PHYS 267 - Probability, Statistics, and Data Analysis for Physics and Astronomy

Course Notes

Made by Richard Dong

Overview of Key Python Functions

`random`

- `random.sample(listof values, number of values)`

`numpy`, import as `np`

- allows us to work with vectors and arrays
- Key Functions:
    - `np.loadtxt("file_name", skiprows=n, usecols(n,m)/usecols=n, max_rows=n)`
        - Allow to load text file into the program and generate data array.
        - skiprows allow us to pick the row number for import
        - usecols allow us to pick the column number for import
        - max_rows allow us to import a certain number of rows
    - `np.array(L)`
        - convert a list L into array, which allows you to manipulate all the values in the array at one without using loops
    - `np.arange(start,end,number of data)`
    - `np.random.seed(n)`
        - fix the result from one trial
    - `np.random.randomn(number of data, sets of data)`
        - randomly generated some normally distributed data
    - `parameters = np.polyfit(x_values,y_values,degree)`
        - for a nth degree model, you will need to set (n+1) parameters
    - `np.linespace(start,end,number of data points)`
    - `np.mean(L)`
    - `np.var(L, ddof=)` , where set ddof=1 for sample and ddof=0 for population
    - `np.std(L, ddof=)` , where set ddof=1 for sample and ddof=0 for population
    - `np.median(L)`
    - `np.sin(n)`
    - `np.cos(n)`
    - `np.pi`

`matplotlib.pyplot`, import as `plt`

- allows us to plot and visualize data
- Object oriented approach:
    - `fig, ax = plt.subplots(figsize=(x,y))`
    - `ax.set_xlim(low,high)`
    - `ax.set_ylim(low,high)`
    - `ax.set_xlabel("x-axis name")`

- `ax.set_xlabel("x axis name")`
- `ax.set_ylabel("y axis name")`
- `ax.plot(x,y,marker=".",markerfacecolor='blue',linestyle="-",color="red", linewidth=,markersize=,label=)`
    - plot red line through blue dots -> line chart
    - marker size determines the size of points
- `ax.errorbar(x,y,yerr=,xerr=,fmt="o",color=,ecolor=,elinedith=,c`

  `lolims=True/False,uplims,xlolims=,xuplims=,label=)`
    - uplims sets the upper limit for y error (so there is only y errors going down)
- `ax.fill_between(x,yupper,ylower,alpha=0.3)`
- `ax.legend(loc="best")`
- `ax.set_title("title name")`

- Using plt to plot directly:
    - `plt.axis("axis name")`
        - for one axis plot only (such as a pie chart
    - `plt.title("title name")`
    - `plt.xlabel("x axis name")`
    - `plt.ylabel("x axis name")`
    - `plt.legend(labels=,title=,loc="best")`
        - generate a legend of the plt with labels and tile at the best location
    - `plt.show()`
- Pie Chart:
    - `plt.pie(data,explode=,labels=list/None,colors=,autopct='%1.xf%`
        - plot a pie chart
        - explot allows to offset any chosen slice
        - start the pie at any angle using startangle
        - label percentage to certain decimal points x using autopct
- Bar Chart:
    - `plt.bar(barclass,barfrequency,color="colour",width=,edgecolor="`
    - `plt.barh(barclass,barfrequency,color="colour",height=,edgecolor`
        - create a horizontal bar plot
- Histogram:
    - `ax.hist(data,bins=number,edgecolor="colour",color="colour",labe`

      `colours)`
    - Frequency histogram:
        - `hist,edges = np.histogram(data,edge values of bins)`
        - `relfreq = hist/float(hist.sum())`
        - `plt.bar(bins[:-1], relfreq, width=8, align='center', color='green')`
        - or, just use `density=True` in the `ax.hist` function

`pandas`, import as `pd`

- allows us to work with *DataFrames* in Python (think of data manipulation through spreadsheets)
- Key Functions:
    - `pd.DataFrame(listof(listof data),columns=(listof col))`
        - allows us to present the data in table with column names of columns

`scipy`

- allows us to access essential scientific algorithms, including ones for basic statistics
    - `from scipy import stats`
- Key Functions:
    - `stats.mode(L)`
        - print out the mode as well as its frequency

