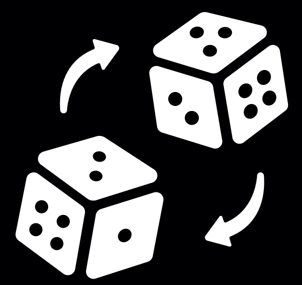
WHAT YOU SHOULD KNOW ABOUT PROBABILITY AND STATISTICS

Assoc. prof. ERALDA GJIKA (DHAMO)

BASIC PROBABILITY CONCEPTS



Understanding outcomes and their likelihoods.
For example, the probability of flipping a coin and getting heads is 0.5.

$${}_{n}P_{r} = \frac{n!}{(n-r)!} \qquad {}_{n}C_{r} = \frac{n!}{r!\,(n-r)!}$$

Mutually Exclusive

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Independent

$$P(A \cap B) = P(A) \times P(B)$$

Do you know that?

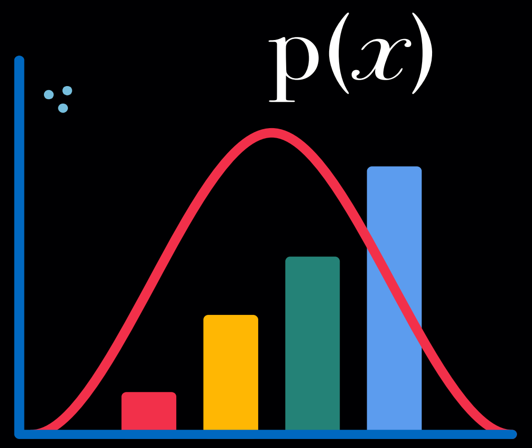
- **Forecasting**: What is the probability of a product being in high demand next month based on past sales data. (Markov Chain)
- **Cyber Threats**: Estimating the likelihood of a system vulnerability being exploited based on historical data of similar incidents.
- **Election Poll Predictions:** Determining the probability of a candidate winning a specific region based on poll data.
- **Medical Predictions:** Assessing the probability of side effects occurring in a clinical trial.

CONDITIONAL PROBABILITY AND BAYES' THEOREM

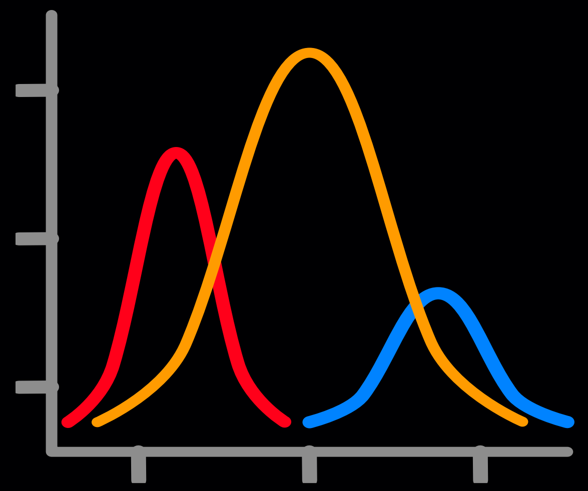
- **Forecasting:** Calculating the conditional probability of increased sales given that a new marketing campaign is launched.
- **Cyber Threats:** Assessing the probability of a successful phishing attack given an observed increase in suspicious emails.
- **Election :** Estimating the probability of a candidate's win given the turnout rates in specific demographics.
- **Medical Predictions:** Evaluating the likelihood of a patient developing a condition given a positive test result.

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

RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS



Discrete/ Continuous



Skewness/ Kurtosis

- **Forecasting:** Modeling monthly sales as a **Normal** distribution to estimate future performance.
- **Cyber Threats:** Using **Poisson** distribution to model the number of daily cyber-attacks.
- **Election :** Modeling poll results as a **Binomial** distribution to predict outcomes.
- **Medical Predictions:** Using a normal distribution to analyze patient blood pressure readings.

$$\mu = E(X)$$

Expectation

$$\sigma^2 = Var(X)$$

Variance

Expectation/Entropy/ Similarity

$$7 = \sum x \, P(X = x)$$

Specific value for **Surprise**.

The probability of observing that specific value for **Surprise**.

$$\text{Entropy} = \sum \log\left(\frac{1}{p(x)}\right) p(x)$$

Surprise

The probability of the **Surprise**.

…and we end up with the equation for **Entropy** that Claude Shannon first published in **1948**.

