Contents

[4.1 Video: 1](#_Toc57841488)

[Categorical Variable Summaries 12:46 1](#_Toc57841489)

[Numeric Variable Summaries 18:54 2](#_Toc57841490)

[4.2 Video: 3](#_Toc57841491)

[Barplot 6:47 3](#_Toc57841492)

[Distributions 21:35 3](#_Toc57841493)

[4.3 Video 4](#_Toc57841494)

[Basics 18:40 4](#_Toc57841495)

[More categorical 28:03 5](#_Toc57841496)

[Other ggplot 45:23 6](#_Toc57841497)

[Homework Notes 7](#_Toc57841498)

# 4.1 Video:

## Categorical Variable Summaries 12:04

Check all that apply. Which of the following would likely be categorical variables?

* Favorite website
* Zipcode
* Hours watching TV in a day
* Number of children

Feedback

* Correct!
* Correct! Even though zipcode has numbers associated with it, the numbers are really just lables.  You can't do math on zipcodes!
* This is likely numeric.
* This is likely numeric.

What does a contingency table display?

* means of each variable within a category
* histograms of observations falling into categories
* counts of observations falling into categories
* total sum for variables within a category

Feedback

* Contingency tables show counts (or percentages) of observations falling into each category.
* Contingency tables show counts (or percentages) of observations falling into each category.
* Correct!
* Contingency tables show counts (or percentages) of observations falling into each category.

What type of object does the table() function return?

* dataframe
* array
* list
* factor

Feedback

* table() returns an array of integer values.
* Correct!
* table() returns an array of integer values.
* table() returns an array of integer values.

Suppose you’ve correctly read the titanic dataset into an object called titanicData and created a contingency table using the following code:

table(titanicData$sex, titanicData$embarked, titanicData$survived)

The levels for each variable are ordered as follows:

* sex (female, male)
* embarked (C, Q, S)
* survived(0, 1)

How do you access the Sex by Survived two-way table for those that embarked at the "S" port?

* tab[2, , ]
* tab[ , 3, ]
* tab[2, 3, ]
* tab[ , 1, ]
* tab[ , , 2]

Feedback

* This would create a two-way contingency table between embarked and survived for males.
* Correct!
* This would create a one-way contingency table for survived for those that were male and embarked on the S port.
* This would create a two-way contingency table between sex and survived for those embarking at the C port.
* This would create a two-way contingency table between sex and embarked for those that survived.

## Numeric Variable Summaries 18:54

Check all that apply. What aspects of a numeric variable’s distribution do we generally try to describe?

* Shape
* Center
* Spread
* Frequency
* Proportion
* Count

Feedback

None

What R function can be used to find the standard deviation of a numeric vector? Just write the function with or without ().

sd, sd()

Feedback

None

What R function did we look at that gives many common sample quantiles and the mean? Just write the function with or without ().

summary(), summary

Feedback

None

# 4.2 Video:

## Barplot 6:47

How can you use base R functions to create a stacked barplot? (That is, what function might you use and what kind of R object should you pass it?)

Feedback

The easiest way to create a barplot using base R functions is to first create a contingency table using the table() function and then pass that as the first argument to barplot(). If the contingency table passed is a two-way table, it will by default be a stacked barplot (which can be changed using the beside = TRUE argument).

## Distributions 21:35

How are a histogram and a kernel smoothed (or density) plot related?

Feedback

A histogram bins a numeric variable and then counts the observations in each bin. A major issue with histograms is that they are highly variable in regard to the selection of the bins. A kernel smoothed plot is similar to a histogram but uses a weighting function to create a smooth curve representation of the distribution.

What is the argument to set the title within most base R plots calls?

* main
* title
* name
* list

Feedback

* Right!
* title is a function that can be used to add a title after the plot is created (but no title should be created within the base R plot call).
* main = “Title of the plot” is used in most base R plot calls.
* main = “Title of the plot” is used in most base R plot calls.

# 4.3 Video

## Basics 18:40

What does the ggplot function itself create?

* A base plotting object with global aesthetics
* A bar graph with the first variable as the x-axis
* The best plot for the data as guessed by the function
* A new dataframe with only the global aesthetics and associated variables

Feedback

* Right!
* This would be the case if a geom\_bar() layer was added and appropriate aesthetics were used.
* The type of plot is specified using a geom or stat layer.
* If you save the ggplot() function output, it does actually return a list with information but not a dataframe.

What symbol is used to add another layer to a ggplot object?

+

Feedback

Remember that ggplot doesn’t quite stay in the tidyverse in terms of syntax. The pipe operator cannot be used with ggplot. The package ggvis will eventually replace the ggplot2 package and use similar syntax as the rest of the tidyverse.

