MICROENGINEERING 110:
Probability and Statistics for Engineers

Prof. Vivek Subramanian

Microengineering

EPFL

INTRODUCTION

About Me

- **Professor of Microengineering**
  - **Joined EPFL in 2018, after spending 18 years as a professor at UC Berkeley**
  - **Spent most of my career working in / with the semiconductor industry**
  - **My research @ EPFL is focused on advanced micromanufacturing for electronics, micromechanics, medical devices, and microrobotics**

  - **Office: MC B4 187 in Microcity (Neuchatel)**
  - **Phone: +41216954265**
  - **Email: Vivek.Subramanian@epfl.ch**

Language of Instruction

- This course will be taught in English
  - Lectures in English
  - Tests in English
  - Exercises in English
- However
  - I understand French pretty well at this point, though my speaking is still poor, so you are welcome to try French if we are struggling to communicate with each other in English

Course Staffing

Student Teaching Assistants:

- **Maxime Charles M Blanpain, maxime.blanpain@epfl.ch**
- **Alice Athénaïs Domitille Marie Lemaire alice.lemaire@epfl.ch**
- **Ismael Tekaya, ismael.tekaya@epfl.ch**
- **Célia Marie Bernadette Lundmark celia.lundmark@epfl.ch**
- **Constance Sophie Hélène Alice Gagneraud constance.gagneraud@epfl.ch**
- **Anatole Ming Debierre anatole.debierre@epfl.ch**

PhD teaching Assistants

- **Kyle Haas kyle.haas@epfl.ch**

Course Logistics

Lectures:

- **Room C01**
- **Pre-recorded videos from previous years will be made available at: https://mediaspace.epfl.ch/channel/MICRO-110+Design+of+experiments/ at the end of each week. These are intended to serve as reference videos for you to use, as a backup to the in-class lectures. You will be responsible for all material covered in class.**
- **I will also be recording videos from CO1 and making them available each week**

Exercises:

- **Weeks with merged lectures / exercises: CO1**
  - **Lecture will be interspersed with exercise. Jupyter notebooks will be used throughout, and the entire 3 hour session will be held in CO1**
- **Weeks with separate exercise: 18h-19h in rooms CO 4-5-6 and 260**
  - **Exercise will be held in computer rooms, and you will use Jupyter notebooks to complete the exercise. Exercises for paper-based solution will also be provided a few days ahead of the exercise session**

Examinations

Examinations:

- **There will be three midterm tests, each worth 10% of the course grade. The tests will be administered via moodle during class hours.**
  - **Test 1: March 27**
  - **Test 2: May 1**
  - **Test 3: May 29**
- **The final examination date will be announced during the semester when available.**

Course Syllabus

Topic	Key material

Course Introduction	• Overview of the course
	• Why do we care about use of statistics?
	  ○ Observation
	  ○ Model building
	  ○ Inference
	• Why do we care about probability?
	  ○ Estimating likelihood

Introduction to statistics	• Mean, Median, Mode, Standard Deviation
	• Population Statistics
	  ○ Graphical Representation
	  ○ Population distributions
	  ○ Mean and standard deviation
	  ○ Sampling

Probability	• Sample spaces and events
	• Properties of probability
	• Discrete Random Variables and Probability
	  • Binomial and Poisson Distributions
	• Continuous Random Variables and Probability
	  • Normal and other continuous distributions
	• Joint Probability Distributions
	  • Covariance and Correlation
	• Point Estimation

Comparison Statistics	• Significance tests
	  • T tests and other hypothesis tests
	  • ANOVA
	• Regression and fitting

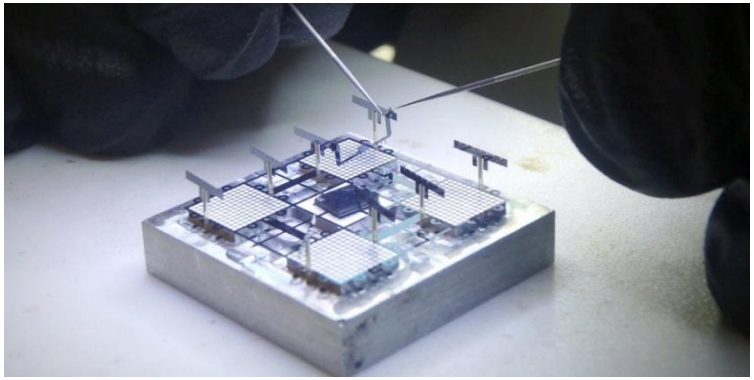
- "Quantitative"
  - Involves measurement
  - Data in numerical form
  - Answers "How much" questions
  - Objective and results in unambiguous conclusions
- "Qualitative"
  - Describes the nature of something
  - Answers "What" or "Of what kind" questions
  - Often evaluative and ambiguous

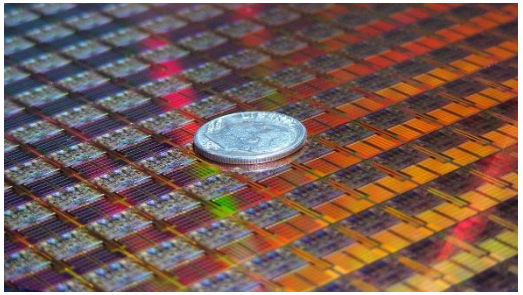
Intro. to Statistics

- **What can Stats do?**
  - **Allow us to draw conclusions from the data**
    - **Group of numbers #1: 6, 1, 8, 3, 5, 4, 9**
    - **Average is 5 1/7**
    - **Group of numbers #2: 8, 3, 4, 2, 7, 1, 4**
    - **Average is 4 ¼**
  - **Allows us to do this objectively and quantitatively**
- **What can Probability do?**
  - **Can help us estimate the likelihood of an event**
  - **Can inform us about the quality of our inferences and estimations**

What do engineers do?

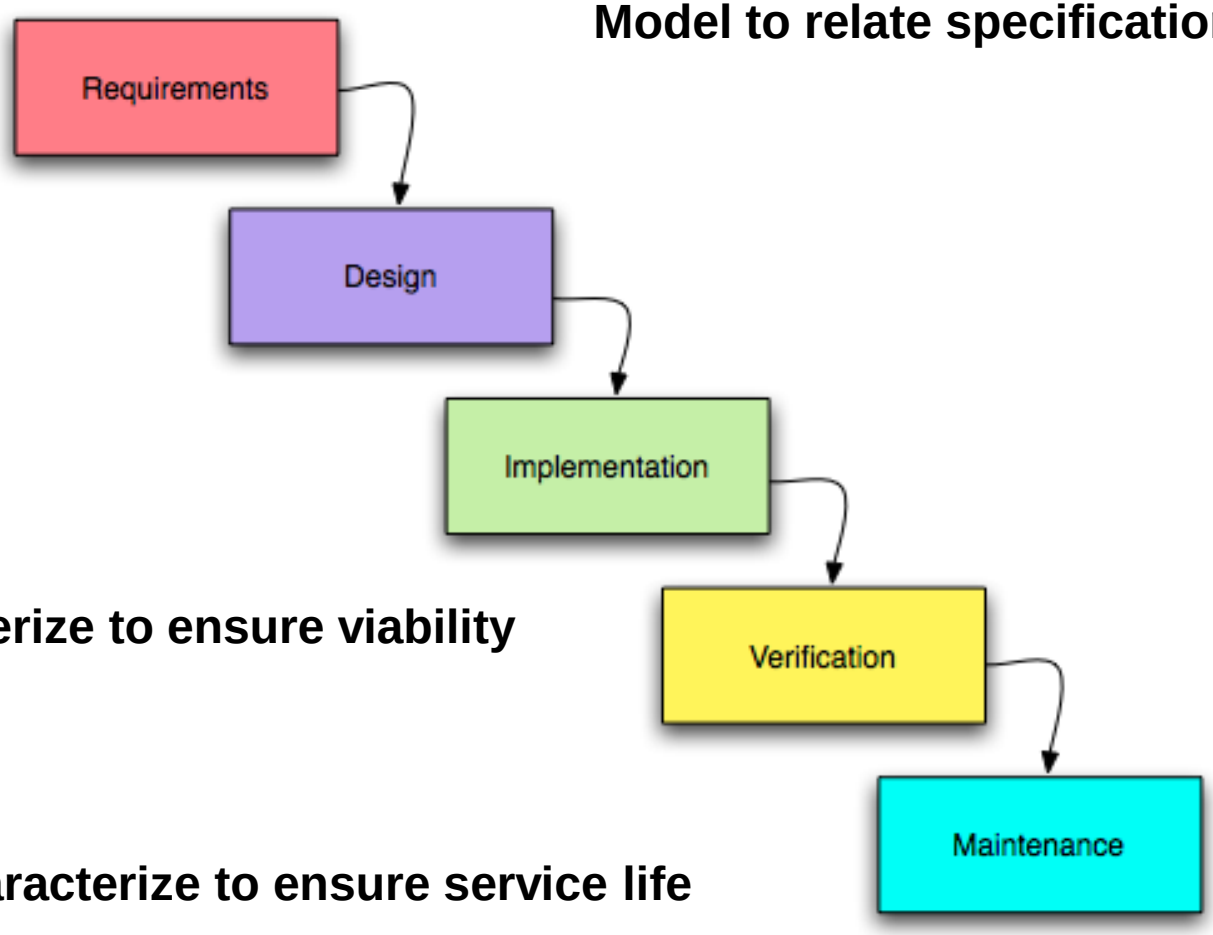
- **We build stuff**





The engineering development cycle

Model to relate specifications to design



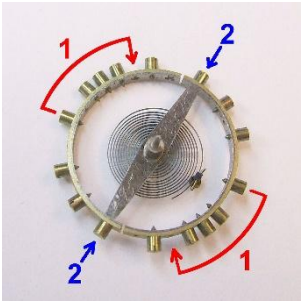
Characterize to ensure viability

Characterize to ensure service life

An example: tuning timing on mechanical watches

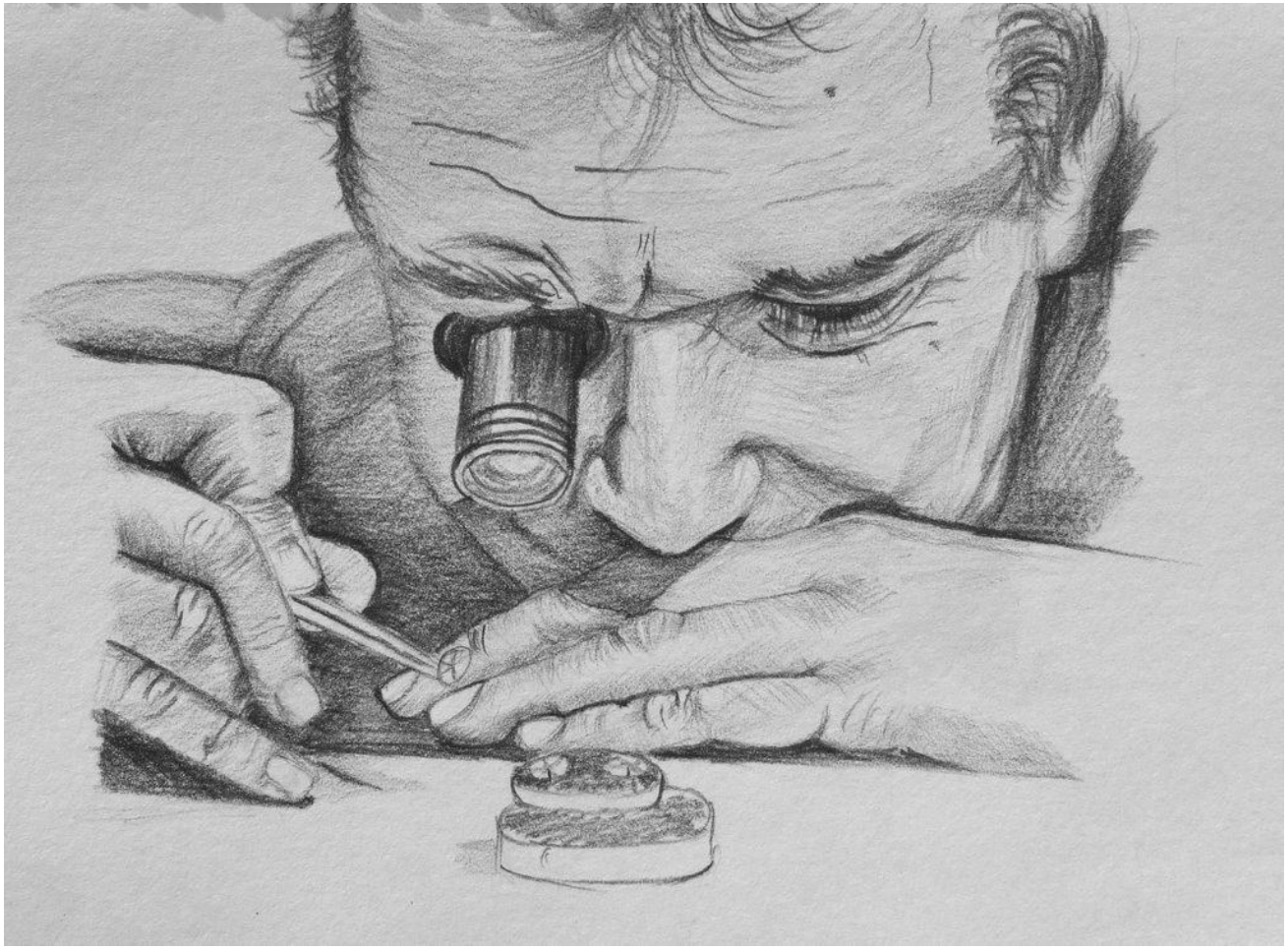
- **A balance wheel's period of oscillation T in seconds, the time required for one complete cycle (two beats), is determined by**
  - **the wheel's moment of inertia I in kilogram-meter^2**
  - **the stiffness (spring constant) of its balance spring κ in newton-meters per radian:**

