Taxonomy of statistics

NRES 710

Fall 2022

Overview of statistical methods

Before we delve into common statistical tests, let's first take a quick walk through the analyses we will cover in this class. As we do this, we will try to put these analysis into a framework - a field guide of sorts – that will help us to make sense of these methods and when to use them.

And before we do that, let's briefly talk about *parametric* vs *nonparametric* statistical tests...

Aside: parametric vs nonparametric

As we saw in the last lecture, **Parameters** are the arguments used to describe probability distributions - they describe the exact shape and location of the distribution. Different probability distributions are associated with different parameters. For example, the normal distribution is described by 2 parameters: mean and standard deviation. The Poisson distribution has only one parameter (the Poisson mean, also known as Lambda). The binomial distribution is described by 2 parameters: size (number of trials) and prob (success probability for each trial).

Parametric statistics In *parametric statistics*, our statistical tests relate directly to the parameters of well-defined probability distributions (e.g., the mean of a normal distribution). That is, we assume *a priori* that our population follows a certain well-defined distribution (e.g., a normal distribution) and we make hypothesis about one or more of these parameters (often the mean of a normal distribution).

Because of the Central Limit Theorem (CLT), we have good reason (in many cases) to assume that our sampling distribution should follow a particular distribution (e.g., z, t, Chi-squared) even if the underlying data don't perfectly match the underlying assumptions of the analysis. However, this is not always a fair assumption – sometimes the assumptions of parametric tests are clearly violated!!

We are fortunate that statisticians have developed a number of tests that do not depend on our data or data summaries following any defined probability distribution. These methods are the so-called **distribution-free** statistical tests, which are often referred to as **non-parametric tests**.

Nonparametric statistics Unlike parametric statistics, *nonparametric statistics* do not require us to make strong assumptions, such as that the underlying statistical population follows a normal error distribution (like in a t-test or linear regression). Using these methods, we can still test similar (but not identical) hypotheses even if our data do not meet standard parametric assumptions

Confusingly, we often use the term "nonparametric" to refer to two different classes of statistical models:
1. Statistical tests that do not require any assumptions about the distribution of data or residuals. These methods are also known as **distribution-free** tests.
2. Statistical models that do not require assumptions about the **shape** of the relationship between two or more variables. These methods include generalized additive models, spline regressions, gaussian processes, and many methods collectively known as **machine learning** (e.g., random forest). These methods are often referred to as **nonparametric regression**.

Independence of data It is important to note that most classical parametric and nonparametric statistics all make one very important assumption – that observations are independent! Violation of this assumption are often called **pseudoreplication** and can wreak havoc with our type I error rates!

Q Given the CLT makes many classical tests robust to non-normal data distributions, why do we need non-parametric statistics?

Okay, let's get back to our main purpose, which is to provide a "field guide" of sorts for determining which statistical analyses to run. This guide will involve characterizing the type of response variable and predictor variable we have at hand – most importantly, whether we have continuous or categorical data for our response and predictor variables.

Continuous response variable

If your response variable is continuous, common classical statistical analyses include: t-tests, ANOVA, linear regression. Each of these tests is associated with non-parametric alternatives. In each case (for the parametric tests, that is!), we are interested in testing if the *population mean* of the continuous response variable is affected by the predictor variables (null hypothesis: nope, there is no effect!). Predictor variables, in turn, can be continuous or categorical (factor variables in R) – but as you suspected, different predictor variable types will lead us to choose different statistical models!

Continuous response, binary categorical predictor If your categorical predictor variable is binary (two levels) and your response variable is continuous, you can use a **two-sample t-test**.

The non-parametric alternative is called the **Mann-Whitney test**.

You can visualize the effect size using a boxplot or barplot with error bars.

Example: Are females larger than males? (null hypothesis: no difference). The response variable (body size) is continuous. The predictor variable (sex) is binary.

Continuous response, multi-level categorical predictor If your categorical predictor has more than two levels, you can use an **ANalysis Of Variance (ANOVA)** followed by *pairwise comparisons* to test which categories differ from one another.

The non-parametric alternative is the **Kruskal-Wallis test**.

You can visualize these relationships with (e.g.) a boxplot or barplot with error bars.

Example Do three different brands of nail polish differ in their durability? (null hypothesis: no effect). The predictor variable (nail polish brand) is categorical, the response variable (time until product wears off) is continuous.

Continuous response, continuous predictor If your predictor variable and response variable are both continuous, you can use **linear regression** analysis.