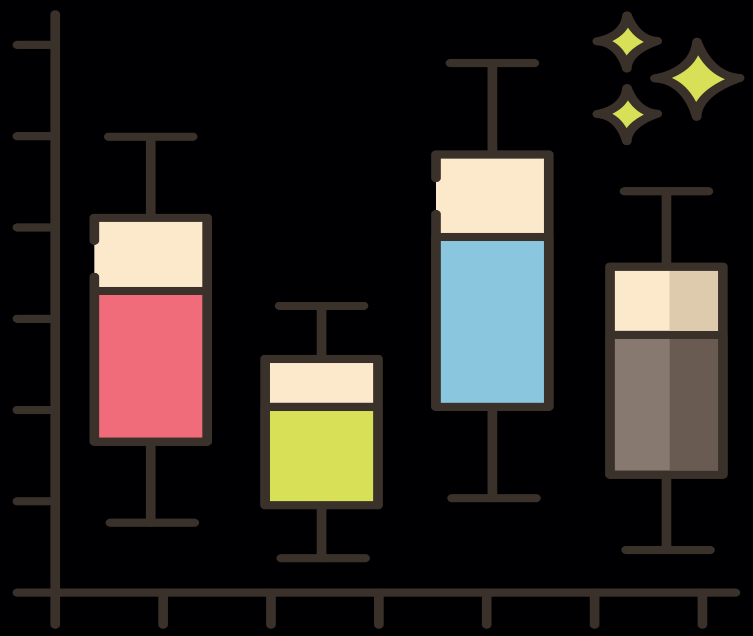
$$\boxed{\text{Entropy} = -\sum p(x)\log(p(x))}$$

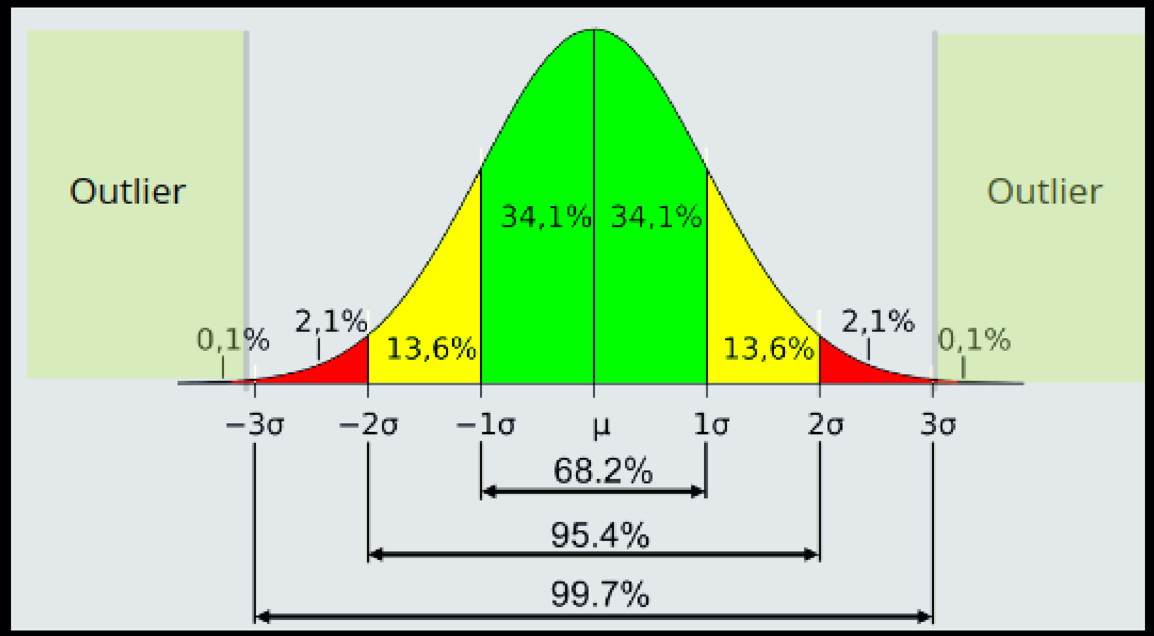
	Heads	Tails
Probability $p(x)$:	0.9	0.1
Surprise: $\log_2\left(\frac{1}{p(x)}\right)$	0.15	3.32

Thus, the **Entropy**, **0.47**, represents the **Surprise** we would expect ***per coin toss*** if we flipped this coin a bunch of times.

$$E(\textit{Surprise}) = (0.9 \times 0.15) + (0.1 \times 3.32) \boxed{= 0.47}$$

DESCRIPTIVE STATISTICS





- **Interquartile Range (IQR):** Use the IQR method to identify outliers. Data points falling outside(Lower fence) Q1−1.5·IQR and Q3+1.5·IQR (Higher Fence) are potential outliers.

- **Z-Scores:** Calculate how many standard deviations a data point is from the mean.
- Points with Z-scores > 3 (or another chosen threshold) might be outliers.

DEALING WITH OUTLIERS

- Removal: In some cases (e.g., sensor errors), it makes sense to exclude outliers.
- Transformation: Apply log or square-root transformations to reduce the impact of outliers.
- Winsorization: Cap outliers to a fixed percentile, such as the 5th and 95th percentiles.
- Replace: Replace outliers with the median, which is less sensitive to extreme values.

