Lecture 12

HYPOTHESIS TESTING

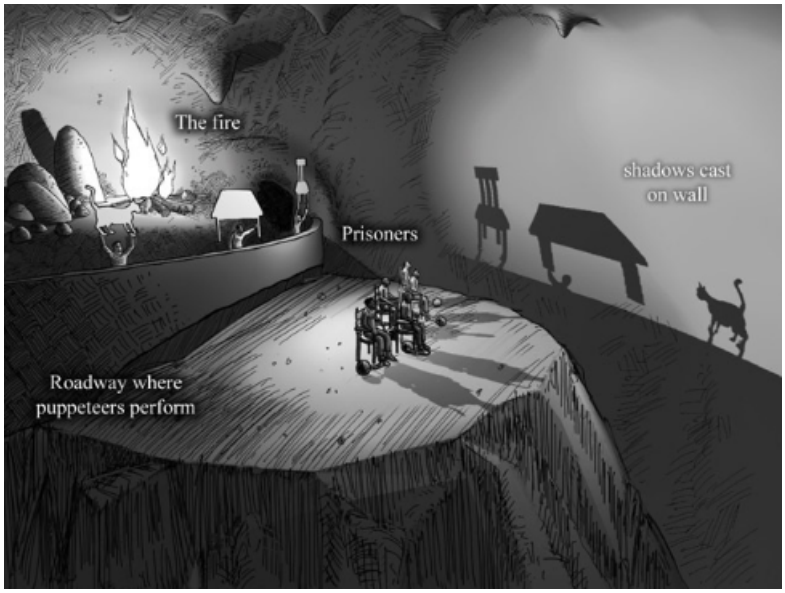
Probability & Statistics

Francisco Richter, Martina Boschi and Ernst Wit

Making informed decisions is a ubiquitous aspect of life, confronting us with questions such as

- "Does a new drug improve health outcomes?"
- "Is a suspect guilty of a crime?"
- "Do environmental factors contribute to disease?"
- "Has a chess grandmaster player engaged in cheating?"

To address these varied queries, a robust statistical framework is essential. Hypothesis testing is the cornerstone of this framework, offering a systematic approach to drawing conclusions from data. This lecture will explore the principles and applications of hypothesis testing, demonstrating its critical role in deciphering data to make informed decisions.



In this exploration, we will see how hypothesis testing acts as a logical tool, akin to a statistical tautology, enabling us to validate or invalidate hypotheses based on empirical evidence. The allegory of Plato's Cave serves as a metaphor for this process, where just like deciphering shadows to uncover a hidden reality, we use data to reveal underlying truths.

1 Hypothesis testing

Statistical hypothesis testing is a key technique of frequentist statistical inference. In order to understand what hypothesis testing means, we describe an analogy. A statistical test procedure is comparable to a criminal trial; a defendant is considered innocent as long as his guilt is not proven. The prosecutor tries to prove the guilt of the defendant. Only when there is enough incriminating evidence the defendant is convicted.

1.1 Modus tollens

The *modus tollens* is the Latin name for a valid argument form and rule of inference. It means "the way that denies by denying". It is the inference that

1. if $P \Rightarrow Q$, i.e. if P implies Q,

2. and the second premise, Q, is false,

3. then it can be logically concluded that P must be false.

For example, we know that if it rains, that then the streets are wet. When we look outside, we see that the streets are dry, so therefore we can conclude that it doesn't rain. The argumentation used in hypothesis testing is something like a *modus tollens*. The only thing is that we are not completely certain about the truth or falsehood of $P \Rightarrow Q$ anymore. We explain this in the following sections.

1.2 Null and alternative hypotheses

At the start of the procedure, there are two hypotheses:

H_0: "the defendant is not guilty", and
H_1: "the defendant is guilty".

The first one is called the *null hypothesis*, and is *for the time being* accepted. The second one is called the *alternative (hypothesis)*. It is the hypothesis the prosecutor tries to prove.

