

# Statistics (I)

---

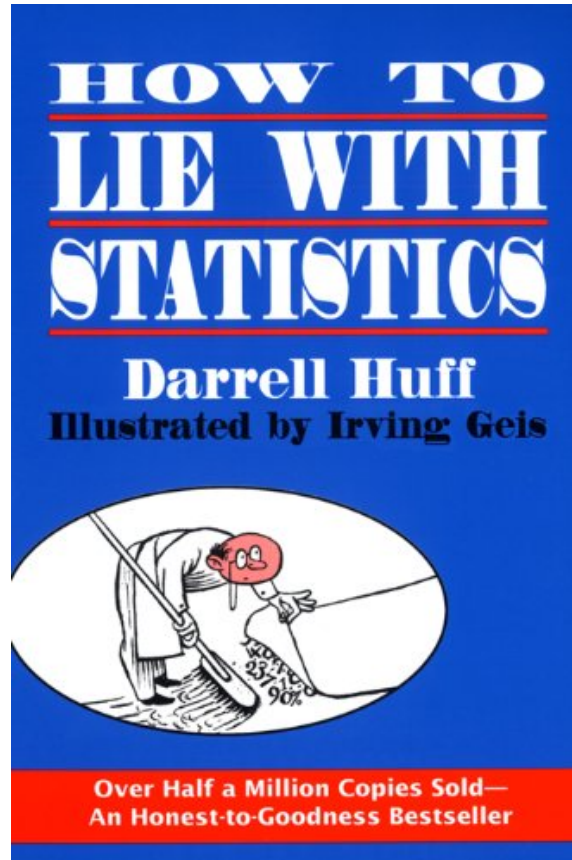
SLIDES BY:

JIANNAN WANG

<https://www.cs.sfu.ca/~jnwang/>

# Why Should you Care?

---



**“ There are three kinds of lies:  
lies, damned lies, and statistics ”**

Would you like to be called a **lying data scientist?**

# Outline

---

**Statistical Thinking**

**Descriptive Statistics**

**Inferential Statistics**

# Outline

---

## Statistical Thinking

Descriptive Statistics

Inferential Statistics

# Statistical Thinking

---

1. Data is just a **sample**
2. Your goal is to infer a **population**
3. Think about how to go “backwards” from the **sample** to the **population**

# Example 1. Image Classification

---

Is it a dog or a cat?



**Dataset:** 1000 images collected from the Web

# Without Statistical Thinking

---

**Treat the 1000 images as the population**

- > Train a model on the data
- > Evaluate a model on the same data
- > **Model accuracy: 95%**

# With Statistical Thinking

---

## What is the population?

- All the images in the Web

## What is your dataset?

- A sample of 1000 images drawn from the Web

## What should you do?

- Split the dataset into a training dataset and a test dataset
- Train the model on the training dataset
- Evaluate the model on the test dataset



# Example 2. Poll Prediction

---

Who will win the election?



**Dataset:** A survey of 1000 people

# Without Statistical Thinking

---

**Treat the 1000 people as the population**

- > Count the number of people who wants to vote for Hillary, e.g., 52
- > Count the number of people who wants to vote for Trump, e.g., 48
- > Hillary will win the election

# With Statistical Thinking

---

## What is the population?

- All the people who will vote in the election day

## What is your dataset?

- A sample of 1000 people before the election day

## Analysis result

Hillary: 52%  $\pm$  3%  
Trump: 48%  $\pm$  2%

Assumption: People have not changed their votes since the time of the poll

# Summary

---

## Statistical Thinking

- Sample, Population and Their Connection
- With vs. Without Statistical Thinking

Descriptive Statistics

Inferential Statistics

# Outline

---

Statistical Thinking

**Descriptive Statistics**

Inferential Statistics

# Descriptive vs. Inferential Statistics

---

## Descriptive Statistics: e.g., Median

- Why? Aim to understand the data
- How? Data summarization, data visualization, etc.