DEALING WITH MISSING VALUES

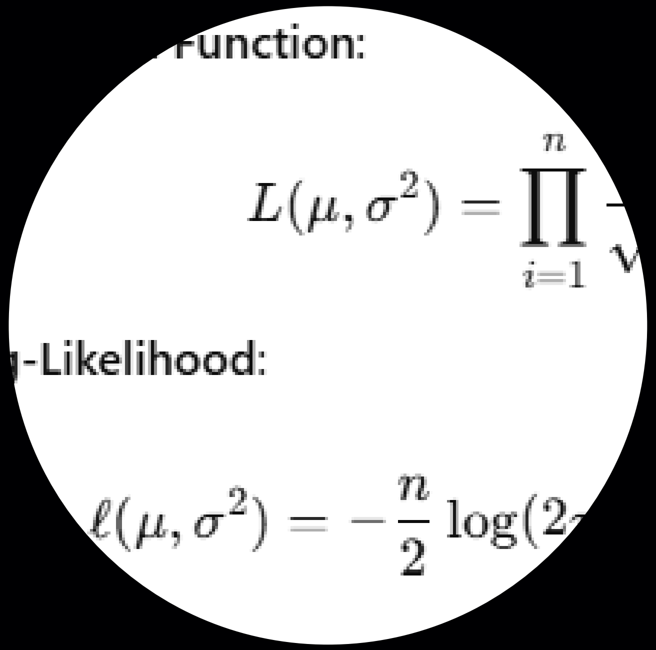
NA Imputation

- **Mean :**
  - Replace missing values with the mean of the non-missing data.
  - Best for symmetric distributions without outliers.
- **Median :**
  - Replace missing values with the median, which is robust to outliers.
  - Useful for skewed distributions or when there are extreme values.
- **Mode :**
  - Replace missing values with the mode, particularly for categorical or ordinal data.
- **Advanced Techniques**:
  - Use measures of spread (e.g., standard deviation) to generate random imputation within a range.
  - Employ regression or machine learning models (e.g., k-Nearest Neighbors) for more sophisticated imputations.

- **Data Normalization/ Standardization:**
  - Use variance measures (e.g., standard deviation) to normalize or standardize data for algorithms that require scaled inputs.
- **Feature Engineering:**
  - Create new variables, such as standardized scores, to highlight significant patterns.
- **Evaluating Data Quality:**
  - High variance in responses or strange clustering around certain location measures can highlight potential data collection issues.

PARAMETER ESTIMATION & HYPOTHESIS TESTING

IS THERE ENOUGH EVIDENCE TO SUPPORT A CLAIM?



Survival Analysis

Estimating the survival time of patients after a treatment.

- Parameter Estimation: Use likelihood methods to fit distributions like Weibull or exponential to survival times.
- Outcome: Helps predict probabilities of survival beyond a certain time and evaluate treatment effectiveness.

Machine Learning

Training models like logistic regression or neural networks.

- Parameter Estimation: **MLE** is used to estimate model parameters by maximizing the likelihood of observed outcomes given the model.
- Outcome: Improves predictive performance of the model.

Variable Importance

- Null Hypothesis (**H0**): The variable's coefficient is equal to zero ($\beta=0$), meaning it has no effect.
- Test Statistic: **t-test** or F-test is used.
- **P-Value**: Determines whether the variable's effect is statistically significant. (<0.05)

Central Limit Theorem and Law of Large Numbers

CLT

When you take a sufficiently large number of samples from a population (regardless of the population's original distribution), the distribution of the sample means will approach a normal distribution (bell curve).

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- **Predictability**: It allows practitioners to assume normality for inferential statistics when working with large data samples.
- **Model Training**: Helps in understanding how averages behave across iterations or samples.
- **Error Estimation**: Central to creating confidence intervals and conducting hypothesis tests in data models.

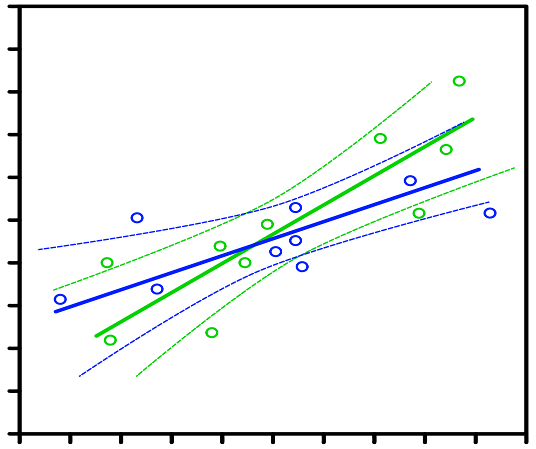
LLN

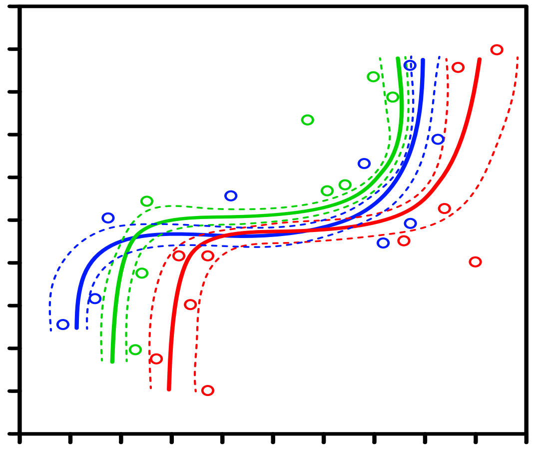
As the sample size (n) increases, the sample mean of independent, identically distributed random variables approaches the true population mean (μ).