- the mode number will be **`stats.mode(L)[0][0][0]`**
  - Uniform Distribution:
    - **`data.stats.pmf(list of x)`**
      - generate the probability for each specific x-value in list of x
  - Binomial Distribution:
    - **`stats.binom.pmf(x,n,p)`** , where p is the probability of success
    - **`stats.binom.cdf(x,n,p)`** , this includes the edge points
    - **`stats.binom.mean(n,p)`**
    - **`stats.binom.var(n,p)`**
    - **`stats.binom.std(n,p)`**
  - Poisson Distribution:
    - **`stats.poisson.pmf(x,mu)`**
    - **`stats.poisson.cdf(x,mu)`** , this includes the edge points
    - **`stats.poisson.mean(mu)`**
    - **`stats.poisson.var(mu)`**
    - **`stats.poisson.std(mu)`**
  - Continuous Uniform:
    - **`stats.uniform.ppf(percent)`** : this is the percent point function and returns a standard deviation multiplier for what value the % occurs at
    - **`stats.uniform.pdf(value)`**
  - Gaussian:
    - **`stats.norm.ppf(percent)`** : this returns the z-score for the percent
    - **`stats.norm.pdf(z-score)`**
      - or **`stats.norm.pdf(value, loc=mu, scale=sigma)`**
    - **`stats.norm.cdf(z-score)`**
      - or **`stats.norm.cdf(value, loc=mu, scale=sigma)`**
    - **`stats.norm.sf(z-score)`**
      - this is 1-cdf()
  - Maxwell:
    - **`stats.maxwell.ppf(percent)`**
    - **`stats.maxwell.pdf(value)`**
  - Four Moments:
    - 1st mean: **`np.mean(data)`**
    - 2nd variance: **`np.var(data)`**
    - 3rd skew: **`stats.skew(data)`**
    - 4th kurtosis: **`stats.kurtosis(L)+3`**

`sklearn.metrics`

- In this course, we use it to calculate the R^2 value for the regression model
- **`from sklearn.metrics import r2_score`** , then use:
  - **`r2_score(actual_y, modelled_y)`**

Other Imports

- **`Statsmodels`** : integrates with NumPy, SciPy and Pandas to explore data, estimate statistical models and perform statistical tests
- **`Seaborn`** : allows us to visualize statistical data (distributions and gradient maps for example)
- **`Patsy`** : allows us to describe statistical models (ie. linear models)

Hypothesis Tests from scipy.stats

Always use `tests_statistics, p_values = scipy.stats.testname()`

	Sample	Measure	Hypothesis Test	Purpose & Conditions	Python Function
Parametric	One Sample Test	Mean	One Sample t-Test	**Purpose:** Check observed mean value of normally distributed data against theoretical reference value **Conditions:** Sample size is small, variance unknown	scipy.stats.ttest_1samp(a, popmean, axis=0, alternative='two-sided')
Parametric	One Sample Test	Mean	Z-Test	**Purpose:** Check observed mean value of normally distributed data against theoretical reference value **Conditions:** Sample size is large, variance known	N/A
Parametric	Two Sample Test	Correlation	Pearson Correlation Coefficient	**Purpose:** Measure linear correlation between two sets of data	N/A
Parametric	Two Sample Test	Mean	Two Group t-Test	**Purpose:** Compare two observed means from independent samples **Conditions:** Sample size is small, variance unknown	scipy.stats.ttest_ind(group1, group2)
Parametric	Two Sample Test	Mean	Paired t-Test	**Purpose:** Compare two observed means from paired, dependent samples **Conditions:** Sample size is small, variance unknown	scipy.stats.ttest_rel(group1,group2)
Parametric	Two Sample Test	Mean	Two Sample Z-Test	**Purpose:** Compare two observed means from independent samples **Conditions:** Sample size	N/A

				is large, variance known	
Non-Parametric	One Sample Test	Mean	One Sample Wilcoxon's Test	**Purpose:** Check observed mean value of normally distributed data against theoretical reference value	scipy.stats.wilcoxon(list of each data - checkValue)
Non-Parametric	One Sample Test	Randomness	Runs Test	**Purpose:** Determine how random your data is	N/A
Non-Parametric	One/Two Sample Test	Distribution	Kolmogorov-Smirnov Test	**Purpose:** Compare an observed distribution to a reference distribution **Conditions:** Data is continuous	N/A
Non-Parametric	One/Two Sample Test	Distribution	Chi Squared Test	**Purpose:** Compare an observed distribution to a reference distribution **Conditions:** Data is binned and represents frequencies	scipy.stats.chisquare(data)
Non-Parametric	Two Sample Test	Correlation	Spearman Rank Correlation	**Purpose:** Test the association between two samples	N/A
Non-Parametric	Two Sample Test	Mean	Mann-Whitney Test	**Purpose:** Compare two observed means from independent samples	scipy.stats.mannwhitneyu(group1, group2)
Non-Parametric	Two Sample Test	Mean	Wilcoxon's Test	**Purpose:** Compare two observed means from paired samples	N/A

In addition, we may also have **One-Way ANOVA**, where we look at whether there are differences between multiple independent groups when there is only one factor afftecting them. We can use `scipy.stats.f_oneway(group1, group2, group3)` .

