Intro to NRES 710

Welcome!

Last compiled: 2024-08-26

## Introductions

Introductions among faculty ([Brian Folt](https://www.unr.edu/nvcfwru/folt-brian)) and students

## Class structure

The most important feature of this class is the [course website](https://brianfolt.github.io/NRES_710/)! <https://brianfolt.github.io/NRES_710/>

We will discuss the course organization and expectations. Please see [course syllabus](syllabus.docx) and [schedule](schedule.pdf) documents, provided separately.

**Office hours:** no office hours, best thing to do is send me an email and find a time to meet; I am happy to meet with you; my office is in KRC 100, enter from the north side of the building.

**Prerequisites:** Students are expected to have a basic understanding of standard statistical concepts and methods, obtained through other coursework. They should be familiar with what a p-value is, an ANOVA, a t-test, these sorts of basic things. But, we are going to cover all of these topics in class – with a practical emphasis.

**Schedule:** The typical schedule for the class will be a **lecture** on Tuesdays and either another lecture or an laboratory **Exercise** on Thursdays. Some days there will be required readings to be read prior to class, so keep an eye on the schedule for those.

**Attendance:** This is a graduate student class and attendance is not mandatory; you are all adults and make your own decisions. Life happens & you will be busy with field work, other obligations, conferences, whatever. It’s okay: if you can’t make class, no reason to tell me.

**Exercises:** The largest part of your grade in the class will be from **Exercises** (60%). These will be every two weeks ago so during the semester, and my goal is have eight different exercises. We will use class time to work on the Exercises; you can work on them individually or in small groups with 2-3 people. I encourage you to work on these with your peers, because a great way to learn material is by communicating it to someone else. However, the goal of the class is to learn how to analyze data and learn those practical skills. So I encourage you to be mindful and make sure that you are learning the material yourselves.

On class days with Exercises, we will work on the material in class. I will be present and available for questions. I will be available to help answer questions and provide guidance on how to complete the Exercises! My hope is that you all will be able to finish the Exercises during classtime, and then I will briefly review your work and give you credit for completing the assignment. However, if you are unable to attend the class session or unable to finish the Exercise during class, you will have six days to complete it and submit it through WebCampus.

However, if at all possible, please attend **Exercise** sessions. My goal for these sessions is that we will all work through the Exercises at the same time. My hope is that you will all be able to finish these important learning assignments during the class session, and then I will review your work and mark you down in the gradebook as having completed the assignment properly. If you can’t make the Exercise class for some reason, that’s okay, you can submit the assignment through WebCampus.

Lab assignments will be up in advanced of the lab. You are welcome to start working on them in advanced, show up to class, get your work checked off, and leave.

**Quizzes:** There will be 6-7 quizzes throughout the semester (20%). They will be given through WebCampus and they will become available the day after the material is presented. Sometimes there will be quick turnaround times on the quizzes and you will only have ~36 hours to finish them. So be mindful of quiz dates and getting them done before they are due.

**Participation:** Students will be graded on **Participation** (20%). As this is a graduate-level class, engagement and participation during class, discussions, exercise labs, and other course-related opportunities is expected and students will be given a grade on their participation.

**Extra Credit:** An extra-credit opportunity is available to students by participating in ‘[StatsChats](https://guides.library.unr.edu/StatsChats)’. StatsChats is an informal weekly gathering to discuss any quantitative questions you may have related to your research (statistics, experimental design, model simulation, analysis, debugging R code, etc.).

Faculty from various departments – Paul Hurtado (math/stats dept), Ken Nussear (geography), Perry Williams (NRES), Kevin Shoemaker (NRES) and others – have hosted informal sessions for grad students in EECB, NRES, Geography etc (and faculty) to discuss data analysis questions. This can be a good opportunity to ask questions about statistics, find out more about what types of data and questions your peers are working with, and help others (thereby reinforcing concepts for yourself!). Keep an eye on that website linked above and if the group identifies a regular meeting time, I will relay that information to you all in class.

