

STAT 140:
Week Zero
Catch Up

Goals for the
Week

Questions

STAT 140: Week Zero Catch Up

09/03/2021

My hope is that after Week 1 (08/30-09/03), you will be able to:

STAT 140:
Week Zero
Catch Up

Goals for the
Week

Questions

- Identify populations and samples, as well as parameters and statistics.
- Describe datasets and identify discrete, continuous, nominal, and ordinal variables.
- Discuss different types of sampling and their advantages and disadvantages.
- Differentiate between anecdotal evidence, observational studies, and experiments.

Labs

STAT 140:
Week Zero
Catch Up

Goals for the
Week

Questions

Question: Are we going to have lectures on Monday and Friday and then practice R on Wednesday?

Yes! That's roughly the schedule I plan on following. Lecture on Monday, practice/lab on Wednesday, and a "catch up" lecture on Friday.

Question: Are the labs going to be completed within class time or will it require us to work on it outside of class as well?

Labs should be completed within class time (assuming I time them correctly!). You will not have to meet outside of class time, but it is a good idea to finish the whole lab so you are prepared for the homework assignments.

Tests and Exams 1

STAT 140:
Week Zero
Catch Up

Goals for the
Week

Questions

Question: I am not sure that I understand the difference between a test and an exam (they are the same in my mind). Could you go over this, please?

Question: Are we going to have tests or exams on using R?

A test will be like a long progress check with short answer questions (multiple choice, true/false, etc.). The tests will not have any questions meant to be answered using R. The exams will be more like a long homework assignment, and will take place in the middle and end of the semester. The exams count for more points than the tests.

Tests and Exams 2

STAT 140:
Week Zero
Catch Up

Goals for the
Week

Questions

Questions: Will exams and tests be open note? I was also wondering how much time we will be given to complete exams/tests would be? Lastly, will they be “take home” or are we completing them in class?

Tests will not be open note. They are like the progress checks in that they should give you an idea of where you are in the course before too much time has passed. The exams will be open note—I would never sit down to write code without my own notes, so I won't ask you to either!

Project

STAT 140:
Week Zero
Catch Up

Goals for the
Week

Questions

Question: Is the final project a semester-long project?

Question: When in the semester will we learn more about and start working on our final projects?

Question: Regarding the project, will we have check ins along the way?

Not quite semester long. It will be completed over roughly a month near the end of the semester. At present, I have some class periods set aside to work on the projects, so you can check in with me at least once before turning the final version in. We will learn more about working on the projects in November.

Evening Help

STAT 140:
Week Zero
Catch Up

Question: I know some classes have PLUMS hours or other times after class where you can get help on homework from a TA or someone. Does this class have anything similar?

Yes!! I am setting up evening help sessions with our TAs, Emma and Anh. When we finalize times, I will post them in the Zoom channel. Hours should start sometime next week, possibly the week after.

Goals for the
Week

Questions

Populations and Samples

STAT 140:
Week Zero
Catch Up

Goals for the
Week

Questions

Question: I could use some review on the difference between populations and samples – it seems like the same group of people could be considered either a population or a sample (depending on how you define your population). Is the difference only in how the data is used (as in, information known about a sample would be used to make conclusions about a larger group, while for a population it wouldn't be)?

This is a very deep question! A sample is a subset of the population—so mathematically speaking, a sample could also be the entire population. But we generally don't think of it that way, and we do consider the sample to be used to draw conclusions about the larger population. If you had the entire population, you wouldn't need statistics! You could just calculate the parameters directly and there would be no uncertainty about them.

Variable Types 1

STAT 140:
Week Zero
Catch Up

Goals for the
Week

Questions

Question: Sometimes I have difficulty in distinguishing variable types, especially for nominal variables and ordinal variables.

Nominal variables have no natural order, whereas there is a hierarchy or natural ordering for ordinal variables. Beyond this week, we will mostly be working with nominal variables.

I am kind of confused between differentiating between numerical and ordinal variables. Looking back on the example, [it seems like] they're sort of numerical, like the class year thing. Since the first [level] is "first year", isn't that just a description of a numerical value?

In this particular example, class year can also be recorded as numerical (e.g., 1, 2, 3, 4). But we don't always know the actual values associated with the levels (like in a five star rating system, or maybe low income (0-\$50k), middle income (\$50k - \$100k), high income (\$100k+)). Then, you would treat the variable as ordinal.

Variable Types 2

STAT 140:
Week Zero
Catch Up

Goals for the
Week

Questions

Question: Where do you draw the line between discrete and continuous variables? For example, step count data covers a very wide range of values, but you can't necessarily track less than one step, so would it be continuous or discrete? This relates to the red blood cell count question also.

It depends on what exactly you are trying to do. Sometimes there is a range of discrete variables that is so wide, we can treat it as continuous—more on that later. Frequently, continuous variables are recorded as discrete, e.g. height. They definitely bleed into each other, and hopefully identification will get easier.

Variable Types 3

STAT 140:
Week Zero
Catch Up

Goals for the
Week

Questions

Question: When you discussed discrete variables, you said that they only took values with “jumps” or “steps,” and the examples we discussed all involved whole numbers of something. If for some reason there were a set of variables whose values could only end in a specific, terminating decimal and had “steps” like 1.25, 2.25, 5.25, etc., would they still count as discrete since they can’t take on any real value the way continuous variables can?

Another good question! It depends on what you are trying to do, but yes, I could envision a situation where there would be jumps or steps with decimal values. Hopefully, we will see examples of this later.

Anecdotal Evidence

STAT 140:
Week Zero
Catch Up

Goals for the
Week

Questions

Question: Anecdotal evidence was briefly mentioned. Was this solely for the purpose of demonstrating that it is not well-suited to statistical analyses, or are there actually instances in which one would use anecdotal evidence?