Introduction to Data Analysis

Data Analysis: process of collecting, modeling, and analyzing data to extract insights and make predictions based on interpreted results

- Methods of data analysis require the application of mathematical statistics

Statistics: based on observations and how we **infer/interpret** such results

- Requires us to understand statistical tests
- Knowledge of probability and uncertainty is required to understand the significance of these statistical tests

Probability: the language of uncertainty that allows us to describe (numerically) how likely an event is to occur or that a proposition is true

- Prior to understanding probability, we have to understand how data is collected and how it is presented and described

Approaches to Data Analysis

1. **Experimental Design**: formulate hypothesis, design experiment and sampling routine
2. **Data Collection**: optimize collection method and carry out data collection
3. **Descriptive Statistics**: generate statistics to summarize/visualize your data
4. **Inferential Statistics**: discuss patterns/differences/characteristics about your data
5. **Estimation**: estimate patterns in population from your sample
6. **Hypothesis Testing**: apply appropriate tests to determine any causative effects or differences between groups; find significance

Sample and Population

When discussing statistics, we must introduce the concept of sampling: when you collect a **sample** (set of data points) from a **population** (large body of measurements)

- The goal is to predict the behaviour of the population by analyzing data from a representative sample

A **variable** is a measurable characteristic that changes, and you can get data points from that variable

- If you can measure 1 variable from your sample, you have **univariate** data
- If you can measure 2 variables from your sample (temperature and location for example), you have **bivariate** data
- If you measure 3+ variables from your sample, you have **multivariate** data

In sample we work with **statistics**, while in poplutation we work with **parameters**

Qualitative Data

Types of Catagorical Data

Qualitative data is commonly known as **categorical data**

Categorical data can be further broken down into:

- **Boolean Data**: only two possible values
- **Nominal Data**: more than two categories are required
- **Ordinal Data**: categories must be *ordered* and have logical sequence

Commonly presented as a statistical table

Frequency and Relative Frequency

Frequency represents the **number of measurements** in each category from a total of n

Relative Frequency is the proportion of measurements in each category

$$f_{rel} = \frac{frequency}{n}$$

Percentage: gives us the percentage of measurements in each category

$$percentage = 100 \times f_{rel}$$

Key Points:

- Sum of all frequencies will be n
- Sum of all relative frequencies will be 1
- Sum of all percentages will be 100%
- Always have a category for **outliers** that you can filter out or include
- Ensure your categories are created such that:
    - 1 measurement falls into 1 category only
    - 1 measurement must fall into any one of your categories (including the “outlier”) one

Pie Charts

useful for visualizing frequency distributions of categories

Considerations

- Overlap? Use a legend
- Too many small slices? Use another type of graph (bar graph)

Bar Chart

commonly used to display categorical (or quantitative) data

Considerations:

- Label all axes,always!
- Add a title!
- Present values in table or on the graph itself

A bar chart does not always have to be vertical

- Can also use a horizontal bar chart

To represent multiple data sets, you may use a 3D bar chart

Creation of Figures

We can approach coding in one of two main ways: functional vs. object-oriented

- **Functional**: use built in functions to create the required figure/axes automatically
- **Object-Oriented**: step-by-step plotting from generating the figure, axes and plotting

As a physicists, always try to use object-oriented

Qunantitative Data

Types of Quantitative Data

Quantitative data is **numerical data** made up of measurements from discrete or continuous variables

Describing numerical data can take place in two critical ways:

- **Graphical Analysis**: help us describe the basic shape of data distributions
- **Numerical Analysis**: generate statistics from the sample data

Considerations when choosing graphical/numerical or both types of analysis:

- If you have a lot of numerical data points, a graph will help show spread
- Too many plots will cause confusion; we need a way to **summarize** the sample data
- For key summaries, use graphs to show key trends and distributions
- For general statistics and numerical measures, keep numbers in the discussion

Different Charts

Pie Chart: useful for displaying breakdowns of **numerical ranges**

Bar Chart: also useful for frequency within **numerical ranges**

Line Chart: useful for trends across time series or along an axis

Histograms and Relative Frequency Histograms

Histograms: bin data into numerical categories **Relative Frequency Histograms**: helps determine distribution across data set

Interpreting Graphs

Distributions of data are determined by their **shape** in a graph

- Look at symmetry, skewness, uni/bi/multimodal distributions

For unimodal distributions, symmetry and skew are easy to spot:

- **positive/right skew**: tail to the right, mean > median > mode
- **symmetrical distribution**: no tail, mean = median = mode
- **negative/left skew**: tail to the right, mean < median < mode

Measures of Centre

Mean: From a set of n measurements, this average is the sum of all measurements divided by n.

$$\bar{x} = \frac{\sum x_i}{n}$$

Median: From n measurements, median m is the value of x in the middle after all values are sorted from smallest to largest

Mode: Most frequently occurring value of x or category

Measures of Spread

Range: Difference between smallest and largest measurements, given as R

Deviation: How far away a singular measurement is away from the mean

- use μ for population mean and $\bar{x}$ for sample mean

$$deviation = (x_i - \bar{x})$$

Sample Variance (s^2):

- Also population variance (σ^2)
- use **sum of squares**
- For population, use N instead of(N–1) and σ^2 and μ

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{N-1}$$

Standard Deviation:

- The positive square root of variance; measures the amount of variation
  - high values: data is far from the mean and the spread is wide
  - low values: data is close to the mean and the spread is narrow

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N-1}}$$

Sample vs. Population

- As scientists, **usually sample** since we are always doing experiments or collecting data based on a subset of the population to make predictions about the population
- Be sure to check your code and syntax to ensure the proper functions or equations are being used!