$$T = 2\pi\sqrt{\frac{I}{\kappa}}$$



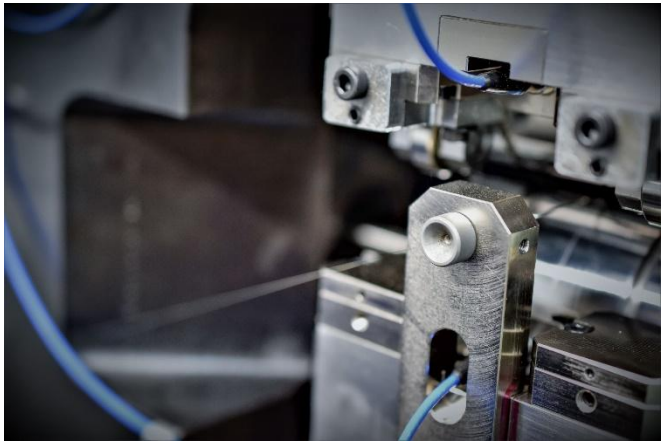
- Springs and balances as produced do not match to the correct frequency.
- Springs are compared with respect to a standard balance and their stiffness is estimated.
- Balance wheels are compared to a standard spring and their inertia is estimated.
- These are paired (“appairage") to get close to the desired frequency of the combination.
- Fine tuning is done on the watch:
  - Spring stiffness is adjusted using the “raquette”. This changes the length of the spring and so its stiffness.
  - Balance wheels usually have screws that tune the moment of inertia.

The problem…

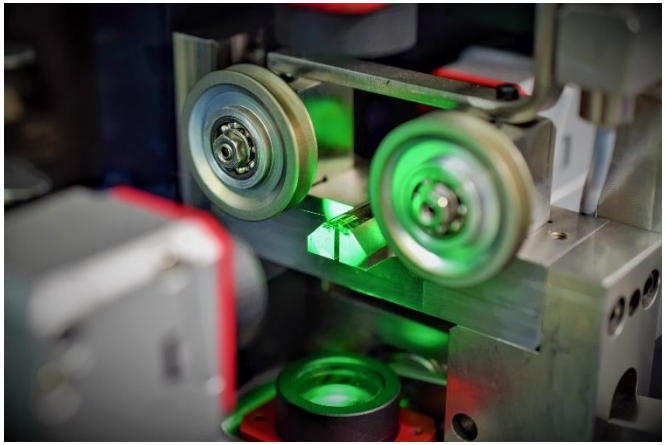


Expensive, difficult to scale

How are Invar-like Hairsprings made?

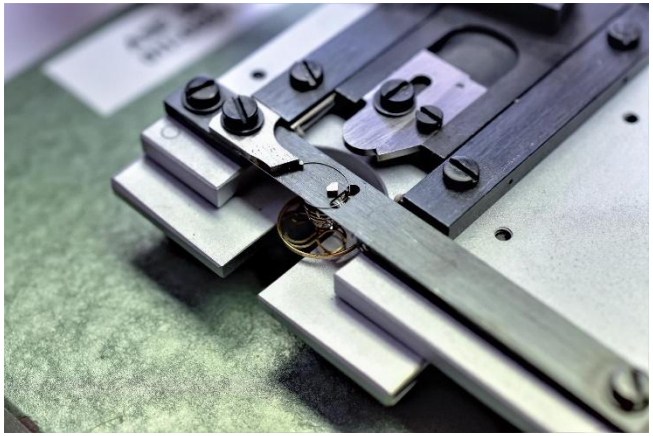


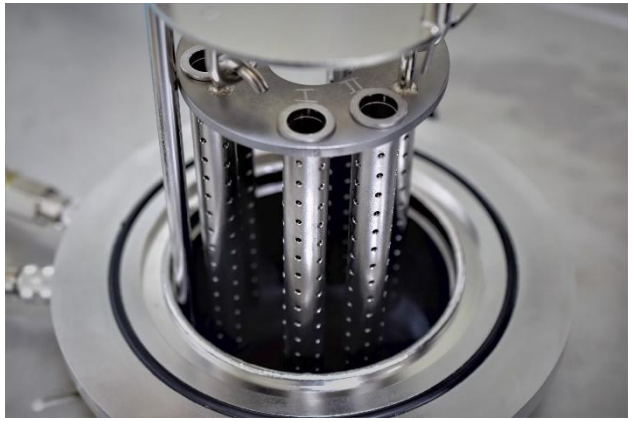
Wire is drawn to set diameter



Spring is coiled and baked

Final shaping

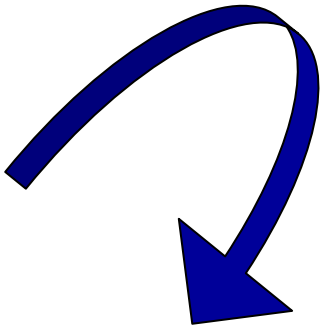


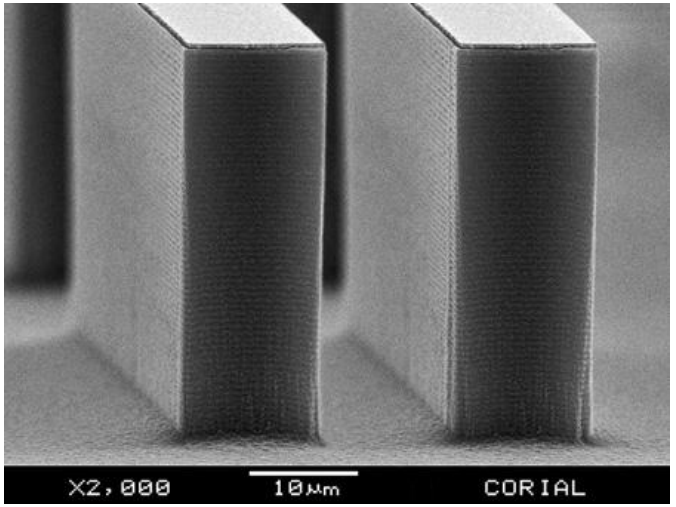


Each step results in variations across the parts

Silicon hairsprings

Plasma chemistry that etches vertically with very precise lateral control



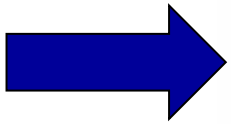


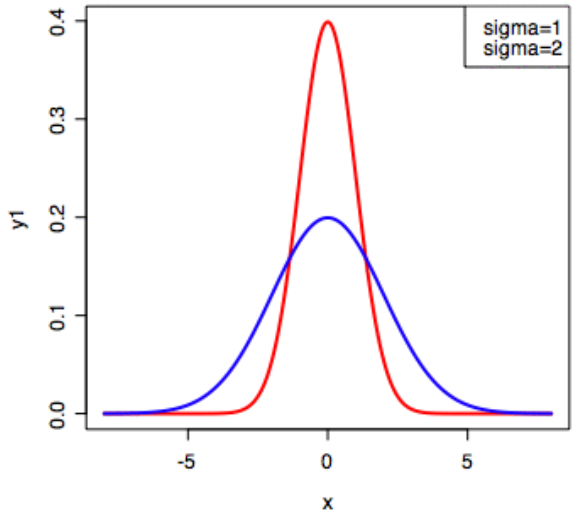
Can produce very tall structures with nm precision