## Inferential Statistics: e.g., A/B Testing

- Why? Aim to use the data (i.e., sample) to learn about a population
- How? Estimation, confidence intervals, hypotheses testing, etc.

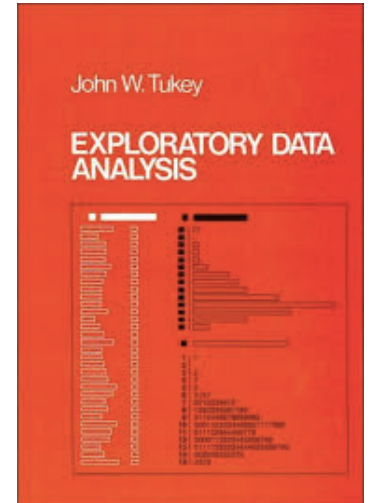
# Exploratory Data Analysis (EDA)

**The process of doing descriptive statistics**



**John W. Turkey**

- Professor at Princeton University
- Founding chairman of the Princeton statistics department in 1965
- Worked on EDA at Bell Labs since 60's
- Wrote a book entitled "Exploratory Data Analysis" in 1977



# EDA is like detective work

---



**From John Turkey**

**“ Exploratory data analysis is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as those that we believe to be there ”**



# Case Study

---

**Is UC Berkeley gender biased?**

	Applicants	Admitted
Men	8442	44%
Women	4321	35%

~~**YES!**~~

# Case Study

---

## Is UC Berkeley gender biased?

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

**NO!**

Women tended to apply to competitive departments with low rates of admission

# Chart Types

---

## Single Variable

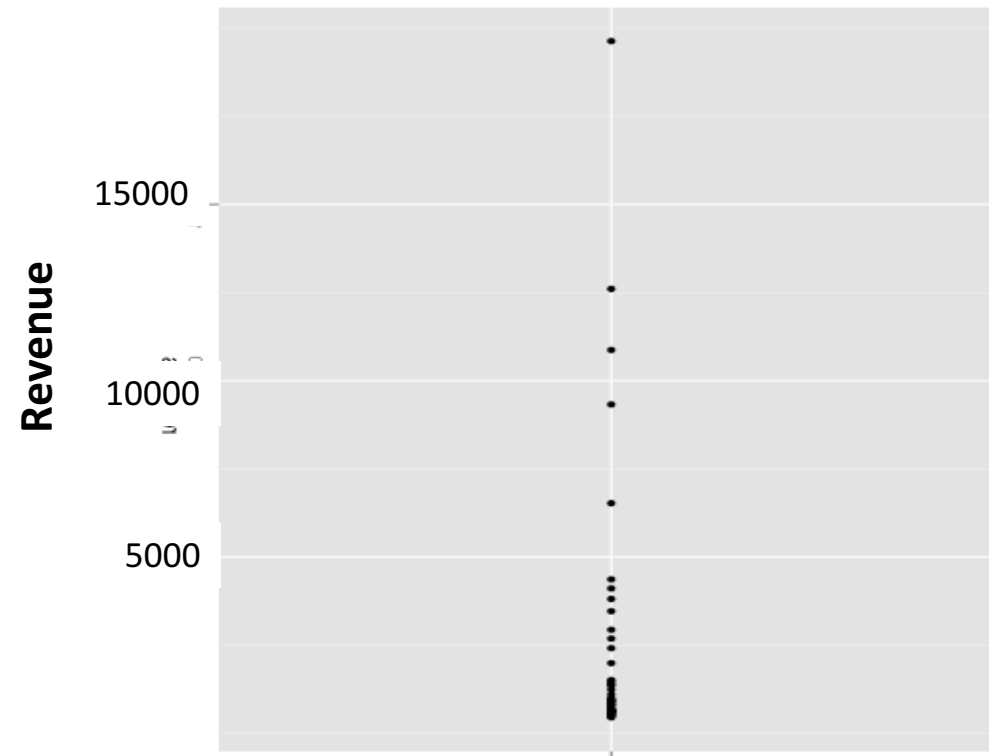
- Dot plot
- Jitter plot
- Error bar plot
- Box plot
- Histogram
- Kernel density estimate
- Cumulative distribution function