Yes. You can read more and discuss anecdotal evidence in your journal, but in general, statisticians frown on using anecdotal evidence. I think the times where it is useful is in journalism or written articles designed for story telling—for example, this article presents statistical evidence on mask wearing but also gives anecdotal evidence to really drive the point home.

Replication

STAT 140:
Week Zero
Catch Up

Goals for the
Week

Questions

Question: Under which circumstance do the researchers have to replicate the study?

When designing your study, you would hope that a study *could* be replicated at any point. Researchers rarely have to replicate the results! But it is generally a sign of good science if you can repeat your study and get consistent results.

Stratified Sampling

STAT 140:
Week Zero
Catch Up

Goals for the
Week

Questions

Question: When taking a stratified sample, is the number of random samples taken from a particular stratum proportional to its representation in the broader population? For example, if a stratum represents a group that comprises twenty percent of the population, will twenty percent of the random samples for the study come from that stratum?

Yes! That is exactly correct. When we say representation, we mean that the proportions from the sample and population match (see next slides for another example).

Clusters and Strata 1

STAT 140:
Week Zero
Catch Up

Question: I'm a little bit confused with the difference between strata and clusters. Can you provide more examples?

Consider this map of a 9-story apartment building.

11*	12
1st Floor	
14	13

21	22*
2nd Floor	
24	23*

31	32
3rd Floor	
34	33

41	42
4th Floor	
44	43

51*	52
5th Floor	
54	53

61	62
6th Floor	
64	63

* = Children in the apartment

71	72
7th Floor	
74*	73*

81	82
8th Floor	
84*	83

91	92*
9th Floor	
94	93*

Clusters and Strata 2

STAT 140:
Week Zero
Catch Up

Goals for the
Week

Questions

- If the apartment manager is trying to estimate how many apartments need their carpet replaced due to wear and tear, they may carry out a cluster sample to make it easy on themselves—say they select two floors and check every apartment on that floor.
- However, if they suspect that apartments with children may have higher wear on the carpets, they might carry out a stratified sample and ensure that 25% of the apartments in their sample have children living in them (same as the entire population).
- A multistage sample would be if the apartment manager was first randomly sampling floors, then randomly sampling apartments on that floor—there are multiple stages of random sampling, so it is a multistage sample.

Sample Characteristics

STAT 140:
Week Zero
Catch Up

Question: Is there any way to include a margin of error based on bias?

Goals for the
Week

Questions

Blind

STAT 140:
Week Zero
Catch Up

Goals for the
Week

Questions

Question: In a double-blind study, do researchers draw conclusions by eventually finding out which study participants were in the treatment group, or do they just look for whatever effect the treatment is supposed to have throughout the sample and assume that if that effect is present in significant quantities, it means that the treatment elicits the intended effect?

They do eventually find out! Researchers find out, and I think (don't quote me on this) part of ethical treatment of study subjects includes letting the subjects know whether or not they got the placebo and if they need to pursue the treatment.

Terminology Review

STAT 140:
Week Zero
Catch Up

Goals for the
Week

Questions

Question: I am a little unclear on the specific differences between blocking and stratifying data. Could we go over a quick summary of clustering, blocking, stratifying, and maybe multistage clustering?

We already saw an example of the difference between clustering, stratifying, and multistage sampling (with the carpets). Blocking is a technique used in experiments, not samples. In an experiment, we would put similar groups in blocks (say, males and females). Then, within those groups, we would randomly assign subjects to the treatment and control groups. If you failed to block first, it would be possible to get a really unbalanced experiment—possibly 70/30 females and males in the control group, and then 30/70 females and males in the treatment group. Then, researchers might identify a treatment effect, but it might really have to do more with sex than the actual drug. To avoid that problem, we block! So, in the end, blocking is similar to stratifying in that the end goal is representation, but the techniques are used in different scenarios (experiments, surveys).

R Help 1

STAT 140:
Week Zero
Catch Up

Question: Do we need to calculate a lot in this course?

Question: Do we need to code in R?

There will be a lot of calculation, and we will hit a point in the course pretty quickly where doing calculations on something like a calculator will be unfeasible. Thus, all of the calculation in this course will be done via R, and we will spend a lot of time developing our R skills.

Question: Is statistics always involved in coding, or is there just some overlap?

There are lots of fields that use coding that aren't statistics. However, to be a statistician, you almost always have to be proficient in some type of code! Like I said, many of the calculations get pretty complicated, pretty fast.

Goals for the
Week
Questions

R Help 2

STAT 140:
Week Zero
Catch Up

Goals for the
Week

Questions

Question: Are there extra resources available for practicing using R?

The Tidyverse labs on the OpenIntro website are a great resource. I've not used them before so I can't comment on all of them, but the first few are very informative (and I'm going to continue working through them as we move through the semester). Another resource is DataCamp, which has some tutorials you can work through.

Question: Where can we go if we are having trouble with homework using R?

In general, the OpenIntro labs and DataCamp, but for specific questions on the homework, you can go to evening help or drop questions in the Zoom Channel! There is a way to include code snippets—let's review this after we work more with RMarkdown.

R Help 3

STAT 140:
Week Zero
Catch Up

Goals for the
Week

Questions

Question: I am confused about the function `level()` and `unique()`.

`levels()` returns all possible values for a categorical variable—also known as a “factor” in R. `unique()` returns only the values that appear in the dataset. Depending on how things are coded, it is possible for a variable in R to take on only some values rather than all possible values (for example, Wednesday, Thursday, and Friday out of all the days of the week).