This is open to any students and faculty that want to improve their their quantitative skills and/or share their knowledge and experience. StatsChats will be weekly on Tuesdays (starting September 3) from 12-1PM in the Lilli Brant reading room in the DeLaMare Science and Engineering Library (third floor). Students will gain +20% credit to their overall class grade **if they attend four (4) out of the ten total StatsChats sessions during the semester**.

**Course website!** A central location for class materials, including lecture and exercise materials, readings, and other useful information.

## General overview

NRES 710 is designed to be an introductory, graduate-level statistics class. “Introductory” because you are not assumed to have much in the way of prior statistical knowledge (although an undergrad course under your belt will be helpful) or prior experience in R programming (although some prior exposure will make the learning curve less steep). “Graduate-level” because once we get rolling, it’ll be full steam ahead! However, another dimension of a graduate-level course is an emphasis on collaborative knowledge development through group discussion and collective learning, and NRES 710 will provide opportunities for this.

This course provides an introduction to both statistics and computer programming in R. With a solid grounding in basic rules of probability and statistics, and armed with some basic computer programming abilities (and some human creativity and ingenuity), you can go extremely far in making sense of data and communicating that understanding with others!

The purpose of this class is two-fold. **First**, I want to teach you many of the **common statistical tests** that you will see in various research journals in your field! My hope is that by avoiding diving into the weeds on theory, we can explore many different statistical tests in simple, easy to understand, and practical ways. However, I will try to teach you some high-level theory about these tests. I want you to understand enough theory to know:

* *why* you are using a statistical tests?
* what are the *assumptions* that the test makes?
* what do the results *mean*?

And the **second** goal of the class is to **learn how to do these tests in the statistical program R**. Why using R…? It’s awesome, flexible, and **free**. Many other schools have their favored softwares for statistics, but most people in ecology and environmental sciences use R. It is free, open source, and growing at an exponential rate in terms of what you can do with it. Packages are being developed daily to produce additional statistical capacity for it. It’s great.

‘Base R’ is the default software built into R that does not including loading any additional packages. Here is a [‘base R’ cheat sheet](base-r.pdf); this is a great reference for most of the basic tasks you will need to perform in R.

Learn to use R scripts, and save your scripts frequently! This is the primary record of what you’ve done and allows you and others to reproduce your workflows.

**No text book.** We will follow material that I am producing and that will be available on the course website and other free resources from the internet. Part of the reason why we have a course website is to provide ~physical documentation of everything we are learning: lectures, code, exercises, resources, etc. So the course website will be the primary resource for learning the material in the class. However, there are additional links on the Syllabus and the the ‘Links’ page on the course website that provide access to other free, openly available information.

This is my first time teaching this class. Given this is my first time teaching this specific material in this format, I welcome frequent and honest feedback on course topics to understand whether course topics are appropriate and/or useful. So, any feedback is welcomed – thank you!

## Grade

* Exercises (60%)
* Quizzes (20%)
* Participation (20%)
* Extra credit (20%)

## Course Schedule

Let’s take a look at the course schedule.

Links everywhere! I will update these weekly to add lecture material after it is presented in class.

Readings! Note the dates associated with readings and be sure to read the papers before the class date indicated.

Code! I hope to make all of the code for lectures available to you before class. My intention with that would be that as I am demonstrating code on the computer, that you could follow along in R on your computer and take notes in your code. I still need to figure out a way to do provide this, so hang tight.

**Data** and **Truth**. In this class, we will generally not use data of unknown sources. One thing that I found challenging about statistics classes is they often give you data, you run an analysis, and afterword you ask yourself: ‘Did I do this right?’ ‘Is this the right answer?’ How do we know?

In this class, we will **generate all of the datasets that we analyze**. We will tell R what ‘truth’ is, and then we will simulate data. Then, when we learn how to use the statistical tests, we will know whether they are working properly, because they should estimate values that we know to be ‘truth’.

The datasets will be fixed-number CSV files. But these will be generated by R-code in the lectures or laboratories. I will provide the R-code that is used to simulate truth, so you can examine it to see what the values are for truth and learn how data simulation works.

Any questions about anything related to how the class will work?

How many of you have used R before?

How many of you have used other statistical softwares: SAS, JMP, SPSS…?