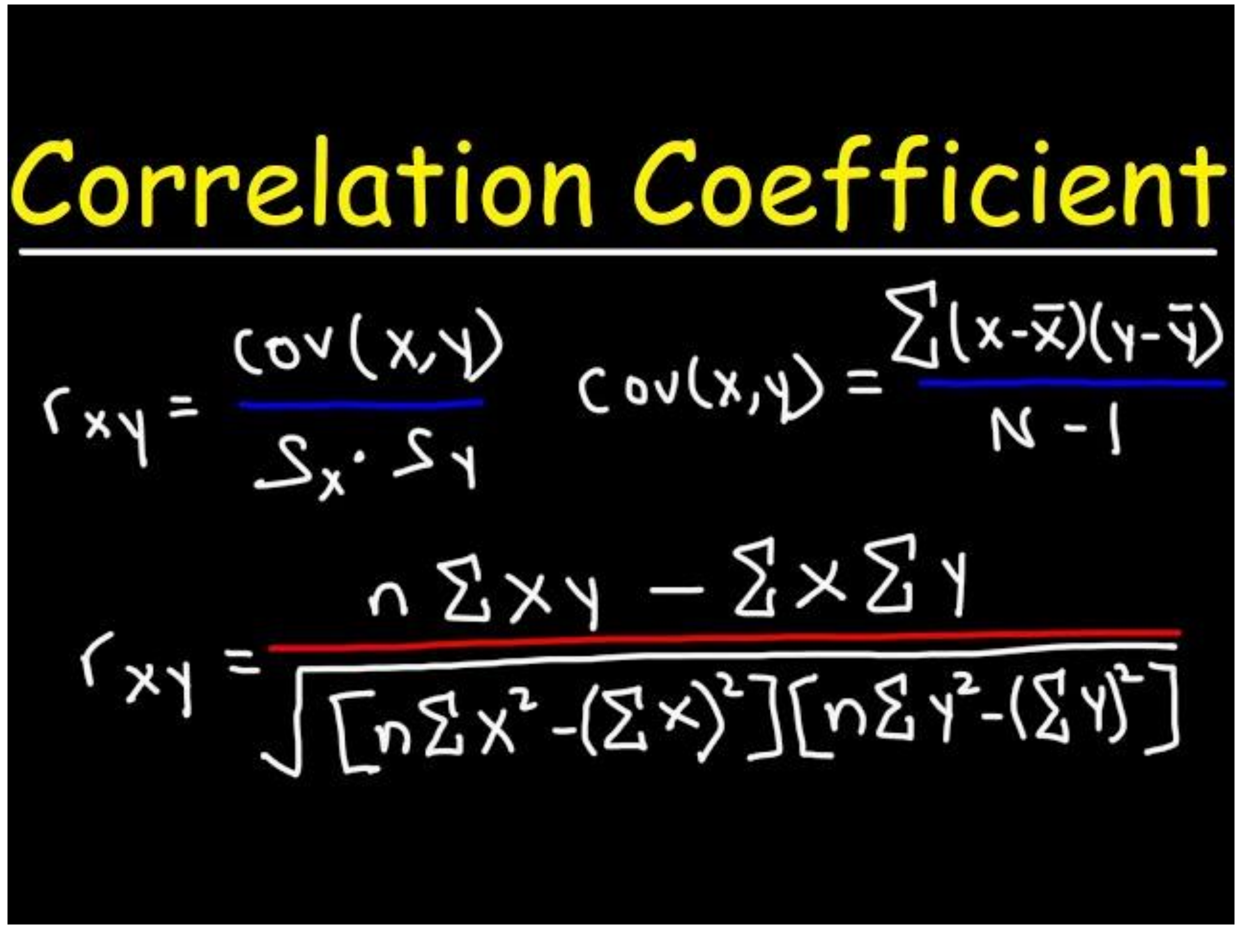
$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{n\to\infty} \mu$$