Makes high quality springs with improved precision

Statistics can help

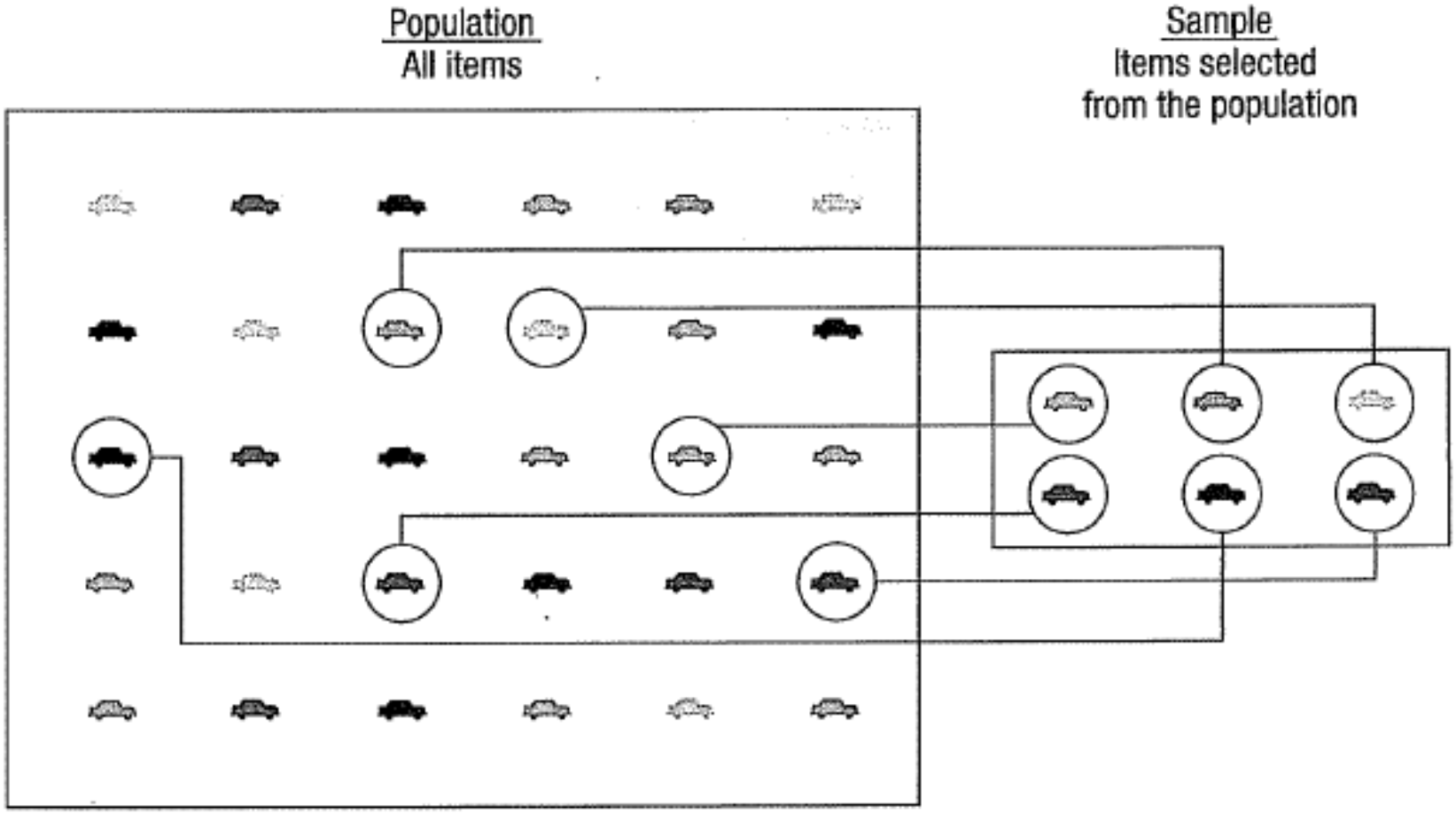
- From metal to ultra-precise silicon



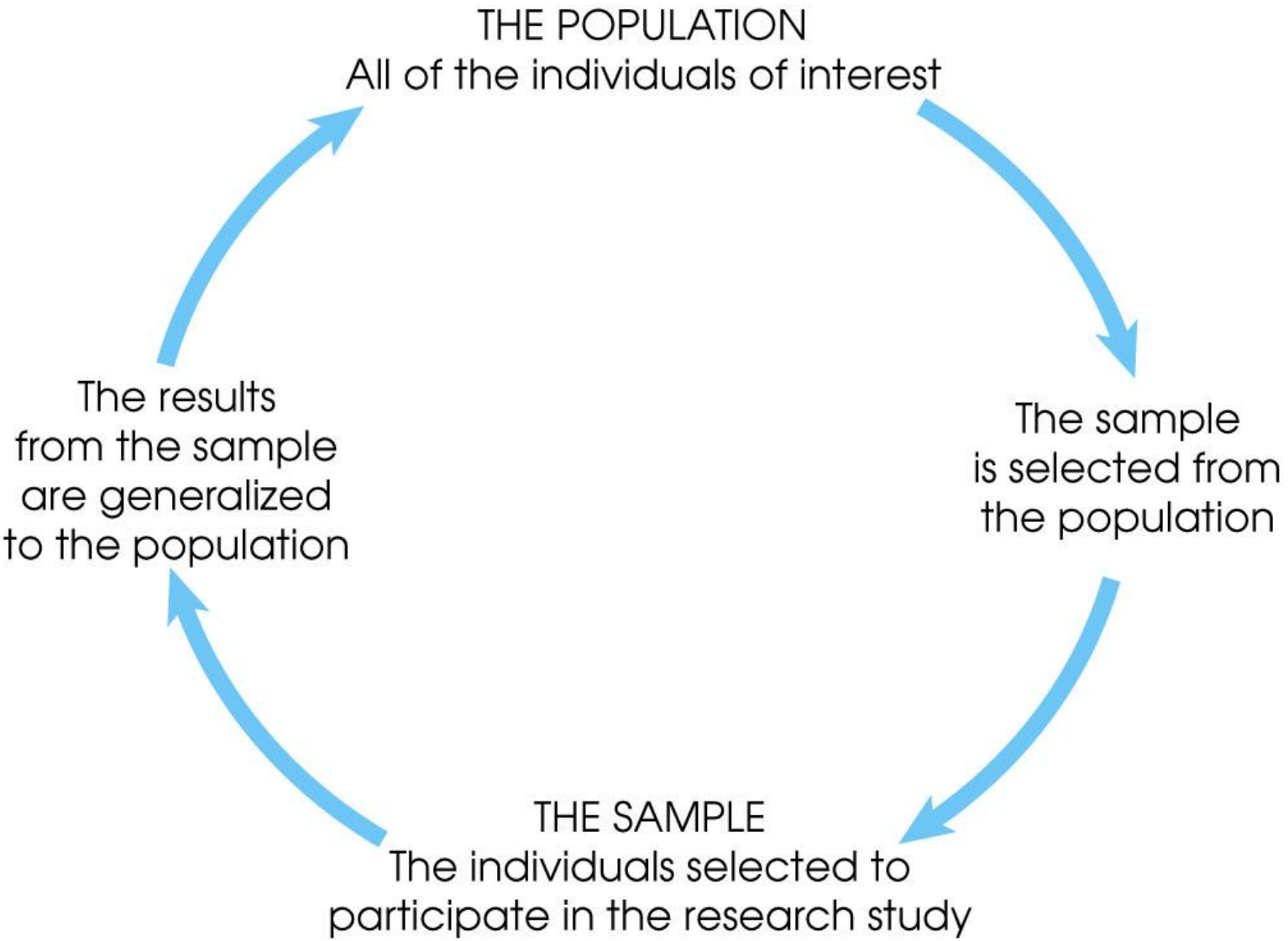


Why is this beneficial?

Sampling – living in the real world



Sampling in inferential statistics



Sampling

- **Size –**
  - **Sample of 5 cards from a deck of 52**
    - **2 of Clubs, 10 of Diamonds, Jack of Hearts, 5 of Clubs, and 7 of Hearts**
  - **What could we conclude about the full deck from this sample about what the full deck looks like without any prior knowledge of a deck of cards?**
  - **Compare this to a sample of 51/52 cards – What could we conclude from this sample?**

Sampling

- **Randomness –**
  - **This time lets use the same 5 card sample, but this time the deck is unshuffled (nonrandom)**
    - **2 of Clubs, 10 of Clubs, Jack of Clubs, 5 of Clubs, and 7 of Clubs**
  - **What would we conclude about the characteristics of our population (the deck) this time versus when the sample was more random (shuffled)?**

Sampling

- **Smaller/less random samples both poorly represent population of entire deck of cards**
  - **Also result in inaccurate inferences about population – poor external validity**

Inferential Statistics

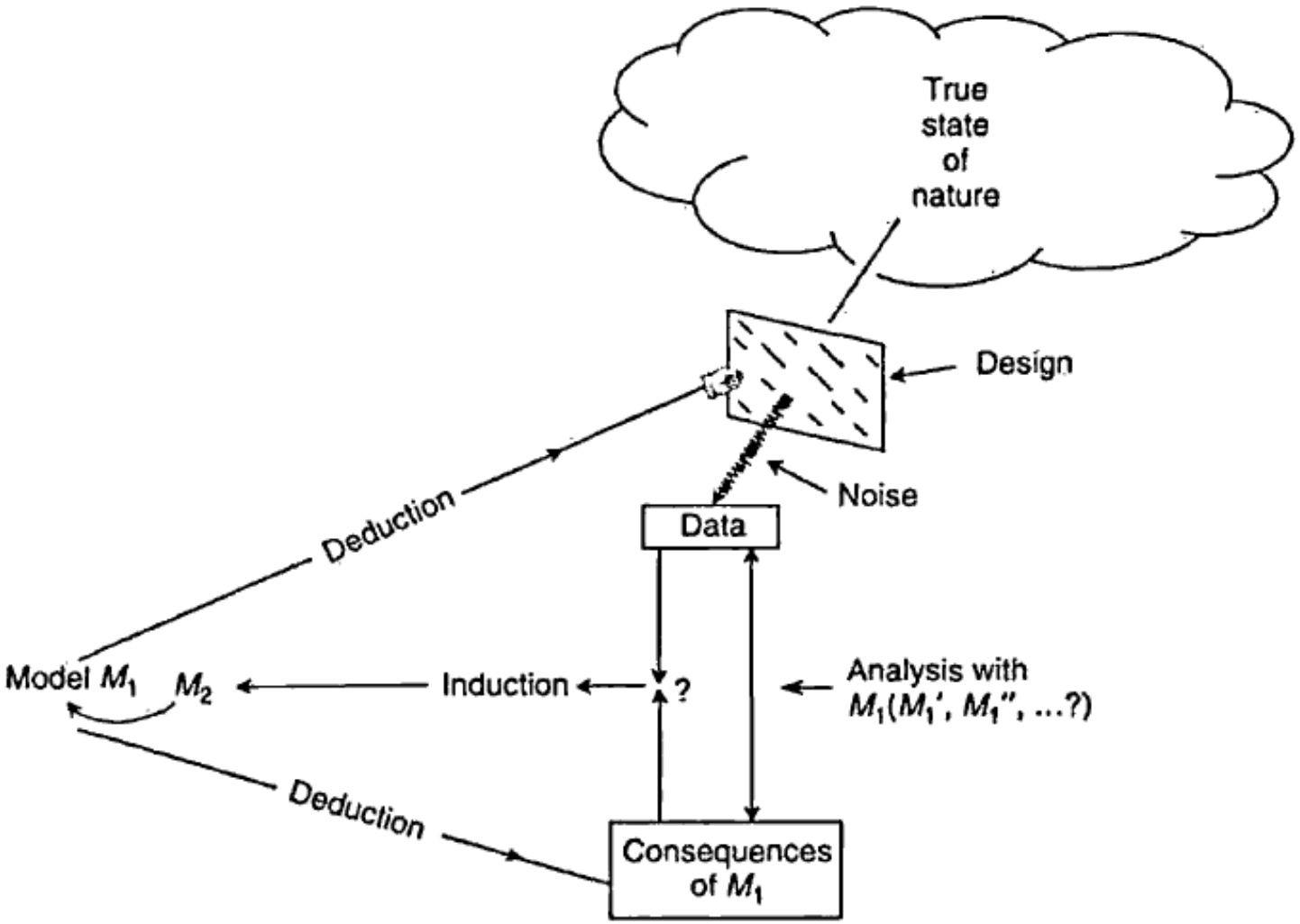
- **Estimation**
  - **e.g., Estimate the population mean weight using the sample mean weight**
- **Hypothesis testing**
  - **e.g., Test the claim that the population mean weight is 70 kg**

Inference is the process of drawing conclusions or making decisions about a population based on sample results

Basic Terminology

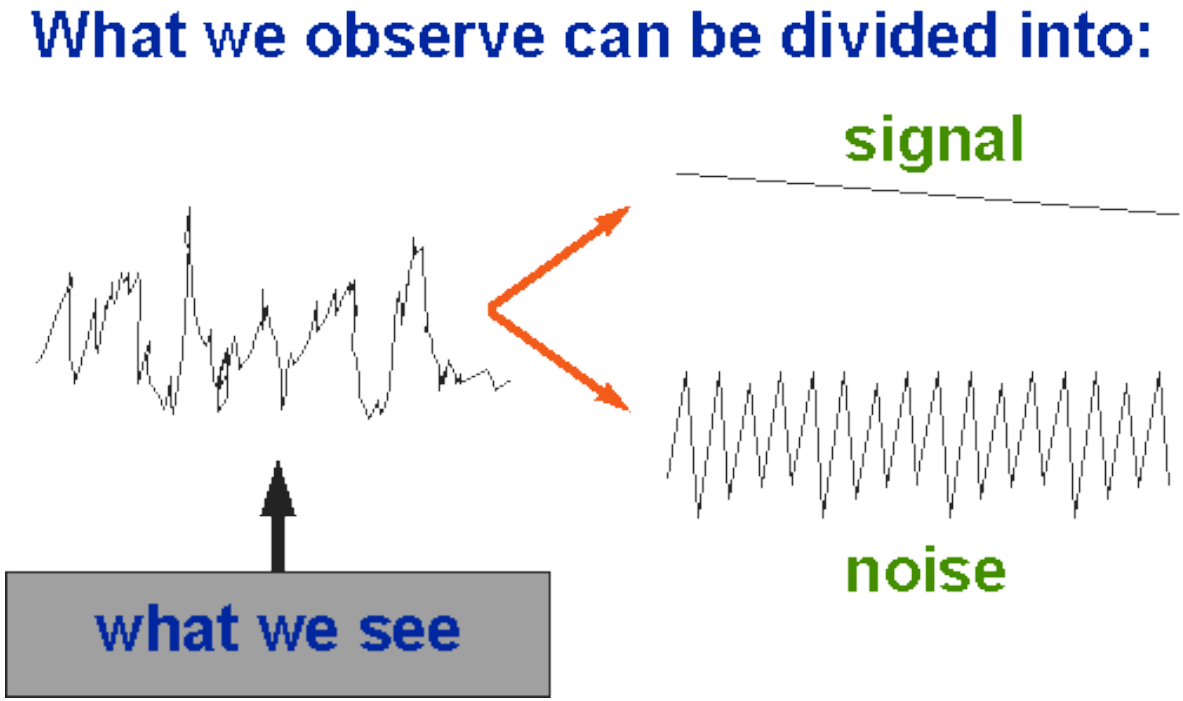
- **Summarizing versus Analyzing**
- **Descriptive Statistics**
- **Inferential Statistics**
  - **Inference from sample to population**
  - **Inference from statistic to parameter**
  - **Factors influencing the accuracy of a sample's ability to represent a population:**
    - **Size**
    - **Randomness**

The tyranny of the real world



“Noise”

- **Effects of things outside our control**

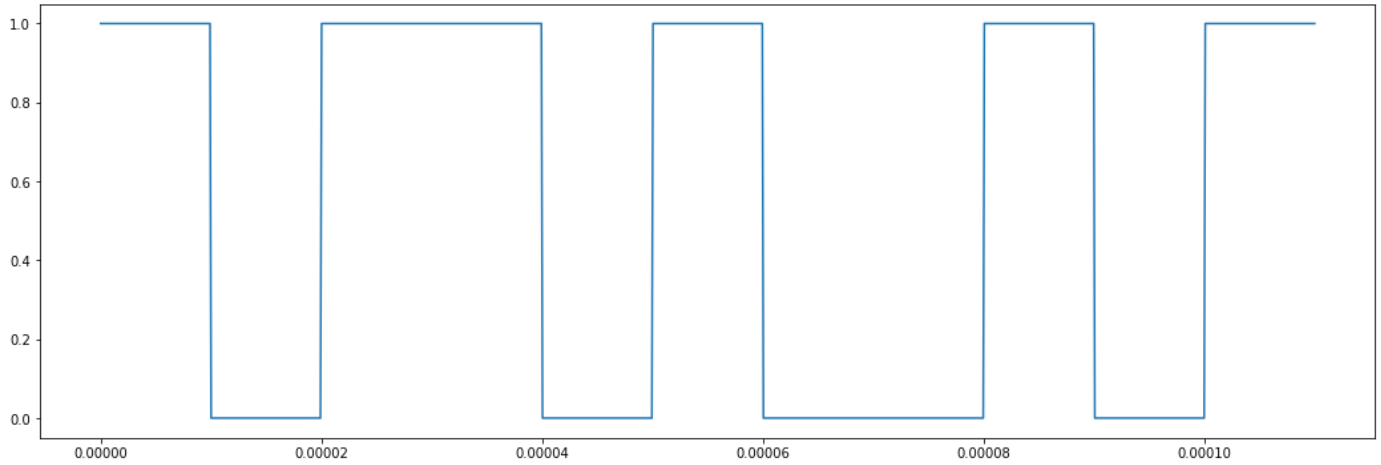


An example: Amplitude Shift Keying

- **ASK is routinely used to encode data in communication**
  - **A more complex version (QAM) is used in cellular and WiFi**

```
bits = np.array([1, 0, 1, 1, 0, 1, 0, 0, 1, 0, 1])
```

