

Mixing Active Learning and Lecturing: Using Interactive Visualization as a Teaching Tool

JSM 2018

Jessica Minnier, PhD & Ted Laderas, PhD

Oregon Health & Science University

 @datapointier

July 29, 2018

Slides available at <http://bit.ly/jsm-minnier>

Setting

OHSU Data Science Institute

- 2 Day workshop
- 3 Hours for "Introduction to Statistics and Data Exploration"
- Aim of DSI: "bring together researchers, librarians, and information specialists for formal training on key topics in data science"

Audience

- Librarians, information scientists, researchers
- Very little mathematical/programming background
- Heterogeneous background in science and research

Goals

Statistical Concepts

- Start with the **didactics**
- Use **interactive visualizations** to illustrate statistical concepts

Data Exploration

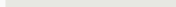

- Empower students to **explore data (no fear!)**
- Encourage understanding of relationships of data

Interactivity

- Interactive plots for exploration of **multi-variable relationships**
- Include some **coding exercises** (as bonus material)

Methods

Approach

- Implement as a  Tutorial, but used with didactic teaching
- : uses Shiny to build interactive R Markdown style workbooks
- Can be deployed as a website, or on student's computer (requires R/Rstudio)

Practicalities

- Categorical data session and continuous data session
- Hosted on github as a package on Github (<https://github.com/laderast/dsiexplore>)
- Hosted workbooks on shinyapps.io for real time interactivity

Interactivity

- Didactic lessons embedded in workbooks with interactive components
- Interactive sliders, dropdown options allow interaction with data filtering and analysis
- Interactive code teaches effect of changing code components on visualizations/analyses

Categorical Data

Ted Laderas and Jessica Minnier
06 November, 2017

Learning Objectives for this Session

What is Exploratory Data Analysis?

What are important associations?

The Data (Whickham)

Let's look at our Outcome

Two Variables Walk Into A Bar

The 2x2 table

The effect of Age on the Data

Not as easy as we thought!

Chi-Squared Test

Your Turn

Your Turn: Calculate your p-value

Missing Values

Be proud of yourself!

References

[Start Over](#)

Learning Objectives for this Session

Please fill out the pre-assessment before you do anything! <https://goo.gl/forms/jRa7TRFWXp5rIF6H3>

Make sure to buddy up, and use your post-its!

At the end of this session you should be able to

- Understand why we need to do Exploratory Data Analysis (EDA)
- Understand statistical association between categorical variables
- Visual ways to identify association
- Basic understanding of Chi-squared statistics
- Understand how to assess missing data

[Next Topic](#)

Categorical Data

Ted Laderas and Jessica Minnier
06 November, 2017

Learning Objectives for this Session

What is Exploratory Data Analysis?

What are important associations?

The Data (Whickham)

Let's look at our Outcome

Two Variables Walk Into A Bar

The 2x2 table

The effect of Age on the Data

Not as easy as we thought!

Chi-Squared Test

Your Turn

Your Turn: Calculate your p-value

Missing Values

Be proud of yourself!

