Statistical Plot

Pattabiraman V

Data Concepts

Vector:

Sequence of data elements of the same type

Each element of the vector are also called components, members, or values

Created in R using c()

Dataframe:

A list of vectors of identical lengths

Example: iris

Variable:

A trait or condition that can exist in different quantities or types

We measure the impacts of *independent* predictor variables on *dependent* response variables

Data Concepts

Continuous Data:

Numeric data which is not restricted to certain values - there are an infinite number of possible values

Discrete Data: Numeric data which is restricted to certain values - for example, number of kids (or trees, or animals) has to be a whole integer

Categorical Data: Data which can only exist as one of a specific set of values - for example, house color or zip code

Binned numeric data (e.g. “between 1 and 2 inches”) is typically categorical

Binary Data: Categorical data where the only values are 0 and 1

Often used in situations where a “hit” - an animal getting trapped, a customer clicking a link, etc - is a 1, and no hit is a 0

Ordinal Data: A type of categorical data where each value is assigned a level or rank

Useful with binned data, but also in graphing to rearrange the order categories are drawn

Referred to in R as “factors”

Data Concepts

Unstructured Data: Data without a strict format, typically composed of text

R used to deal with unstructured data by converting it to factors; while this isn't necessary anymore, some functions still require text data to be in factor form

Data Distribution: How often every possible value occurs in a dataset

Usually shown as a curved line on a graph, or a histogram

Normal Distribution: Data where mean = median, 2/3 of the data are within one standard deviation of the mean, 95% of the data are within two SD and 97% are within 3.

Many statistical analyses assume your data are normally distributed

Many datasets - especially in nature - aren't

Data Concepts

Unstructured Data: Data without a strict format, typically composed of text

R used to deal with unstructured data by converting it to factors; while this isn't necessary anymore, some functions still require text data to be in factor form

Data Distribution: How often every possible value occurs in a dataset

Usually shown as a curved line on a graph, or a histogram

Normal Distribution: Data where mean = median, 2/3 of the data are within one standard deviation of the mean, 95% of the data are within two SD and 97% are within 3.

Many statistical analyses assume your data are normally distributed

Many datasets - especially in nature - aren't

Skewed Distribution: Data where the median does not equal the mean

A left-skewed distribution has a long tail on the left side of the graph, while a right-skewed distribution has a long tail to the right

Named after the tail and not the peak of the graph, as values in that tail occur more often than would be expected with a normal distribution

Statistical Terms

Estimate: A statistic calculated from your data

Called an estimate as we are approximating population-level values from sample data

Synonynm: metric

Hypothesis Testing: Comparing the null hypothesis (typically, that two quantities are equivalent) to an alternative hypothesis

The alternative hypothesis in a two-tailed test is that the quantities are different, while the alternative hypothesis in a one-tailed test is that one quantity is larger or smaller than the other

Almost never used in business, as the important question is usually not *does x cause y* but *can x predict y*

p Value: The probability of seeing an effect of the same size as our results given a random model

High p values often mean your independent variables are irrelevant, but low p values don't mean they're important - that judgement requires a rational justification, and examining the effect size and importance. Otherwise you're just equating correlation and causation.

The 0.05 thing is from a single sentence, taken out of context, from a book published in 1925. There's no reason to set a line in the sand for "significance" - 0.05 means that there's a 1 in 20 probability your result could be random chance, and 0.056 means it's 1 in 18. Those are almost identical odds.

Some journals have banned their use altogether, but others still will only accept "significant" results

Statement from the American Statistical Association:

A p value, or statistical significance, does not measure the size of an effect or the importance of a result. By itself, a p value does not provide a good measure of evidence about a model or a hypothesis.

"Robust" A term meaning an estimate is less susceptible to outliers

Means are not robust, while medians are, for instance.

Statistical Terms

Regression:

A method to analyze the impacts of independent variables on a dependent variable

ANOVA and models are both types of regression analyses

General Linear Model:

Formulas representing the expected value of a response variable for given values of one or more predictors

The typical y = mx + b format of model

Sometimes abbreviated GLM; R uses lm() to construct these

Generalized Linear Model:

Depending who you ask, these may or may not be linear models - they tweak the normal formula in one way or another to measure outcomes that general linear models can't address

In this course, we'll only be using **logistic models**

Sometimes abbreviated GLM; R uses glm() to construct these

Estimates and Statistics

N

The number of observations of a dataset or level of a categorical.

In R, run nrow(dataframe) or length(Vector) to calculate.

To calculate by group, run count(Data, GroupingVariable)

Examples: nrow(iris), length(iris$Sepal.Length), count(iris, Species)

Mean: The average of a dataset, defined as the sum of all observations divided by the number of observations.

In R, run mean(Vector) to calculate.

Example: mean(iris$Sepal.Length)

Trimmed Mean: The mean of a dataset with a certain proportion of data not included

The highest and lowest values are trimmed - for instance, the 10% trimmed mean will use the middle 80% of your data

mean(Vector, trim = 0.##)

mean(iris$Sepal.Length, trim = 0.10)

Variance: A measure of the spread of your data.

var(Vector)

var(iris$Sepal.Length)

Standard Deviation: The amount any observation can be expected to differ from the mean.

sd(Vector)

sd(iris$Sepal.Length)

R- Statistic code

```
nrow(iris)
length(iris$Sepal.Length)
mean(iris$Sepal.Length)
mean(iris$Sepal.Length, trim = 0.10)
var(iris$Sepal.Length)
sd(iris$Sepal.Length)
sd(iris$Sepal.Length)/sqrt(length(iris$Sepal.Length))
mad(iris$Sepal.Length)
median(iris$Sepal.Length)
```

R- Statistic code

Details

```
min(iris$Sepal.Length)
max(iris$Sepal.Length)
max(iris$Sepal.Length) - min(iris$Sepal.Length)
quantile(iris$Sepal.Length, c(0.25, 0.5, 0.75))
IQR(iris$Sepal.Length)
cor(iris$Sepal.Length, iris$Sepal.Width, method = "pearson")
cor.test(iris$Sepal.Length, iris$Sepal.Width, method =
"pearson")
lm(Sepal.Length ~ Species, data = iris)
model <- lm(Sepal.Length ~ Species, data = iris)
anova(model)
model <- lm(Sepal.Length ~ Species, iris)
summary(model)
```

Example Plot

Consider the Orange dataset, which is automatically included in R. Note that the O is capitalized!

1. Look at Orange using either head or as.tibble() (you'll have to run library(tidyverse) for that second option). What type of data are each of the columns?
2. Find the mean, standard deviation, and standard error of tree circumference.
3. Make a linear model which describes circumference (the response) as a function of age (the predictor). Save it as an object with <-, then print the object out by typing its name. What do those coefficients mean?
4. Make another linear model describing age as a function of circumference. Save this as a different object.
5. Call summary() on both of your model objects. What do you notice?
6. Does this mean that trees growing makes them get older? Does a tree getting older make it grow larger? Or are these just correlations?
7. Does the significant p value prove that trees growing makes them get older? Why not?

Thank you