



scan for
slides

Seize the Data: Around the World of Statistics in Five Days

Ashley Mullan
Vanderbilt University
WSDS 2026

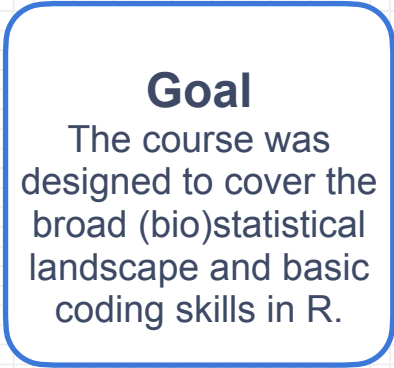




The course was offered for incoming 9th/10th and 11th/12th grade students.



The course ran for one week over the summer with five six-hour days.



The course was designed to cover the broad (bio)statistical landscape and basic coding skills in R.

- Math and data background
- Willingness to engage with course
- Application interests

- Math and data background
- Willingness to engage with course
- Application interests

- Math and data background
- Willingness to engage with course
- Application interests

- Long days
- Accelerated schedule
- Dedicated project time
- Little to no homework

- Math and data background
- Willingness to engage with course
- Application interests

- Long days
- Accelerated schedule
- Dedicated project time
- Little to no homework

- Balance between broad scope and enough detail
- Balance between foundations and application



Daily Topics

Each day covered statistical concepts or skills (bio)statisticians and data scientists use frequently.



Daily Topics

Each day covered statistical concepts or skills (bio)statisticians and data scientists use frequently.

1. Types of variables, well-defined research questions, common study designs, and limitations of measurement



Each day covered statistical concepts or skills (bio)statisticians and data scientists use frequently.

1. Types of variables, well-defined research questions, common study designs, and limitations of measurement
2. Data summarization, wrangling, and visualization in R

Daily Topics

Each day covered statistical concepts or skills (bio)statisticians and data scientists use frequently.

1. Types of variables, well-defined research questions, common study designs, and limitations of measurement
2. Data summarization, wrangling, and visualization in R
3. Regression mechanics, interpretation, and diagnostics



Each day covered statistical concepts or skills (bio)statisticians and data scientists use frequently.

1. Types of variables, well-defined research questions, common study designs, and limitations of measurement
2. Data summarization, wrangling, and visualization in R
3. Regression mechanics, interpretation, and diagnostics
4. Navigating the scientific literature, presentation design, and giving critical feedback

Daily Topics

Each day covered statistical concepts or skills (bio)statisticians and data scientists use frequently.

1. Types of variables, well-defined research questions, common study designs, and limitations of measurement
2. Data summarization, wrangling, and visualization in R
3. Regression mechanics, interpretation, and diagnostics
4. Navigating the scientific literature, presentation design, and giving critical feedback
5. Public speaking and statistics in the media



Activities

- We usually began with a **journaling-style activity** where students **responded to questions** that would come up in lecture.



Activities

- We usually began with a **journaling-style activity** where students **responded to questions** that would come up in lecture.
- Longer lectures were broken up with **small-group activities**, where students would **apply** the newly introduced concepts.



Activities

- We usually began with a **journaling-style activity** where students **responded to questions** that would come up in lecture.
- Longer lectures were broken up with **small-group activities**, where students would **apply** the newly introduced concepts.
- Each topic would conclude with another **journaling-style activity**, where students would jot down a few **takeaways**.



-

-

Activities

- We usually began with a **journaling-style activity** where students **responded to questions** that would come up in lecture.
- Longer lectures were broken up with **small-group activities**, where students would **apply** the newly introduced concepts.
- Each topic would conclude with another **journaling-style activity**, where students would jot down a few **takeaways**.
- All coding topics were introduced with **sample code** and **practice questions** that could be answered by **tweaking** the sample code.
- Groups for each topic's activity were **randomized**.
- Days 2-5 began with a **friendly competition** to recap past topics.



Examples: Journaling Intro

What is/are ...

- Data
- Similarity/
difference
- Assumptions
- “Strong”
conclusions
- Reliable sources



Why do we ...

- Slide 6 of 18

What can go wrong during ...

