MATH 1052H - S62 - Lecture 11

Some Extra Ideas

We are done the material which will be graded for MATH 1052H - but there's so much more to statistics! This lecture will just quickly go through some ideas that are common, but slightly outside the scope of 1052H, so at least you will know their names if you ever happen to see them again. Or know that you've seen them … and maybe that'll be enough to let you look them up.

Simple Hypothesis Tests

Basic Comparative Hypothesis Tests

We saw a number of “basic” hypothesis tests: one mean, one proportion, two means (independent, pooled), two means (independent, not pooled), and two means (paired). There are many, many more.

Chi-Squared Tests

You saw a Chi-squared test of independence and goodness-of-fit; there are also chi-squared tests for one and two variances (like means).

Tests on Medians

You can compare the median of one population to a null hypothesis, or two medians to each other.

Non-Parametric (Rank) Tests

Recall that we had a lot of assumptions built into our tests - the data being approximately normal, homoskedastic residuals, etc., etc.. Sometimes you don’t have this! What do you do?

This leads to an entire area of statistics known as “non-parametric”: that is, not based on parameters. These have many less assumptions. Let’s look at a couple of examples.

Mann-Whitney Test (or Wilcoxon Rank Sum Test)

This is a non-parametric test for the hypothesis “two samples were drawn from the same distribution”. It makes absolutely no assumptions about the distribution, and relies only on the **relative ranks** of the observations in the combined sample.

Runs Test (or Wald-Wolfowitz Test)

This is a test of randomness (similar to goodness-of-fit in some ways). It compares the lengths of runs of the same value in a sample to what would be expected in a random sample, and similar to Mann-Whitney, has no assumptions whatsoever.

Linear Models (ANOVA)

ANOVA Models

We saw one- and two-way ANOVA (the latter with interaction). You can easily scale ANOVA up to n-way ANOVA (just keep adding factors!). Here are some of the variations you might see in a more advanced setting.

ANOVA: Repeated Measures

Repeated measures design is a research design that involves multiple measures of the same variable taken on the same or matched subjects either under different conditions or over two or more time periods.

This is quite common in experimental psychology, where multiple measurements might be taken **per subject**, with common stimuli and setting.

One neat thing that comes from this design is the ability to break down variability into within-subjects and within-treatments - this gives a lot more control over variance for these kinds of experiments.

ANOVA: Factorial

We saw this in the simplest case as the two-way ANOVA with interaction. As you scale this, it is known as a **factorial design** (or a fully crossed design). For this to work properly, you typically require the experimental units to take on all possible combinations of all possible levels across all of your factors. Often this is done with factors that are restricted to two levels only, in which case we sometimes call this a 2^k factorial: k different factors, each with two levels.

ANOVA: MANOVA

MANOVA stands for Multivariate ANOVA, and is used when you design your experiment and there is more than one response variable. It’s pretty messy, so if you get to this point, prepare yourself for some fairly nasty algebra and learning some new techniques …

Non-Parametric ANOVA

There is also a non-parametric version of one-way ANOVA, called the Kruskal-Wallis Test by Ranks. It is used for comparing two or more independent samples, without assumptions.

Linear Models (Regression)

Generalized Linear Models (GLMs)

In our last example of Lecture 10, we briefly discussed the logistic regression model. There is an entire world of models that all share one idea: rather than modeling

$$y \sim \beta_0 + \sum_{j=1}^{k} \beta_j x_j$$

instead we might want to model

$$\eta(y) \sim \beta_0 + \sum_{j=1}^{k} \beta_j x_j .$$

That is, we have a linear combination of predictors … but it models a function of our observed y variable, not y itself. These are known as **Generalized Linear Models** (not General!), and you can do an entire course on them!

Generalized Additive Models (GAMs)

An extension of GLMs are Generalized Additive Models (GAMs), which relax one further thing: instead of having x_i's as predictors, we allow smooth functions of those predictors to be used instead. These are very powerful, but complex and more difficult to use.

Mixed Effects Models

Very common in ecological applications (biology, environmental science) are GLMMs and GAMMs: Generalized Linear/Additive **Mixed** Models. These models extend the previous two in order to deal with mixed effects modeling - heterogeneity in residuals and variance structure, nested data (random intercepts and slopes, versus fixed in the previous), violations of independence, zero-inflated or truncated models for counts, and Generalized Estimating Equations.

All of these are more reasonable models for the actual, real-world data found in these fields. If you are interested in this specific topic, portions of these models are taught in the fourth-year BIOL course at Trent on statistics and computation.

Time Series

What is a Time Series?

Time series are sequentially ordered observations of the realizations of a random process. What that means for us is … they are not independent! A ridiculous amount of a first course in statistics pretends the entire world is independent, and it’s not.

Time Series are specifically non-independent in a temporal sense: the observation at time t is not independent of the observation before it (or some observations before it).

AR, MA, ARMA and ARIMA

The four most classic and common models for attempting to model time series are the AutoRegressive (AR), Moving Average (MA), AutoRegressive Moving Average (ARMA) and AutoRegressive Integrated Moving Average (ARIMA) models. The details are well beyond the scope here, but now you’ve seen those words, and know vaguely that they are attached to data that has a temporal component.

The Spectrum

The power spectrum, spectrum, or spectral density are all ways of describing the frequency contributions in a time series. They require Fourier transforms, and can be fairly complex … again, now you’ve seen them.

Conclusion

Lots of Models …

Cast your mind back to when we covered linear regression for the first time. I tried to give you a hint then about the complexity and sheer scope of modeling as a framework. Statistical modeling is incredibly varied, and there are hundreds (if not thousands or tens of thousands) of **models** commonly used in science, all based on some form of statistical framework.

Please never be afraid to learn: you have a strong foundation now, and should know enough R that you can learn to use additional packages and code! There's over 13,000 packages on CRAN, many of which have implementations of statistical models and modeling frameworks - just sitting there waiting for you to use them!

Final Note

And remember my note: when you get to the point where you don’t know what to do next, because you’ve never seen that kind of data before … consult a statistician! We’re friendly, and we will help - it’s what we do for science and scientists.