Math Virtual Learning

AP stats / Analyzing Bivariate Data

May 1st , 2020

Lesson: May 1st, 2020

Objective/Learning Target:
Students will review the tools to describe and analyze bivariate data.

Review #1

If quartiles $Q_1 = 50$ and $Q_3 = 70$, which of the following must be true?

1. The median is 60
2. The mean is between 50 and 70.
3. The standard deviation is at most 20.

Review #2

A study of weekly hours of television watched and SAT scores reports a correlation of r= -1.18. From this information, we can conclude that…

a. Students who watch more TV tend to have lower SAT scores
b. The fewer the hours in front of a TV, the higher a student's SAT scores
c. There is little relationship between weekly hours of television watched and SAT scores
d. There is strong negative association between weekly hours of television watched and SAT scores, but it would be wrong to conclude causation
e. A mistake in arithmetic has been made.

Answers

#1. This is a trick question. All three are False. The median will be somewhere between 50 and 70 but can be 50 or 70 as well. There is no guarantee of where though. A single outlier could cause the mean to be outside the two values and cause a much larger standard deviation

#2. r is only defined over the range -1 to 1. With r = -1.18, there is clear evidence of a mistake.

Bivariate data

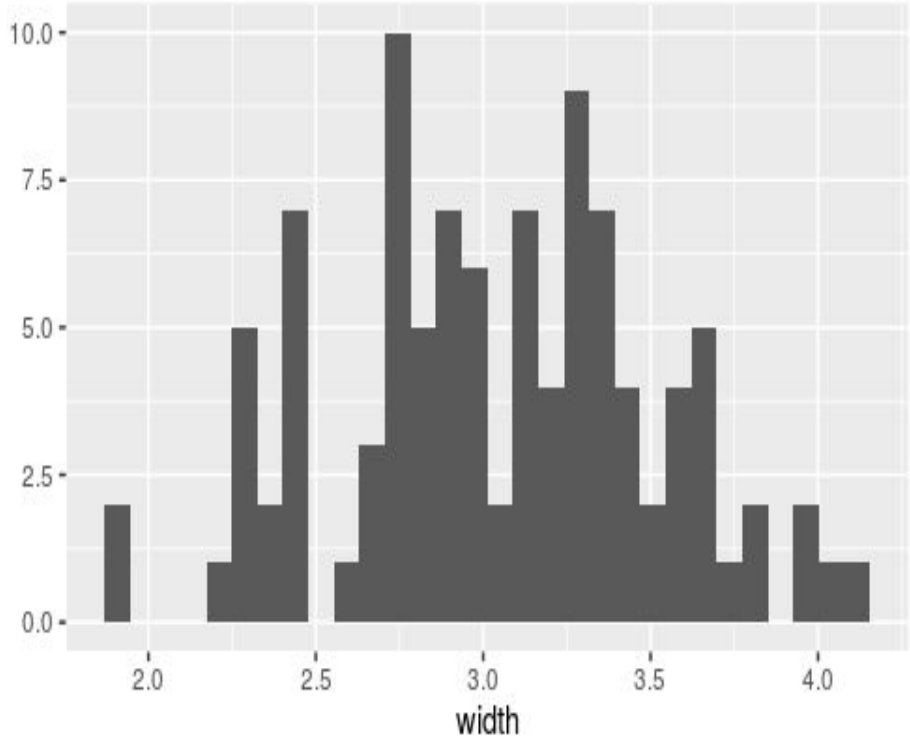
Sometimes, it does not make sense to view a single variable at a time. Especially when those variables have some sort of relationship. We can all think of variable that change together (e.g. shoe size vs height, hours studying vs test score, size of car vs fuel efficiency).

In this review we want to cover how those variables interact. View the following video to refresh your knowledge of bivariate data.