  - **Baseband signal (i.e., the direct encoding of the digital information)**

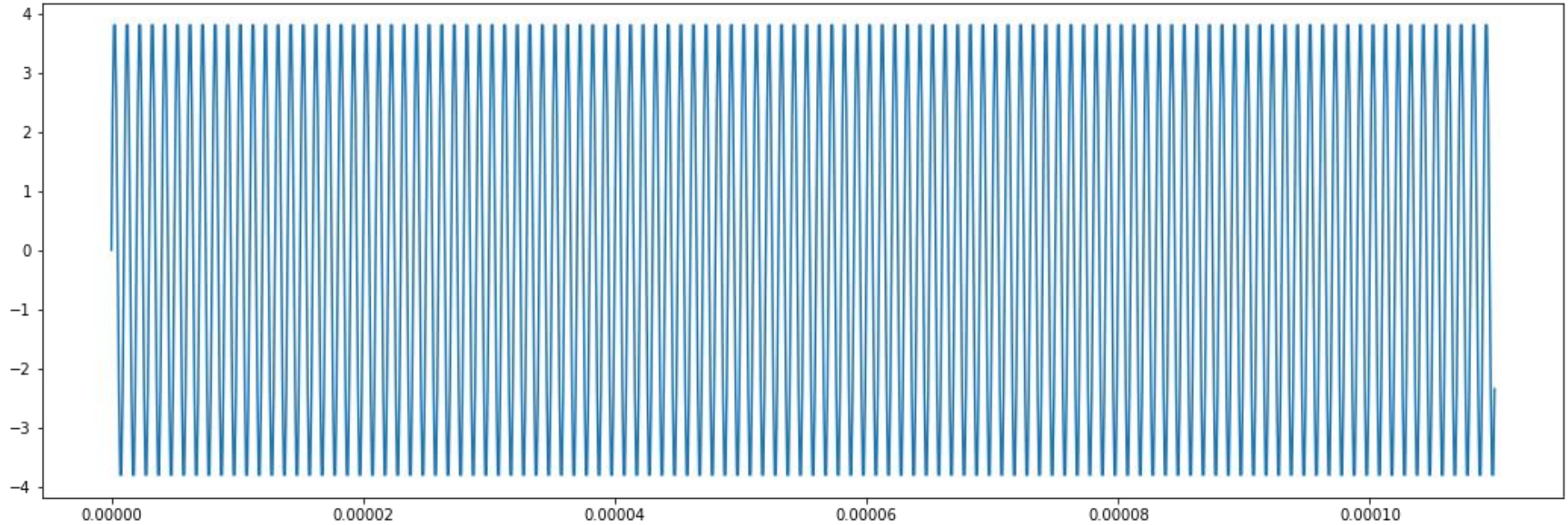


- **What are some possible sources of error in decoding the baseband signal?**

Carrier Signal

- **To transmit, we modulate onto a carrier at a specific frequency**

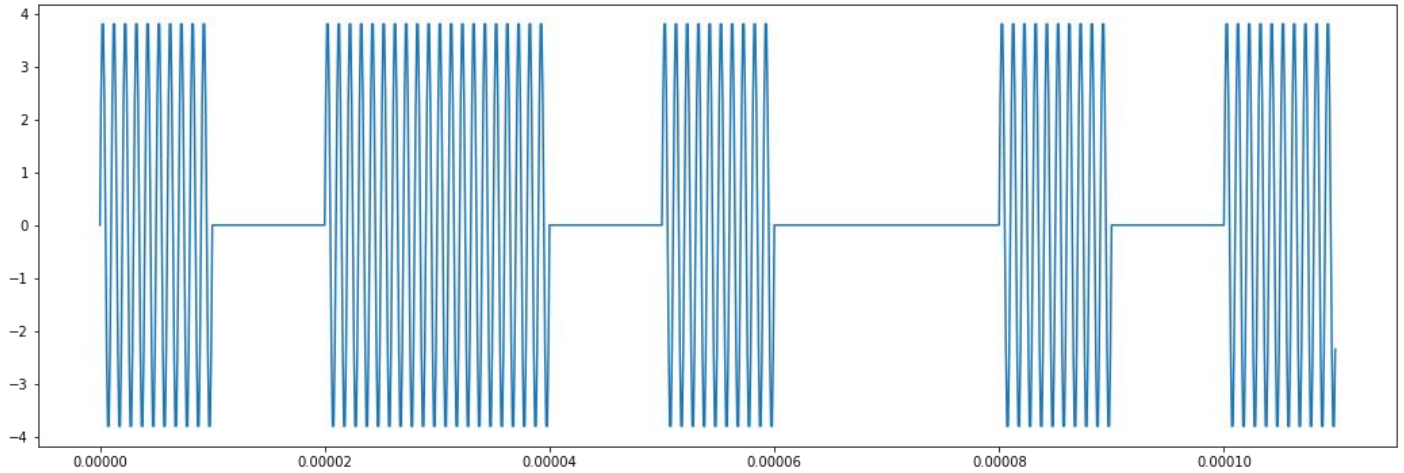
```
Ac - Amplitude of the carrier signal
fc - Frequency of the carrier signal
tc - Time variable of the carrier the signal
xc - Carrier signal row vector
xc = Ac * np.sin(2.0 * np.pi * fc * tc)
plot_signal(tc, xc)
```



ASK Modulation

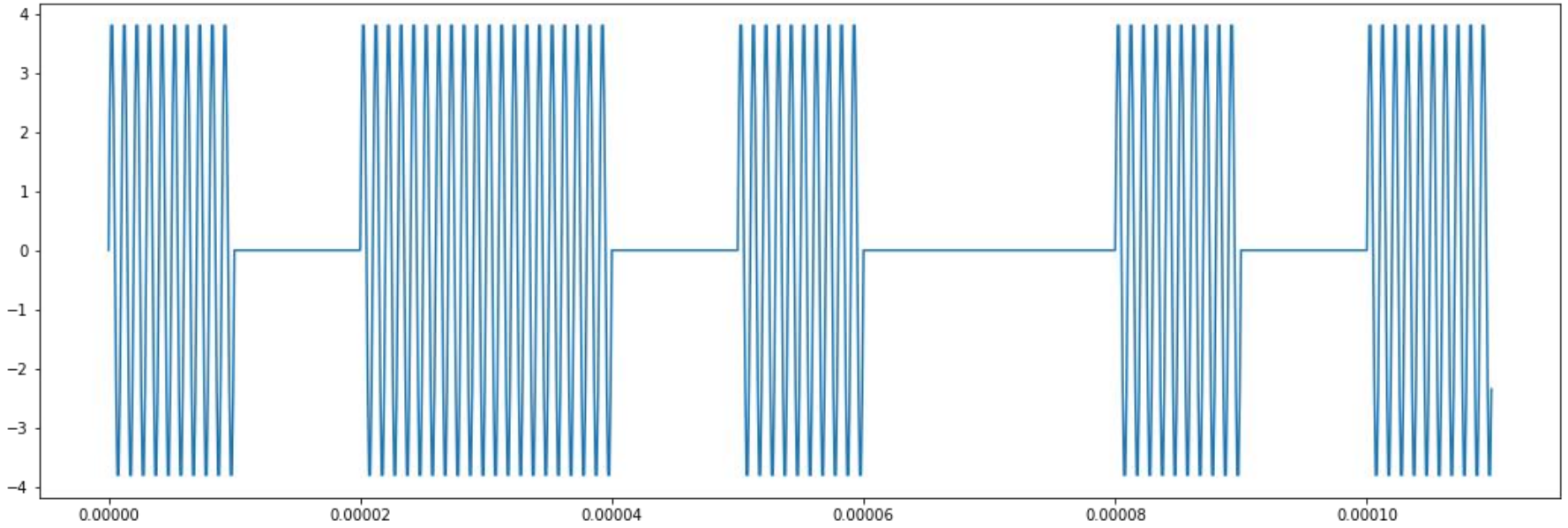
- **Simple:**

```
modulated_signal = x * xc
plot_signal(tc, modulated_signal)
```