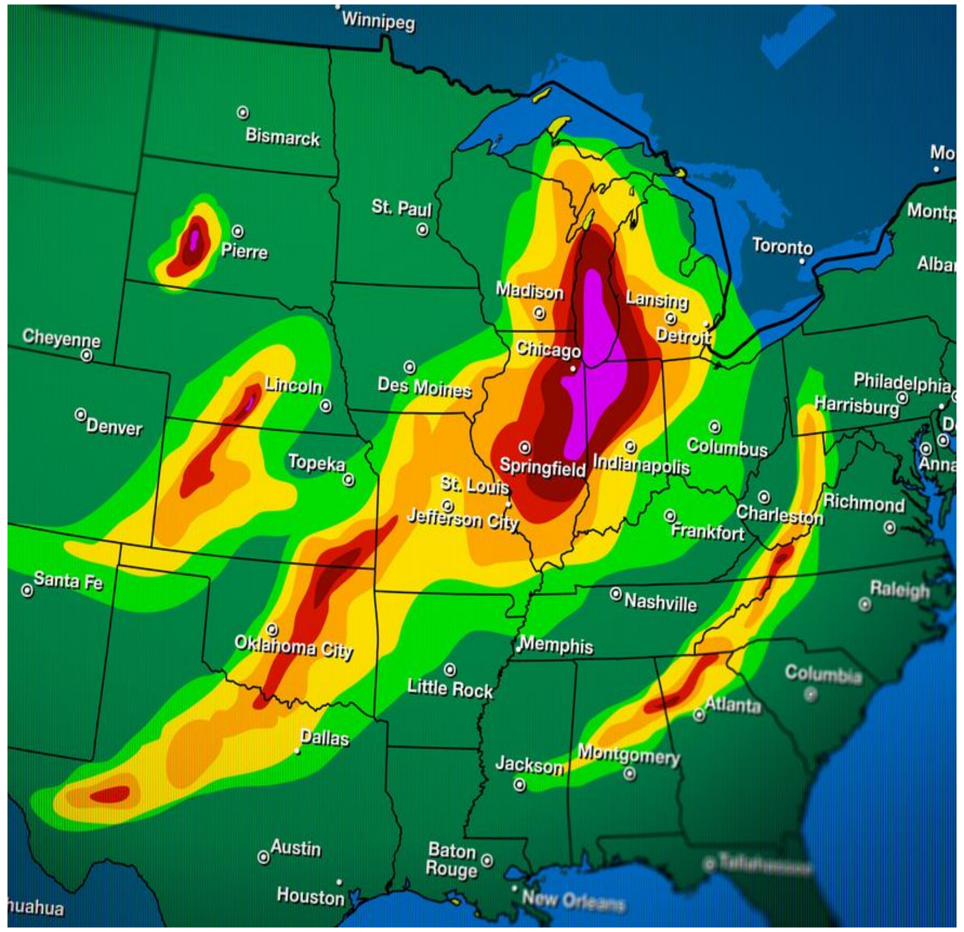
- **Consistency**: It guarantees that as more data is gathered, averages and predictions based on the data become more reliable and representative of the true underlying patterns.
- **Validation**: Ensures that machine learning models trained on larger datasets are less likely to overfit and more likely to generalize well to new data.
- **Monte Carlo Simulations**: Fundamental for simulations where repeated sampling is used to estimate probabilities and expected values.