References

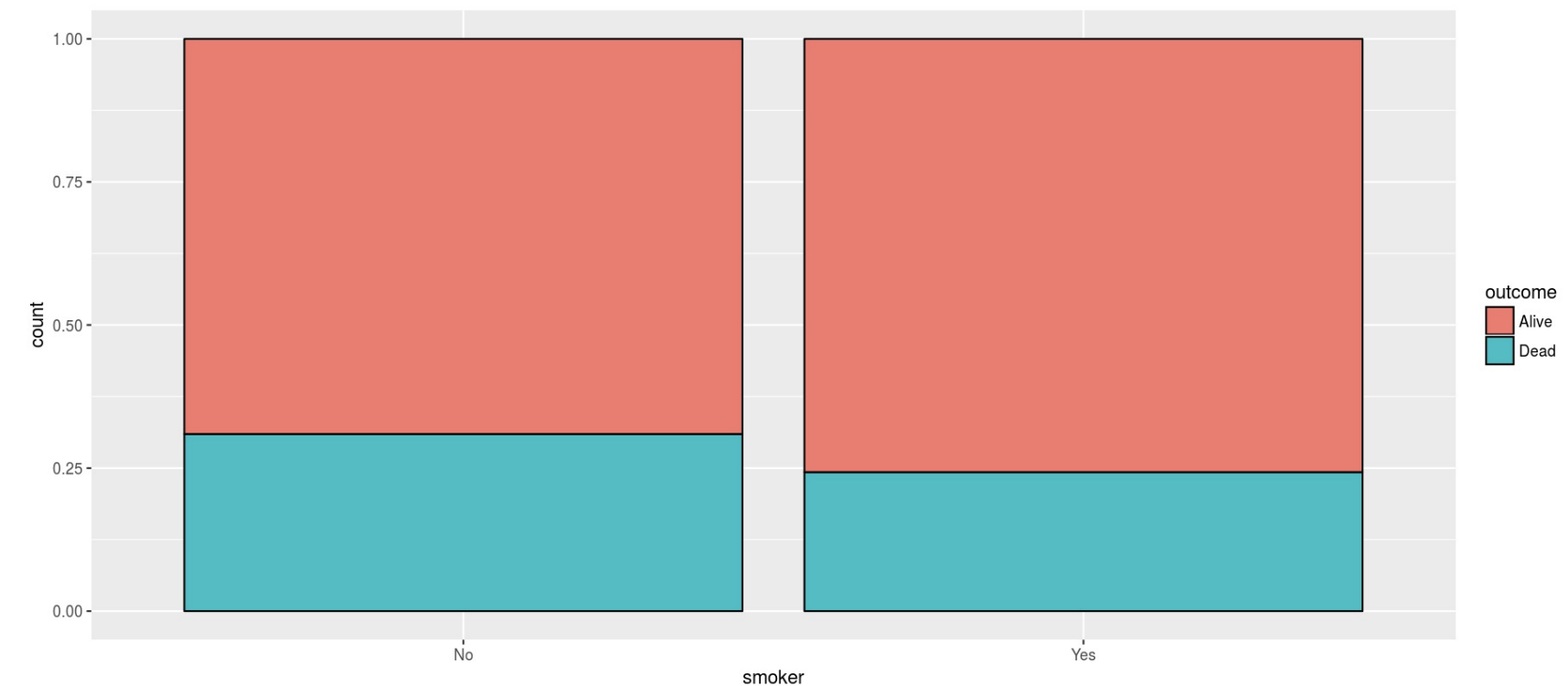
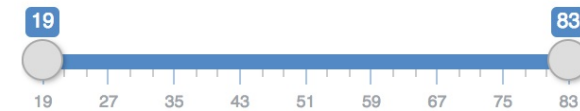
Start Over

Not as easy as we thought!

So as you get older, you're more likely to die. This may be messing up our overall results!

Let's ask the question again, with a younger group: are smokers under 60 more likely to die than non smokers?

Age Cutoff



For patients who are under 60, is smoking associated with death?

☐

Yes, the proportion of smokers who die is greater than the proportion of non-smokers who die for those patients younger than 60 years.

☐

No, the proportion of smokers who die is smaller than the proportion of non-smokers who die for those patients younger than 60 years.

Submit Answer

Categorical Data

Ted Laderas and Jessica Minnier
06 November, 2017

Learning Objectives for this Session

What is Exploratory Data Analysis?

What are important associations?

The Data (Whickham)

Let's look at our Outcome

Two Variables Walk Into A Bar

The 2x2 table

The effect of Age on the Data

Not as easy as we thought!

Chi-Squared Test

Your Turn

Your Turn: Calculate your p-value

Missing Values

Be proud of yourself!

References

[Start Over](#)

P-values

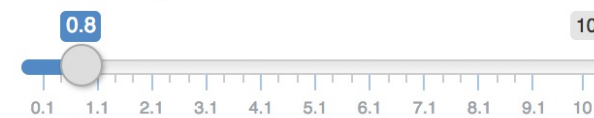
There is a straightforward interpretation to the *p-value*, and it has to do with how unique or rare our case is compared to our distribution of randomly generated cases.

So the *p-value* is interpreted as the probability that we will see a random case with the same exact statistic or higher.