Measures of Relative Standing

z-score: distance between observation and mean based on standard deviation

- 95% of observations lie within 2 standard deviations from the mean (|z-score| < 2)
- 99.7% of observations lie within 3 standard deviations from the mean (|z-score| < 3)
- Helps us determine whether data is considered **an outlier** (>2,>3)

$$z - score = \frac{x - \bar{x}}{s}$$

percentile: when n meansurements are ordered based on magnitude, the pth percentile is the value of x that is greater than p% of the measurements and less than (100-p)%

- Q1 = 25th percentile = lower quartile
- Q2 = 50th percentile = median
- Q3 = 75th percentile = upper quartile

Probability

Origins of Probability

Real life is unpredictable in many cases, due to:

- Incomplete Knowledge
- Large Numbers
- Sensitivity to Initial Conditions
- Open Systems

Probability can be defined in one of two ways:

- The **frequency** with which unpredictable events occur – **"Frequentist Approach"**
- The **degree of belief** that some hypothesis is correct – **"Bayesian Approach"**

This create two branches of probability that we will discuss

This create two branches of probability that we will discuss

Events and Sample Space

We collect data through an **experiment** (flipping a coin)

- The outcome is called a **simple event** – heads, tails are the possible outcomes
- The set of possible outcomes is called the **sample space** S = {heads, tails} of size 2, where you list the simple events
- An **event** is a collection of simple events - A is an event where you roll a die and get a value > 3 (there is more than 1 possible answer to get a value > 3)

If you have experiments with stages (ie. 3 coin tosses), you can create a **tree diagram** to help visualize the sample space

Probability of Event A

$$P(A) = \frac{n_A}{N}$$

Where n_A is the frequency of event A, and N is the total number of events

mn Counting Rules

There are m possible outcomes for the 1st event, and n possible outcomes for the 2nd event, then the total number of possible values are given by mn

Extended mn Rule: For i events, just multiply all number of possible outcomes

When to use: When trying to figure out how many outcomes are possible **without** worrying about order or groups

Permutations and Combinations

refer to ways in which objects from a sample space can be selected to form subsets

Permutations

Use when order matters

$$nPk = \frac{n!}{(n-k)!}$$

Combinations

When order of group does not matter

$$nCk = \frac{n!}{(n-k)!k!} = \frac{nPk}{k!}$$

Event Relations

Union: $A \cup B$

- either events A or B can occur
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- **Union for Disjoint / Mutually Exclusive Events**
  - either A or B can occur, but no overlap
  - $P(A \cup B) = P(A) + P(B)$

Intersection: $A \cap B$

- Both events A and B can occur

- $P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$

Complement: A^c

- When A does not occur
- $P(A^c) = 1 - P(A)$

Conditional Probabilities

the likelihood of an event occurring based on the **occurrence of a previous event**

- Probability of A given B is $P(A|B)$
- It is the fraction of P(B) that intersects with A

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B|A)}{P(B)}$$

The second formula is known as **Bayes' Rule**

Events are **independent** if and only if:

$$P(A \cap B) = P(A)P(B)$$
and
$$P(B|A) = P(B)$$

Otherwise, they are considered **dependent**.

Discrete and Continuous Probability Distributions

Random Varaibles

We call X a random variable to represent **any variable that varies or changes** depending on the **outcome** of the experiment being measured

If the possible outcomes are listed out using **whole numbers**, we have a **discrete random variable**

- **Finite**: fixed number of possible values
- **Countably Infinite**: possible values can be listed out, but not easily (as there are infinitely many)

If the possible outcomes can be described using an **interval of real numbers**, we have a **continuous random variable**

- **Uncountably Infinite**: too many possible values to list or count, but all are measured with high precision

Baysian vs. Frequntists Views on Probability

Frequentist:

- Probabilities are interpreted as long-run frequencies
  - goal is to create procedures with frequency guarantees
- Parameters are **fixed constants** and probability statements are about **procedures**

Bayesian:

- Probabilities are interpreted as subjective degrees of belief
  - goal is to state and analyze those beliefs

- Parameters are **random variables** and probability statements are about those parameters
- Here, we choose a **probability density** (the “prior” distribution) that expresses our beliefs about a parameter before we see any data
  - Then we choose a **statistical model** that reflects our beliefs about the data given the prior
  - After observing our data, we update our beliefs and calculate the **posterior distribution**

Probability Distributions

A **probability distribution** is a mathematical **function** that gives the **probabilities of occurrence** of different possible outcomes of an experiment