1.3 Two kinds of errors

The hypothesis of innocence is only rejected when the evidence is really strong, because one doesn't want to convict an innocent defendant. The error of convicting an innocent person is called an *error of the first kind*. The testing procedure should be set up in such a way that the occurrence of this error is rare. The reverse error, i.e. acquitting a person who committed the crime, called the *error of the second kind* is considered less serious. As a consequence of this asymmetric behaviour, the frequency of occurrence of the second type error is often rather large.

	H_0 is true truly not guilty	H_1 is true truly guilty
Accept Null Hypothesis Acquittal	Right decision	Wrong decision Type II Error
Reject Null Hypothesis Conviction	Wrong decision Type I Error	Right decision

- **Type I Error (False Positive)**: Occurs when the null hypothesis is true, but we incorrectly reject it. Denoted by α (alpha), also known as the significance level.
- **Type II Error (False Negative)**: Occurs when the null hypothesis is false, but we fail to reject it. Denoted by β (beta). The power of a test is defined as $1-\beta$ and represents the probability of correctly rejecting a false null hypothesis.

1.4 Test statistic

A criminal trial can be regarded as either or both of two decision processes: guilty vs not guilty or evidence vs a threshold ("beyond a reasonable doubt"). What we wish is to judge the defendant, but really we are judging the performance of the prosecution (which bears the burden of proof). Similarly in hypothesis testing, these decisions are always made using some (hopefully) clever summary of the data. This summary is called a *test-statistic*. There are many summaries of the data possible: one might show nothing (e.g. "the suspect is a woman"), whereas another (e.g. "the suspect has blood-stained clothes") might give some idea that the that the suspect is guilty.

1.5 P-value

We now attempt to reformulate the *modus tollens*, where

- P $=$ "null hypothesis is true"
- Q $=$ "not observing this value of the test statistic"

If $P \Rightarrow Q$ was certain, then if the null hypothesis is true, one would *never* observe this value of the test statistic. BUT we do observe exactly this value of the test-statistic, and therefore the null hypothesis must be false.

However, often $P \Rightarrow Q$ is not certain at all. If the suspect is not guilty, it is not certain that she has no blood on her clothes. She could have tried to help the victim immediately afterwards. We want to replace the certain statement,

$$P \Rightarrow Q$$

by an uncertain statement: "the probability that Q happens if P is true", $P(Q|P)$. If this value is almost 1, then the implication is almost certain and then we can tolerate the *modus tollens* argument. However, if this is far from 1, then the implication is too uncertain to be able to use the *modus tollens*.
In actual fact, one does not use $P(Q|P)$, but $P(\text{not } Q|P) = 1 - P(Q|P)$. This value is called the *p-value*, i.e.,

> **p-value**: probability of observing a value for the test statistic that is at least as extreme as the value that was actually observed, assuming that the null hypothesis is true.

P-values close to zero, therefore, mean that the implication $P \Rightarrow Q$ is almost certain and that we can use the *modus tollens* to reject the null hypothesis.

1.6 Significance level

One use of hypothesis testing is deciding whether experimental results contain enough information to cast doubt on conventional wisdom. The question is what is enough? Should the probability of the implication be 99% (i.e. the p-value = 0.01) or 99.9% (i.e. the p-value=0.001) before we accept the validity of the *modus tollens*? In most fields of science and society, one uses 95% as the cut-off – i.e. a cut-off of 0.05 for the p-value). This is the result of an arbitrary choice by the inventor of hypothesis tests, R.A. Fisher, in the 1930s. The cut-off is called the *significance level*.
In statistics, a result is called *statistically significant* if it is unlikely to have occurred by chance alone, according to a pre-determined threshold probability, the significance level. The phrase "test of significance" was coined by Ronald Fisher: "Critical tests of this kind may be called tests of significance, and when such tests are available we may discover whether a second sample is or is not significantly different from the first."
A result that was found to be statistically significant is also called a positive result; conversely, a result that is not unlikely (i.e. quite likely) under the null hypothesis is called a negative result or a null result.
The probabilities of the two types of errors are determined by the distributions of the test statistic under the null and alternative hypotheses.
If Z is our test statistic:

- The probability of a Type I error, $P(\text{Reject } H_0|H_0 \text{ is true}) = \alpha$.
- The probability of a Type II error, $P(\text{Fail to Reject } H_0|H_0 \text{ is false}) = \beta$.

