Minor in AI

18 Feb 2025
Class Notes

From Chances to Choices: Mastering Probability Distributions & Hypothesis Testing

Syllabus:

- Understand Probability Distributions: Explain different types of probability distributions (e.g., uniform, normal, binomial, Poisson) and their real-world applications.
- Identify Key Properties: Describe the characteristics (mean, variance, skewness) of common probability distributions.
- Formulate Hypotheses: Differentiate between null and alternative hypotheses in statistical testing.
- Perform Hypothesis Testing: Apply basic hypothesis testing techniques (e.g., t-test, chi-square test) to make data-driven decisions.
- Interpret Results: Analyze p-values and confidence intervals to draw meaningful conclusions from statistical tests.
- Connect Concepts to AI & ML: Recognize the role of probability distributions and hypothesis testing in machine learning and data science.

Distributions:

1. Uniform Distribution

It is a distribution where all outcomes have equal probability. Every value in a given range is equally likely to occur. It is flat, constant probability across the range. Uniform sampling ensures each data point has an equal chance of being included in your sample. This distribution does not have peaks or clustering of values.

How to identify?

- Check if the values are evenly spread across the range.

Examples:

- Rolling a fair die: Each face of a fair die has an equal chance of landing face up (1/6 probability).
- Random number generators: These algorithms are designed to produce sequences of numbers that are uniformly distributed within a given range.
- Drawing a card from a well-shuffled deck: Each card in the deck has an equal chance of being drawn.
- Lotteries: In a fair lottery, each ticket has an equal chance of being the winning ticket.

Why do we need random numbers?

- It is Foundation of many algorithms. Many ML algorithms rely on randomness, whether for initializing weights in neural networks, shuffling data, or sampling. Uniform distributions are the go-to for generating those random numbers. It provides a fair starting point. You're not biasing your algorithm towards certain values from the outset.
- Weights in neural networks are often initialized with random numbers from a uniform distribution within a certain range. This helps the model learn effectively.
- When building AI models for things like traffic flow or customer behaviour, uniform distributions can help model events that have an equal chance of occurring.

2. Normal Distribution (Gaussian)

It is a bell-shaped distribution where most values cluster around a central mean, and fewer values appear as you move away from the center. It is symmetric around the mean, peaks at the mean, tails off on both sides. Most values are near the mean, extreme values are rare. The normal distribution, often called the "bell curve" or Gaussian distribution, is one of the most common probability distributions in statistics. It describes many natural phenomena where data tends to cluster around a central mean, with fewer and fewer data points occurring further away from that mean.

How to Identify:
Data follows a symmetric bell curve.

Examples:
- Test scores (most students score near the average, few score very high/low).
- Check out student height and weight data set analysis here: https://www.kaggle.com/datasets/burnoutminer/heights-and-weights-dataset
- Similar to blood pressure, heart rates in a population often show a normal distribution.
- Size of organisms: In many species, size follows a normal distribution.
- Many machine learning algorithms assume that the data follows a normal distribution, or that the errors in the model are normally distributed. This assumption simplifies the mathematics and makes it easier to analyze the model. For example, linear regression models often assume that the errors are normally distributed.
- Understanding normal distributions helps in data analysis, feature engineering, model building, and interpreting results.
- To simulate in spreadsheet: =NORM.INV(RAND(), mean, standard_dev)

3. Binomial Distribution

Number of successes in n trials of a yes/no experiment. **Example**: Flipping a coin 10 times and counting heads.

When is it used?

The binomial distribution is used in many situations where you have repeated independent trials with two possible outcomes:

Key Differences from Normal Distribution:

- **Discrete:** The binomial distribution is discrete, meaning the number of successes can only be whole numbers (0, 1, 2, ...). The normal distribution is continuous.
- **Shape:** The binomial distribution can be skewed, especially if p is close to 0 or 1. As n gets larger and p is closer to 0.5, it starts to look more like a normal distribution.

A factory produces light bulbs. The probability of a bulb being defective is 0.02. You take a sample of 50 bulbs. What's the probability that exactly 2 are defective?

In a spread sheet cell, type =BINOM.DIST(2, 50, 0.02, FALSE) – refer video for more reference