- We use $p(x)$ for each value of x for random variable X

Types of Probability Distribution

- **Probability Mass Function (PMF)**: gives the probability that a **discrete** random variable is exactly equal to some value
- **Probability Density Function (PDF)**: gives the probability that a **continuous** random variable falls within a particular range of values (versus taking on one exact value)
  - Given by the **area** under the density function but above the horizontal axis
- Common characteristics:
  - $0 \leq p(x) \leq 1$: individual probability much be between 0 and 1
    - individual probability is 0 for continous probability distributions
  - $\sum p(x) = 1$: all probabilities must add up to 1
    - the area must be 1 for continous probability distributions
- **Comulative distribution functions**: provide the probability that X takes on a value less than or equal to x

Discrete Probability Distributions

Uniform Probability Distribution

distribution where PMF is a constant value; every value has equal chance --> flat curve

$$p(x) = \frac{1}{n}$$

Binomial Probability Distribution

distribution where you have n identical trials, each with only 1 of 2 possible outcomes (p, success or q = 1-p, failure)

Values of p and q are consistent from trial to trial; trials are independent

Poisson Probability Dirstribution

Distribution for events that occur an average μ number of times over a certain period of time or space

Events must occur randomly and independently of one another

Continuous Probability Distributions

Uniform, Continuous Probability Distribution

For c is a constant:

$$f(x) = c$$

Exponential Probability Distributions

$$f(x) = \frac{1}{\mu} e^{\frac{-x}{\mu}}$$

Normal/Gaussian Probability Distribution

Naturally occuring distribution affected by population mean μ and standar deviation σ

- μ locates the **centre** of the distribution
  - Distribution must be symmetric around the mean
- σ determines the **shape** of the distribution (height, width of curves)
  - large value increases spread and reduce height

Stadarided normal distribution means that the normal distribution has $\mu = 0$ and $\sigma = 1$

- Any normal distribution can be standardized by converting its values into z-scores and plot the z-score distribution.
- They will tell us how many standard deviations from the mean each value lies
- This allows us to calculate the probability of certain values occurring and to compare different data set

$$X = \mu + z\sigma$$

If...

- X < μ, z<0
- X > μ, z>0
- X = μ, z=0

Summarizing Quantities

Coefficient of Variation (CV, RSD)

Relative standard deviation

- It is a dimensionless ratio of the standard deviation and the mean
- Useful in expressing the precision and repeatability of experiments

$$CV = \frac{s}{\bar{x}}$$

Percentiles

Indicate value below which a given % of observations fall

- integrate the area under probability function to find the probability of values falling in

  between a and b

$$P_{int} = \int_a^b p(x)\, dx$$

Expeted Values

The generalization of a weighed average of a random variable

- We say the the exptected value of X is $E(X)$, μ_X , or μ

For PMF:

$$E(X) = \mu_X = \sum xp(x)$$

For PDF:

$$E(X) = \int_{\infty}^{\infty} xf(x)\,dx$$

We can also calculate variance:

$$\sigma^2 = E(X^2) - \mu^2$$

Moment Generating Functions (MGF)

Uniquely determines the distribution of a random variable

Four Moments of a Probability Distribution

- 1st Raw Moment: Mean
- 2nd Central Moment: Variance
- 3rd Standardized Moment: Skew
  - Level of asymmetry that deviates from a normal distribution
    - The direction of skew comes from whichever tail is longer
    - **Positive** skew means longer tail on the **right** → right-skewed
- 4th Standardized Moment: Kurtosis
  - The peakedness of the distribution
    - The peakedness comes from the distribution of tails which affects how sharp a peak is
      - **Leptokurtic**: K>0, very sharp peak
      - **Normal**: K=0
      - **Platykurtic**: K<0, very flat peak

Central Limit Theorem

When selecting a random sample from a population, the numerical measures from the **sample** are called **statistics** (ie. mean, median, etc.)

- The **sampling distribution** of a statistic is the probability distribution for the possible values of that statistics when random samples of size n are repeatedly drawn from the population

The Central Limit Theorem states that in general conditions, the sums and **means of random samples** of measurements from a population tend to have an approximately **normal distribution**

- We can say that the sampling distribution of the mean is:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- If the population has a normal distribution, the sampling distribution of $\bar{x}$ will be **exactly** normally distributed regarless of n
- If the population distribution is non-normal, the sampling distribution of $\bar{x}$ will be **approximately normal** when n is large ($n \geq 30$)

Increasing N reduces sampling error, and allows us to make a good estimate of the population mean

- can use it to predict **parameters** of a population like standard deviation and mean

Normal (Gaussian) Distribution in Real Life

A “normal” distribution does take place under “normal” circumstances, especially after applying CLT

Examples of normal distributions in real life

- Height of the population
- Birth weight of babies
- Shoe sizes
- Test scores (usually)
- Blood pressure for men vs. women
- Rolling a dice
- Coin toss (probability of heads for all tosses)
- Random motion of particles
- Concentration of a specific ion within the human body
- Measurement error from experiments