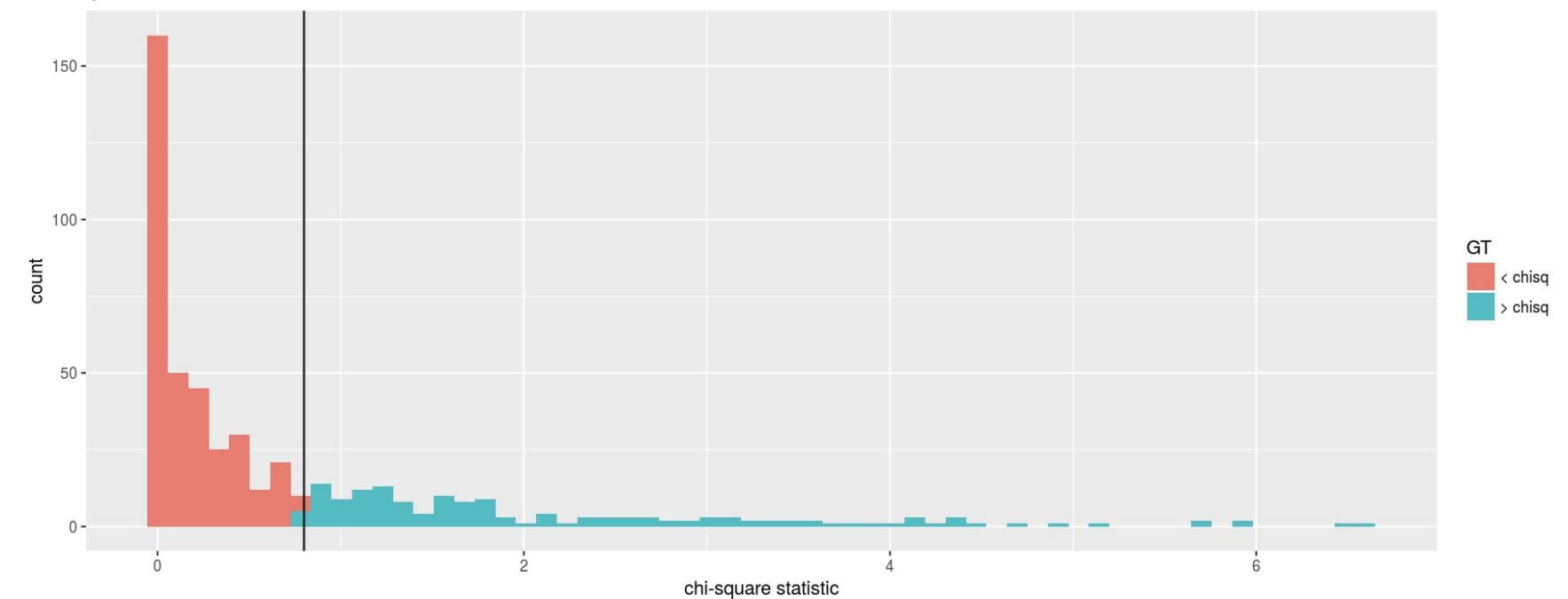
For example, if we had 10,000 random cases, and our p-value was 0.2, that means that of our 10000 cases, we would expect to see $10000 * 0.2 = 2000$ random cases with our statistic or greater.

Try adjusting the value of the chi-square statistic and see how many random cases are expected to have that statistic or higher.

slide to adjust statistical cutoff



Chi-square statistic: 0.8
% above chi-square: 30.4
p-value: 0.304



Continuous Data

Ted Laderas and Jessica Minnier
11/06/2017

Learning Objectives for this Session

EDA with continuous variables

How do we assess associations
between two continuous variables?

What is a factor that may be associated
with BMI?

Missingness and suspicious data
elements

T Test

Explore other factors

Advanced Topic: Linear Models -
multiple predictors/associations

Your Turn

Resources and extra practice

Post-session survey

Correlation

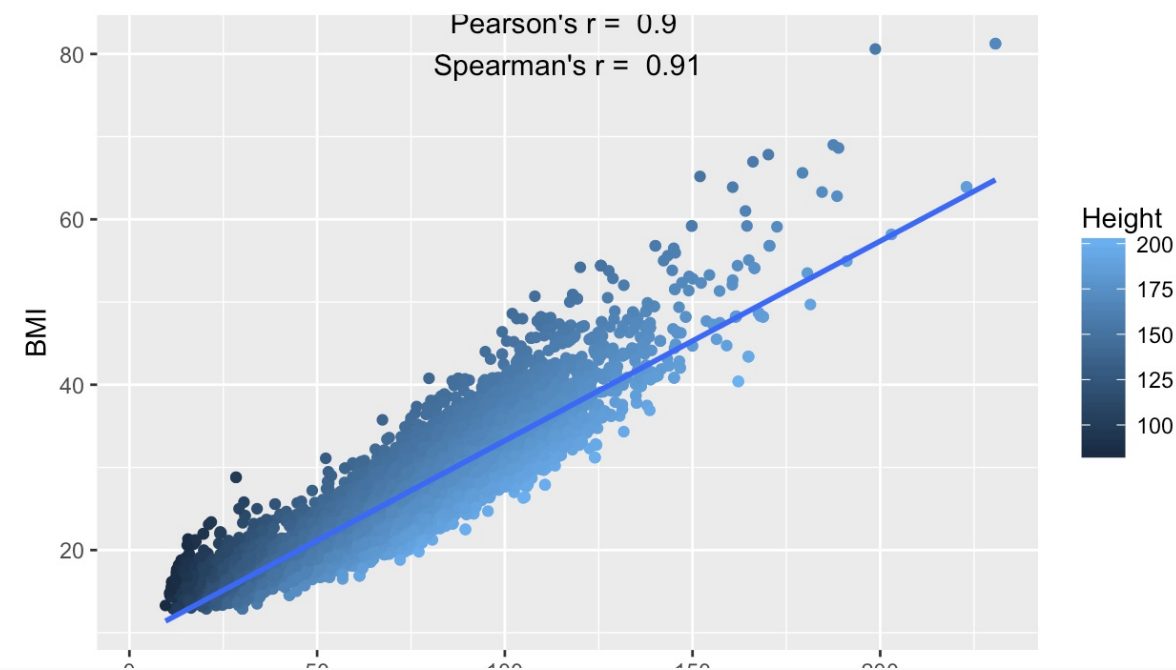
A simple statistical quantification of the association of two continuous variables is the **Pearson's Correlation Coefficient** (often labeled r).