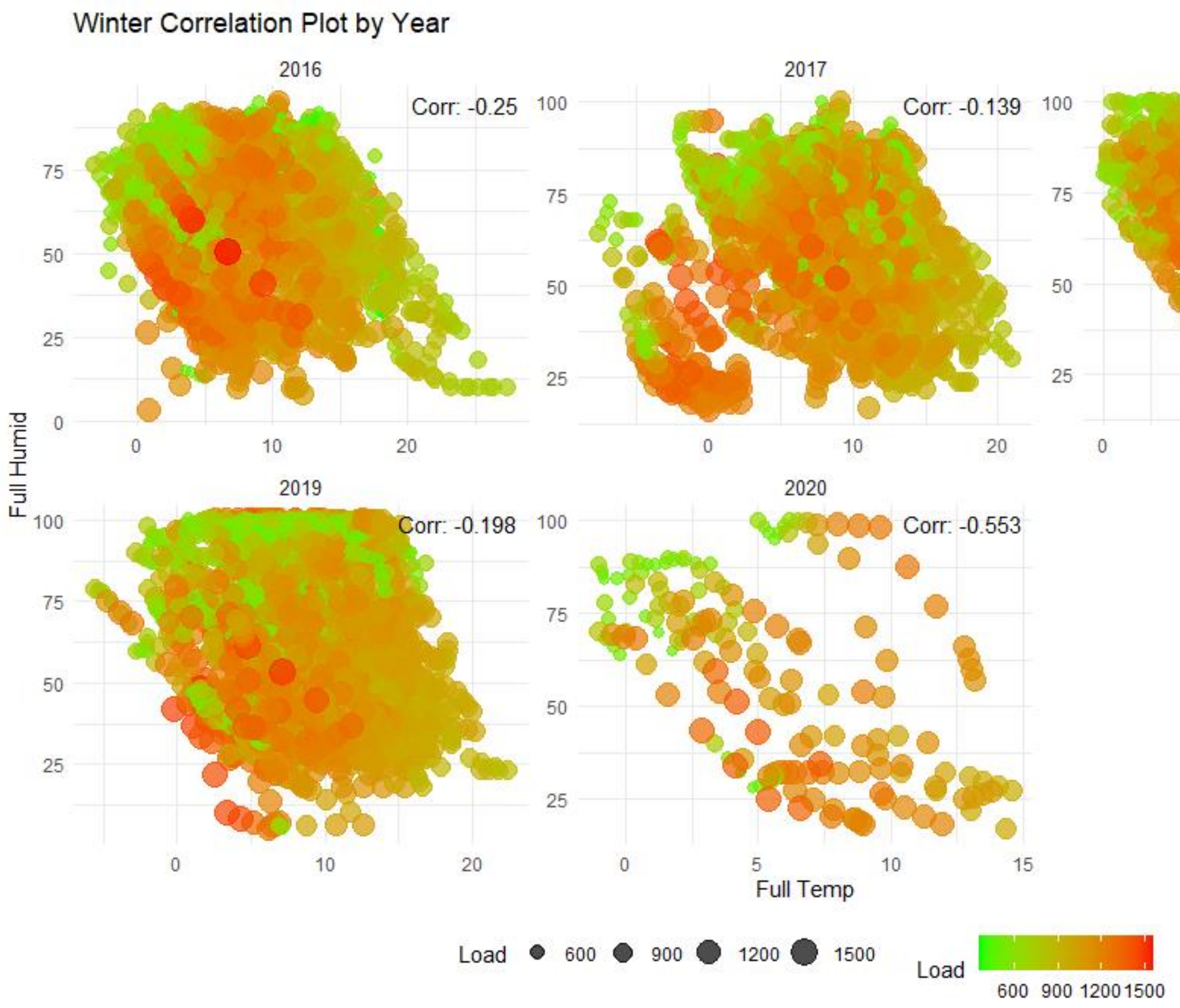
CORRELATION

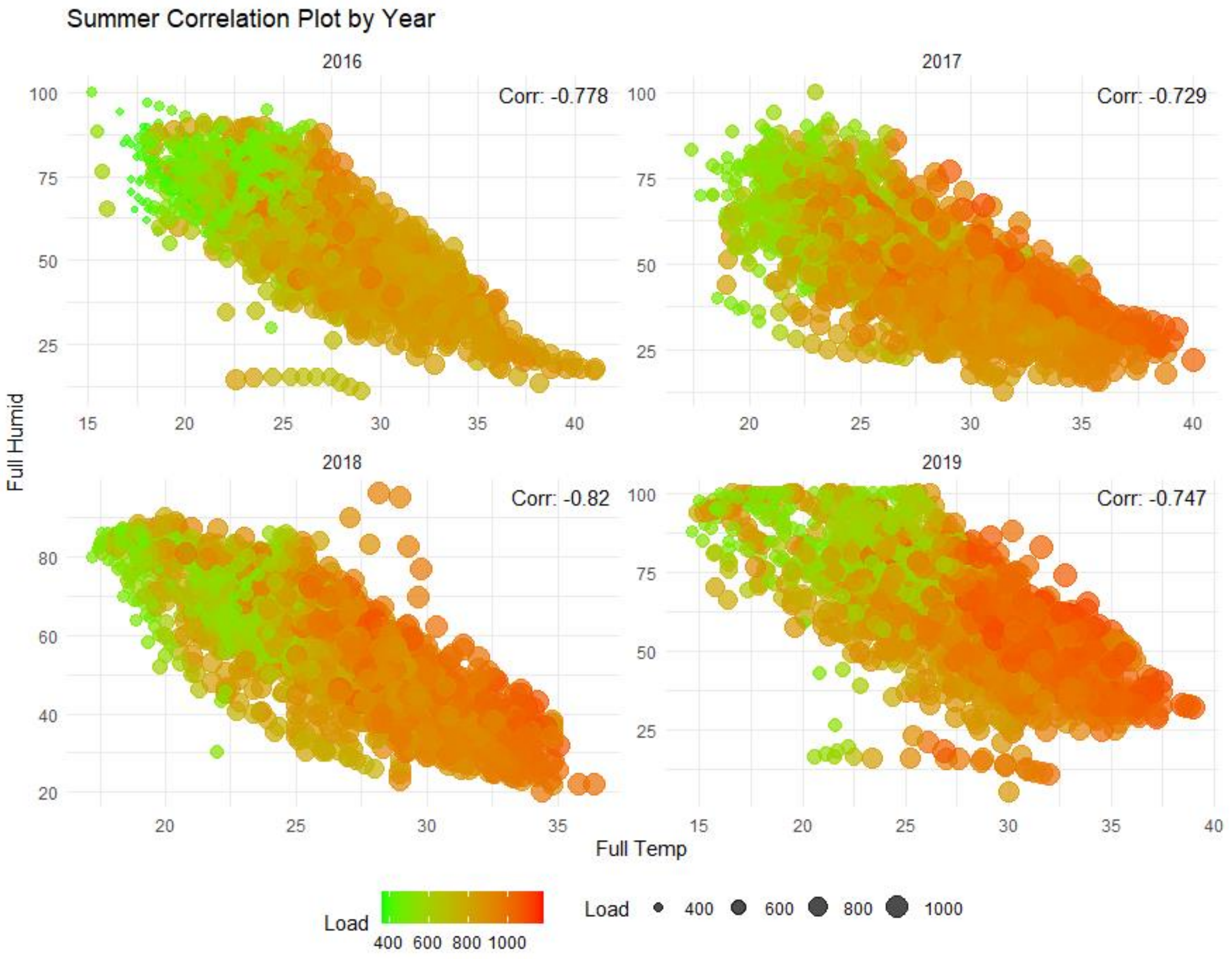






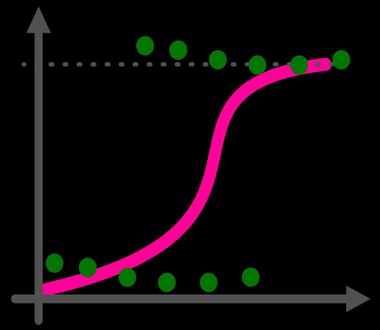






Key Differences

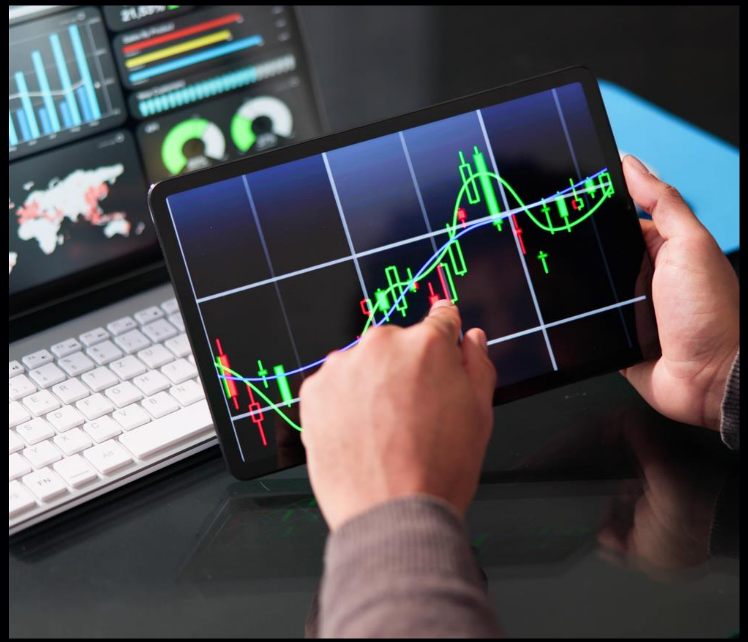
Aspect	Correlation	Causation
Definition	Statistical association between variables.	One variable directly causes a change in another.
Implied Relationship	Variables move together but not necessarily cause one another.	A direct cause-effect relationship exists.