From UC Berkeley “Introduction to Data Science”

# Dot plot

---

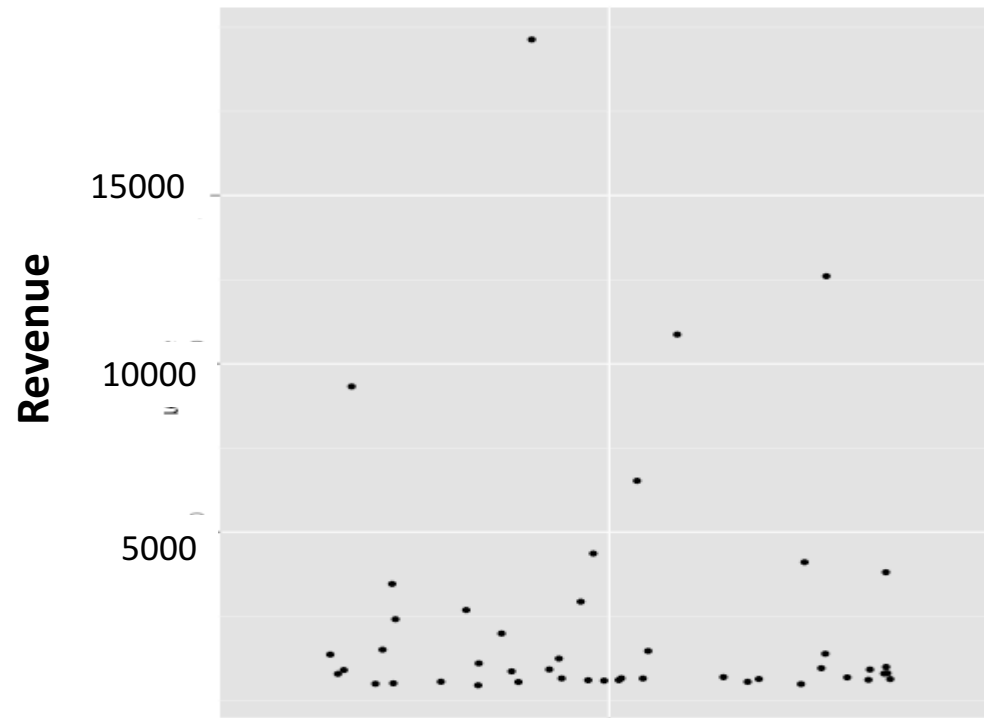
Revenue
2000
11000
5400
204944
32244
1232
...



# Jitter plot

Noise added to the x-axis to spread the points

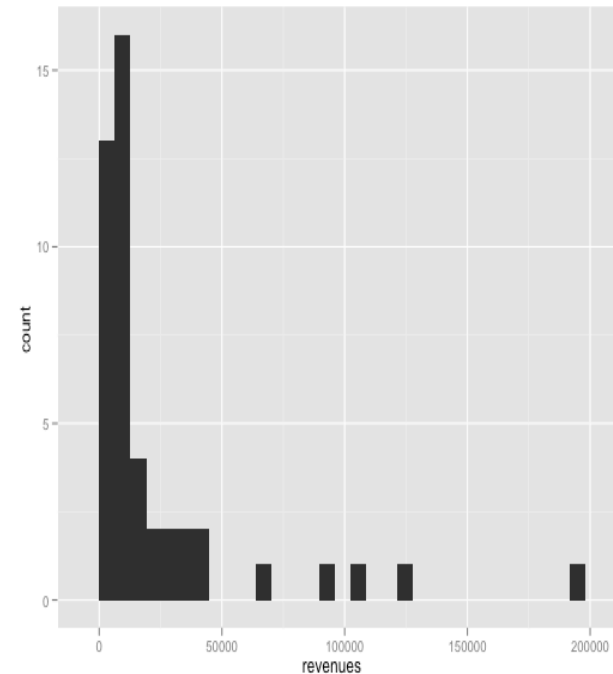
Revenue
2000
10000
5400
204944
32244
1232
...