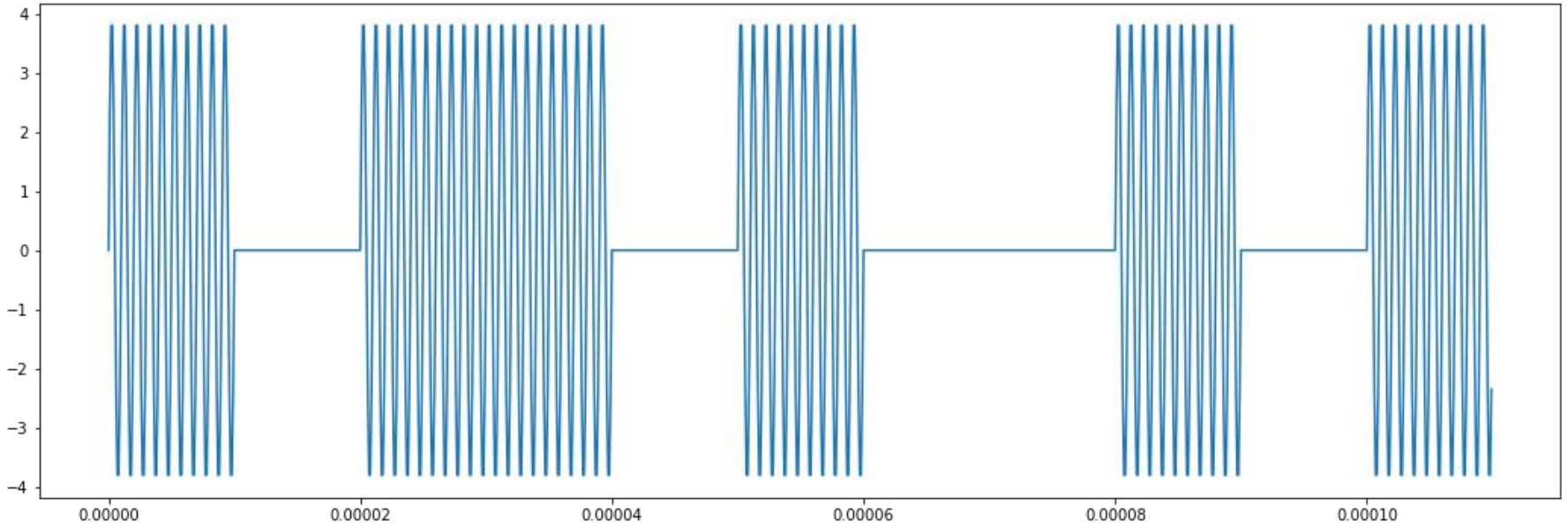


- **What are some possible sources of error?**

Discussion – Effect of noise around zero



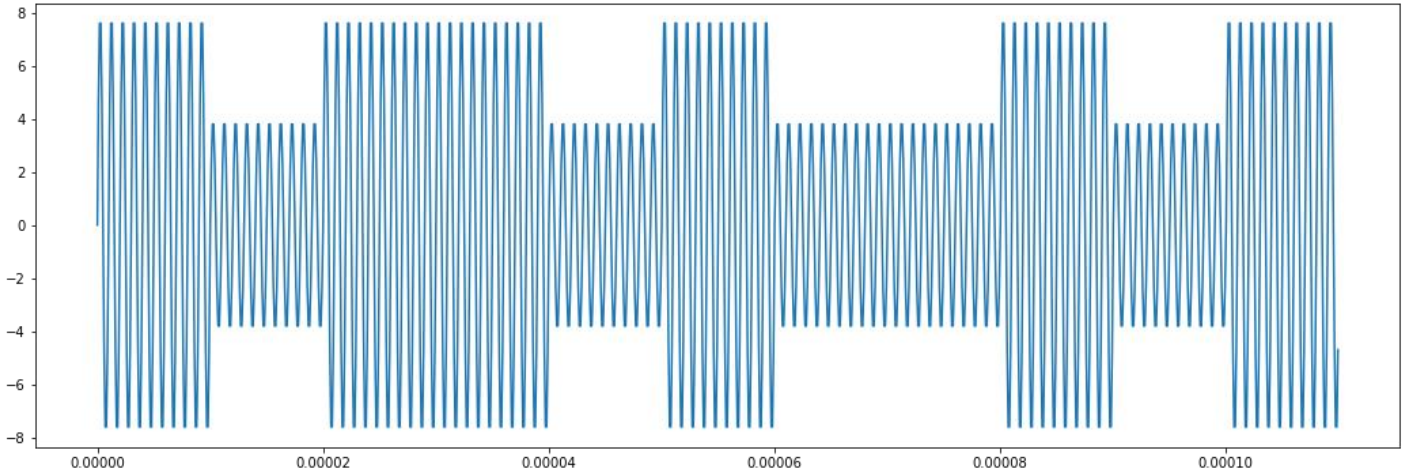
Discussion – Effect of attenuation



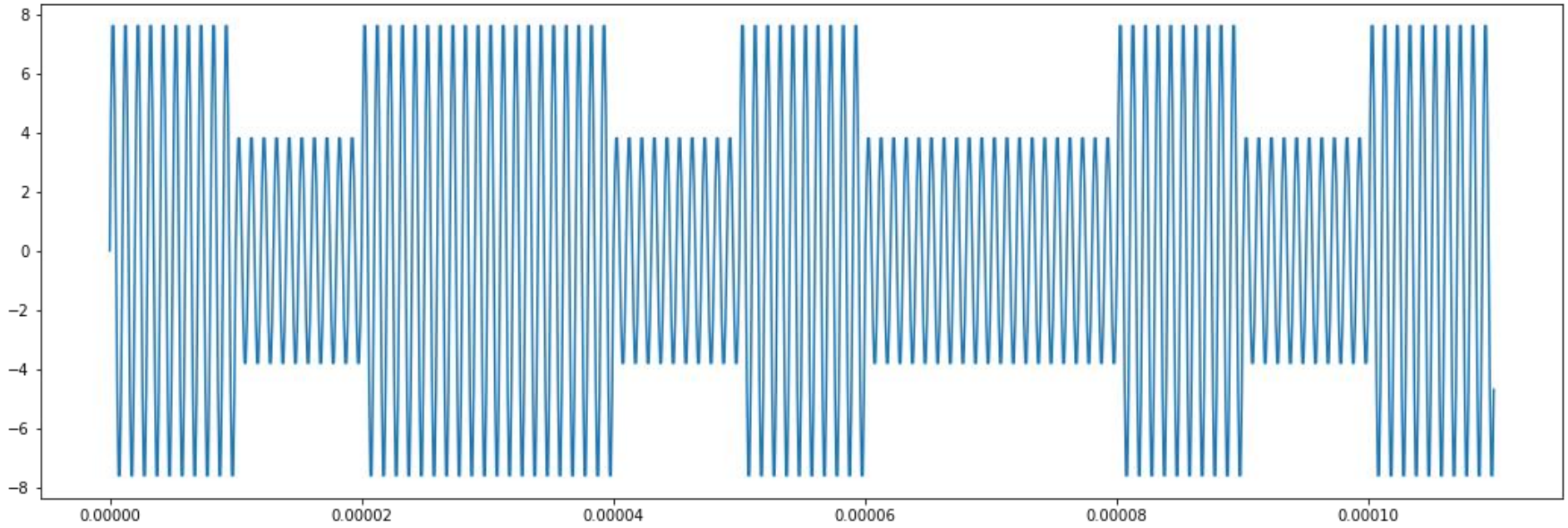
ASK Modulation

- **2 level:**

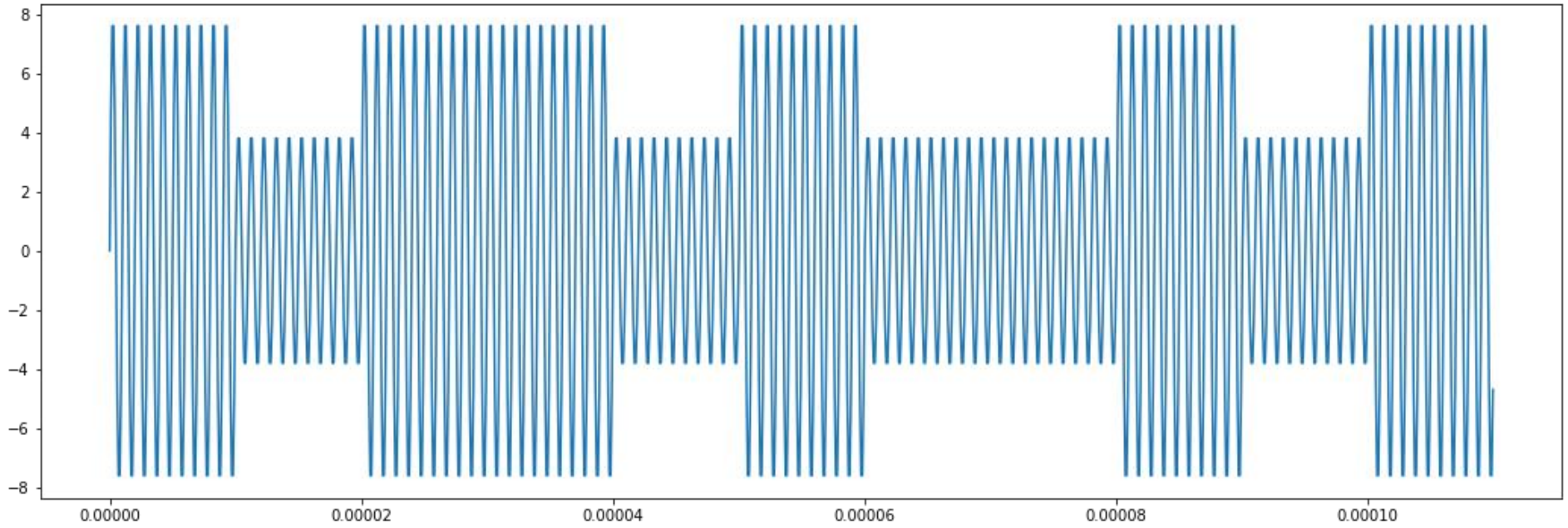
```
A1 - Amplitude of the carrier signal for 1
A2 - Amplitude of the carrier signal for 0
A1 = 8
A2 = 4
for i in range(N):
  if bits[i] == 1:
    sig = A1 * np.sin(2.0 * np.pi * fc * tc)
    ask_modulated_signal = np.append(ask_modulated_signal, sig)
  else:
    sig = A2 * np.sin(2.0 * np.pi * fc * tc)
    ask_modulated_signal = np.append(ask_modulated_signal, sig)
```



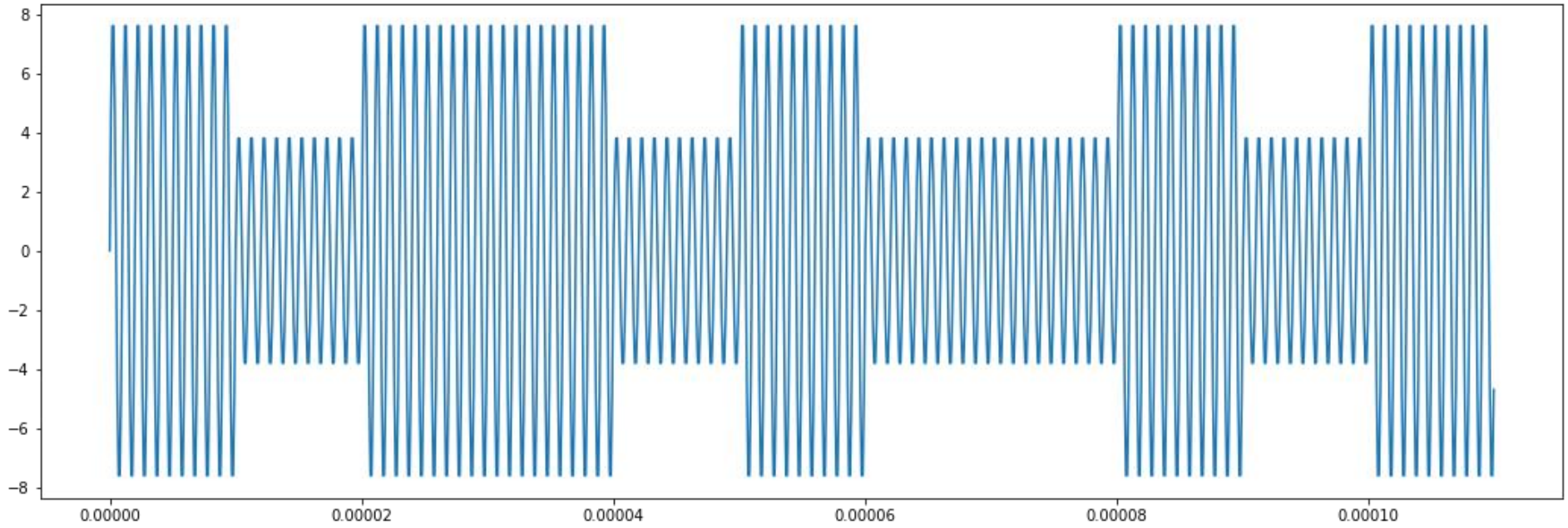
Discussion – Effect of attenuation



Discussion – Probability



Discussion – Where to draw the threshold? Conditional Probability



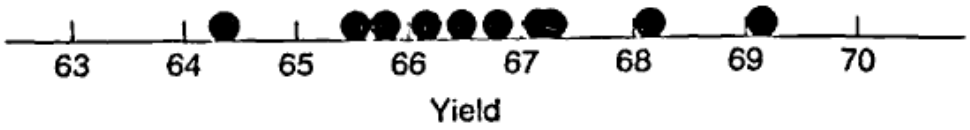
ANALYZING POPULATION DATA

Replicates

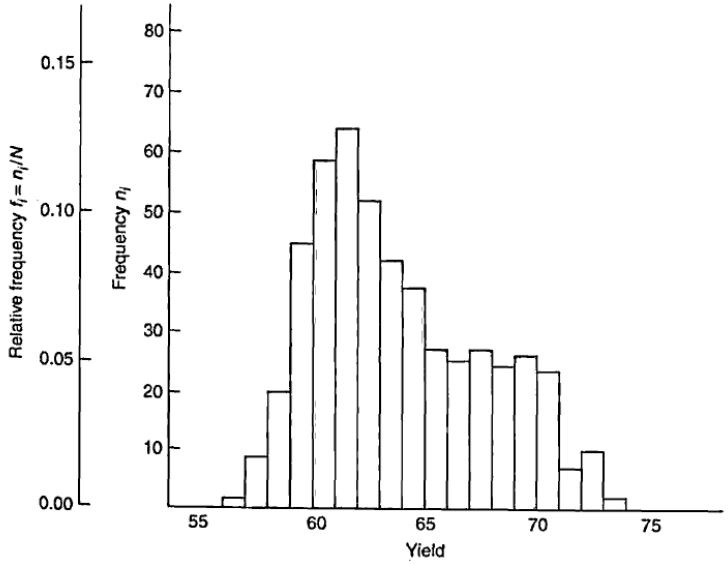
- **Replicates are collections of data that is expected to be nominally identical.**
- **Example:**

