Next Steps

NRES 710

Fall 2020

We've reached the end of the course, congratulations!!

But this is not the end of your statistical training - learning statistics and modeling (informally or formally) is an ongoing part of your growth as a scientist!

The types of analytical approaches you decide to learn (informally or formally) depends on a lot of factors:

- the type of data you will be analyzing (both responses and predictors)
- whether your data are primarily coming from controlled experiments or uncontrolled field-based studies
- whether you are interested in understanding mechanisms or making predictions (or both)
- whether you have one predictor variable or multiple predictor variables.
- whether you have one response variable or multiple response variables.
- whether you are testing hypotheses or generating hypotheses
- whether you have key sources of non-independence like phylogenetic dependencies or spatiotemporal dependencies.
- whether you align philosophically more with Bayesian or frequentist inference

Here are some final thoughts to help you think through your next steps as a data analyst!

Bayesian vs frequentist

This course has been entirely grounded in frequentist inference- which is a bit strange for me, as someone who tends to gravitate toward Bayesian inference!

In both Bayesian inference and frequentist inference, the truth is unknown. But the two approaches differ in how they treat probability and uncertainty.

In frequentist inference, the truth is fixed but unknown, and incomplete sampling prevents us from knowing the full truth. Sampling variance is the only source of uncertainty. Therefore there is uncertainty associated with sample statistics- but no uncertainty associated with population parameters. It is not appropriate to think of the truth itself (the population parameter) as uncertain.

In Bayesian inference, probability is treated as a *degree of belief*, and not as a reflection of sampling uncertainty. In this way, we treat the population parameters themselves as uncertain, which reflects incomplete knowledge of the truth. We can therefore assign probability distributions to the population parameters- more "peaked" distributions represent more precise knowledge and flatter, spread-out distributions represent less precise knowledge (often due to insufficient data).

Many of you are likely to find Bayesian statistics more intuitive than frequentist statistics- I certainly do!

For example, you can interpret confidence intervals as (e.g.) "I am 95 confident that the true value is between this lower bound and this upper bound". This is the way most of us WISH we could interpret confidence intervals.

NOTE: in Bayesian inference, confidence intervals are often called 'credible intervals' to differentiate them from frequentist confidence intervals.

Parametric vs nonparametric tests

If you are running statistical tests and you want to be sure your results are not compromised by a failure to meet parametric assumptions, then you might want to dive deeper into nonparametric statistics. The most powerful and flexible family of nonparametric tests are called **permutation tests**. These tests rely on dissociating your response and predictor variables by "scrambling" the information, repeating this procedure hundreds of times, and seeing if your observed level of signal exceeds the expected level of signal from these "scrambled" datasets. If so, you can reject your null!

Learning to bootstrap

Bootstrap methods allow you to generate confidence intervals and sampling distributions for pretty much any test statistic you'd like, without relying on known distributions like the t-distribution, F-distribution, Chi-square distribution, etc. The bootstrap techniques simply involve repeatedly running the analysis on datasets sampled with replacement from your actual dataset. Learning to bootstrap is well worth the time for almost any aspiring data analyst.

Machine learning

Machine learning techniques are very diverse approaches for detecting signals in datasets. Machine learning methods tend to impose far fewer assumptions about the data than classical statistical and modeling approaches. For example, **random forest** methods enable detection of non-linear responses and complex interactions without having to specify these non-linear functions and interactions in advance. We just provide the algorithm with a response variable and a bunch of predictor variables and let the random forest algorithm find the signals. In addition, supervised machine learning methods like random forest allow you to rank your predictor variables in order of their importance (degree of useful predictive power).

Machine learning methods are often useful for performing variable selection (identifying a small set of important predictor variables), and for identifying important non-linear responses or interactions.

Supervised machine learning (e.g., random forest) involves giving the algorithm a bunch of observations with known response and known predictor variables and letting the algorithm find the relationships between the response and predictor – just like multiple linear regression.

Unsupervised machine learning (e.g., clustering algorithms) involve giving the algorithm a bunch of observations, each associated with a bunch of measurements, and allowing the algorithm to identify key differences and similarities (patterns) in the data. For example, the algorithm my divide the observations into groups that share certain similarities. These groups (clusters) can then enable researchers to identify key 'hidden' similarities between observations.

Multivariate statistics

Multivariate statistics involves trying to interpret multiple response processes at the same time. For example, community ecologists often want to understand the presence/absence and population dynamics of multiple species in a single ecological community. These researchers therefore are dealing with muliple response variables at the same time - for example, they may wish to model the presence or absence of species A, B, C, and D as a function of three environmental covariates (x1, x2 and x3) and inter-species competitive exclusion. Such an analysis would be called a **multivariate analysis**.

I will try to post more on multivariate statistics later - and definitely in the next iteration of this class!

Dimensionality reduction

Often classified as a type of multivariate statistical analysis, **principal components analysis (PCA)** and other dimensionality reduction methods (e.g., non-metric multidimensional scaling, or NMDA) are useful for taking a large set of variables and trying to summarize the variation among observations according to a smaller set of derived variables (often known as "axes" – for example, "PCA axis 1"). This can enable visualization of similarities and differences among observations – eg., by plotting PC1 vs PC2.

Some analysts will run a PCA on all their observations (e.g., all 100 predictor variables) in order to reduce their set of inter-correlated predictor variables from a very large number to a set of just a few (often 2 or 3) uncorrelated predictor variables. This is one way of dealing with **multicollinearity**. However, it introduces a new problem- how to interpret your new variables!

Last words

I will try to add to this 'next steps' lecture in the future. In the meantime, I wish you all the best as you continue your journey as creative, effective data analysts.

As always, your feedback is very much appreciated.

Thanks for bearing with me this semester!!!

–End of lectures–