# Histogram

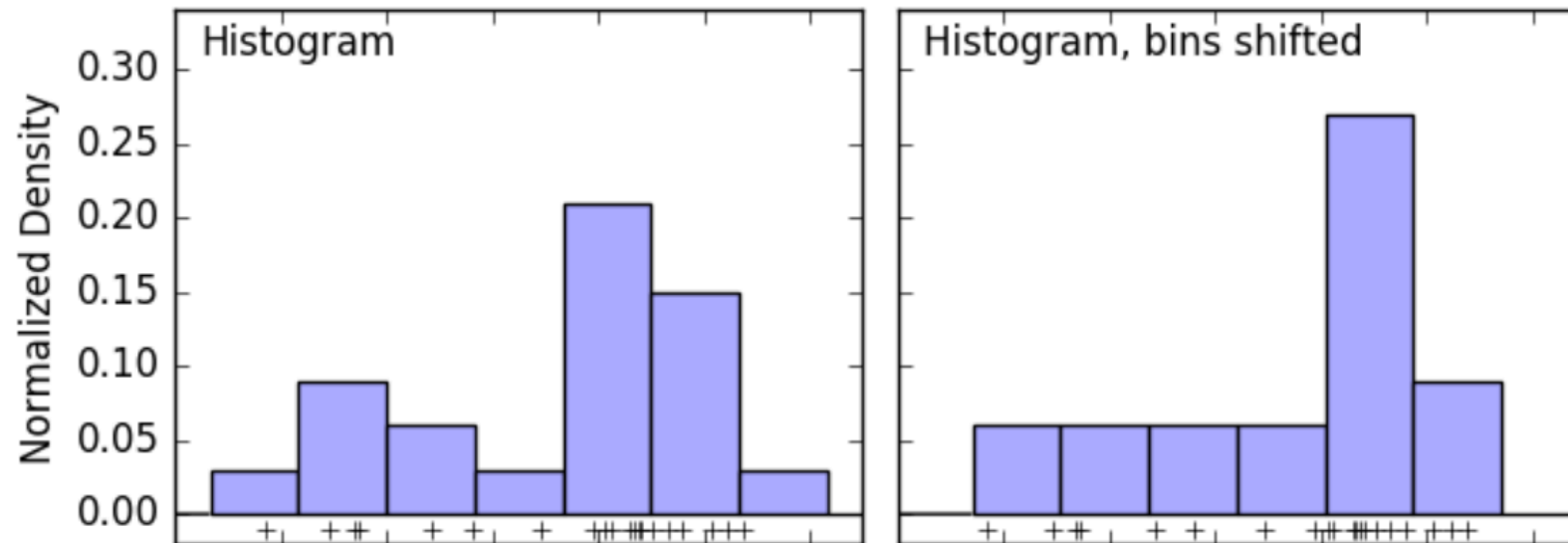
## Probability Density Functions

Revenue
2000
100000
5400
204944
32244
1232
...