66.7 64.3 67.1 66.1 65.5 69.1 67.2 68.1 65.7 66.4

- **These may be visualized in various ways:**
  - **Dot diagram**

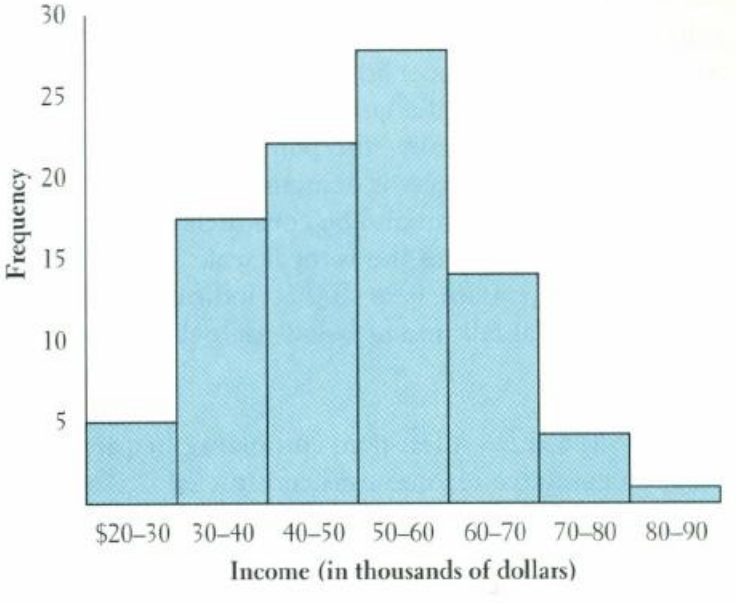


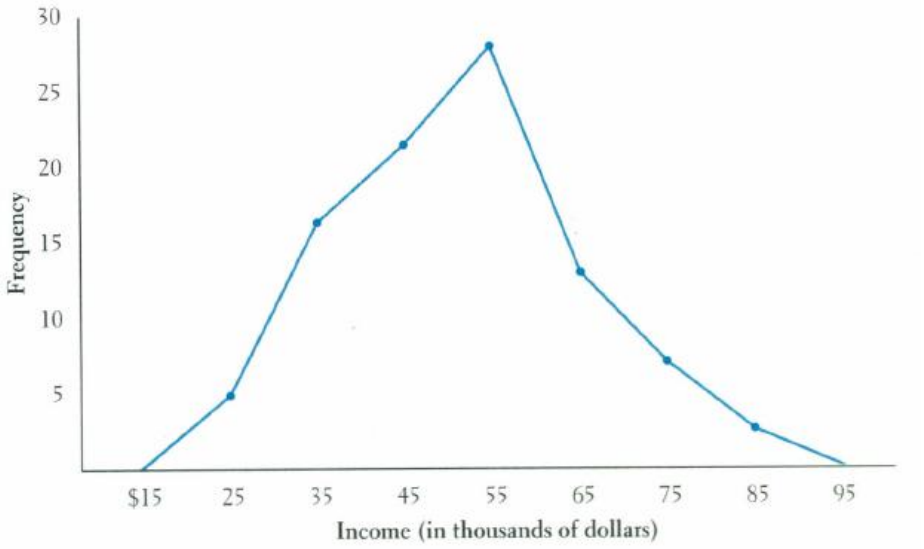
  - **Frequency distribution (e.g. for 500 observations)**



Histogram

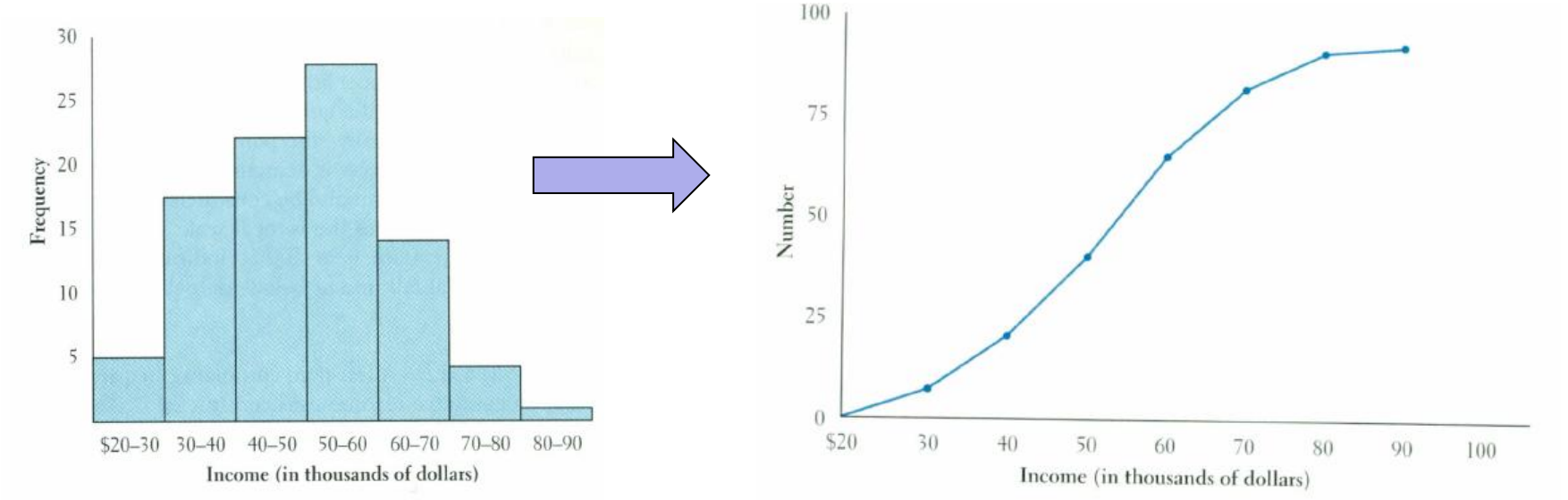
- **The adjacent bars indicate that a numerical range is being summarized by indicating the frequencies in arbitrarily chosen classes**
- **Also sometimes displayed as a “frequency polygon”**





Ogive or cumulative frequency plot

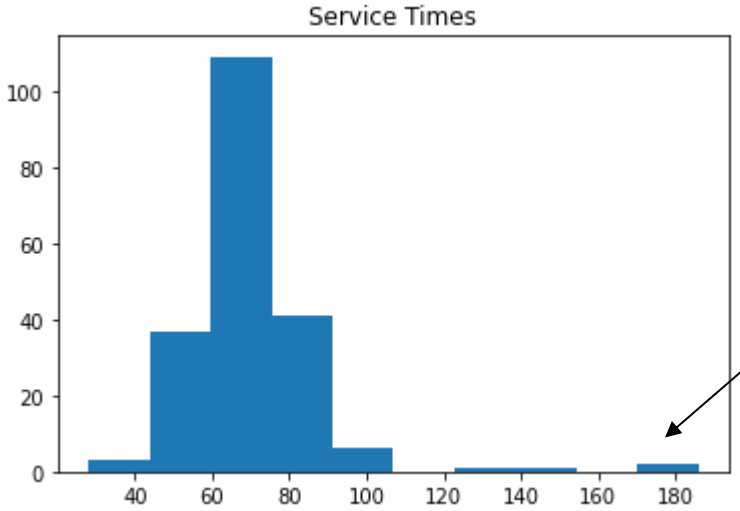
- **Lists cumulative population at or below the x-axis value**



Example: Times on phone waiting for service

- **Definitions**

```
data = np.array([45,62,52,72,91,88,64,65,69,59,70,…])
data = data.reshape(len(data), 1) # Prepare the data for
pandas
df = pd.DataFrame(data=data, columns=['Service Times']) #
Transform data to pandas DataFrame
print('Data: \n', df)
df.hist(['Service Times'], grid=False)
plt.show()
```



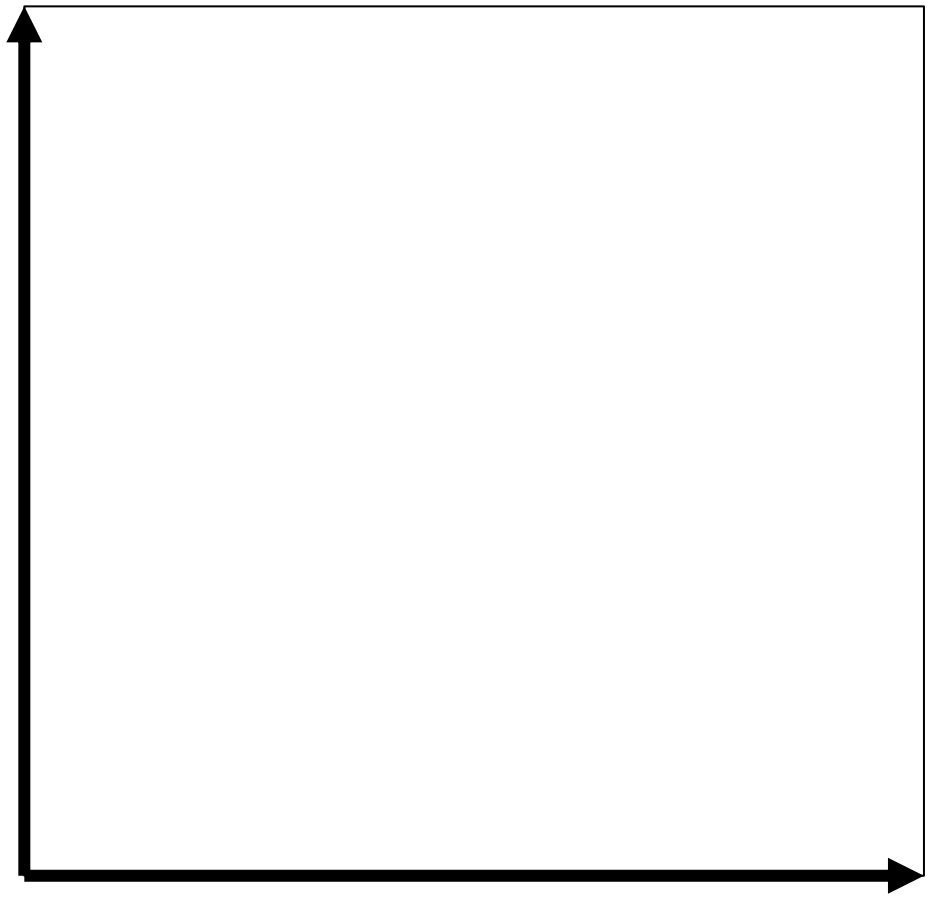
What are these?

Baking the best Pizza

- Pizza tastes best when baked at a very high temperature (e.g., 375°C)
- However, variations and dependencies (e.g. on dough uniformity) get larger at high temperature
- Consider two options (275°C and 375°C). Come up with a histogram and a cumulative distribution plot for the taste score for each condition. Assume 20 taste scores per condition

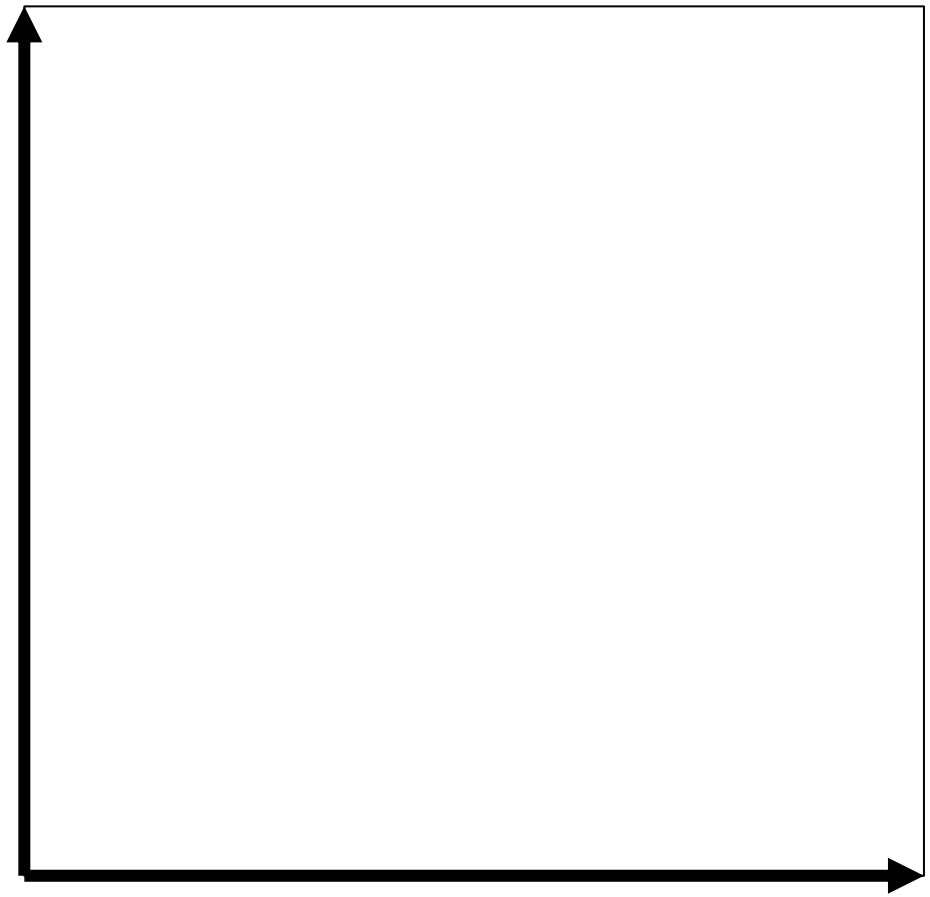
Pizza statistics - Histogram

Score	Frequency @ 275°C	Frequency @ 375°C
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		



Pizza statistics – Ogive

Score	Frequency @ 275°C	Frequency @ 375°C
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

