



 [learn-co-curriculum](#) / [dsc-statistical-measures-lab](#) Public 0 stars  32 forks Star Watch ▾

<> Code

 Issues Pull requests Actions Projects Security Insights solution ▾

...

This branch is [4 commits ahead](#), [5 commits behind](#) master. Contribute ▾

hoffm386 execution count normalization ...

on Jan 25, 2021

 5[View code](#) README.md

Statistical Measures - Cumulative Lab

Introduction

Another section down! Let's pull together the statistical measures learned so far to analyze a dataset and produce some business recommendations.

Objectives

You will be able to:

- Recall the concepts and applications of measures of central tendency and dispersion
- Practice applying and interpreting measures of central tendency and dispersion
- Recall the concepts and applications of covariance and correlation
- Practice applying and interpreting covariance and correlation

Your Task: Sales Data Analysis and Advertising Recommendations



Photo by [Scott Graham](#) on [Unsplash](#)

Business Understanding

Imagine you work for a company that sells widgets¹ and your boss has asked you to look into the sales data across your media markets for this year. She wants to know:

1. What sales volume do we have in a typical market?
2. How variable are sales across markets?
3. If we have 25k more dollars to spend in advertising per market, should we spend it on TV, radio, or newspaper ads?

¹Here we are using the [second definition](#) of widget: "an unnamed article considered for purposes of hypothetical example"

Data Understanding

For this lab we will be using a popular dataset known as the "Advertising Dataset". It comes from [An Introduction to Statistical Learning with Applications in R](#) by G. James, D. Witten, T. Hastie and R. Tibshirani. We have downloaded this dataset for you and stored it in this repository.

This dataset contains four lists. Each number in each list represents the value for that list in a given market. The four lists are:

1. `sales` : the number of widgets sold (in thousands)
2. `tv` : the amount of money (in thousands of dollars) spent on TV ads
3. `radio` : the amount of money (in thousands of dollars) spent on radio ads
4. `newspaper` : the amount of money (in thousands of dollars) spent on newspaper ads

So, for example:

- the **third number** from each list represents the value of `sales`, `tv`, `radio`, and `newspaper` in **one** market,
- the **fourth number** from each list represents the value of `sales`, `tv`, `radio`, and `newspaper` in **another** market,

and so on.

Requirements

1. Sales Data Summary

Write code that describes the number of markets a given list has records for, as well as the sales numbers for the markets with the minimum and maximum sales.

2. Typical Sales Volume

Use a measure of central tendency to describe a "typical" market's sales.

3. Dispersion of Sales Volume

Use a measure of dispersion to describe how variable sales are across markets.

4. Correlations between Advertising Expenditure and Sales

Calculate the correlation between TV, radio, and newspaper ad spending and widget sales.

5. Where to Spend Additional Dollars

Use the findings from step 4 to make a recommendation.

Sales Data Summary

In the cell below, we've opened up the dataset and loaded it into lists named `sales`, `tv`, `radio`, and `newspaper`.

```
import pandas as pd

data = pd.read_csv("advertising.csv", index_col=0)

sales = list(data["sales"])
tv = list(data["TV"])
radio = list(data["radio"])
newspaper = list(data["newspaper"])

# display the first 10 sales amounts
sales[:10]

[22.1, 10.4, 9.3, 18.5, 12.9, 7.2, 11.8, 13.2, 4.8, 10.6]
```

Replace `None` with appropriate code so that this cell prints out the correct information. For this part, you only need to use the `sales` variable.

Reminder: Replace `None` with code that **calculates** the answer. Don't calculate the answer by hand and then replace `None` with the number of your answer!

```
num_markets = len(sales)
min_sales = min(sales)
max_sales = max(sales)

print(f"""
This dataset contains records for {num_markets} markets

The fewest sales for any market was {min_sales} thousand widgets

The most sales for any market was {max_sales} thousand widgets
""")

This dataset contains records for 200 markets

The fewest sales for any market was 1.6 thousand widgets

The most sales for any market was 27.0 thousand widgets
```

Run this code to create a histogram of all sales data:

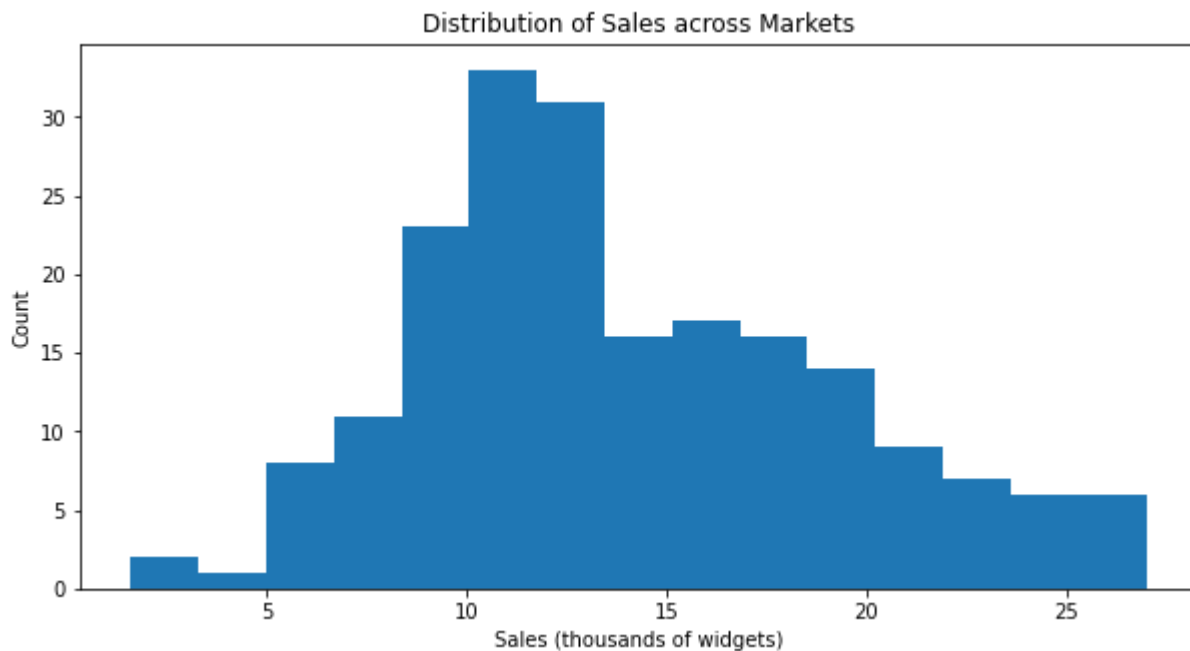
```
import matplotlib.pyplot as plt
%matplotlib inline

fig, ax = plt.subplots(figsize=(10, 5))

ax.hist(sales, bins=15)

ax.set_xlabel("Sales (thousands of widgets)")
ax.set_ylabel("Count")

ax.set_title("Distribution of Sales across Markets");
```



Typical Sales Volume

Now we should be able to address the first business question: *What sales volume do we have in a typical market?*

That sounds like a question to be answered by a **measure of central tendency**.

Reminder: the three measures of central tendency we've introduced are:

- Mean
- Median
- Mode

Choose the measure that seems most reasonable to you (it's a judgment call — there isn't always a single correct answer!) and complete the cell below, using NumPy or SciPy to compute the measure.

```
import numpy as np
from scipy import stats

# You can make an argument for either of these, really

# OPTION 1
measure_central_tendency = np.mean(sales)

print(f"""
Typical sales volume is {measure_central_tendency} thousand widgets

I chose mean as the relevant measure because that is normally
what people think of when they think of an average/typical example
""")

# OPTION 2
measure_central_tendency = np.median(sales)

print(f"""
Typical sales volume is {measure_central_tendency} thousand widgets

I chose median as the relevant measure because it is useful when
data is skewed (which this data is, as we can observe from the
plot above)
""")

# Mode makes less sense here, because these aren't categorical
# and there are so many different values present. The mode is
# 9.7 thousand, which is the sales number for only 5 out of the
# 200 records, and doesn't really line up with the peak of the
# histogram. It's hard to argue that 9.7 thousand is really
# "typical" here.

Typical sales volume is 14.0225 thousand widgets

I chose mean as the relevant measure because that is normally
what people think of when they think of an average/typical example

Typical sales volume is 12.9 thousand widgets

I chose median as the relevant measure because it is useful when
```

data is skewed (which this data is, as we can observe from the plot above)

Dispersion of Sales Volume

Now that we have a number to represent the typical sales volume, let's answer: *How variable are sales across markets?*

That sounds like a question to be answered by a **measure of dispersion** (also known as a measure of spread).