# Limitation of Histogram

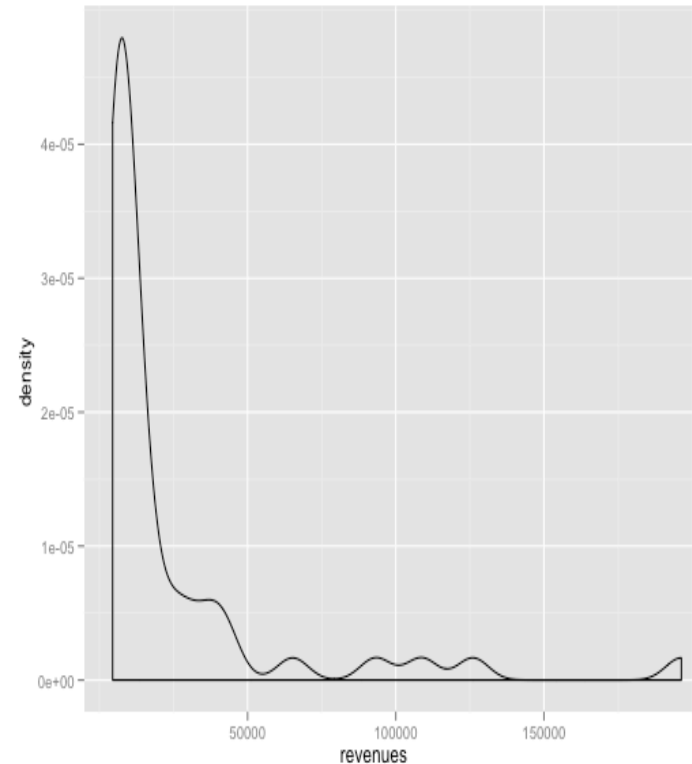
The choice of binning can have a big effect on the resulting visualization



# Solution 1: Kernel density estimate

Like a smooth histogram

Revenue
2000
10000
5400
204944
32244
1232
...