You can visualize the effect size using a scatterplot with a regression line and confidence "ribbon".

If you just want to know if two variables are correlated but you are not interested in modeling one (the response) as a function of another (the predictor) you can run a **Pearson correlation test**.

The non-parametric alternative to a Pearson correlation test is a **Spearman rank-correlation test**.

If you don't want to assume a linear relationship, you can use polynomial regression, spline fits (GAM), or machine learning methods.

Example: What is the relationship between tree diameter and age- can tree diameter be used to effectively predict the age of a tree? (null hypothesis: nope, there is no relationship). The predictor variable (age) is continuous, and the response variable (tree diameter) is also continuous.

Continuous response, both continuous and categorical predictors If you have one categorical variable that you hypothesize may be influencing a continuous response variable– but there is also a "pesky" continuous variable that you also suspect is influencing your continuous response – you can use **Analysis of Covariance (ANCOVA)**

More generally, if you have a set of continuous and categorical predictor variables that you hypothesize may be influencing your continuous response variable, you can use **multiple linear regression**.

You can visualize these relationships using boxplots and scatterplots with regression lines (and confidence intervals).

Practically speaking, linear regression and ANOVA/ANCOVA are two sides of the same coin. Both involve modeling the mean of a continuous response as a function of one or more predictor variables. Both assume an underlying normally distributed population.

You can run all of these analyses using the workhorse of linear modeling in R, the `lm()` function. Basically, if you have a continuous response variable, you can use the `lm()` function!

Linear regression is parametric, and assumes the residuals (differences between observed data and predictions) are normally distributed. More generally, it assumes that the error distribution for the underlying population is normally distributed.

The non-parametric alternative might be something like a **distribution-free regression tree** analysis (e.g., `party()` in R, which makes no assumptions about the distribution of residuals or the shape of the relationship) or a **generalized additive model (GAM)** (which assumes the residuals follow a defined distribution, but makes not assumption about the shape of the functional response).

Discrete (count) response variable

With a discrete count response, you can either assume a continuous response variable and use the same techniques as you would for a continuous response (e.g., linear regression). If this is not justifiable, you can use **generalized linear models (GLM)** with an error distribution that follows a **discrete** probability distribution such as a Poisson distribution (see previous lecture).

This type of relationship can be visualized with a scatterplot and regression line (with confidence band)

Generalized Linear Models (GLM) are a widely used parametric class of models that enable researchers to perform regression analysis while *relaxing* the assumption that residuals must be normally distributed. In GLM, residual error can be distributed according to discrete distributions like Bernoulli (binary), Binomial, Poisson and Negative Binomial – or according to non-normal continuous distributions like the Gamma or Exponential distributions.

Nonparametric options include classification/regression trees and GAM.

Example: does the number of eggs produced per year by a desert tortoise depend on the availability of food in the prior year?

Categorical response variable

If your response variable is categorical (factor variable, ordinal variable, binary variable) then your choice of statistical methods changes:

Categorical response variable, categorical predictor If both your response variable and your predictor variable are categorical, then you can use a **chi-squared test** or a **Fisher exact test** to test for an association between the two variables.

In this case, it can be informative to summarize your data as a *contingency table* like we did with the 'lady tasting tea' example. That is, we make a table that summarized the number of observation in each unique category of our response and predictor variables.

Example: test for an association between salamander color morph (melanistic vs wild-type) and mating behavior ('sneaker' vs territory holder).

Categorical response variable, continuous predictor If your response variable is binary (true/false, two levels) and your predictor variable is continuous, you can use **logistic regression** (which is a type of *generalized linear model (GLM)*).

Example: Does the probability of site-occupancy for American pika depend on elevation?

If your categorical response variable has more than two levels, you can use **multinomial logistic regression** (for categorical responses) or **ordinal logistic regression** (for ordinal responses). I don't plan to explore these methods in this class, but it can still be useful to know they exist!

When the taxonomy breaks down

Not all statistical tests fall into this neat hierarchy. Take a one-sample t-test for example. Imagine you are asking the question: Is the sample mean equal to 32? What exactly is the predictor variable?

Another example might be a normality test. We want to test if our data are normally distributed – again, what is the predictor variable in this case?

Discussion: final projects

For the remainder of the class period, let's discuss your final projects. Take a minute to think through your data set and try to determine: what is your response variable, what is your predictor variable? Are these data continuous or discrete? Based on the taxonomy, what main analysis or analyses do you anticipate using for your final projects???

–go to next lecture–