} \\ &= \operatorname*{argmax}_{\sigma^2} \left\{ \operatorname*{argmax}_{\mu_c} \sum\nolimits_i \log \phi(y_i; \mu_c, \sigma^2) + \operatorname*{argmax}_{\mu_t} \sum\nolimits_i \log \phi(z_i; \mu_t, \sigma^2) \right\}\end{aligned}$$

Taking the first term,

$$\sum\nolimits_i \log \phi(y_i; \mu_c, \sigma^2) = \frac{-m}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum\nolimits_i (y_i - \mu_c)^2$$

which is maximised over μ_c at $\hat{\mu}_c(\boldsymbol{y}, \boldsymbol{z}) = \bar{y}$, regardless of the value of σ^2 or of $\boldsymbol{z}$. The argument for the MLE of μ_t proceeds the same way, and we end up with $\hat{\theta}(\boldsymbol{y}, \boldsymbol{z}) = (\bar{y}, \bar{z}, \hat{\sigma}^2(\boldsymbol{y}, \boldsymbol{z}))$, where we have not bothered to compute the MLE for σ^2. Then applying the result from Q2b we derive $\hat{\psi}(\boldsymbol{y}, \boldsymbol{z}) = \bar{z} - \bar{y}$.

4. The growth rate is not constant with age, and so it makes sense to incorporate the age of each child in the model for the change in height. How might you modify the above statistical model to allow for this? [5 marks]

Answer. Let $x_1, \ldots, x_m, x_{m+1}, \ldots x_{m+n}$ be the ages of the children at the start of the experiment. One simple model would be to include a common age effect using

$$f(\boldsymbol{y}, \boldsymbol{z}; \mu_c, \mu_t, \alpha, \sigma^2) = \prod_{i=1}^{m} \phi(y_i; \mu_c + \alpha(x_i - \bar{x}), \sigma^2) \cdot \prod_{j=1}^{n} \phi(z_j; \mu_t + \alpha(x_{m+j} - \bar{x}), \sigma^2)$$

where $\bar{x}$ is the arithmetic mean of all of the ages. In this model, $\mu_t - \mu_c$ is still the treatment effect, in the sense that if you contrasted the treatment with the control for one child of age x, the difference in the expectations would be $\{\mu_t + \alpha(x - \bar{x})\} - \{\mu_c + \alpha(x - \bar{x})\} = \mu_t - \mu_c$. However, the MLE of $\mu_t - \mu_c$ would be different to before, because the model is different.

5. Ideally, the estimated treatment effect could be used to decide whether or not to make milk available in the school. Briefly consider the difficulties in generalising from the experiment to the school.

   This question does not have a simple answer. Its purpose is to illustrate that even in this rather simple situation, generalising from an experiment on a sample to an inference about a population is hard.