Evidence Required	Statistical data (e.g., correlation coefficient).	Strong evidence, often experimental.
Examples	Ice cream sales and drowning.	Smoking and lung cancer.



REGRESSION ANALYSIS

01

Using logistic regression to predict the likelihood of a company reaching its quarterly sales target based on advertising spend.



02

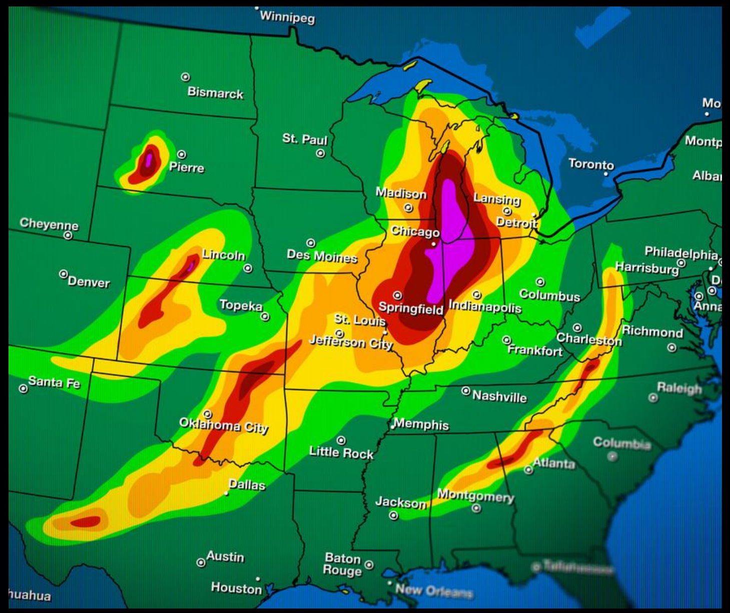
Applying logistic regression to model the probability of a security breach based on **detected anomalies** in network traffic.