Correlation = a quantity measuring the extent of interdependence of variable quantities

Pearson's correlation coefficient: a measure of the linear correlation between two variables

- Note that this is quantifying a *linear* relationship.
- Value between -1 and +1, with 0 denoting no linear correlation
- We can visually represent the linear relationship with a line through the scatter plot.
- If the relationship is relatively curved or exponential Pearson's correlation will not capture this relationship.
- An alternative might be the **Spearman's correlation** which essentially is the Pearson's correlation of the *ranks*. This evaluates *monotone* relationships.

Question: How well does the line "fit" the data?



Continuous Data

Ted Laderas and Jessica Minnier
11/06/2017

Learning Objectives for this Session

EDA with continuous variables

How do we assess associations
between two continuous variables?

What is a factor that may be associated
with BMI?

Missingness and suspicious data
elements

T Test

Explore other factors

Advanced Topic: Linear Models -
multiple predictors/associations

Your Turn

Resources and extra practice

Post-session survey

Start Over

✓ Correlation explorer

Now you can try to get a feel for what correlation (linear and non-linear) looks like. Try a few pairs:

(For fun sometime, play the “guess the correlation” game at guessthecorrelation.com)

X-axis

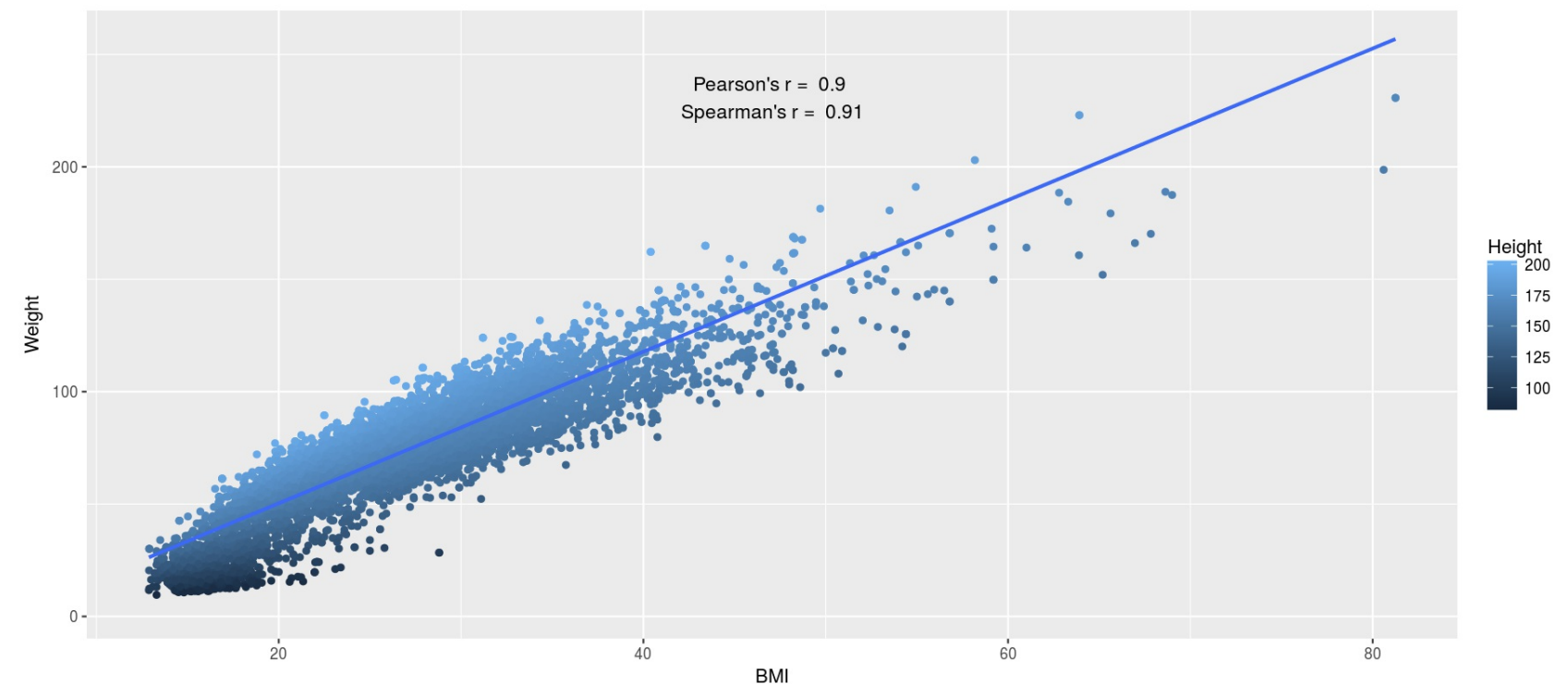
BMI

Y-axis

Weight

Color

Height



Continuous Data

Ted Laderas and Jessica Minnier
11/06/2017

Learning Objectives for this Session

EDA with continuous variables

How do we assess associations
between two continuous variables?