What geom layer below would be used in creating a bar plot?

* geom\_bar()
* geom\_pillar()
* geom\_bp()
* geom\_bar\_plot()

Feedback

* Right!
* The cheat sheet is usually the best place to find the most commonly used plot layers.
* The cheat sheet is usually the best place to find the most commonly used plot layers.
* The cheat sheet is usually the best place to find the most commonly used plot layers.

## More categorical 28:03

What is the argument that allows you to create side-by-side bar plots that represent a two-way table?

* position = “dodge”
* position = “fill”
* position = “jitter”
* position = “stack”

Feedback

* Correct!
* position = “dodge” implies the two parts of the bar ‘miss’ each other or should be side-by-side.
* position = “dodge” implies the two parts of the bar ‘miss’ each other or should be side-by-side.
* position = “dodge” implies the two parts of the bar ‘miss’ each other or should be side-by-side.

TRUE/FALSE: There is only one function that can be used to create a bar plot using ggplot.

* TRUE
* FALSE

Feedback

* There is generally a stat layer or a geom layer that could be added to create the same plot.
* Correct!

Suppose I had the following code (after reading in the titanic dataset into titanicData):

ggplot(data = titanicData, aes(x = embarked)) + geom\_bar(aes(color = as.factor(survived)), lwd = 2)

What layer could be used to change the labels that show up in the automatically created legend?

* scale\_color\_discrete(labels = …)
* labels = ...
* bar\_labels(...)
* geom\_bar\_layers(labels = …)
* scale\_bar\_discrete(labels = …)

Feedback

* Right!
* Since we want to change the automatically generated labels associated with the color aesthetic where that aesthetic is set to a categorical variable (factor), we want to use scale\_color\_discrete() and set the labels with a character vector.
* Since we want to change the automatically generated labels associated with the color aesthetic where that aesthetic is set to a categorical variable (factor), we want to use scale\_color\_discrete() and set the labels with a character vector.
* Since we want to change the automatically generated labels associated with the color aesthetic where that aesthetic is set to a categorical variable (factor), we want to use scale\_color\_discrete() and set the labels with a character vector.
* Since we want to change the automatically generated labels associated with the color aesthetic where that aesthetic is set to a categorical variable (factor), we want to use scale\_color\_discrete() and set the labels with a character vector.

Where is a good place to find the commonly modified aesthetics for a ggplot graph?

* Cheat sheet
* Help file
* All aesthetics apply to any graph

Feedback

* Right!
* As with much of the tidyverse, the help files tend to be less useful as they rely on vignettes and cheat sheets.
* The cheat sheet is the best place to go for commonly used aesthetics on each geom layer.

## Other ggplot 45:23

When does something you specify about a ggplot plot (generally) belong in an aes() function?

* when that attribute will depend on the data
* all attributes of the plot should go in the aes
* aes is only used on layers, so only when you are building a layer
* when you are creating a new variable to be used in the plot

Feedback

* Right!
* Most of the time, attributes that you want constant for the entire plot are set outside of the aes(). For instance, you can set fill = “red” in a geom\_bar() layer to fill all the bars with red.
* You can set aes() values in the base ggplot() call as well.
* Most of the time, attributes that you want constant for the entire plot are set outside of the aes(). For instance, you can set fill = “red” in a geom\_bar() layer to fill all the bars with red.

Which variable would be on the x-axis according to the code below?

ggplot(CO2, aes(x = uptake)) +

geom\_histogram(color = "blue", fill = "red", size = 2, binwidth = 3, aes(x = conc))

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

conc

Feedback

As local aesthetics override the global ones defined in the ggplot() function call, we would have conc on the x-axis.

What is the idea of faceting and how do facet\_grid() and facet\_wrap() differ?

Feedback

Faceting is the idea of creating the same graph at levels of one categorical variable or at the combinations of the levels of two categorical variables. Using facet\_grid(), you specify the levels that you want on the x- and/or y-axes explicitly. Using facet\_wrap(), we allow the ggplot2 package to determine how many plots to create in each row.

# Homework Notes

Why do the following two pieces of code yield different results? (Hint: you may want to run them!)

c(TRUE, 1, "hi")

c(c(TRUE, 1), "hi")

Feedback

These differ due to how the implicit coercion occurs. When you try to combine TRUE, 1, and “hi” all at once, R must convert everything to characters since vectors must have elements all of the same type and character is the most flexible. This means we get “TRUE”, “1”, and “hi” in the resulting vector.

When we combine TRUE and 1 first using c(TRUE, 1), TRUE is implicitly coerced to a number (1). The vector 1, 1, is then combined with “hi”, which again causes coercion to characters. This means we get “1”, “1”, and “hi” in the resulting vector.