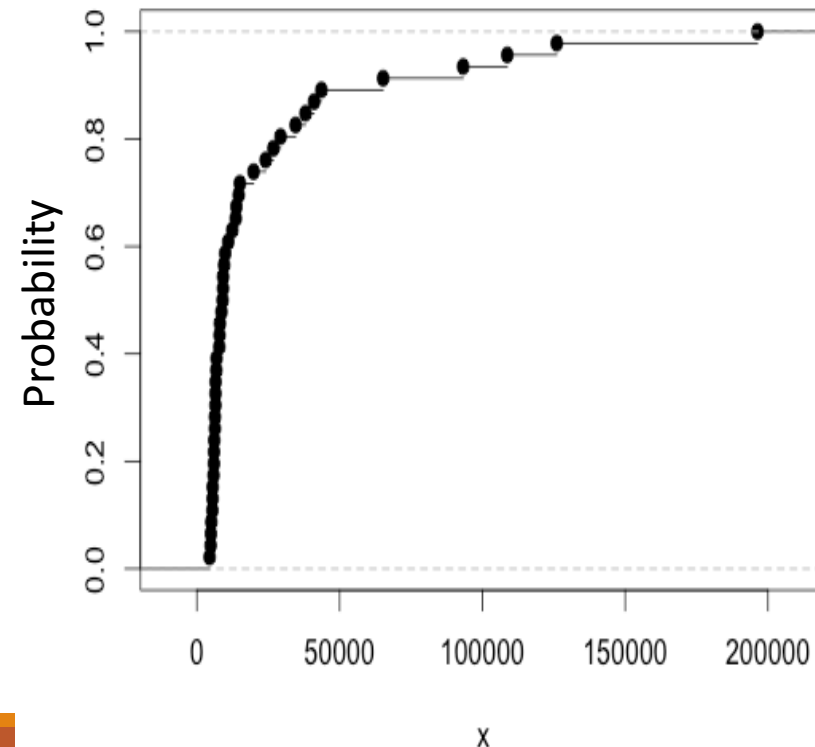




# Solution 2: Cumulative distribution function

## Integral of the histogram

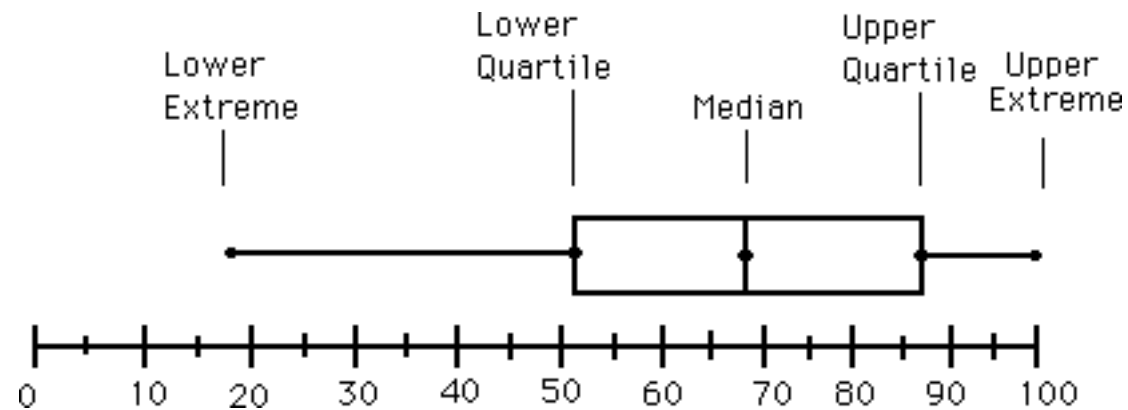
Revenue
2000
10000
5400
204944
32244
1232
...



# Box Plot

## A graphical form of 5-number summary

- Min, 25% Quartile, Median, 75% Quartile, Max

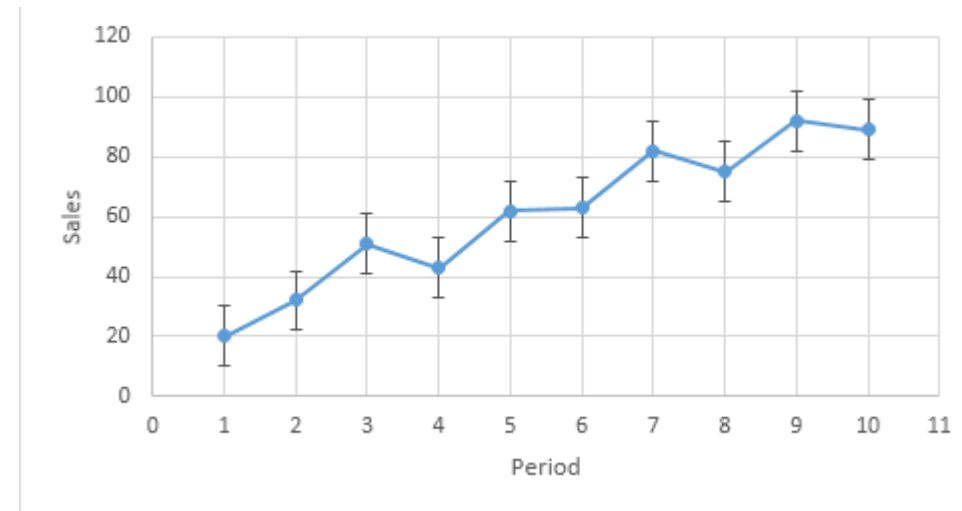
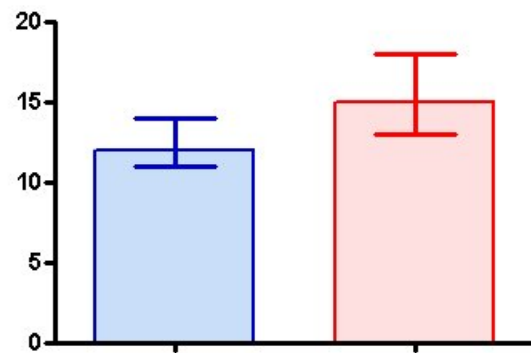


# Error Bars

Usually based on confidence intervals (CI).