What is a factor that may be associated
with BMI?

Missingness and suspicious data
elements

T Test

Explore other factors

Advanced Topic: Linear Models -
multiple predictors/associations

Your Turn

Resources and extra practice

Post-session survey

[Start Over](#)

✓ Correlation explorer

Now you can try to get a feel for what correlation (linear and non-linear) looks like. Try a few pairs:

(For fun sometime, play the “guess the correlation” game at guessthecorrelation.com)

X-axis

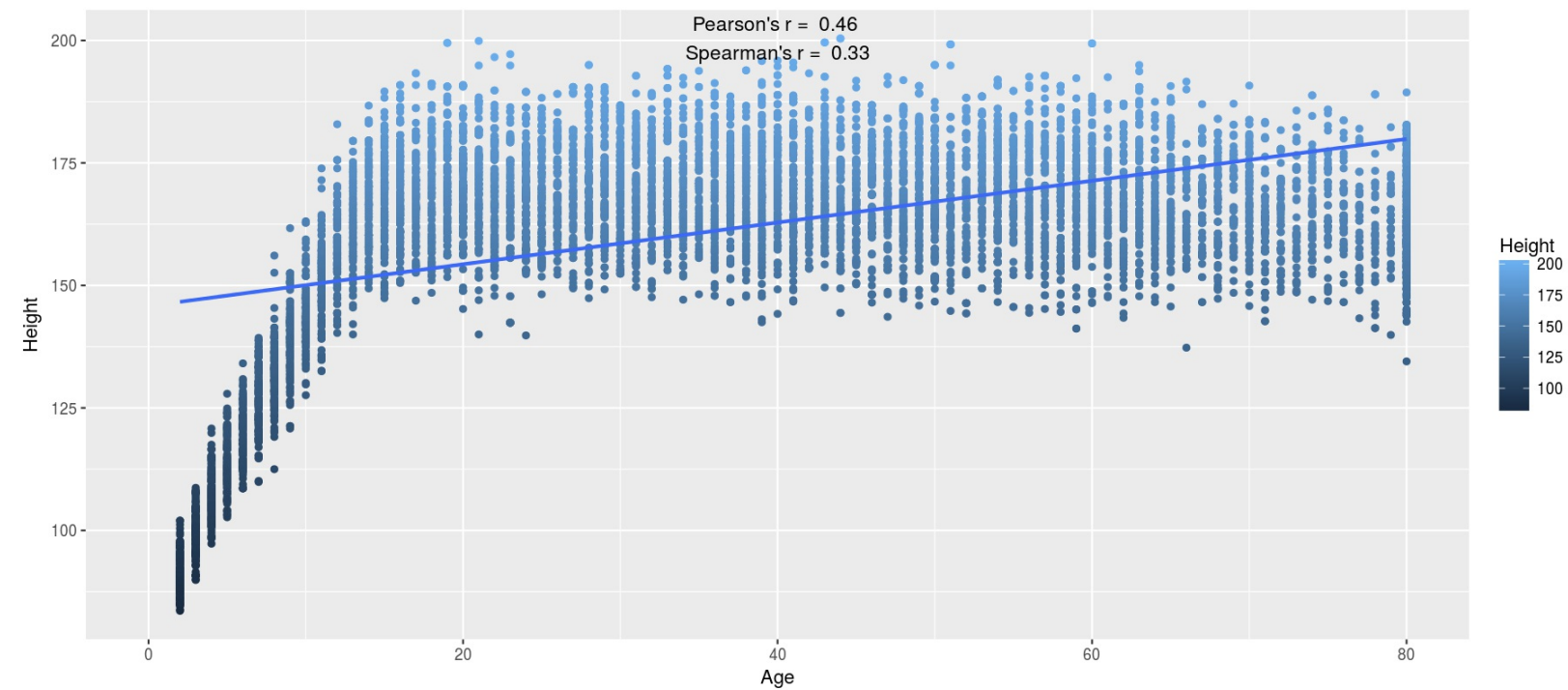
Age

Y-axis

Height

Color

Height



Continuous Data

Ted Laderas and Jessica Minnier
11/06/2017

Learning Objectives for this Session

EDA with continuous variables

How do we assess associations
between two continuous variables?

What is a factor that may be associated
with BMI?

