All models are wrong

Richard Torkar

April 2021

All models are wrong, but some are useful,[1] is something you will hear a statistician say at least once in their lifetime. The saying, attributed to George E. P. Box, is good because it is true. When we design our models we assume things. The assumptions are not always correct, and even if they were, our lack in understanding the underlying data generation model will always make our models approximate. Hence, we should be prepared to constantly refine our models step-by-step to improve them when our knowledge improves.

At many universities they teach frequentist statistics in the first stats course. The approach is often taught by presenting a flowchart that researchers/students can follow to get to the 'correct' model, which they then use to test hypotheses.

There are two problems with this approach. First, each of these 'models' have assumptions and often these are not stressed enough. Second, the approach is systematic but unfortunately not very flexible. What happens when you face a problem that does not have a 'correct' model one can use?

In our case, you will be taught to design models in a very flexible, math-like, notation. It will be your fantasy that will limit what models to design. Since the model design is very explicit it will force you to also be explicit concerning your assumptions. This, I believe is the best we can do, and shouldn't we always try to do our best?

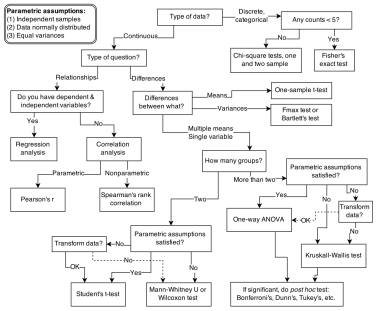
As such you will see that common things in statistics like, p-values, hypothesis testing, etc. is something we don't think about much in this course. We see them as voodoo concepts that generally speaking don't belong to the sciences (Sect. 1.2.1 in the book stresses this and I urge you to read it).

The tools we have at our disposal are:

1. Bayesian data analysis grounded in Bayes' theorem.
2. Model comparison grounded in information theory.
3. Multilevel models employing partial pooling.
4. Graphical causal models implemented as directed acyclic graphs.

The above might sound like black magic and I fully understand that most, if not all of you, have no clue what the above means. That's OK. You will understand at the end of this course, if not before.

[1] G. E. P. Box. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976. doi: 10.1080/01621459.1976.10480949



Yes, all of these tests you see in the figure is a 'model'.

Bayesian data analysis is founded on Bayes' theorem which we use to calculate the posterior probability distribution. The theorem simply states that the posterior probability distribution is proportional to the likelihood times the prior belief.

$\propto$ means *proportional to*.

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

The posterior probability distribution (PPD) is the end goal with our Bayesian analysis. Once we have it we can ask it questions and it will give us answers, but beware that it will provide rubbish answers if you're not careful!

So, in math speak we get a PPD like this: Given a *prior belief* that a probability distribution function is $p(\theta)$ and that the observations x have a likelihood $p(x|\theta)$, then the posterior probability is defined as,

First, θ is pronounced: **thee**-tə (**th** as in <u>th</u>ink). Second, $p(\theta)$ is *probability of* θ, while $p(x|\theta)$ is the probability of x given θ, i.e., conditional on θ.

$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}$$

where $p(x)$ is the normalizing constant and is calculated as (it's not needed for this course, but if you want to please read Thompson's book on calculus, which you can find in the file area),[2]

[2] S. P. Thompson. *Calculus made easy: being a very-simplest introduction to those beautiful methods of reckoning which are generally called by the terrifying names of the differential calculus and the integral calculus*. Macmillan, London, 2nd ed edition, 1944

$$p(x) = \int p(x|\theta)p(\theta)\, d\theta$$

for continuous θ, or by summing $p(x|\theta)p(\theta)$ over all possible values of θ for discrete θ.[3]

[3] Refresh your mind: $\int$ is sort of a long S and you can think of it as "the sum of", while $d\theta$ means the sum of all the little bits of θ (since we're dealing with continuous and not discrete values.

Model comparison allows us to choose between models. We use cross-validation (CV) and information criteria (IC) to do so. More complex models will generally make better predictions, they have more parameters, so we penalize too complex models to make sure we don't learn too much from the data (i.e., overfitting).

"With four parameters I can fit an elephant, and with five I can make him wiggle his trunk." — John von Neumann

In summary, CV and IC provides us with three important things:

1. Useful expectations of predictive accuracy.
2. An estimate of the tendency of a model to overfit.
3. Spot highly influential observations.

In this course we use mostly WAIC and LOO information criteria for model comparison.

Multilevel models are not a new thing, however they fit Bayesian data analysis very nicely. Handling uncertainty on many levels (by adding models in models) allows us to feed the uncertainty through the levels. However, there is another advantage with multilevel models: partial pooling.

Partial pooling is the key technology, it allows us to,

1. adjust estimates for repeat sampling,
2. adjust estimates for imbalance in sampling,
3. study variation, and
4. avoid averaging.

GRAPHICAL CAUSAL MODELS, or structural causal modeling, is a concept Judea Pearl was awarded the Turing Award for in 2011. We will make use of graphs (directed acyclic graphs) to model our belief concerning causality. A statistical model is a beautiful association engine, but is never enough if we want to discuss causality (which we do!) In short, we require a causal model with which to design both the collection of data and the structure of our statistical models. This way we are explicit about our assumptions.

WHEN YOU MAKE A CHOICE, you also delimit future possibilities. If you marry someone and get a child it means that another child might not be born. Imagine if you could play out all the paths in your life? That is basically what Bayesian inference is about.

First, we count the possibilities. Then, we combine with other information. Finally, we go from counts to probabilities. This is the language we use,

- possible explanations of the data are the parameters,
- the relative number of ways that a value can produce the data is a likelihood,
- the prior plausibility of any specific value is the prior probability, and
- the new, updated plausibility of any specific value is the posterior probability.

In Chapters 2.4–2.5 McElreath explains how a Bayesian model works when taking the above into account. Read it carefully; it's the mechanics and mechanics matters.

Later in the course we will make use of a technique called Markov chain Monte Carlo (MCMC). The reason is that it is superior to other techniques when the models become more and more complex. However, in the beginning of the course we will use other techniques: grid approximation and, in particular, quadratic approximation.

Quadratic approximation is often equivalent to the maximum likelihood estimate and its standard error.

SAMPLING TO SUMMARIZE. Once we have a PPD we can summarize it in different ways. It's important that you know these ways,

1. intervals of defined boundaries,

2. questions about intervals of defined probability mass, and
3. questions about point estimates.

SIMULATE FROM THE MODEL. Generating implied observations from a model is useful for many reasons. What you do is to generate fake data and perform model checking (did the software work and is the model adequate), i.e., posterior predictive checks.

WHY IS NORMAL SO... NORMAL? The short answer is the central limit theorem: Any process that adds together random values from the same distribution converges to a normal distribution (if you multiply small numbers the same thing happens since it is approximately the same as addition, and if you multiply large deviates they produce a Gaussian (Normal) distribution in the log scale).

The earliest version of the central limit theorem is the de Moivre-Laplace theorem.

When we talk about the distributions we use (i.e., what we assume about the underlying data generation process that produced the data we have in our hands) we fall back on *ontological* and *epistemological* arguments. We should think about these arguments before we begin designing our models.

Ontological arguments could be, e.g., the Gaussian distribution is a common pattern in nature. Processes that add together fluctuations are Gaussian. However, nature comes with many patters, not only the Gaussian!

The normal distribution is often called Gaussian after Carl Friedrich Gauss.

Epistemological arguments are rooted in information theory and the concept of maximum entropy, i.e., we want to pick a distribution that allows the data to tell us its story.