- Data collection
- Data analysis
- Communication of results
- Peer review

Examples: Middle Activity

- Translate **abstract ideas** into **concrete examples**



Examples: Middle Activity

- Translate **abstract ideas** into **concrete examples**
- Go through **previously introduced material** and **align** it with the **theme** of the lecture



Examples: Middle Activity

- Translate **abstract ideas** into **concrete examples**
- Go through **previously introduced material** and **align** it with the **theme** of the lecture
- Take **standard curriculum** and **re-examine** with a data lens



Examples: Middle Activity

- Translate **abstract ideas** into **concrete examples**
- Go through **previously introduced material** and **align** it with the **theme** of the lecture
- Take **standard curriculum** and **re-examine** with a data lens
- Motivate **code** as a way to **streamline computation**



Examples: Middle Activity

- Translate **abstract ideas** into **concrete examples**
- Go through **previously introduced material** and **align** it with the **theme** of the lecture
- Take **standard curriculum** and **re-examine** with a data lens
- Motivate **code** as a way to **streamline computation**
- Discuss **pros and cons** of competing **approaches**



Examples: Middle Activity

- Translate **abstract ideas** into **concrete examples**
- Go through **previously introduced material** and **align** it with the **theme** of the lecture
- Take **standard curriculum** and **re-examine** with a data lens
- Motivate **code** as a way to **streamline computation**
- Discuss **pros and cons** of competing **approaches**
- Consider **consequences** of analytic **choices**



Examples: Code

```
`{r}  
df1 |> pull(age) |> table() #count students in each age  
  
df1$age |> table() #another option for the same table  
  
table(df1$age) #a third option for the same table  
  
table(df1$age, df1$gender) #count students in each age by gender  
`
```

Code examples provided **multiple syntax options** to empower students to develop a **style** that made sense to them and explore **packages** that streamline common tasks.

Examples: Code

3. What percent of students in `df1` are seniors?

```
```{r}
#YOUR CODE HERE
```
```

4. Use the `is.na()` function and the `table()` function to count how many heights are missing in `df2`.

```
```{r}
#YOUR CODE HERE
```
```

5. Write some code to figure out a fact about `df1` that we haven't asked for yet!

```
```{r}
#YOUR CODE HERE
```
```

Practice questions varied in **difficulty** and **flexibility** to give students different **opportunities to succeed**.



Examples: Code

```
```{r}
penguins |> #start with the penguins data
 ggplot(aes(x = species)) + #aesthetics control the variables you plot
 geom_bar() #the "bar" geom makes a bar chart, which counts discrete variables
```
```

```
```{r}
penguins |> #start with the penguins data
 ggplot(aes(x = species)) + #aesthetics control the variables you plot
 geom_bar(fill = "blue",
 color = "blue3") + #make a bar chart and change the colors
 labs(x = "Penguin Species", #labels make your graph easier to read
 y = "Number of Penguins",
 title = "Titles can help share a main point, like our many Adelie penguins!",
 subtitle = "Cool pun here",
 caption = "Data come from the palmerpenguins package.") +
 theme_bw() + #themes change the starting blank plot. they are theme_XX()
 theme(plot.title = element_text(hjust = 0.5), #optional, center title
 plot.subtitle = element_text(hjust = 0.5)) #optional, center subtitle
```
```

Examples: Code

4. Use the `lm()` function to fit linear regression models to each sample. Use the age, gamer, and sex variables. Interpret each coefficient. I have provided some starter code for you!

```
```{r}
#| label: linreg_model_fits
#| eval: false

model1 <- NULL #replace the NULL with your model
#example: model1 <- lm(y ~ age + height, data = df)

model1 |>
 tidy(conf.int = TRUE, conf.level = 0.95) |> #summarize the model
 mutate(across(!term, \(x) round(x, digits = 3))) #round to 3 digits
```
```

5. You may have noticed some weirdly named coefficients. This is R is automatically mapping our binary variables to be coded as zeroes and ones! For example, it is reading a man as a "1" and a woman as a "0" within our data. What if you want to switch things up? Use `mutate(is_woman = ifelse(sex == "W", 1, 0))` to add a recoded version of `sex`, and then refit the models you did in Question 4 with this new variable instead of `sex`. What do you notice?

Modeling questions provided **starter code** and invited students to tweak **covariates** and **encoding** options.