Missingness and suspicious data
elements

T Test

Explore other factors

Advanced Topic: Linear Models -
multiple predictors/associations

Your Turn

Resources and extra practice

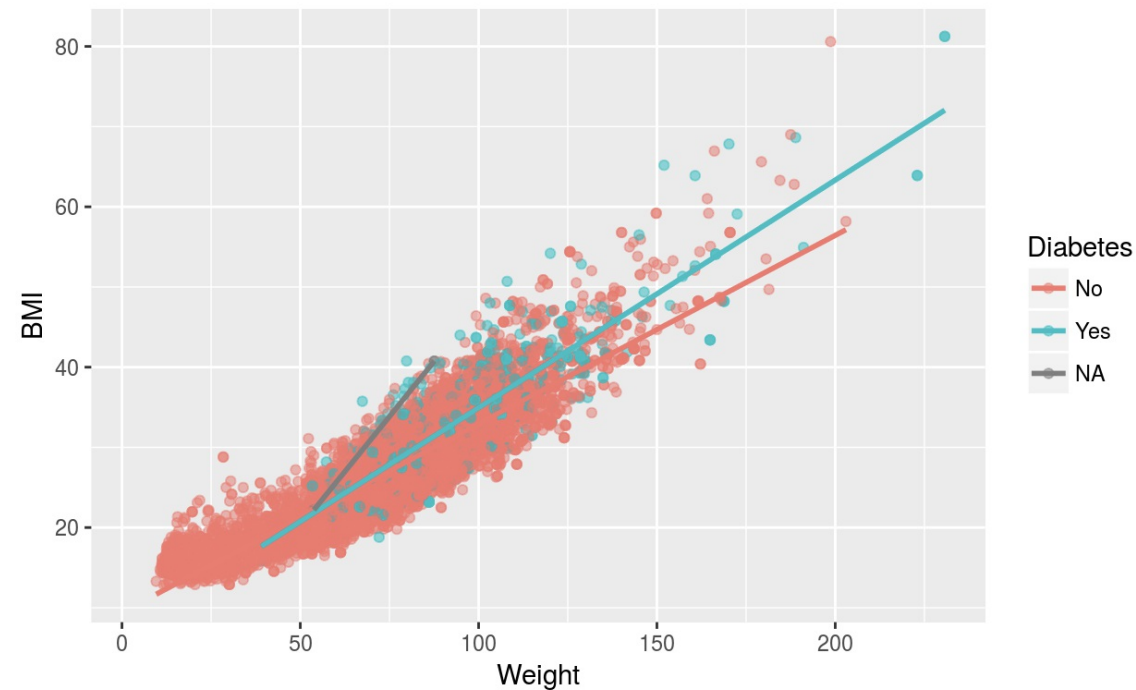
Post-session survey

[Start Over](#)

✓ Practice Coding

If you want to practice coding a scatter plot, try editing `ggplot2` code below to show a scatter plot of `Age` vs `Height`, colored by `Gender`:

```
Code Start Over Solution Run Code  
1 # edit the ggplot code after x= and y= and color= to change the axes and the  
2 # color  
3 NHANES %>% ggplot(aes(x = Weight, y = BMI, color = Diabetes)) + geom_point(alpha = 0.5) +  
4   stat_smooth(method = "lm", se = FALSE)
```



[Previous Topic](#)

[Next Topic](#)

Continuous Data

Ted Laderas and Jessica Minnier
11/06/2017

Learning Objectives for this Session

EDA with continuous variables

How do we assess associations
between two continuous variables?

What is a factor that may be associated
with BMI?

Missingness and suspicious data
elements

T Test

Explore other factors

Advanced Topic: Linear Models -
multiple predictors/associations

Your Turn

Resources and extra practice

Post-session survey

Start Over

However! The T-test is pretty robust to slight violations of the normality assumption, especially since we have a large sample size

- statistics side note: thanks to the Central Limit Theorem, our test is still *valid* as in we preserve our type I error; for a nice explanation of this see [this Stats Geek blog post](#) and [Lumley T, et al 2002](#))

So, let's run a t-test (yay!) to assess the difference in means of BMI comparing diabetics and non-diabetics:

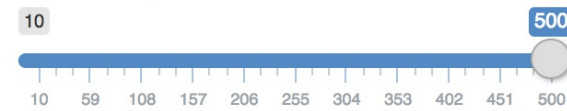
Difference in Means	Means No	Means Yes	T Statistic	P Value
-6.4	26.16	32.56	-20.83	3.9e-78

Note the p-value is extremely small. This is because we have a very large sample size and the difference in means is pretty large.