Bivariate Data

Bivariate data

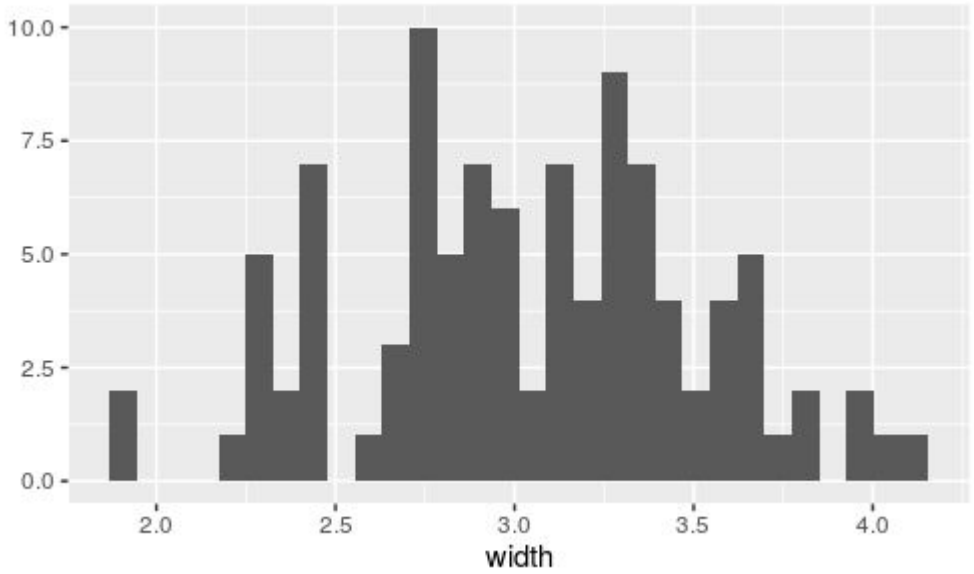
As a brief example of when bivariate data might be useful, we will look at trees. Since this data is randomly generated, let's call them truffula trees. A common practice in forestry is to measure the diameter of trees at breast height, meaning you walk up to the tree and measure at about chest height for most people. To the right is a histogram of these diameters.



Diameter of Truffula trees

We see that we have an approximately normal distribution, centered near 3 meters. But the values range all the way down below 2 and up past 4 meters. So the typical diameter varies quiet a bit.

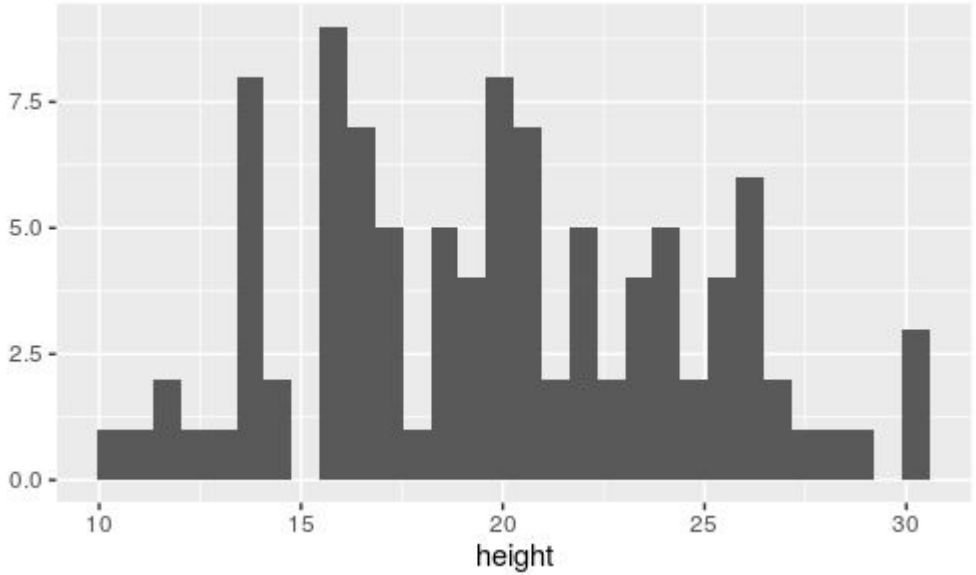
Now let's look height.