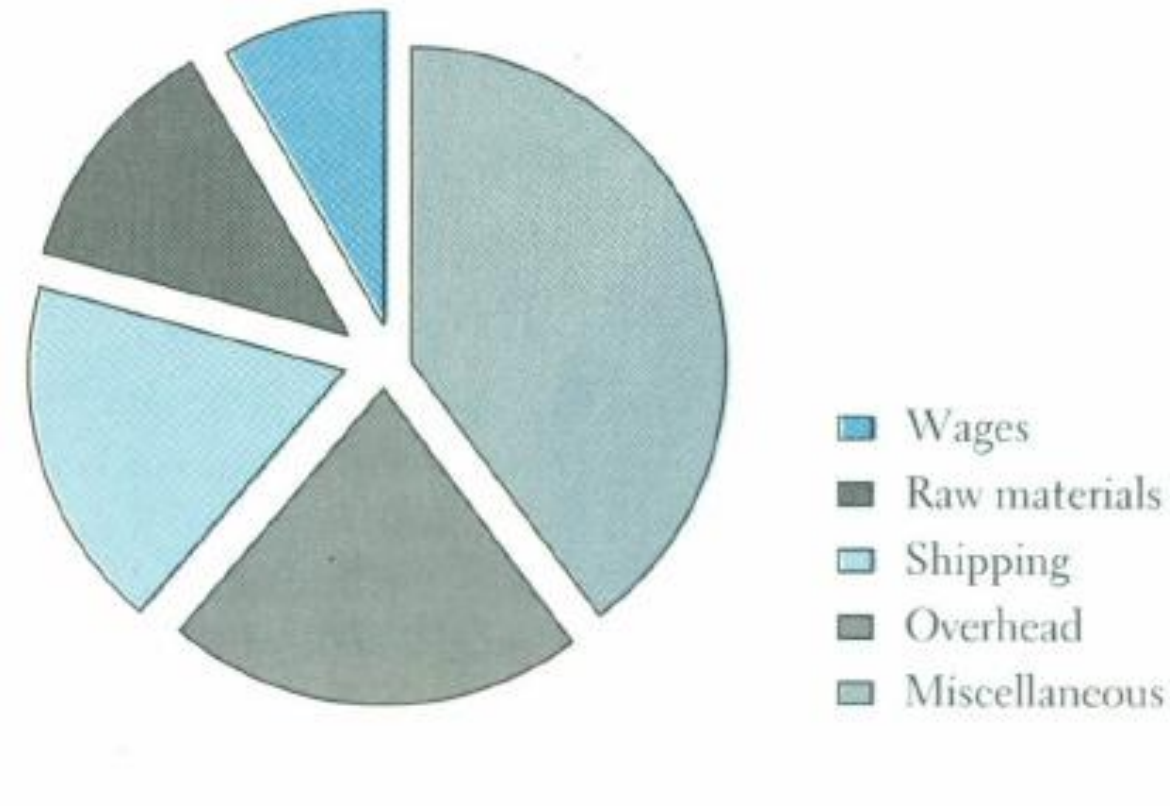


Questions to consider

- What are the advantages and disadvantages of each temperature?
- Which one would you pick for your store, and why?
- Which one would you pick if baking at home, and why?

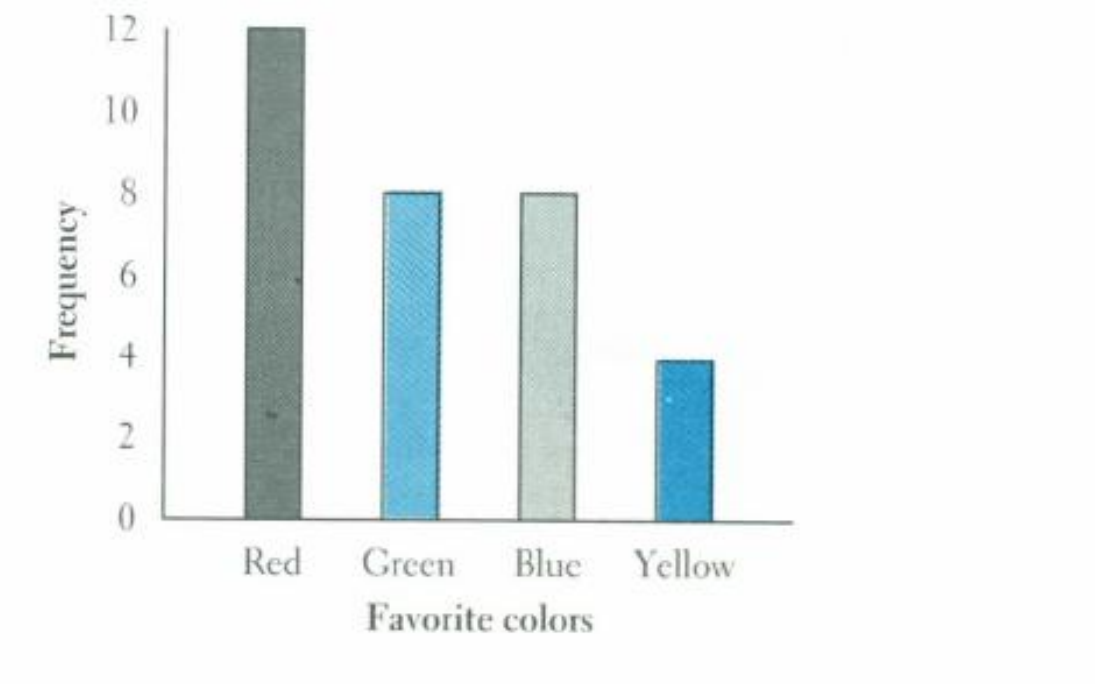
Pie Chart

- **The pie chart is an effective way of displaying the percentage breakdown of data by category.**
- **Useful if the relative sizes of the data components are to be emphasized**



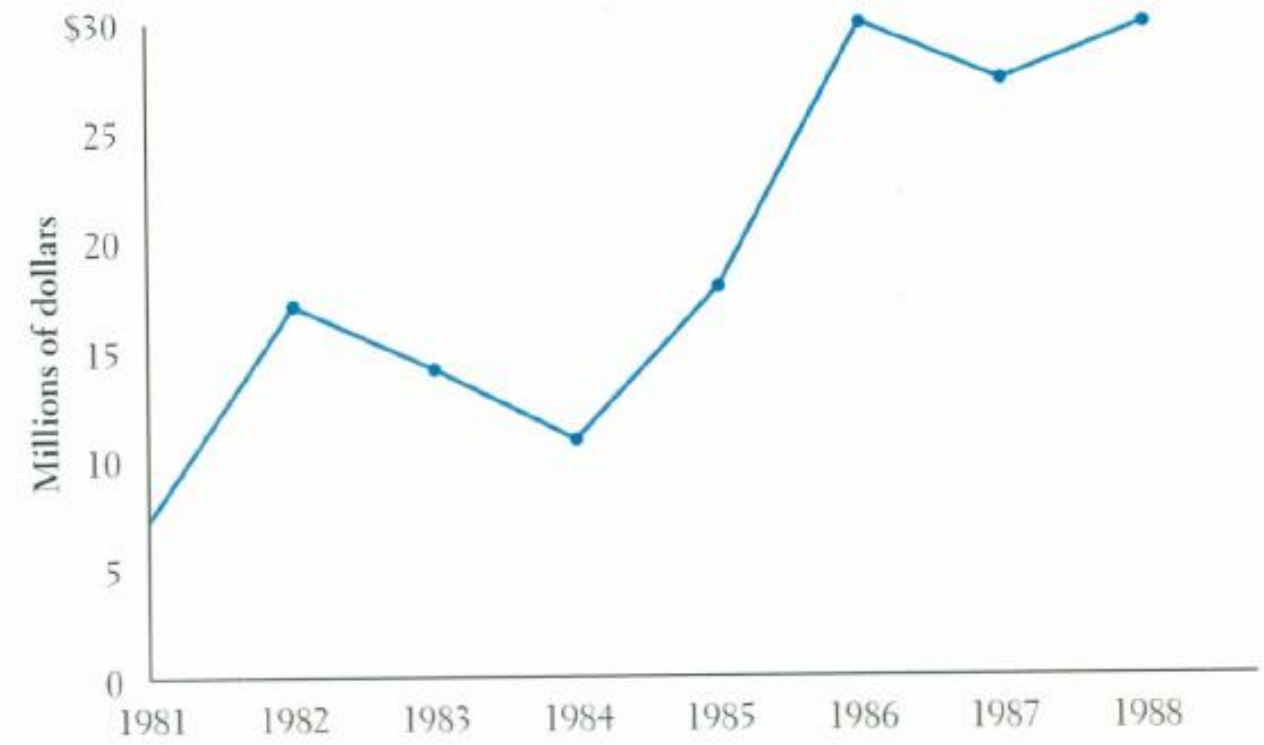
Bar chart

- **Another common method for graphically presenting nominal and ordinal scaled data**
- **The bars are separated, and this is why such a graph is frequently used for nominal and ordinal data – the separation emphasize the plotting of frequencies for distinct categories**



Time Series Graph

- **The time series graph is a graph of data that have been measured over time.**
- **The horizontal axis of this graph represents time periods and the vertical axis shows the numerical values corresponding to these time periods**



Measures of Central Tendency

- **These measures tap into the average distribution of a set of scores or values in the data.**
  - **Mean**
  - **Median**
  - **Mode**

The Mean

- The “mean” of some data is the average score or value, such as the average age of a student or average weight of professors that like to eat donuts.

- Inferential mean of a sample: $\bar{y}=(\sum y)/n$
- Mean of a population: $\eta=(\sum y)/N$

- The main problem associated with the mean value of some data is that it is sensitive to outliers.
  - Example, the average weight of political science professors might be affected if there was one in the department that weighed 600 pounds.

Donut-Eating Professors – the problem with outliers

Professor	Weight		Weight
Schmuggles	165		165
Bopsey	213		213
Pallitto	189		410
Homer	187		610
Schnickerson	165		165
Levin	148		148
Honkey-Doorey	251		251
Zingers	308		308
Boehmer	151		151
Queenie	132		132
Googles-Boop	199		199
Calzone	227		227
	194.6		248.3

Question: How can I reduce the impact of outliers?

The Median

- Because the mean average can be sensitive to extreme values, the median is sometimes useful and more accurate.
- The median is simply the middle value among some scores of a variable.

Professor	Weight
Schmuggles	165
Bopsey	213
Pallitto	189
Homer	187
Schnickerson	165
Levin	148
Honkey-Doorey	251
Zingers	308
Boehmer	151
Queenie	132
Googles-Boop	199
Calzone	227
	194.6

Rank order and choose middle value.

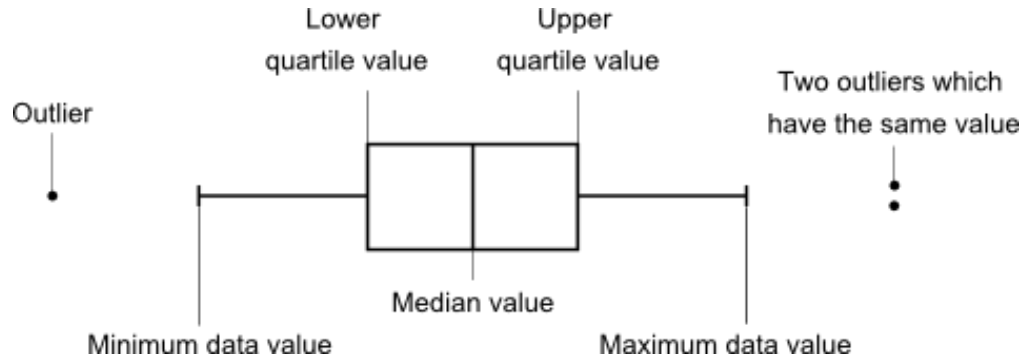
If even then average between two in the middle

Weight
132
148
151
165
165
187
189
199
213
227
251
308

- If we know the median, then we can go up or down and rank the data as being above or below certain thresholds.
- You may be familiar with standardized tests. 90th percentile, your score was higher than 90% of the rest of the sample

Quartiles and box plots

- **We can extend our concept of the median (50% point) and develop upper and lower quartiles. These can then be used to generate a box plot**

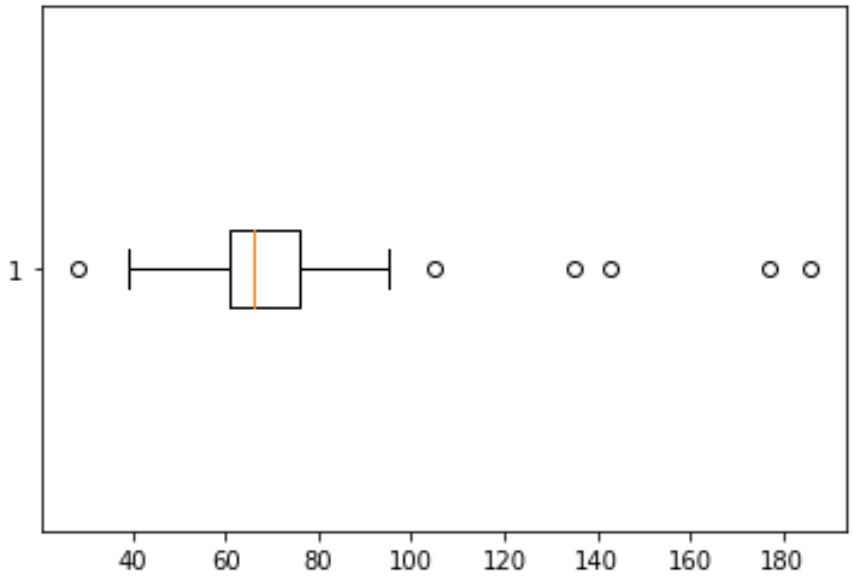


- **There is no rigorous mathematical definition for what exactly is or isn't an outlier, however there are a few tests and criterions that can be applied. These include Chauvernet's criterion, Peirce's criterion, Grubb's test for outliers and Dixon's Q-test.**
  - **Interquartile range: IQR = Q3 – Q1**
  - **Lower outlier(s) < Q1−(1.5×IQR)**
  - **Upper outlier(s) > Q3+(1.5×IQR)**

Box plots

```
Q1 = df.quantile(0.25)
Q3 = df.quantile(0.75)
IQR = Q3 - Q1
#print(IQR)
# We print the outliers here
mask = ((df < (Q1 - 1.5 * IQR)) |(df > (Q3 + 1.5 * IQR)))
filtered_data = (mask*df).to_numpy()
print("Outliers:")
for datum in filtered_data:
    if(datum != 0):
        print(datum)
plt.boxplot(df, vert=False)
plt.show()
```

Outliers:
[177]
[135]
[143]
[186]
[28]
[105]



The Mode

- The most frequent response or value for a variable.
- Multiple modes are possible: bimodal or multimodal.