Maxwell-Boltzmann Probability Distribution

describes the statistical distribution of particles in a system among different energy levels

- It is officially considered a “chi distribution” that takes in to account a set of independent random variables, each following a standard normal distribution
- Since the MB distribution is defined and used for describing particle speeds in idealized gases, there are **three independent random variables** (x, y, z components to velocity) and thus three degrees of freedom from Euclidean 3D space

Log-Gaussian

- Sometimes, if a distribution does not look Gaussian, we can take the logarithm to form a Log-Gaussian distribution

Error Analysis

Importance of Error Analysis

- No matter what measurement is taken, there is room for error, no matter how small
  - Evaluating this error and uncertainty is called error analysis
- Every measurement taken must also include an estimate of the **level of confidence** associated with the value presented
  - Allows others to judge the quality of the experiment
  - Allows for meaningful comparisons with other similar experiments / values or a theoretical prediction
- Prior to experimental design, we have to understand how to report measurements and uncertainty in measurements
  - Allows us to check results and decide if a **scientific hypothesis** is confirmed or refuted due to the significance of your results

Types of Error

Random Error

statistical fluctuations (in either direction) in measured data due to limitations in the precision of the measurement device

- Examples:
  - enviromental factors
  - instrumentation limitations

- physical variations
- Fix: Reduce contribution of error by **averaging over large sample sizes**

Systematic Error

reproducible inaccuracies (in the same direction) that causes bias in measured data

- Examples:
  - unclear definition of measurement
  - missed parameters or factor in meansurement
- Fix: None; they are difficult to detect and cannot be fixed by increasing sample size

Human Error

errors related to poor technique and understanding

- Not considered an error; must be fixed or corrected prior to moving forward

Accuracy vs. Precision in Measurements

Any meansurement is reported as $meansurment = best\ estimate \pm uncertainty$

Accuracy

how close your measurement is to the **true value**

- Commonly reported as **relative error**

$$relative\ error = \frac{measured\ value\ -\ expected\ value}{expected\ value}$$

Precision

how consistent your measurements are; reliability / **reproducibility** of your result

- Commonly reported as **relative (fractional) uncertainty**
- We call $\pm\delta x$ the **absolute uncertainty of a measurement x**

$$relative\ uncertainty\ = |\frac{uncertainty}{measured\ quantity}|$$

Error Propagation

The exact formula for **propagation of error** for a function f(x,y,z) relates each variable and their standard deviation

$$\sigma_f^2 = (\frac{\partial f}{\partial x})^2\sigma_x^2 + (\frac{\partial f}{\partial y})^2\sigma_y^2 + (\frac{\partial f}{\partial z})^2\sigma_z^2$$

- Addition, subtraction and logarithmic equaions lead to an **absolute standard deviation** where we use σ_f
- Multiplication, division, and exponential equations lead to **relative standard deviations**, where we use $\frac{\sigma_f}{f}$

Type	Example Function	Standard Deviation (σ_f)
Addition or Substraction	$f = x + y - z$	$\sigma_f = \sqrt{\sigma_x^2 + \sigma_y^2 + \sigma_z^2}$
Multiplication or Division	$f = \frac{xy}{z}$	$\frac{\sigma_f}{f} = \sqrt{\left(\frac{\sigma_x}{x}\right)^2 + \left(\frac{\sigma_y}{y}\right)^2 + \left(\frac{\sigma_z}{z}\right)^2}$

Exponential	$f = x^c$	$\frac{\sigma_f}{f} = c\left(\frac{\sigma_x}{x}\right)$
Logarithmic	$f = \log x$	$\sigma_f = 0.434\frac{\sigma_x}{x}$

Standard Error

The **population standard deviation** σ shows the distribution within a sample of what we are measuring

We call the standard deviation of a **statistic** the "standard error of the estimator"

- The term "estimator" is used because the statistic is used to infer details about the population's parameter
- In other words, how precise the estimator is

In many cases, we look at the **averaged value (mean)**

- If we want to look at the **precision of the mean**, we can calculate the standard deviation of the mean, which is traditionally called the **standard error of the sample**
  - standard deviation of the sample devided by the square root of the sample size
  - So the standard error, by definition, is the standard deviation of $\bar{x}$ which is simply the square root of the variance

$$SE = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Errors Bars

Error bars represent the variability of data and uncertainty in a reported measurement

- Often represent 1 standard deviation of uncertainty, 1 standard error or a particular confidence interval
- When reporting, make sure you state what kind of error you have used for your error bars!