Reminder: the measures of dispersion we've introduced are:

- (Average) absolute deviation
- Variance
- Standard deviation
- Interquartile range

Choose the measure that seems the most reasonable to you, and write up your answer in the cell below, following the format from the previous question (first calculating the measure, then explaining your answer).

```
# Of these options, variance is probably the least useful,  
# since the units are thousands of widgets squared. That is  
# challenging for a business audience to interpret
```

```
# OPTION 1
```

```
measure_dispersion = np.mean(np.absolute(sales - np.mean(sales)))
```

```
print(f"""
```

```
A typical market differs from the mean market by  
{round(measure_dispersion, 3)} thousand widgets
```

```
I chose average absolute deviation as the measure because:
```

- it represents the variability in a single number
- the units are understandable
- compared to standard deviation, it is less impacted by outliers. If higher dispersion is fine (i.e. we aren't worried about consistency), this gives us a metric that isn't unnecessarily skewed

```
""")
```

```
# OPTION 2
```

```
measure_dispersion = np.std(sales, ddof=1)
```

```
print(f"""
A typical market differs from the mean market by
{round(measure_dispersion, 3)} thousand widgets

I chose standard deviation as the measure because:
- it represents the variability in a single number
- the units are understandable
- outliers are more represented, because we find the sum
  of squares then take the square root. If higher
  dispersion is "bad" (i.e. we want consistency), this
  gives us the more realistic/pessimistic metric
""")

# OPTION 3
p25 = np.percentile(sales, 25)
p75 = np.percentile(sales, 75)

measure_dispersion = p75 - p25

print(f"""
The 25th percentile market had sales of {p25} thousand
widgets, and the 75th percentile market had sales of {p75}
thousand widgets

The difference between those markets is {round(measure_dispersion, 3)},
meaning that the middle 50% of markets is spread across
this range (inter-quartile range)

I chose quartiles (and IQR) as the measures because:
- it represents more information about the data through
  three separate numbers
- unlike average absolute deviation or standard deviation,
  these numbers represent actual values in the dataset, not
  averaged values
- it represents the skew of the dataset, showing that the
  25th percentile is closer to the median (less spread out)
  than the 75th percentile is (more spread out)
""")

A typical market differs from the mean market by
4.28 thousand widgets

I chose average absolute deviation as the measure because:
- it represents the variability in a single number
- the units are understandable
- compared to standard deviation, it is less impacted by
  outliers. If higher dispersion is fine (i.e. we aren't
```


worried about consistency), this gives us a metric that isn't unnecessarily skewed

A typical market differs from the mean market by 5.217 thousand widgets

I chose standard deviation as the measure because:

- it represents the variability in a single number
- the units are understandable
- outliers are more represented, because we find the sum of squares then take the square root. If higher dispersion is "bad" (i.e. we want consistency), this gives us the more realistic/pessimistic metric

The 25th percentile market had sales of 10.375 thousand widgets, and the 75th percentile market had sales of 17.4 thousand widgets

The difference between those markets is 7.025, meaning that the middle 50% of markets is spread across this range (inter-quartile range)

I chose quartiles (and IQR) as the measures because:

- it represents more information about the data through three separate numbers
- unlike average absolute deviation or standard deviation, these numbers represent actual values in the dataset, not averaged values
- it represents the skew of the dataset, showing that the 25th percentile is closer to the median (less spread out) than the 75th percentile is (more spread out)

Correlations between Advertising Expenditure and Sales

Now that we have a general understanding of the distribution of the sales data, we can start to answer: *If we have 25k more dollars to spend in advertising per market, should we spend it on TV, radio, or newspaper ads?*

(Eventually we will learn more sophisticated multivariate modeling techniques that will allow us to simulate the impacts of different choices here such as *given TV spending of \$ x_1 and radio spending of \$ x_2 , how would increasing newspaper spending by 25k impact \$ y ?*, but for now we will just use the tools we have learned so far.)

In order to make this recommendation, let's find the **correlation between each advertising medium and the associated sales**.

(Recall that *covariance* is the numerator of the correlation formula, and that we typically use correlation rather than just covariance because its magnitude is more interpretable.)