Professor	Weight
Schmuggles	165
Bopsey	213
Pallitto	189
Homer	187
Schnickerson	165
Levin	148
Honkey-Doorey	251
Zingers	308
Boehmer	151
Queenie	132
Googles-Boop	199
Calzone	227

What is the mode?

Answer: 165

Important descriptive information that may help inform your research and diagnose problems like lack of variability.

Analysis of service time data example

```
print("Sample mean: ", mean(df))
```

Sample mean: Service Times   69.345

```
print("Sample median: ", median(df))
```

Sample median: 66.0

```
r = 0.05 # r value for trimming
print("Sample trimmed mean: ", stats.trim_mean(df, r))
```

Sample trimmed mean: [67.88333333]

```
print("Sample mode: ", stats.mode(df))
```

Sample mode: ModeResult(mode=array([61]), count=array([13]))

Population vs Sample Averages

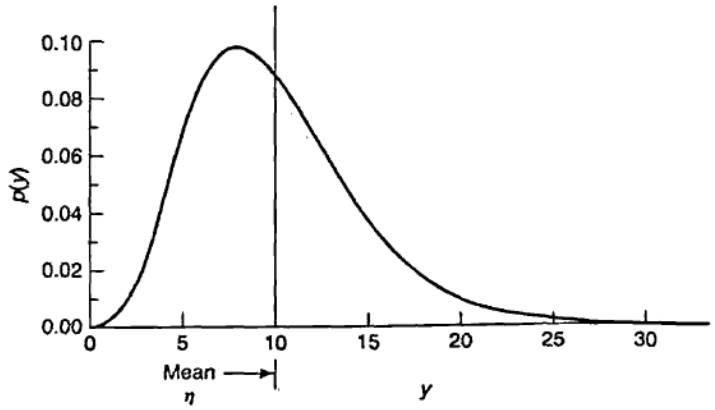
- **Population mean (eta), indicated by:**

$$\eta = \frac{\sum y}{N}$$

- **Sample mean (y-bar), indicated by:**

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n} = \frac{\sum y}{n}$$

- **Both represent the "point of balance" of a distribution**



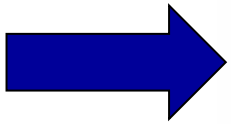
The mean of the population is also called the *expected value* of y or the *mathematical expectation* of y and is often denoted as $E(y)$. Thus $\eta = E(y)$.

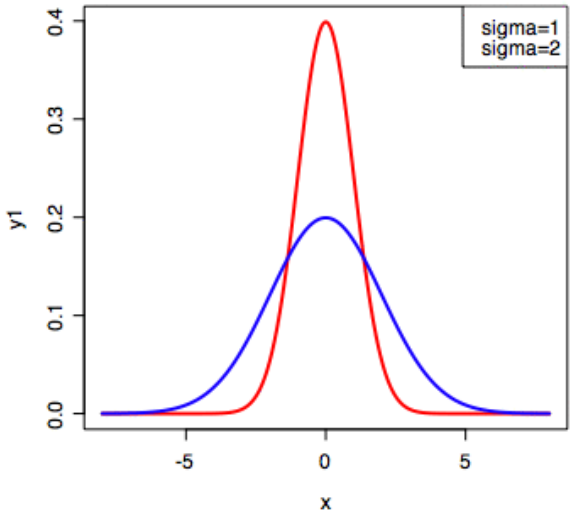
Measures of Dispersion

- **Measures of dispersion tell us about variability in the data.**
- **Basic question: how much do values differ for a variable from the min to max, and distance among scores in between. We use:**
  - **Range**
  - **Standard Deviation**
  - **Variance**

Reminder: Why dispersion is important

- From metal to ultra-precise silicon





Why is this beneficial?

The Range

- **r = h – l**
  - **Where h is high and l is low**

- **In other words, the range gives us the value between the minimum and maximum values of a variable.**

- **Understanding this statistic is important in understanding your data, especially for management and diagnostic purposes.**

- **Other ranges include IQR (Inter-Quartile Range)**

```
print("Upper sample quartile: ", stats.mstats.mquantiles(df, prob=[0.75]))
print("Lower sample quartile: ", stats.mstats.mquantiles(df, prob=[0.25]))
print("Interquantile range: ", (stats.mstats.mquantiles(df, prob=[0.75]) -
stats.mstats.mquantiles(df, prob=[0.25])))
```

Upper sample quartile: [76.]
Lower sample quartile: [61.]
Interquantile range: [15.]

The Standard Deviation and Variance

- **A standardized measure of distance from the mean.**

$$S = \sqrt{\frac{\sum(X - \overline{X})^2}{(n-1)}}$$

$\sqrt{\ }$ =square root
$\sum$=sum (sigma)
X=score for each point in data

$\overline{X}$=mean of scores for the variable
n=sample size (number of observations or cases

- **Variance is just S^2**

$$S^2 = \frac{\sum(X - \overline{X})^2}{(n-1)}$$

Analysis of service time data example

```
print("Sample variance: ", stats.tstd(df)**2)
print("Sample standard deviation: ", stats.tstd(df))
```

Sample variance: [309.31253769]
Sample standard deviation: [17.58728341]

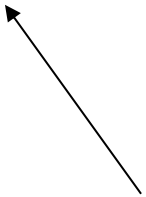
Sample vs population variance and standard deviations

- **Population:**

$$\sigma^2 = E(y-\eta)^2 = \frac{\sum(y-\eta)^2}{N}$$

- **Sample:**

$$s^2 = \frac{\sum(y-\bar{y})^2}{n-1}$$



The reason for this is explained next

Residuals and degrees of freedom

- The residuals are the deviations of the individual data points from the mean, i.e.:

$$y_1 - \bar{y}, y_2 - \bar{y}, \ldots . y_n - \bar{y}$$

- It necessarily follows that the sum of the residuals is always 0

$$\sum (y - \bar{y}) = 0$$

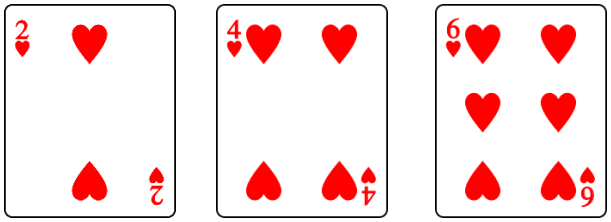
- Note that for any set of n residuals, there are only n-1 independent constraints (since the mean plus the n-1 residuals fully define the data set)

- This is called the number of degrees of freedom, i.e. $\upsilon = n - 1$

- This is why we use n-1 in the sample standard deviation. Otherwise, we would effectively "double-count" one piece of data.

An example to show this (not a proof!)

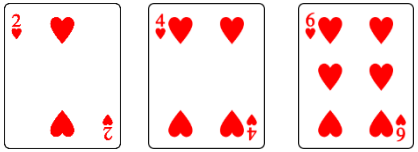
- **Consider a population of 3 cards**

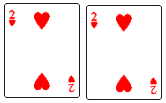
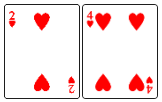
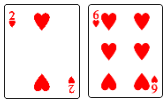
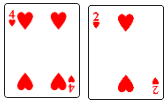
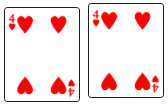
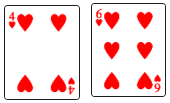
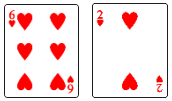
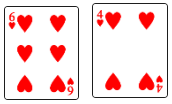
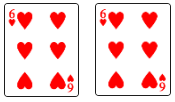


- **Population mean:** $\eta = \frac{2+4+6}{3} = \frac{12}{3} = 4$
- **Population variance:** $\sigma^2 = \frac{(2-4)^2+(4-4)^2+(6-4)^2}{3} = \frac{8}{3}$

- **Suppose we were to take many repeated samples of 2 cards from this population. For an accurate sample variance, we would expect:** $E(s^2) = \sigma^2$
  - **i.e., the sample variance should be "unbiased", i.e., in a complete set of random samples, it should match the population**

An example to show this (not a proof!)

- Population 
- 9 possible samples of 2 cards

Sample	$\bar{y}$	s^2 (unbiased, i.e., n-1)
	2	0
	3	2
	4	8
	3	2
	4	0
	5	2
	4	8
	5	2
	6	0

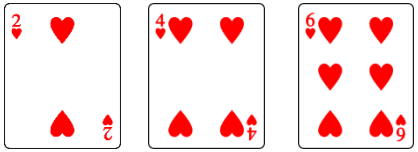
$$E(\bar{y}) = \frac{2+3+4+3+4+5+4+5+6}{9} = 4$$

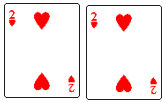
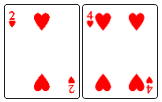
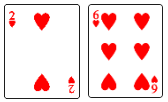
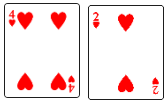
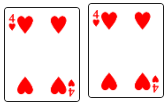
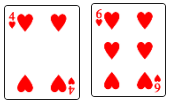
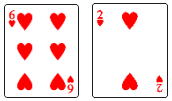
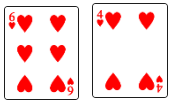
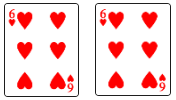
i.e., $E(\bar{y}) = \eta$

$$E(s^2) = \frac{0+2+8+2+0+2+8+2+0}{9} = \frac{8}{3}$$

i.e., $E(s^2) = \sigma^2$

An example to show this (not a proof!)

- Population 
- 9 possible samples of 2 cards

Sample	$\bar{y}$	s^2 (biased, i.e., n)
	2	0
	3	1
	4	4
	3	1
	4	0
	5	1
	4	4
	5	1
	6	0

$$E(\bar{y}) = \frac{2+3+4+3+4+5+4+5+6}{9} = 4$$

i.e., $E(\bar{y}) = \eta$

$$E(s^2) = \frac{0+1+4+1+0+1+4+1+0}{9} = \frac{4}{3}$$

i.e., $E(s^2) \neq \sigma^2$

A proof

$$
\begin{aligned}
\mathrm{E}\left[\sigma^2 - s^2_{\text{biased}}\right] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\right] \\
&= \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n}\left((x_i^2 - 2x_i\mu + \mu^2) - (x_i^2 - 2x_i\overline{x} + \overline{x}^2)\right)\right] \\
&= \mathrm{E}\left[\mu^2 - \overline{x}^2 + \frac{1}{n}\sum_{i=1}^{n}(2x_i(\overline{x}-\mu))\right] \\
&= \mathrm{E}\left[\mu^2 - \overline{x}^2 + 2(\overline{x}-\mu)\overline{x}\right] \\
&= \mathrm{E}\left[\mu^2 - 2\overline{x}\mu + \overline{x}^2\right] \\
&= \mathrm{E}\left[(\overline{x}-\mu)^2\right] \\
&= \mathrm{Var}(\overline{x}) \\
&= \frac{\sigma^2}{n}
\end{aligned}
$$

This is true for truly random samples from an uncorrelated population… you'll learn more about this soon

So, the expected value of the biased estimator will be

$$\mathrm{E}\left[s^2_{\text{biased}}\right] = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2$$

So, an unbiased estimator should be given by

$$s^2_{\text{unbiased}} = \frac{n}{n-1}s^2_{\text{biased}}$$