Given a significance level α, the critical value Z_α is the value for which:

$$P(Z > Z_\alpha|H_0 \text{ is true}) = \alpha$$

And for the power of the test:

$$P(Z > Z_\alpha|H_1 \text{ is true}) = 1 - \beta$$

In hypothesis testing, we aim to minimize both Type I and Type II errors. However, in practice, there is a trade-off:

$$\text{As } \alpha \text{ decreases, } \beta \text{ increases (and vice versa).} \tag{1}$$

Thus, the goal is often to determine an acceptable level of α (common choices are 0.05, 0.01) and design the test to minimize β for that α. Another approach is to maximize the power of the test, which is $1 - \beta$, given a fixed α.
The choice of α should reflect the relative seriousness of the two types of errors in a particular testing context. For example, in medical testing, a Type I error (falsely diagnosing a healthy person as sick) might be deemed more serious than a Type II error (falsely diagnosing a sick person as healthy), leading to a choice of a smaller α.

Example: A Clairvoyant Card Game

Lets consider a more realistic example. A person (the subject) is tested for clairvoyance. He is shown the back of a randomly chosen playing card 25 times and asked to guess its suit. The number of correct guesses (hits) is denoted by X.

As we seek evidence of clairvoyance, we start by assuming the null hypothesis that the person is not clairvoyant. The alternative hypothesis is that the person is (to some degree) clairvoyant. Under the null hypothesis, the subject can only guess. For each card, the probability of any single suit is $1/4$. Under the alternative hypothesis, the probability of a correct guess, p, differs from $1/4$. Thus, the hypotheses are:

$H_0 : p = 1/4$ (just guessing)
$H_1 : p \neq 1/4$ (true clairvoyance).

We call this a two-sided test, meaning that *a priori* it is logically possible for p to be either greater than or less than $1/4$. This is typically the case in practical situations, so most tests are two-sided. The main consequence of a two-sided test is in the calculation of the p-value: the exceedance probabilities are multiplied by 2 (see below).

If the test subject correctly predicts all 25 cards, it seems reasonable to reject the null hypothesis and consider him clairvoyant. With only 5 or 6 hits, however, there is no reason to doubt the null hypothesis. But what about 12 hits, or 17 hits? We must determine the critical number c of hits at which we consider the subject to be clairvoyant. How do we determine this critical value c?
In practice, one decides how strict to be by choosing how often one is willing to make a Type I error (a false positive). For example, with $c = 25$:

$$\text{Type I error} = 2 \times P(X = 25 \mid H_0 \text{ is true}) = 2 \times \left(\frac{1}{4}\right)^{25}.$$

This probability is extremely small.
The probability of a false positive is thus the probability of randomly guessing all 25 cards correctly, times 2 (since the test is two-sided). If we are less strict and choose $c = 10$, we have:

$$\text{Type I error} = 2 \times P(X \geq 10 \mid H_0 \text{ is true}) = 2 \times \sum_{k=10}^{25} \binom{25}{k} (0.25)^k (0.75)^{25-k} \approx 0.14.$$

So, $c = 10$ yields a much larger probability of a false positive.
Before conducting the test, one typically decides on a desired probability of a Type I error. Common values range from 1% to 5%. Depending on this chosen error rate, we calculate c. For example, if we select an error rate of 1%, then c is determined by:

$$\text{Type I error} = 2 \times P(X \geq c \mid H_0 \text{ is true}) \leq 0.01.$$

From all c values satisfying this inequality, we choose the smallest one. Why? Because remember there is another error, the Type II error (false negative), which we also want to minimize. In our example, this leads us to select $c = 12$.
What if the subject did not guess any cards correctly at all? Getting zero hits can also be considered unusual. The probability of guessing incorrectly once is $q = 1 - p = 3/4$. Using the same reasoning, the probability of calling all 25 cards wrong is:

$$\text{p-value} = 2 \times P(X = 0 \mid H_0 \text{ is true}) = 2 \times (0.75)^{25} \approx 0.0014.$$

This is about a 1 in 1000 chance and would be evidence for dismissing H_0 in favor of H_1. While it may seem strange, this result suggests the subject consistently avoids guessing correctlyalso a form of clairvoyance.