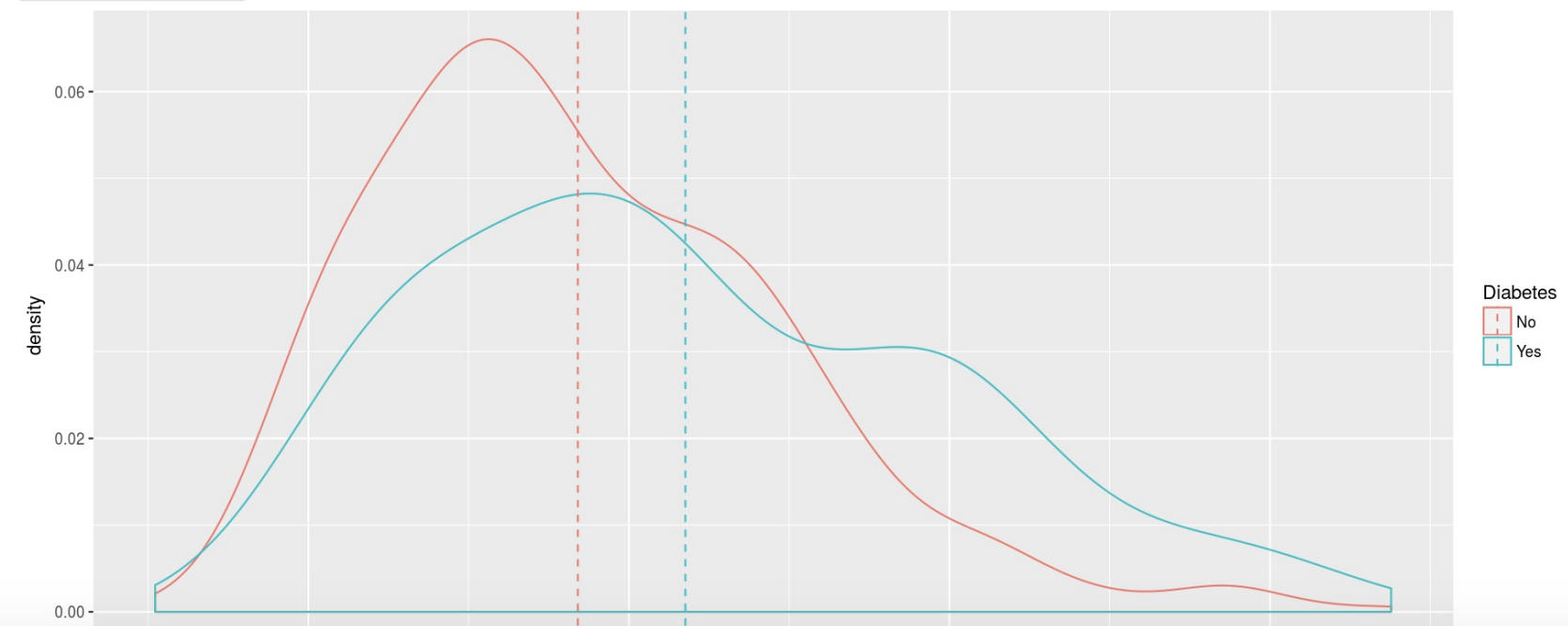
Smaller sample size

What happens if we have a much smaller sample size? We can examine the effect of sample size by randomly sampling a subset of the data. Look at our test statistic and p-value, as well as the difference in means.