95% CI means *95% of points are in the range*

Not necessarily symmetric



# Chart Types

---

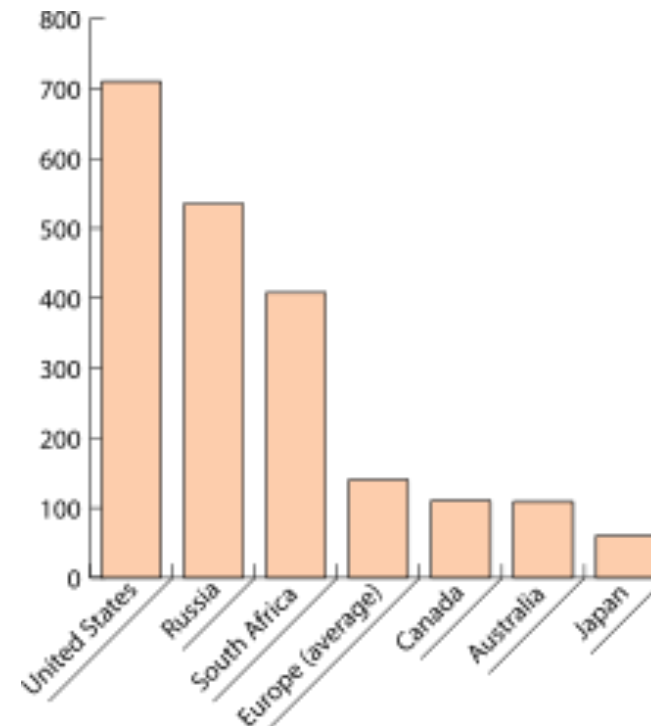
## Two or more variables

- Bar chart
- Scatter plot
- Line plot
- See more at <https://pandas.pydata.org/pandas-docs/stable/visualization.html#plotting-tools>

# Bar Plot

One variable is categorical

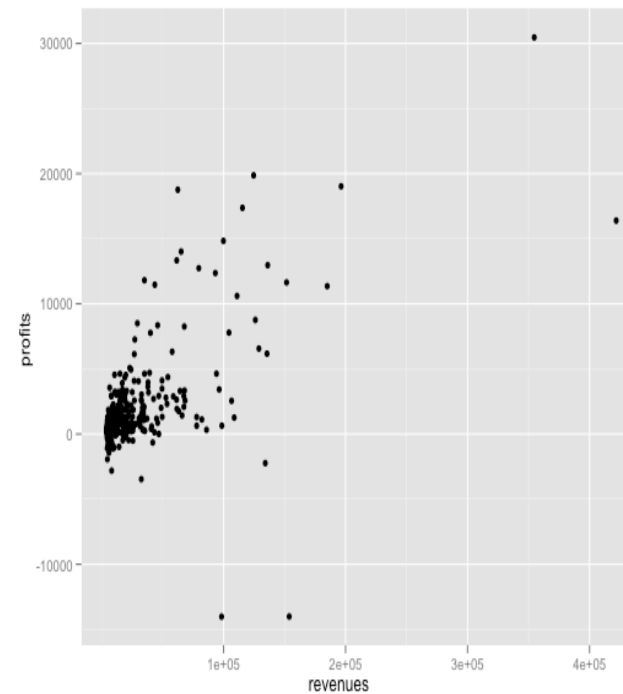
Region	Revenue
US	720
Russian	540
South Africa	400
Canada	120
...	



# Scatter Plot

Variables are both numerical

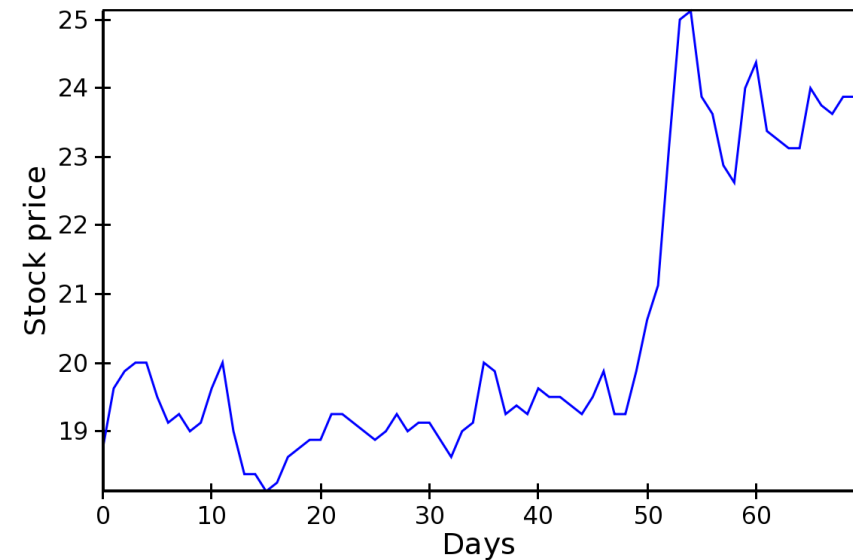
Revenue	Profit
20000	1000
45000	450
50234	-200
34522	900
	...