In the following cell, compute the correlation between `sales` and `tv`, `radio`, and `newspaper` using NumPy.

```
tv_corr = np.corrcoef(sales, tv)[0][1]
radio_corr = np.corrcoef(sales, radio)[0][1]
newspaper_corr = np.corrcoef(sales, newspaper)[0][1]

print("Correlation of Sales and TV Ad Spending:", tv_corr)
print("Correlation of Sales and Radio Ad Spending:", radio_corr)
print("Correlation of Sales and Newspaper Ad Spending:", newspaper_corr)
```

```
Correlation of Sales and TV Ad Spending: 0.7822244248616061
Correlation of Sales and Radio Ad Spending: 0.5762225745710551
Correlation of Sales and Newspaper Ad Spending: 0.22829902637616528
```

Which type of ad spending has the highest correlation?

```
print("TV has the highest correlation with sales")
```

```
TV has the highest correlation with sales
```

Let's also plot out each of the ad types vs. sales:

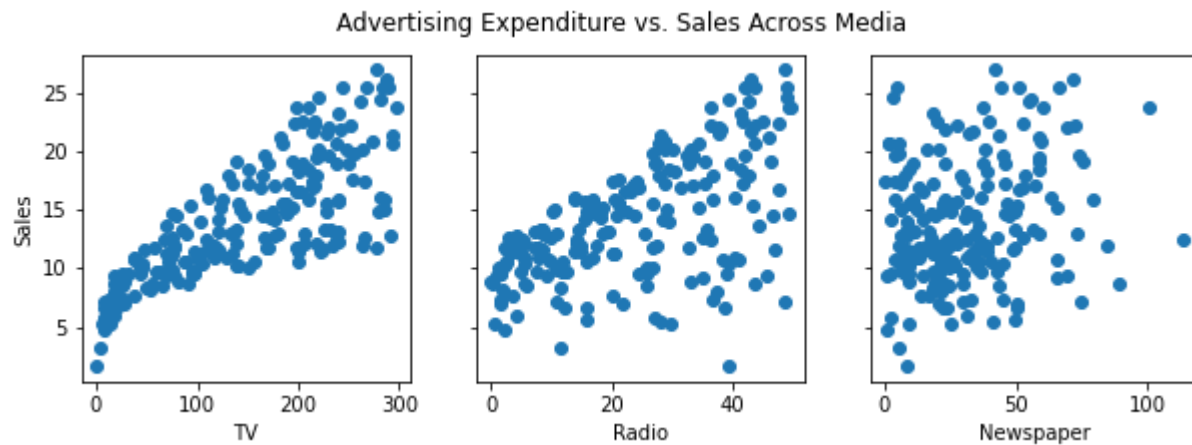
```
fig, (ax1, ax2, ax3) = plt.subplots(ncols=3, figsize=(10,3), sharey=True)

ax1.scatter(tv, sales)
ax2.scatter(radio, sales)
ax3.scatter(newspaper, sales)

ax1.set_xlabel("TV")
ax2.set_xlabel("Radio")
ax3.set_xlabel("Newspaper")

ax1.set_ylabel("Sales")

fig.suptitle("Advertising Expenditure vs. Sales Across Media");
```



Where to Spend Additional Dollars

Based on the correlation numbers and a visual inspection of those plots, make a recommendation to your boss about where to spend 25k extra dollars per market and why.

```
print("""
Based on correlation alone, TV spending clearly has the highest
correlation with sales. The scatter plot for TV looks closest
to a straight line as well, which aligns with the idea that there
is an association between increased TV spending and increased
sales. While we can't necessarily assume that correlation =
causation here, it seems reasonable to guess that increased TV ad
spending in those markets has led to increased sales.
```

```
(You could also make an argument that radio would be the better
place to spend the money, even though the correlation is lower and
the scatter plot is farther from a straight line. It looks like we
are spending less than half as much on radio right now as we're
spending on newspaper, and less than 1/6 as much as on TV. So a
targeted expenditure of 25k on radio might go farther than the
same amount spent on TV.)
""")
```

Based on correlation alone, TV spending clearly has the highest correlation with sales. The scatter plot for TV looks closest to a straight line as well, which aligns with the idea that there is an association between increased TV spending and increased sales. While we can't necessarily assume that correlation = causation here, it seems reasonable to guess that increased TV ad spending in those markets has led to increased sales.

(You could also make an argument that radio would be the better

place to spend the money, even though the correlation is lower and the scatter plot is farther from a straight line. It looks like we are spending less than half as much on radio right now as we're spending on newspaper, and less than 1/6 as much as on TV. So a targeted expenditure of 25k on radio might go farther than the same amount spent on TV.)

Conclusion

In this cumulative lab, you practiced analyzing sales and advertising data in order to make a business recommendation. Unlike some other labs, there was more ambiguity and we asked you to make some judgment calls in order to use data science concepts for a business audience. In the rest of the course, you will continue building your technical skills and technical communication skills!

Releases

No releases published

Packages

No packages published

Languages

● Jupyter Notebook 100.0%