


Statistical Data Analysis

Mark Andrews

Psychology Department, Nottingham Trent University

 @xmjandrews

 mark.andrews@ntu.ac.uk

What is the goal of statistical data analysis?

- ▶ In any statistical data analysis, we begin with data \mathcal{D} .
- ▶ Our goal is create a probabilistic model of the distribution of which our data is a random sample.
- ▶ This distribution is known as the *population* or the *true data generating process* or the *true probabilistic generative model*.
- ▶ We aim to produce a model of this distributon.
- ▶ This model is a mathematical model of some phenomenon in the world or in reality (the underlying that lead to the data).
- ▶ Like any mathematic model or scientific model is allows us in principle to explain and to predict.

Probabilistic generative models

- We begin by assuming or proposing a probabilistic generative of \mathcal{D}

$$\mathcal{D} \sim f_0(\mathcal{D}; \theta_0)$$

- This is a model of where \mathcal{D} from, and is also a model of any other possible data drawn from the population of which \mathcal{D} is a random sample.
- Here f_0 is a probabilistic model with a particular functional form, e.g. a normal linear model, a normal linear multilevel model with random slopes, etc. For example, if

$$\mathcal{D} = x_1, x_2 \dots x_n \quad x_i \in (0, \infty),$$

then one possibility would be to have

$$\log(x_i) \sim N(\mu, \sigma) \quad \text{for } i \in 1, 2 \dots n$$

Probabilistic generative models

- ▶ The functional form of f_0 defines a *probabilistic model family*.
- ▶ The θ_0 signifies the unknown variables, e.g. parameters and latent variables etc.
- ▶ We assume that the values of θ_0 are *fixed* but *unknown*.
- ▶ Here, we may also be assuming that we have a set of *explanatory* variables. These are treated as fixed and known.
- ▶ Our aim is to infer the values of θ_0 . At which point we have now determined our probabilistic model of \mathcal{D} and of the population of which \mathcal{D} is a random sample.

Inference

- ▶ There are at least three general approaches to inference.
 - ▶ Sampling theory based inference
 - ▶ Likelihood based inference
 - ▶ Bayesian inference

Sampling theory based inference

- ▶ We calculate *estimators* of θ_0 using the data \mathcal{D} .
- ▶ We then calculate the *sampling distribution* of these estimators for given values of θ_0 .
- ▶ From this, we may calculate p-values and confidence intervals for the true values of θ_0 .

Likelihood based inference

- We calculate the likelihood function

$$L(\theta_0|\mathcal{D}) \propto P(\mathcal{D}|\theta)$$

- From this, we may calculate ranges of values of θ_0 for which there is non-trivial evidence.

Bayesian inference

- ▶ For Bayesian inference θ_0 , we must first provide a probability distribution of the possible values of θ_0 . This would specify the range of values that θ_0 could take in principle, and their relative probabilities.
- ▶ By making this addition, we are assuming an expanded probabilistic generative model:

$$\mathcal{D} \sim f_0(\mathcal{D}; \theta_0),$$

$$\theta_0 \sim h_0(\theta_0; \Omega_0)$$

where $h_0(\Omega_0)$ is a probability distribution (over the possible values of θ_0) and is parameterized by Ω_0 which is fixed and *known.

- ▶ The probability distribution $h_0(\Omega_0)$ is known as the *prior*, and Ω_0 is known as the hyper-parameter.

Bayesian inference

- ▶ Having specified $h_0(\Omega_0)$, we may use elementary rules of probability calculus to calculate the following:

$$P(\theta_0|\mathcal{D}, \Omega_0) = \frac{f_0(\mathcal{D}; \theta_0)h_0(\theta_0; \Omega_0)}{\int f_0(\mathcal{D}; \theta_0)h_0(\theta_0; \Omega_0)d\theta_0}$$

- ▶ This is the *posterior* distribution over the possible values θ_0 given the observed data θ_0 .
- ▶ It tells us the probable values of θ_0 given that (by assumption) θ_0 was sample from $h_0(\theta_0|\Omega_0)$ and then \mathcal{D} was sampled from $f(\mathcal{D}|\theta_0)$.

Having inferred θ_0

- ▶ Having inferred θ_0 , in general, we have a range of values of θ_0 for which there is support or evidence (according to one definition or another).
- ▶ We now have our model $f(\mathcal{D}; \theta_0)$ determined. Or rather, we have a set of plausible models of this kind.
- ▶ We may now reason with this, make predictions with this, and so on. Just as we would with any mathematical model or scientific model generally.

Model evaluation

- ▶ Regardless of our approach to inference, all of our conclusion are contingent upon our assumptions.
- ▶ In general, and especially, this includes our assumed probabilistic model (family) of the data $f_0(\theta)$.
- ▶ Infinitely many probabilistic models are possible as alternatives, each with more or less parameters and other unknown variables.
- ▶ Likewise, we may include more or less explanatory variables.
- ▶ We need to choose between these alternatives, or consider their relative support.

Model evaluation

- ▶ One general approach to model evaluation is to ask if the predictions of the (inferred) model make sense.
- ▶ Does the inferred model predict data like the data we've already obtained? In the Bayesian context, these are known as *posterior predictive checks*.
- ▶ More generally, does our inferred model predict future data, i.e. other data from the *population*?

Model evaluation

- ▶ Without future new data to use, whether our alternative models predict this data may seem impossible.
- ▶ We may use cross-validation here. Cross validation essentially *holds out* some of the current data and treats it as new data from the population that the can use to evaluate our current models?
- ▶ Many of the *information criteria* model evaluation metrics, such as Akaike Information Criterion (AIC) and Watanbe Akaike Information Criteria (WAIC) can be justified as approximations to cross validation.
- ▶ Some Bayesian approaches also use Bayes factors to evaluate models. Bayes factors, however, evaluate the relative average evidence provided by the data for the models' parameters, where the averaging is done over the *priors* over the parameters.