## What is science?

**Science is the search for *truth* through the accumulation of *facts*. It is a method of answering questions.**

* The terms *facts* and *truth* are often used synonymously.
* **Truth** is the way things really are, what’s really happening, what we want to know about how the world works.
  + The problem is that we can never really know truth. We can seek to understand it and try to know it, but we can never know truth.

Truth exists, and we want to know it, but we can never really know it because of two problems:

1. **Sampling error** – there are always errors in measurements.

* How tall is Dr. Brian Folt? If everyone in the class tried to measure me to the mm, how many different measurements would we get? 25!
  + Error will happen because I slouch, people have to stand on a chair, shaky hands, etc., etc.
  + This is all sampling error!
  + **Q:** How can we decrease sampling error? Standardize procedure, measurers, tools…
* We can decrease error by standardizing measurers and measurement methods. And then we can take an **average** to get a pretty good **estimate**.

1. **Process error** – **underlying features that contribute to variation in whatever is being measured**. This is harder to understand, and we might benefit from using a slightly more complicated example.

**Q:** On the UNR campus, who is taller: **Students** or **Faculty**?

* Walking around campus, we might make an observation that Students are taller than Faculty. They are young, growing, and have great posture. Faculty are old, aren’t growing, have bad posture (me!)
* Students are maybe a little taller on average…? But there is overlap. Not all students are the same size, and not all faculty are the same size. There is variation in each of these groups that causes overlap in size. This overlap makes it hard to know exactly how much taller students are than faculty. What causes this overlap?
* **Process error**. Process error is underlying processes that contribute to the outcome on an individual by individual basis. What things might cause height to vary among students and faculty?
  + Diet
  + Genetics
  + Behavior
  + Age
  + Etc.

**Process error** – **underlying features that contribute to variation in whatever is being measured**.

* If every faculty member was the same height and every student was the same height, it would be really easy to measure the difference. But they aren’t! There are other processes that contribute to height outside of whether someone is a student or faculty member, which makes it difficult to come up with a good estimate of height.

This is due to process error. Process error is often thought of as **‘Noise’** in the system. Process error can be due to systematic systematic processes that we might be able to measure, or random variation.

Statisticians often chalk up process error as ‘random noise’. But this is noise is often driven by underlying processes! We can create predictive models to understand whether coins will land as heads or tails, to predict what clothes we will wear in the morning (optimizing style, comfort, etc.).

**Q:** Why does process error make it difficult for us to understand truth? For example, would we test for height differences between Students and Faculty? What would our experiment be?

* Sample!! We might ask 100 students and 100 faculty how tall they are, and it might give us an estimate of an effect. Why might this not be accurate?
  + People lie?
  + Small sample size? OR, sample size is not *representative*
* Our sample may not be representative, due to random chance and process error. If by some random chance, we don’t get any samples from the basketball team, then we estimate for student height might be LOW. Or, if we only sample the gymnastics team, also low.
* With small sample sizes, those underlying process errors can influence our ability to measure real effects.
* But, if our sample is representative, then we should be able to get pretty close to ‘truth’ – although there will always be some error.

**Q:** How could we ensure TRUTH?

If we sample the entire population, we can approximate truth! But it’s often impractical to get the whole population. We often don’t have access to the entire population, and we must resort to taking samples.

So we have talked about **truth**, what about **facts**?