Example: Power Plant's Effect on Cancer Incidence

We aim to determine if a power plant influences the incidence of a specific cancer, measurable through a continuous scale like biomarker levels. Our linear regression model is:

$$Y_i = \beta_0 + \beta x_i + e_i$$

Here, Y_i represents cancer severity or biomarker levels, x_i indicates proximity to the power plant, and e_i is the error term. The coefficient β measures the power plant's effect.
Our hypothesis test is:

- $H_0 : \beta = 0$ (No effect)
- $H_1 : \beta \neq 0$ (Some effect)

The least squares estimator for β is:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

A possible test statistic T is defined as:

$$T = \frac{\hat{\beta} - \beta_0}{\text{SE}(\hat{\beta})}$$

If H_0 is true, the p-value is:

$$\text{p-value} = P(|T| \geq |t_{obs}|)$$

Where T is the t-statistic under H_0, and t_{obs} is the observed value.
The Central Limit Theorem ensures that, even if e_i is not normally distributed, $\hat{\beta}$ will approximate a normal distribution for large samples, allowing the use of standard t-tests.

Permutation Test

A permutation test is a non-parametric statistical method used for hypothesis testing, particularly to determine if two or more samples originate from the same distribution. The process is delineated as follows:

1. The null hypothesis H_0 posits no effect or difference between the groups, e.g., equal means.
2. All observations from the different groups are pooled together, creating a comprehensive pool of data.
3. This pooled data is then randomly permuted and redistributed into new groups, each matching the size of the original groups.
4. For each permutation, a test statistic is calculated. This statistic could be any measure of interest, depending on the research question:

$$T = f(Y^{(1)}) - f(Y^{(2)})$$

5. This process is repeated numerous times (e.g., 10,000 times) to generate a distribution of the test statistic under H_0.
6. The actual test statistic calculated from the original, unpermuted data is then compared to this distribution.
7. The p-value is determined by the proportion of permuted test statistics that are as extreme or more extreme than the observed statistic:

$$p\text{-value} = \frac{\text{Number of permutations with } T \geq \text{ observed } T}{\text{Total number of permutations}}$$

A low p-value (typically less than 0.05) indicates that the observed difference is unlikely due to chance, leading to a rejection of H_0. This test is valued for its versatility and robustness, as it does not make assumptions about the distribution of the data.

Example: Teaching methods

Consider the scenario of evaluating two teaching methods in terms of their effectiveness. Our objective is to use a hypothesis testing framework to ascertain if the observed differences in student pass rates between two schools are statistically significant.
Hypotheses:

- Null Hypothesis (H_0): There is no significant difference in the pass rates between the two schools.
- Alternative Hypothesis (H_1): There is a significant difference in the pass rates between the two schools, attributable to the teaching methods.