Total Sample Size



Take a Sample

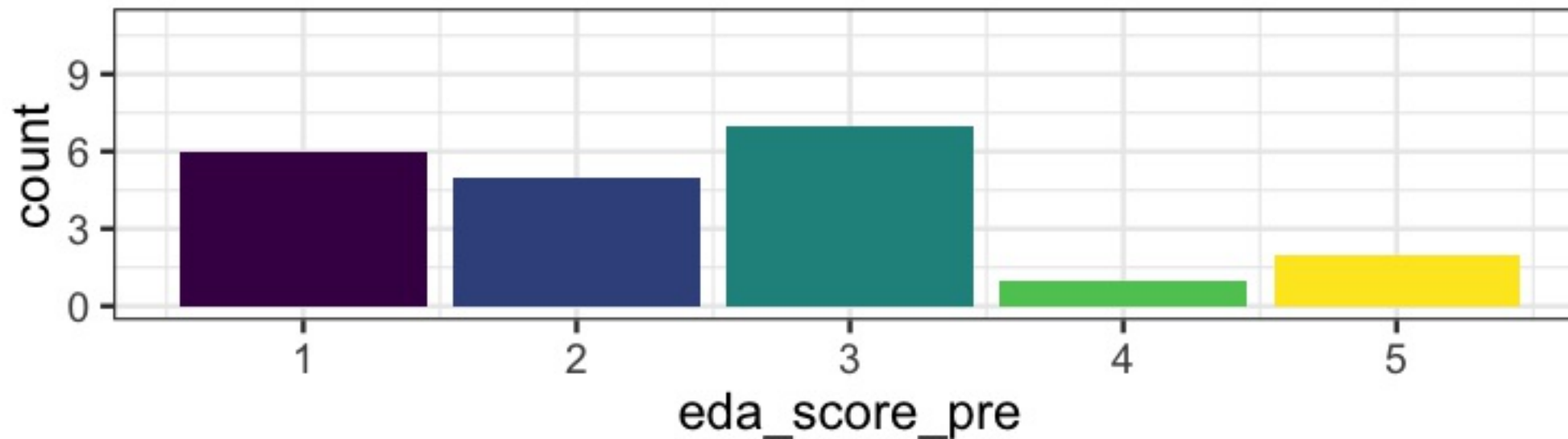


Conclusions & Results

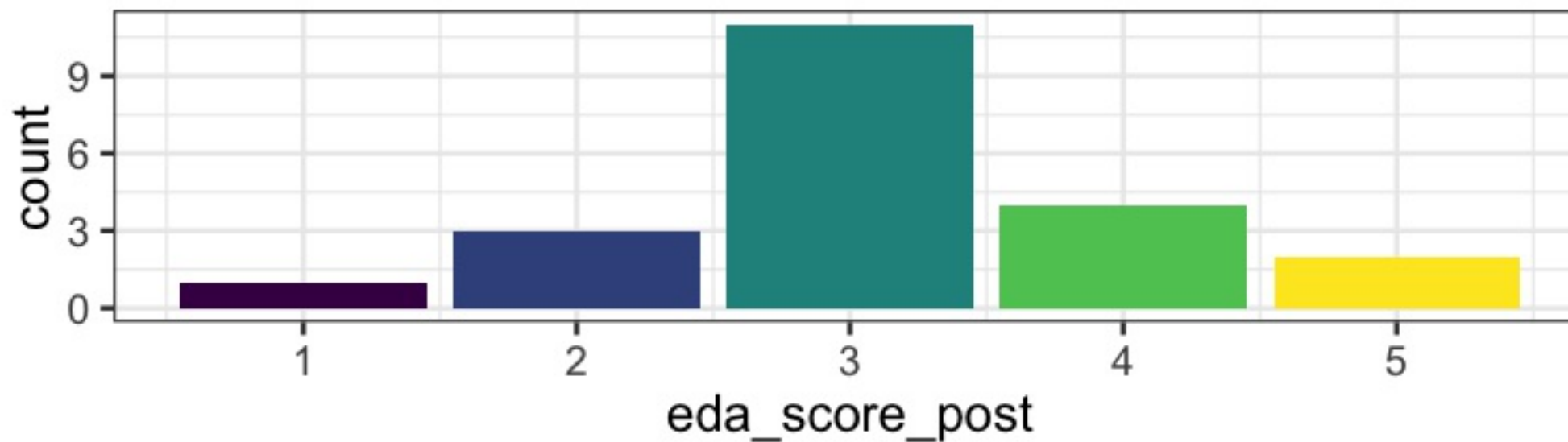
- LearnR package + Shiny in R → interactive workbooks
- Students were empowered to learn
- Students liked the visualizations
 - “Very well done and methodical treatment - the sliders were great!”
- Students felt engaged with the subject
 - “Explanation of key statistical concepts was effective and really made me want to learn more.”
- Pre/Post-workshop survey: 95% of learners felt they gained practical knowledge (n=22)

Survey Results

Please rate your level of ability for EDA prior to this session



Please rate your level of ability for EDA after this session



Impact

Pros:

- Accessible to beginners
- Mathematical concepts are more memorable
- Sparks discussions
- Empowers and engages students in scientific discovery/analysis

Cons:

- Advanced students may require more challenging activities
- Visualizations must be tested for effectiveness
- Requires programming skills to implement

Future Work and Adaptations

- Expand materials with more advanced statistical concepts
- Longer workshops \rightarrow more interactive material, more topics
- Determine which interactive explorations are most effective

Introduction to Visualization/Data Literacy




- Extension of this work: <https://tladeras.shinyapps.io/dataLiteracy/>
- [HMSP410, Health Informatics](#) for OHSU-PSU School of Public Health (co-taught by Ted Laderas and Bill Hersh)

Further Information

- eCOTS e-poster: <https://www.causeweb.org/cause/ecots/ecots18/posters/3-03>
- Categorical Data: <https://tladeras.shinyapps.io/categoricalData/>
- Continuous Data: https://minnier.shinyapps.io/ODSI_continuousData/
- LearnR package: <https://rstudio.github.io/learnr/>
- DSIexplore LearnR package: <https://github.com/laderast/DSIExplore>

Thank you!

Ted Laderas, PhD  [laderas](#),  [laderast](#)  <https://laderast.github.io/>

Contact me:  [minnier-\[at\]-ohsu.edu](mailto:minnier-[at]-ohsu.edu),  [datapointier](#),  [jminnier](#)

Slides available at <http://bit.ly/jsm-minnier>

Code for slides available at https://github.com/jminnier/talks_etc

Slides created via the R package [xaringan](#) by [Yihui Xie](#) with the metropolis theme