Experimental Design and Hypothesis Testing

Statistical Inference and Types of Tests

Statistical Inference:

The process through which **inferences (predictions) about a population** are made based on statistics calculated from a sample of data from the population

- Statistics are from the sample, parameters for from the population

This means that there are two approaches to statistical inference:

- **Hypothesis Testing**: making a **decision** about the value of a parameter based on a preconceived idea about its value
  - we have to first **come up with a hypothesis**, a value for the hypothesis and its null, then reject or accept the null to make conclusions
- **Parameter Estimation**: estimating or **predicting the value** of a parameter
  - we have to look at **estimators** as well as the **maximum likelihood** of getting certain values

Experimental Design

The process by which a hypothesis is investigated

- Involves deciding which factor is the **independent variable** (to be manipulated) and which one is the **dependent variable**.
  - Note that the **independent variable** is called a **factor** and its manipulations are referred to as **factor levels**
- We adjust the factor to see its effects on the dependent variable to determine where there is a causal relationship
- We apply **statistical test** to reject / accept the null hypothesis

Steps for proper experimental design

1. Understand and consider all **variables** and their relationships to one another
2. Present a testable **hypothesis** specific to what you are looking for
3. Design an **experiment or sampling routine** to collect data and manipulate the independent variable
4. Apply appropriate **statistical tests**
5. Analyze results for **significance** and check for optimizations. Repeat 3-5 if required.
6. Summarize, present and discuss your **findings and conclusions**

Controls

Controls help reduce or isolate the effect of external factors on your study

- If you are only interested in your independent variable, controls help prevent your data from being affected by other factors

Types of Control

- **Experimental Control**: controlling the environment around the experiment (temperature, humidity, etc.)
- **Procedural Control**: run the experiment on a negative control group and experimental group to make sure any factors arising from the procedure itself can be eliminated
  - **Placebo Effect**: when the control group exhibits effects when there should be none
- **Temporal Control**: observing two groups prior to manipulating factors
  - Good for experiments that run for long periods of time
- **Statistical Control**: Instead of adjusting the environment, we record the environment's settings and analyze their effect afterwards

Null Hypothesis

- Hypothesis Testing: the act of testing an assumption regarding a population's parameter
  - Start off with the formation of a hypothesis H1
- Each main hypothesis has a contradictory, **null hypothesis** Ho
  - Our goal is to reject or accept the null hypothesis because it is easier to do so than the alternative hypothesis H1
  - Why?
    - Null is testing a mean value $\mu =?$, while alternative is testing that $\mu \neq ?$, which means that the population parameter can be smaller, greater, or different
    - Easier to disprove a mean that is not one value than is many values
- Rejecting the null hypothesis concludes that our hypothesis is likely true.

One vs. Two Tailed Test of Hypothesis

One Tailed Test

- Test to see if the parameter is significantly greater OR less than X, but not both

- Hypothesis: $\mu < X$ or $\mu > X$

Two Tailed Test

- Test to see if the parameter is significantly greater or less than X, in either direction
  - Hypothesis: $\mu \neq X$

Approach to Hypothesis Testing

1. Create a null hypothesis $H0$ and assign it a value
2. Create a hypothesis $H1$ which is your alternative hypothesis, and determine whether it requires one- or two-tailed hypothesis testing
3. Determine a test statistic and its P-value
4. Determine rejection regions and test P-value
5. Make conclusions from your results on significance and confidence levels

Test Statistics

- A single number calculated from the sample data that we can use for our hypothesis test
  - We assign the null hypothesis the value of the test statistic, $\bar{x}$
- The goal is to test this mean value $\bar{x}$ against the population mean μ, which we get from our null hypothesis

  - we can use a **z-score**:
$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$
  - This will give us how many standard deivation away from the population mean that $\bar{x}$ is, and allow us to find the p-value of $\bar{x}$
- We can construct **rejection regions** based on a chosen α level of **significance** which correlates to a (1-α) level of **condifence**
  - usually we use confidence levels of 95% and $z_{\alpha/2}$ is used for critical values for two tailed tests

P-Values

- The P-value is the probability of seeing a select set of data if the null hypothesis is true
  - With the z-score of the test statistic and the desired confidence inverval, we can easily find the probability of obtaining certain z-scores in our distribution
- For a given probability density function of p(x), to test the null hypothesis, the p-value for X>x is:
$$P(> x) = \int_{x}^{\infty} p(x)\, dx$$
- If the p-value is **0.05 or lower** for a confidence level of 95%, it means that we can reject his null hypothesis and say with the same level of confidence that the hypothesis is true
- Rejecting the null hypothesis does not prove that the hypothesis is true, just gives us a high likelihood that it is so

Acceptable Errors

- We know that in reality, our hypothesis can either be true or false
  - But with hypothesis testing and rejection of the null hypothesis, our statements can only go so far
- The worst thing that can happen is if you make a mistake and reject/accept the null hypothesis when it should have been the opposite. There are two ways this can happen, each outlining the type of error made
  - **Type I Error**
    - when null hypothesis is really true, but the statistical tests lead you to believe it

is false
    - This is called a **false positive** and very damaging to the conclusion
  - **Type II Error**
    - when null hypothesis is really false, but the statistical tests lead you to believe the null hypothesis is true
      - This is called a **false negative**. It is less damaging, because your hypothesis lives to see another day and you can run the tests again

Experimental Design and Hypothesis Testing

Estimation Theory

Branch of statistics that deals with **estimating values of parameters** based on measured data with a random component

- **Estimator**: an attempt to approximate unknown parameters using measurements

General methods to approxiamte those values:

- **Probalistic Approach**: assume the measured data is **random** with a probability distribution based off of key parameters
  - This is the focus of this course
- **Set-Membership Approach**: assume the measured data vector belongs to a set that depends on a parameter vector

Why estimation theory?

