Bayes class 11: hypothesis testing and GoF

Sean van der Merwe

# Bayesian Hypothesis Testing

In the classical hypothesis testing framework:

We assume that the world is boring. There is no change, no difference, no impact, all coefficients are zero, all models are the same. We observe something interesting. We ask ourselves, “how likely is it to observe something at least this interesting under the assumption that the world is boring?”

If the answer is large then we don’t conclude anything and continue to assume our gloomy outlook.

If the answer is small then we take it as evidence against the null hypothesis of boringness and conclude that there is at least one interesting thing going on.

Problems:

Evidence against the null hypothesis is not necessarily evidence for the alternative. If the alternative is at least as unlikely to have produced the interesting result as the null then we are bound to make false conclusions.

It does not take practical significance into account at all. Statistical significance is established first and then practical significance is considered as an entirely separate step.

It does not account for any prior knowledge or similar tests already constructed. This can create bias in both directions very easily.

Can you think of a case where prior tests would favour the null and where it would favour the alternative?

Bayesian approaches:

Two popular Bayesian alternatives are:

1. Regions of posterior equivalence. This is where we explicitly bring in a measure of practical significance and calculate a posterior probability of being within or outside the region, or of posterior regions overlapping an interesting amount.
2. Model comparison. This is where we build models that fit each (of possibly many) reasonable hypotheses and see which is best supported by the data.

How would you test equality of means and equality of variances at the same time when your data is grouped?

# Goodness of Fit (GoF)

One of the popular classes of hypothesis tests is those that try to ask whether ‘a model fits the data’. At least that’s what some people would like to ask, but that is not actually a testable or sensible question. In reality, data comes from the true data generating process (complicated) while the model is much simpler and thus cannot truly ever fit the data perfectly.

A better question is, “Assuming that the model is an accurate approximation of the data generating process, how surprising is the discrepancy between the observed data and what would be expected based on model predictions?”

We would like to summarise this discrepancy using a p-value. What do we need in order to calculate a p-value?

## Independent and identically distributed observations

For distributions with no parameters, like the U(0,1), goodness of fit is very easy. There are many options for statistics (AD and KS are the most popular) and the distribution of that statistic is only dependent on sample size.

For distributions that are truly location-scale invariant we can often find statistics which have distributions that are also invariant to the location and scale of the original data. These statistics merely need to account for the sample size and the fact that the parameters must be estimated, not the actual values of the parameters. Examples include most normality tests (like Shapiro-Wilk) and exponentiality tests.

For most distributions though, goodness of fit statistics have distributions that depend on both the sample size and the values of the unknown parameters, meaning that they cannot be established exactly. This implies that we must use a parametric bootstrap procedure to get a p-value. So the procedure is then: 1. Estimate parameters, 2. Get statistic, 3. Simulate many new samples using parameter estimates, 4. For each new sample estimate those parameters, 5. For each new sample get the statistic, 6. Get proportion of new statistics that exceed (more extreme/surprising) than the original statistic (and call the final result a p-value).

How does this change in a Bayesian model fit?

## Regression models

In a regression model, every observation follows a different distribution. Sometimes people say that the residuals are i.i.d., but that only applies to the theoretical residuals, not the observed residuals. Testing the observed residuals in an ordinary regression model for normality actually tests for a lot of things, but normality is surprisingly low on the list of things it tests for. Thus, we need to adjust for this if we want to check goodness of fit.

Previously, people attempted using the parametric bootstrap approach for such problems too, but it is very computationally intensive, and unstable.

Luckily there is a clever new approach that works rather well: DHARMa. The principle is simple:

1. Get the predictive distribution for each observation (or a simulated sample from it).
2. Get the empirical CDF value of each observation according to its own distribution (proportion of simulations less than the observation).
3. Test whether these empirical CDF values come from a U(0,1) distribution (as they should if the model was a perfect fit).

While this approach works for both frequentist (e.g. GLM) models and Bayesian models, it flows naturally from any Bayesian regression model. It is essentially equivalent to the notion that x% prediction intervals should cover x% of the observations, regardless of the value of x.

Further, a U(0,1) qqplot can help identify where the model and the data most disagree, highlighting possible outliers. How might an outlier present itself on a qqplot?