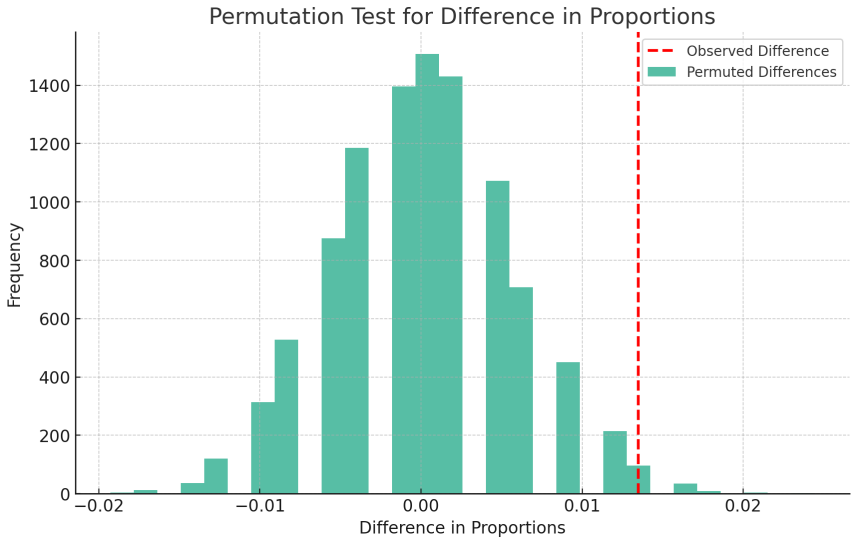


Figure 1: Histogram of permuted differences in pass rates with the observed difference indicated.

The data includes results from two schools: School A, employing a new teaching method with 773 students (of whom 18 passed), and School B, using a traditional method with 1123 students (of whom 11 passed). The permutation test, chosen for its robustness in non-parametric settings, is used to compare the pass rates.
The permutation test produced an observed difference in pass rates of approximately 0.0135. The p-value for this observed difference is 0.0398. Since this p-value is below the conventional threshold of 0.05, we reject the null hypothesis in favor of the alternative hypothesis. This result suggests that the difference in pass rates between the two schools is statistically significant and may be attributed to the difference in teaching methods.

Example: Is Nakamura Cheating?

Following recent allegations of cheating by Vladimir Kramnik against Hikaru Nakamura, this analysis aims to statistically evaluate Nakamura's winning streaks. Kramnik's accusations, made in November 2023, were based on Nakamura's exceptional performance, including a 45.5 out of 46 win rate. Chess.com's investigation found no evidence of cheating, but we aim to independently assess the statistical likelihood of such performance (source).
The hypotheses tested are:

- H_0: The occurrence of long winning streaks is due to chance.
- H_1: The occurrence of long winning streaks is too improbable to be due to chance, indicating potential cheating.

Using the ELO rating system, the probability of Nakamura winning a game is calculated, taking into account his high performance in recent games:

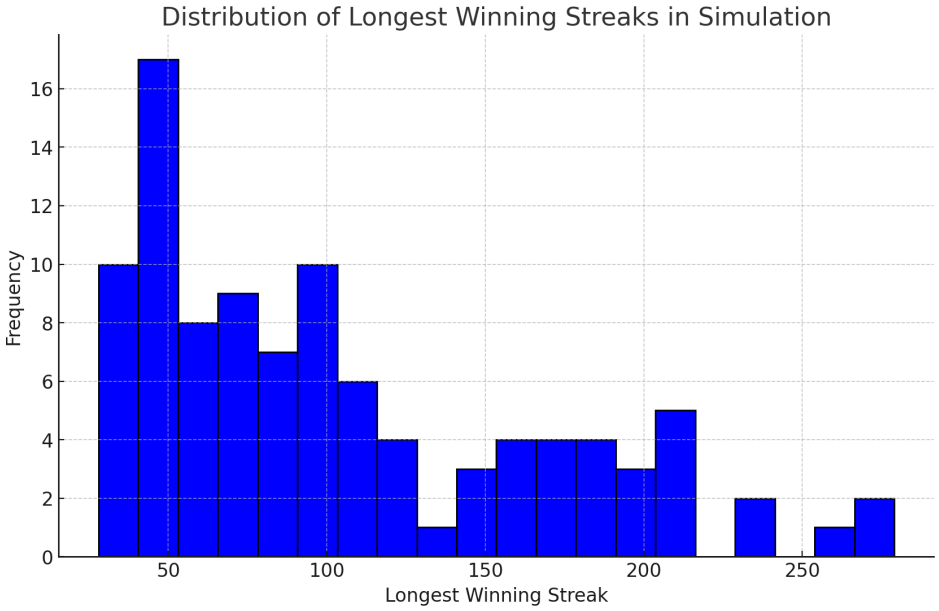
$$P(\text{Win}) = \frac{1}{1 + 10^{(\text{Opponent ELO} - \text{Nakamura ELO})/400}}$$

The simulation parameters are as follows:

- Number of games per simulation: 25,000
- Number of simulations: 100
- Nakamura's ELO (adjusted for recent performance): 3300
- Opponent's ELO range: 2700 to 3100
- Threshold for significant winning streak: 40 games

The key statistic in this analysis is the longest winning streak. The simulation results, which are analyzed to understand the likelihood of Nakamura's winning streaks occurring by chance, include the mean and median longest winning streaks, as well as the frequency of simulations with winning streaks above the 40-game threshold.

