STA130 Study Guide

[Sections]

[Different types of data in R]

Data type	Symbol	Description	Examples
Integer	int	Numbers (w/ decimals)	1, - 2, 3, - 4
Double	dbl	Numbers (w/ or w/o decimals)	2, 2.02, 22222
Logical	lgl	TRUE or FALSE	TRUE, FALSE
Character	chr	Words, surrounded by quotation marks	“I”, “love”, “stats”
Factor	fct	Looks like “character” type, but only take values from a prespecified list	If continents list →can’t take “blue”

[Logical operators]

Operator	Syntax
equal	==
not equal	! =
less than (less than or equal to)	< (<=)
greater than (greater than or equal to)	> (>=)
not	!
and	&
or	|

Vectors

- c() function combines single elements into a vector
- is. functions to check the data type of a vector (e.g. is.numeric(), is.character())

Coercion

- R switches between data types automatically for certain operations
- ex) sum(c(TRUE, FALSE)) becomes sum(c(1, 0))
  - Counts the number of values of TRUE in a vector

Data frames

rows	Individual observations / records
columns	Variables

read_csv()	● Load data in R ○ the resulting object type is called a “tibble”

glimpse()	● Tells how many rows and columns there are ● Listing out the column names ● Tells what their data type is ● Giving a peak at the first few values
head()	● Shows what the top couple rows of the data look like
Pipes %>%	● Make it easy to apply functions to our data, step-by-step

[Types of Variables]

Categorical Variables	**Nominal variable**	Unordered descriptions (ex: turtle, butterfly, snail)
	Ordinal variable	Ordered descriptions (ex: unhappy, ok, awesome)
	Binary variable	Only 2 mutually exclusive outcomes (ex: yes, no)
Quantitative (numerical) Variables	**Continuous variable**	Measured data, can have infinite values within possible range (ex: 3.2 inch, 34.16g)
	Discrete variable	Observations can only exist at limited values, often counts (ex: 7, 5, 9)

[Visualizing and describing the distribution of a **quantitative(numerical) variable**]

Histogram	
	● **Height** of each **bar** counts the # of values of the variable which fall within the corresponding **bin** ● **Horizontal axis**: numerical (no gaps between bins) ● **Vertical axis**: the number of values which fall within each bin
Features	● **Shape**: overall pattern of the values of the variable ● **Center**: describes a ‘typical’ value of the variable ● **Spread**: describes how concentrated the values of the variables are
Shape	**Skewness** ● Symmetric (not skewed) ● Left-skewed ● Right-skewed **Modality** ● Unimodal ● Bimodal ● Multimodal ● Uniform
Creating a histogram in R	● Replace **blue part** ggplot(data = **data_name**, aes(x = **numerical variable**)) + geom_histogram(color = “**black**”,

	fill = “**gray**”,
	bins = **#**) +
	labs(x = “**horizontal axis name**”)

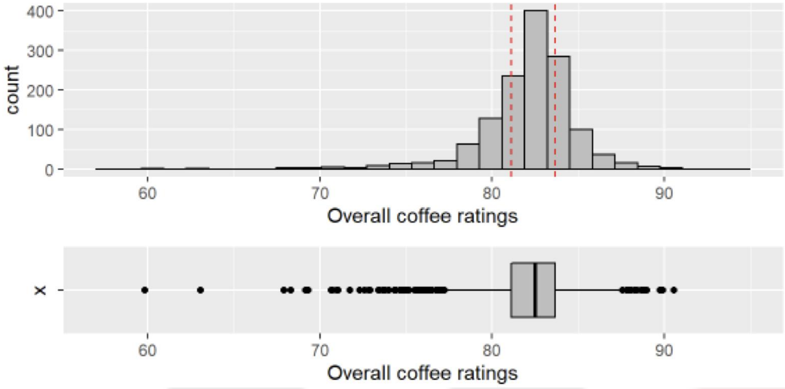
mean / median / modality	● Left skewed ○ mean < median < mode ● Right skewed ○ mode < median < mean ● Symmetric ○ mode ≈median≈mean

[Measuring the center & spread of a numerical distribution]

Center		**Spread**	
Mean	$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$	Variance	$s^2 = \frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ Average squared distance between the values and their mean
Median	Half of the values are smaller/larger than the median	Standard Deviation	$s = \sqrt{s^2}$

Boxplot	
	● Summarizes the distribution of a quantitative variable using five statistics ● Plot unusual observations (outliers)
Features of the plot	● **Median**: line in the middle of the box ● **First quartile (Q_1)**: one quarter (25%) of the data values are smaller than it ● **Third quartile (Q_3)**: three quarters (75%) of the data values are smaller than it ● **Inter-quartile range (IQR)**: Q_3 - Q_1, 50% of the values ● Whiskers extend within 1.5 x IQR ● **Outliers**: points beyond the whiskers
Creating a boxplot in R	● Replace **blue part** ggplot(data = **data_name**, aes(x = **“”,** y = **numerical variable**)) + geom_boxplot(color = “**black**”, fill = “**gray**”) + labs(y = “**vertical axis name**”) ● When **comparing distributions** across different categories ggplot(data = **data_name**, aes(x = **categorical variable,** y = **numerical variable**)) + geom_boxplot(color = “**black**”, fill = “**gray**”) + labs(y = “**vertical axis name**”, X = “**horizontal axis name**”)

[Histogram vs. boxplot for the same distribution]



[Visualizing and describing the distribution of a **categorical variable**]

Barplots	
	● One bar for each category ● Height of a bar →# of values of the variable which fall within the corresponding category ● Arbitrary widths (should all be the same) ● Gap between each bar ● Arbitrary order of the bars ● **Shape or center don't make sense**
Creating a barplot in R	ggplot(data = **data_name**, aes(x = **categorical variable**)) + geom_bar(color = "**black**", fill = "**gray**") + labs(x = "**horizontal axis name**") ● **Flipped version** ggplot(data = **data_name**, aes(x = **categorical variable**)) + geom_bar(color = "**black**", fill = "**gray**") + labs(x = "**horizontal axis name**") + coord_flip()

[Tidy Data]

Dataset is tidy because:

- ☐ Each value has its own cell
- ☐ Each observation has its own row
- ☐ Each variable has its own column

Tidy data examples	Not Tidy data examples