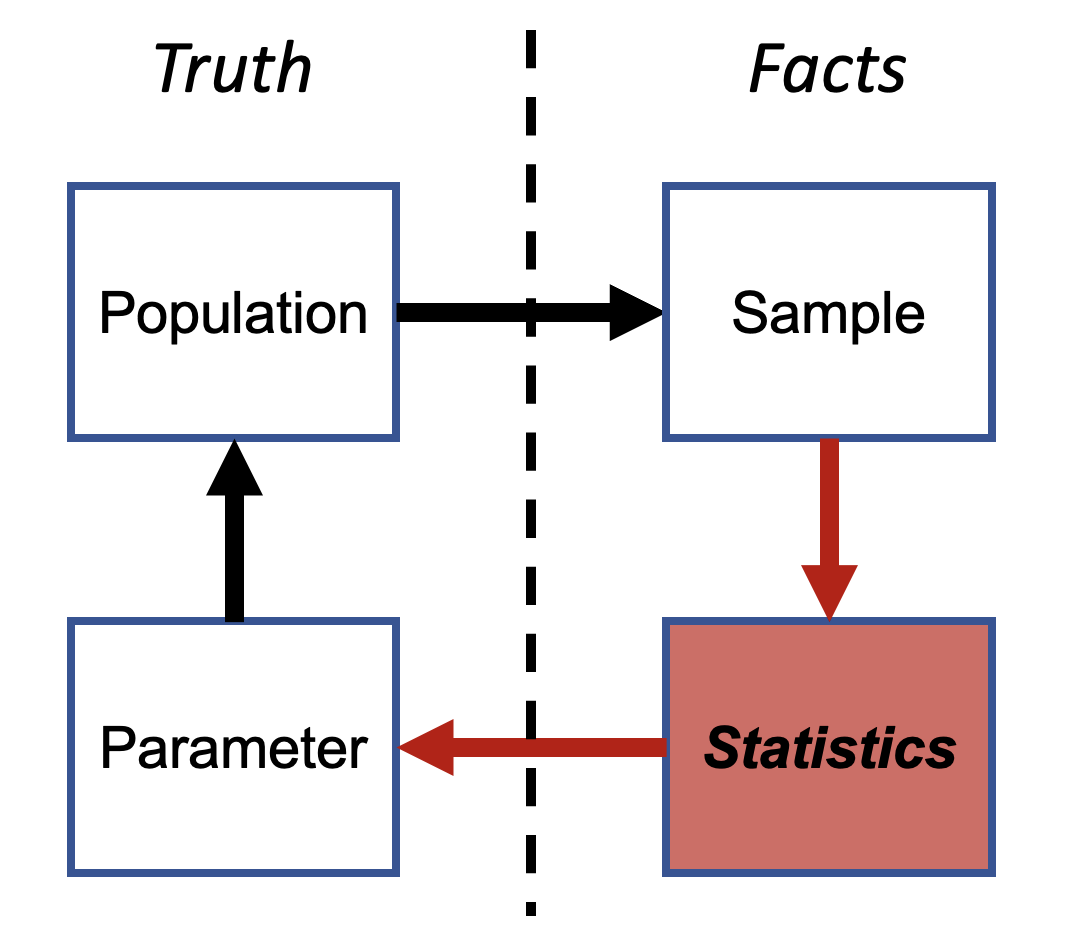
**Facts – estimates of truth** from our sample

* Going back to our original definition, science is the search for truth through the accumulation of facts. We want to understand how the world works – and we do that by collecting data and accumulating facts. Data are facts.
* Measurements. Facts might be accurate, inaccurate; they might be reliable and repeatable, or they might not be reliable and repeatable.
* We go out and we collect data, and these data are facts.

How does ‘statistics’ come into this? **Statistics is the method (math) by which we use facts (data) to estimate truth.**

* We collect data (facts) and use statistics to calculate an average (a statistic), which is an estimate of truth, based on facts.
* You can’t do science without statistics!

We can use statistics as a process to make inference about a property of a *population* (a parameter) from a representative *sample*. In the figure below, the vertical line separates the stuff we can observe and measure directly (*facts*; right side) and the stuff we can’t observe but want to make inference about (*truth*; left side).



Do you see the ‘p-value’ anywhere in here?

* P-values are not statistics. They are not estimates of truth.
* P-values have their value and their place in this field. But, it’s my opinion and others that most statistic classes emphasize P-values way too much, in my opinion. That’s why we are going to read papers about on this topic before class on Thursday.

In we are going to learn how to use statistics to estimate truth. To do this, I am going to make my own data! That way we can know whether the statistical methods we are using are estimating truth properly! We will learn about what a p-value is and how it is used, but try to de-emphasize p-values and place more emphasis on measure effects and uncertainty around those effects. I also mentioned that truth is impossible to know, so how can we know if we are doing a good job of estimating it? We will SIMULATE DATA to know truth, so that we can then fit statistical analyses and ensure that we are correctly estimating truth.

[–go to next lecture–](lecture_2.html)