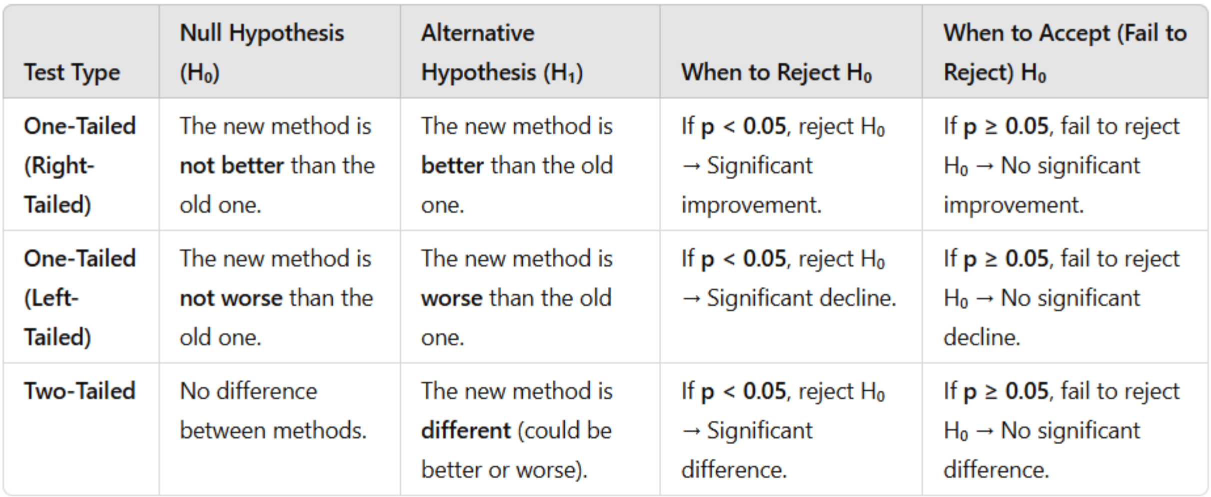
4. Poisson distribution

The Poisson distribution is a discrete probability distribution that describes the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known average rate and independently of the time since the last event. It's useful for modelling rare events. Here are some examples:

Examples:
- Number of phone calls received by a call center per hour: If a call center receives an average of 10 calls per hour, the Poisson distribution can be used to calculate the probability of receiving, say, 15 calls in an hour, or 5 calls in an hour.
- Number of cars passing a specific point on a highway per minute: Traffic flow can often be modeled using the Poisson distribution. If the average is 20 cars per minute, you can calculate the probability of 25 cars passing in a minute, or only 10.
- Number of customers entering a store per hour: Similar to call centers, customer arrivals at a store can often be modeled using the Poisson distribution.
- Number of typos on a page: If a typist makes an average of 2 typos per page, the Poisson distribution can be used to calculate the probability of a page having 0 typos, 1 typo, 3 typos, etc.
- Word Count Distributions: The distribution of word counts in a document or corpus can sometimes be approximated by a Poisson distribution, especially for less frequent words. This can be useful in tasks like text classification or topic modeling.
- Spam Detection: The number of certain keywords or phrases in an email (which might be relatively rare) could be modeled using a Poisson distribution. A high count of such keywords might suggest spam.

Hypothesis and Tests:

Refer to the video on how we built the premise for it.



Test Type	Null Hypothesis (H_0)	Alternative Hypothesis (H_1)	When to Reject H_0	When to Accept (Fail to Reject) H_0
One-Tailed (Right-Tailed)	The new method is **not better** than the old one.	The new method is **better** than the old one.	If **$p < 0.05$**, reject H_0 → Significant improvement.	If **$p \geq 0.05$**, fail to reject H_0 → No significant improvement.
One-Tailed (Left-Tailed)	The new method is **not worse** than the old one.	The new method is **worse** than the old one.	If **$p < 0.05$**, reject H_0 → Significant decline.	If **$p \geq 0.05$**, fail to reject H_0 → No significant decline.
Two-Tailed	No difference between methods.	The new method is **different** (could be better or worse).	If **$p < 0.05$**, reject H_0 → Significant difference.	If **$p \geq 0.05$**, fail to reject H_0 → No significant difference.

What Does the 5% Significance Level ($\alpha = 0.05$) Mean?

The significance level (α) is the threshold we set to decide when to reject H_0.

Why 5%?

- $\alpha = 0.05$ means we allow a 5% chance of mistakenly rejecting H_0 (Type I Error).
- In other words, we accept a 5% risk of saying "something is happening" when actually it's just random noise.

If $p < 0.05$ (below the threshold):

- The observed result is so rare under H_0 that we reject H_0 and say there's likely a real effect.

If $p \geq 0.05$ (above the threshold):

- The observed result is not rare enough → Fail to reject H_0 (not enough evidence to prove an effect).

Definitions:

Hypothesis Testing

Hypothesis testing is a way to check if a claim about data is true. It compares observed results with what we expect. If the difference is significant, we reject the initial assumption. It helps in decision-making by analyzing whether an effect is real or just due to chance.

Confidence Interval

A confidence interval gives a range of values where we expect the true value to be. For example, a 95% confidence interval means we are 95% sure the actual value falls within that range. It helps estimate uncertainty in data and is commonly used in statistics.

P-Value

The p-value tells us how likely it is to see our results if the initial assumption (null hypothesis) is true. A small p-value (like below 0.05) suggests strong evidence against the null hypothesis, meaning the observed effect is likely real and not just due to random chance.

T-Test

A t-test is used to compare the means of two groups to check if they are significantly different. It helps in determining if a change or difference between groups happened due to chance or an actual effect. It's commonly used in experiments and research studies.

Null Hypothesis

The null hypothesis assumes no real effect or difference exists. It is the starting point of hypothesis testing, stating that any observed changes are due to chance. If data strongly contradicts it, we reject the null hypothesis and accept an alternative explanation.

Alternate Hypothesis

The alternate hypothesis is the opposite of the null hypothesis. It suggests there is a real effect, difference, or relationship between variables. If enough evidence is found against the null hypothesis, we accept the alternate hypothesis, indicating the observed change is meaningful.

Note:

Refer to video for more explanation.

The class also referred to

- NetLogo simulation tool
- Jamovi for statistical analysis