# Line Plot

One variable is ordinal

Days	Price
1	15.34
2	17.12
3	18.56
4	19.21
...	...



# Summary

---

## Statistical Thinking

### **Descriptive Statistics**

- Descriptive vs. Inferential Statistics
- Exploratory Data Analysis
- Chart types

## Inferential Statistics



# Outline

---

Statistical Thinking

Descriptive Statistics

## **Inferential Statistics**

- Estimation (this lecture)
- Hypothesis Testing (next lecture)
- Regression (next lecture)

# Estimation

---

## Problem statement

- Estimate a numerical value associated with a population

## Examples

- Estimate the percentage of the people in the US who will vote for Trump
- Estimate the median annual income of all households in the US

# Example: Median Annual Income

---

**How to estimate the median annual income of all households in the US?**

- Randomly select 10,000 households from the US
- Report their median annual income: 50,000USD
- **BUT, we need to report something like**

50,000 ±500 USD

# A Naïve Solution

- Randomly select 10,000 households from the US
- Report their median annual income

Repeat this process for  
100 times

50,000    49,600    50,200    ...    49,200

You have to survey 1,000,000 million households in total 😞

# A Smart Solution: Bootstrapping

---

## Key Idea: Resampling

- Sample with replacement from the original data sample

**Population:** 1, 1, 8, 2, ... 3, 3

**Sample:** 3, 8, 1, 8, 3

**Resample:** 8, 3, 3, 3, 1

# A Smart Solution: Bootstrapping

---

- Randomly select 10,000 households from the US
- Draw a resample from the 10,000 households
- Report the median annual income of the resample

Repeat this process for  
100 times

You do NOT need to survey any new household. 😊

# Notes on Bootstrapping

---

Start with a large random sample (at least 30)

Replicate the resampling procedure as many times as possible (more than 1000 times)

Does not work for min/max

# Conclusion

---

## Statistical Thinking

- Sample, Population and Their Connection
- With vs. Without Statistical Thinking

## Descriptive Statistics

- Descriptive vs. Inferential Statistics
- Exploratory Data Analysis
- Chart types

## Inferential Statistics

- Estimation and Bootstrapping

### Assignment 4: EDA and Bootstrap

#### Objective

Statistics play a vital role in data science for (at least) two reasons. First, it can be used to **Explore** data. Second, it can be used to **Analyze** data. In this assignment, you will learn about EDA and statistical inference through the following:

1. Be able to perform EDA on a single column (i.e., univariate analysis)
2. Be able to perform EDA on multiple columns (i.e., multivariate analysis)

**Due next Monday**



# Plan for a 1-year Data Strategy

---

Team 1. SFU President Office

Team 2. BC Government

Team 3. Justin Trudeau Campaign Team

Team 4. Vancouver Hockey Team

Team 5. BC Children's Hospital

Team 6. Data Science Startup

# Final Project

---

## Companies

- Vancity. Property value prediction with market data
- WeRobotics. Automated Feature Detection of Aerial Imagery from South Pacific

## SFU Profs

- Department of Linguistics. Topic modeling and visualization of news comments
- School of Business. Machine learning to detect misstated financial statements
- Department of Earth Sciences. A prototype Canadian Natural Hazards Database

## Previous Cohorts

- Samsung Electronics Canada. Predictive Maintenance on IOT devices
- Terramera. Fall Detection using wearable sensor data
- Best Buy Canada. Predicting Soccer games and tournaments