Height of Truffula trees

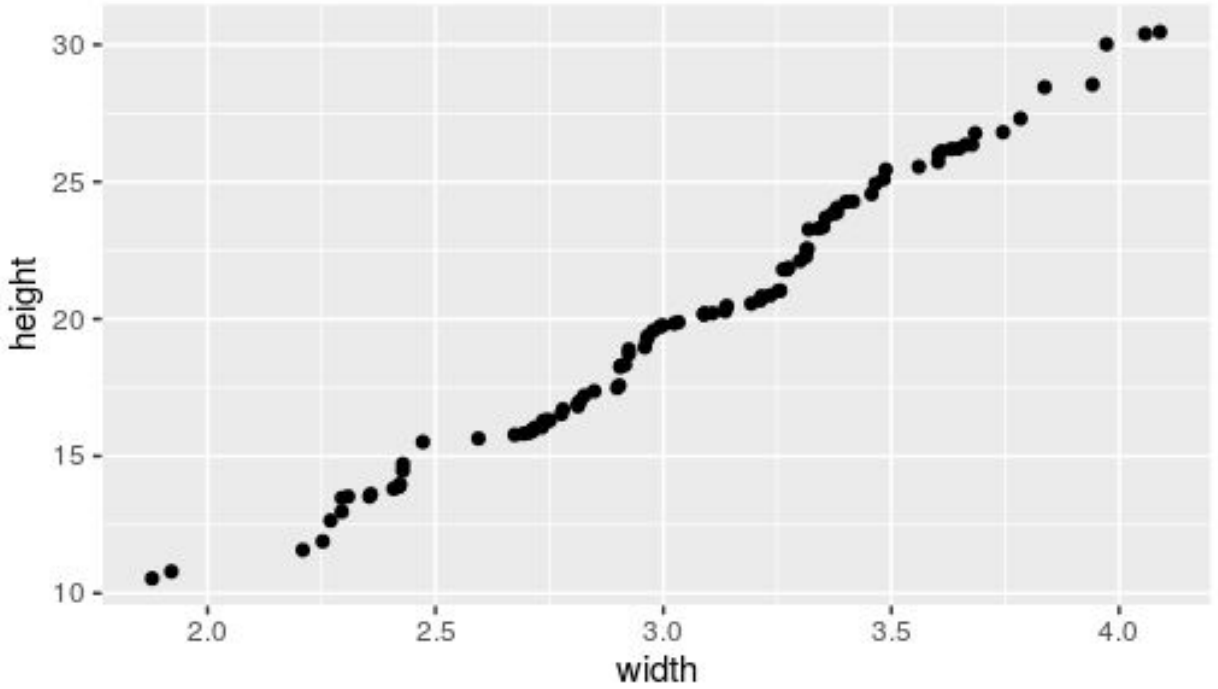
The heights are also a little normal, maybe a little less so than the diameters. They are centered at about 20 meters, but vary from 10 to more than 30 meters. There is a big difference between a 10 meter and 30 meter tree…

Is there a better story here?



Comparing diameter and height

We see that when the diameter and height of a tree are plotted together we see a clear relationship. Bigger diameter trunks belong to taller trees. We could now use an LSRL to create a model of this relationship



Linear model

To the right is a typical print out from creating a LSRL model for bivariate data. Can you write and interpret the model?

```
Call:
lm(formula = height ~ width, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.02443 -0.58151  0.03677  0.44866  2.13395

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.2009     0.4193  -24.33   <2e-16 ***
width         9.9051     0.1362   72.73   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6482 on 98 degrees of freedom
Multiple R-squared:  0.9818,    Adjusted R-squared:  0.9816
F-statistic:  5290 on 1 and 98 DF,  p-value: < 2.2e-16
```

Linear model

$$\widehat{height} = -10.2 + 9.9\ Diameter$$

This model indicates that there is an estimated 9.9 meter increase in height per 1 meter increase in Diameter. There is also an estimated height of -10.2 meters for a zero diameter tree. The intercept although interpreted here, doesn't have a true interpretation here. Diameter zero is outside the range of the data.

What values of Diameter should we estimate the height with our model? Remember to not extrapolate.

Extra practice

Free Response Problem

Answer Document