03

Predicting the **likelihood** of a patient having a specific disease based on their medical history and test results.

04

Climate Predictions: Modeling **temperature changes** over time to predict climate trends.



LEARN MORE

01 Support Vector Machines (SVMs)

Relies on geometry and optimization principles, separating classes by maximizing the margin between them, which can be thought of as a form of hypothesis testing for decision boundaries.

02 Random Forests

Uses concepts of resampling (bootstrap) and descriptive statistics like majority voting or averaging for classification or regression, making it a robust ensemble model.

03 Gradient Boosting Machines (e.g., XGBoost, LightGBM)

Extends residual analysis from regression, iteratively improving predictions by minimizing the residuals (errors) in a stepwise manner.

04 Neural Networks

Builds on linear regression and probability by modeling complex, non-linear relationships between inputs and outputs and using probability-based loss functions like cross-entropy.

05 Hidden Markov Models (HMMs)

Grounded in probability theory, specifically the laws of conditional probability and Bayes' theorem, for modeling time-series or sequential data.

06 Bayesian Networks

Extensively uses Bayes' theorem and conditional independence to model probabilistic relationships between variables in a network structure.

Time for some MAGIC tricks!

ERROR

69%	80%	93%
Accuracy	Accuracy	Accuracy

Schools of thought for Statistics

- **Frequentist (Classical)**:

  Probability is the long-run frequency of events in repeated experiments.

- **Bayesian**:

  Probability represents a degree of belief or uncertainty, updated with new evidence.

- **Likelihoodist** :

  The likelihood function measures how well models explain observed data.

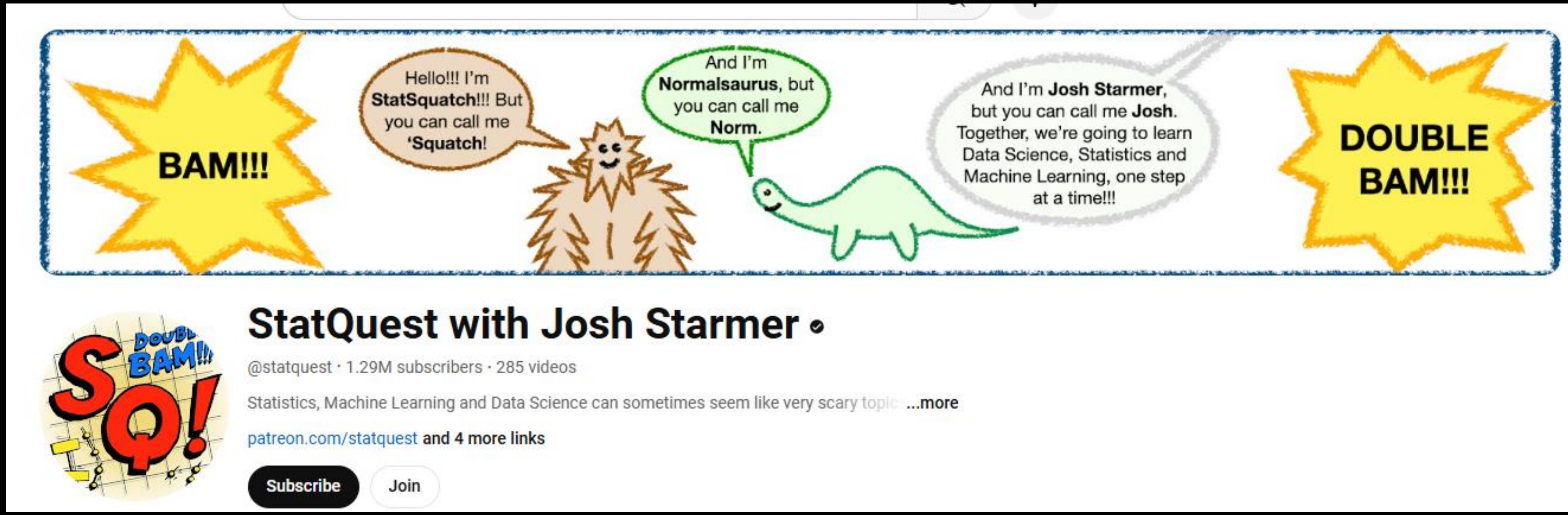
- **Fisherian** :

  Emphasizes experimental design and significance testing for measuring evidence.

- **Nonparametric** :

  Avoids strong assumptions about the underlying population distribution.

RESOURCES - Web



Josh Starmer

StatQuest

Matt Dancho

Business Science

Business Science

RESOURCES - Books

Mathematics for ML

https://mml-book.github.io/

Statistical Learning (R/Python)

https://www.statlearning.com/

JOIN THE CONVERSATION

EGjika

https://github.com/EGjika/

Google Sites

https://sites.google.com/view/eralda-gjika-dhamo/bio

https://www.linkedin.com/in/eralda-gjika-71879128/

THANK YOU

ERALDA GJIKA (DHAMO)
DECEMBER 2024