- Allow us to take our sample of measured data as an input to produce an **estimate** of parameters with some level of confidence
- Allow us to **infer** the value of unknown parameters in a statistical model
- Help to **understand the behaviour of population** with the help of a small sample

Types of Estimators

- **Estimator**: rule for calculating an estimate of a given parameter based on observed data
  - **Estimator**: "rule"; **Estimand**: "quantity of interest; **Estimate**: "result"
- **Maximum Likelihood Estimator (MLE)**: estimate parameter of an **assumed probability distribution** given some observed data
  - Process of maximizing a likelihood function for which the observed data is most probable for the statistical model chosen
- **Bayes Estimator**: minimizes the **posterior expected value** of a **loss function**
  - Posterior distribution consists of prior distribution and observed data
  - Loss function (usually quadratic) is the loss incurred in estimating a parameter's value
- **Method of Least Squares**: typical in regression analysis; **minimizes sum of squares of residuals** to get the best fit for a set of data points
- **Markov Chain Monte Carlo (MCMC)**: class of algorithms to sample from a probability distribution; algorithm runs until Markov chain reaches equilibrium

Maximum Likelihood Estimation (MLE)

A method that determines values for the parameters of a statistical model (ie. linear)

- Answers: which are the best parameters for my model?

- Ex: for a linear model: y=ax+b, where a is the parameter in the model, given some **postulated claim** about b (which can be considered as noise / the value of y when x = 0)

No matter which model is chosen, we use θ to be a vector of all parameters

- Ex: for a linear model: $\theta = (a, b)$
- Our goal with MLE is to select parameters θ that make observed data most likely (ie: maxmimize the likelihood)
- We must make the assumption that the data we use to estimate the parameters will be n **independent and identically distribution (IID)** samples

Likelihood

We have assumed our data are IID so they must all share the same PMF (discrete) or PDF (continuous)

- We can use $f(X|\theta)$, a probability distribution function, to refer to this shared distribtuion

Likelihood means the **joint (overall) probability** of the data (discrete) or the joint probability density of the data (continuous)

- Since we have assumed each data point is independent, the likelihood of all our data is the product of the likelihood of each data point

$$L(\theta) = \prod_{i=1}^{n} f(X_i|\theta)$$

With MLE, we need to choose values of θ that maximize $L(\theta)$

- We can use the notation $\hat{\theta}$ to represent the best choice of values for our parameters
  - the argmax of a function is the value of the domain at which the function is maximized

$$\hat{\theta} = argmax_{\theta}\ L\theta$$

We can then take the log on both side since log is monotonic, which gives us:

$$LL(\theta) = log(L(\theta)) = log(\prod_{i=1}^{n} f(X_i|\theta)) = \sum_{i=1}^{n} log(f(X_i|\theta))$$

Where to find $\hat{\theta} = \{x_0, x_1, \ldots\}$, we can the the partial derivative of the $LL(\theta)$

Example: for a normal distribution, we could have $X_i = N(\mu = \theta_0, \sigma^2 = \theta_1)$

Linear Regression

- linear approach for modeling the relationship between dependent and independent variables; commonly used for predictive analysis and modeling
- In most cases, we want to use linear regression to search for a **best-fit line** to a given (observed) data set (x_i, y_i)
  - We are interested in the parameters k and d that help to minimize the sum of squared residuals
    - the residuales (ϵ_i) is the differences between observed and perdicted values

$$y_i = kx_i + d + \epsilon_i$$

- Since linear regression is solved to minimize the square of sum of residuals, it is commonly referred to as **Ordinary Least-Squares (OLS) regression**

- Note that for linear and OLS regression, we assume all variability to lie in the residuals

Coefficient of Determination (R^2)

- We don't just have to have a linear model; we can have higher-order regressions based on what degree of function we are trying to fit (quadratic, exponential, etc)
- For each model, we can determine the coefficient of determination (R^2)
  - It is a statistical measure in a regression model that determines the proportion of variance in the dependent varaible, that is explained by the independent variable
  - In other words, it is the sum of squares (SS) by the proposed model divided by the total sum of squares

$$R^2 = 1 - \frac{SS_{model}}{SS_{total}}$$

- This tells us:
  - Relation to unexplained variance as R2 tells us variance of models' errors compared to the data's total variance
  - Goodness of fit - high R2 = better fit