-

Project

- Students were introduced to **six datasets** on Day 1, and they **ranked them** by interest level to help the instructor make groups.
- In **groups** of three or four, students designed **research questions** about their data and answered them with **descriptive statistics**, **visualizations**, and **regression**.



Project

- Students were introduced to **six datasets** on Day 1, and they **ranked them** by interest level to help the instructor make groups.
- In **groups** of three or four, students designed **research questions** about their data and answered them with **descriptive statistics**, **visualizations**, and **regression**.
- Each team **combined** their questions into a **presentation** that introduced the data and their **findings**.



Project

- Students were introduced to **six datasets** on Day 1, and they **ranked them** by interest level to help the instructor make groups.
- In **groups** of three or four, students designed **research questions** about their data and answered them with **descriptive statistics**, **visualizations**, and **regression**.
- Each team **combined** their questions into a **presentation** that introduced the data and their **findings**.
- Presentations were given on the **final day of class**, and the **students in the audience** were invited to **share strengths** they observed in each presentation.



Overall Learning Goals

- I hoped students would leave this course with the instinct to **explore and question** real data they encounter.
- Both the **hands-on activities** and the **project** kept this goal in mind.

Disease risk in a state, like TN!

| Measurement | Good | Bad (or possibly bad) |
|--|---|--|
| Counts of confirmed cases in each county at the end of the month (numerical, discrete) | <ul style="list-style-type: none">- Helps policymakers decide how to send out resources- Consistent timing, can help you track trends over time- You have a good baseline for the lowest number | <ul style="list-style-type: none">- Counties with big populations will have more cases- We miss people that are sick but haven't been diagnosed |
| Percent of folks over 18 that report having the disease in 2018 in each county (numerical, continuous) | <ul style="list-style-type: none">- Helps you compare counties more fairly- Consistent timing- You have a good baseline for the lowest | <ul style="list-style-type: none">- We rely on folks telling the truth- We set an arbitrary time frame, it could be thrown off by seasonality or another external event |

| Bias | Example | Consequence |
|--------------|---------|-------------|
| Selection | | |
| Non-response | | |
| Voluntary | | |
| Convenience | | |

- This policy had **mixed results**.

Slide 14 of 18

Watch Out For:

- Open-ended activities can **run longer** than expected



Watch Out For:

- Open-ended activities can **run longer** than expected
- Students may come up with research questions that require **advanced data wrangling**



Watch Out For:

- Open-ended activities can **run longer** than expected
- Students may come up with research questions that require **advanced data wrangling**
- Tech installation challenges may be **different on each machine**, causing some students to speed through setup and some to fall behind the group



Watch Out For:

- Open-ended activities can **run longer** than expected
- Students may come up with research questions that require **advanced data wrangling**
- Tech installation challenges may be **different on each machine**, causing some students to speed through setup and some to fall behind the group
- Choose the **level of depth** for your treatment of model diagnostics wisely



Watch Out For:

- Open-ended activities can **run longer** than expected
- Students may come up with research questions that require **advanced data wrangling**
- Tech installation challenges may be **different on each machine**, causing some students to speed through setup and some to fall behind the group
- Choose the **level of depth** for your treatment of model diagnostics wisely
- Make sure you have **examples** ready to go for any activities where you ask students to **look for resources**



Students Really Enjoyed

- Working through code **examples**
- The group **project**
- Learning from **peers**
- **Recapping** each topic
- Support from the **teaching team** during activity time
- Having **access** to all course materials



Takeaways

- Giving **open-ended instructions** but having examples in your back pocket gives new statistics students the **flexibility** to explore the things that interest them most



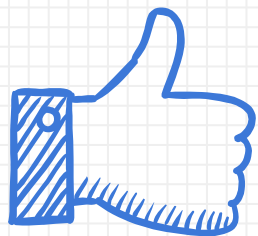
- Slide 17 of 18

Takeaways

- Giving **open-ended instructions** but having examples in your back pocket gives new statistics students the **flexibility** to explore the things that interest them most
- Providing space for students to **apply definitions** and abstract concepts to **their own interests** sparks thoughtful discussion
- Starting with the data landscape before jumping straight into coding builds **student confidence**



- Slide 17 of 18



THANKS!

Any questions?



ashley.e.mullan@vanderbilt.edu



[ashleymullan.github.io](https://github.com/ashleymullan)