A p-value is calculated to determine the statistical significance of the results, comparing the observed longest winning streak against the distribution from the simulations. The conclusion will be based on whether this p-value is lower than the typical significance level of 0.05.
A histogram is included to visually represent the distribution of the longest winning streaks in the simulations, providing a clear visual interpretation of the data.



2 Decision Theory Principles

Decision theory is an interdisciplinary study involving psychology, economics, statistics, and philosophy, focusing on the nuances of decision-making, particularly in situations of uncertainty. The primary goal of this field is to dissect the decision-making process and enhance its effectiveness. At the heart of decision theory lies the concept of expected utility (EU), which represents the cumulative utility of all potential outcomes, each weighted by its probability.

2.1 Theory Foundations

Definition 1 (Expected Utility)**.** *The Expected Utility (EU) is formulated by the equation:*

$$EU = \sum_{i=1}^{n} p_i u(x_i), \tag{2}$$

where p_i represents the probability of outcome x_i, and $u(x_i)$ is the utility correlated with that outcome.

The utility function $u(x)$, pivotal in decision theory, quantifies an individuals preferences among different outcomes. Consider a decision scenario such as choosing between receiving \$100 today or \$110 in a year. The utility function captures the individual's preference for immediate versus delayed rewards, thus influencing the decision-making process.

Example. Consider an individual facing the decision to invest in a stock. The utility function $u(x)$ could represent the satisfaction gained from different monetary outcomes. Let's say the stock has a 50% chance of doubling the investment (gain) and a 50% chance of losing half of it (loss). If the individual's investment is \$100, the expected utility could be calculated as follows, assuming a simple utility function where the utility is equal to the monetary value:

$$\begin{aligned} EU &= 0.5 \times u(\$200) + 0.5 \times u(\$50) \\ &= 0.5 \times \$200 + 0.5 \times \$50 \\ &= \$125. \end{aligned}$$

□

In this framework, we define loss as the inverse of utility.

Definition 2 (Loss)**.** *The loss associated with an outcome is the negative utility, defined as:*

$$L(x_i) = -u(x_i). \tag{3}$$

These foundational elements are crucial in statistical decision theory, guiding the pursuit of optimal decisions. They find applications across diverse fields, from public policy analysis to balancing potential benefits and costs in various sectors.

One of the earliest applications of decision theory, which is closely related to game theory, is Pascal's Wager. The French mathematician Blaise Pascal proposed a decision problem to determine whether one should believe in God. The wager is simple: either God exists or He doesn't. Pascal argues that believing in God is the most rational choice, given the possible outcomes.

	God Exists	**God Doesn't Exist**
Believe in God	Infinite reward (Heaven)	Finite loss (wasted time in worship)
Don't Believe in God	Infinite loss (Hell)	Finite gain (time saved)

To quantify the utilities in Pascal's Wager, let's assign symbolic values to the outcomes. Let U_{Heaven} represent the utility of achieving Heaven, U_{Hell} for Hell, U_{Worship} for the finite loss due to worship, and $U_{\text{Time Saved}}$ for the finite gain of time saved by not worshipping. Assuming the existence and non-existence of God are equally probable, let p represent the probability, such that $p = 0.5$.

Calculating Expected Utilities

1. *Believing in God:* The expected utility for believing in God, denoted as EU_{Believe}, is a combination of the utility of going to Heaven and the loss due to worship. Thus, it is given by:

$$EU_{\text{Believe}} = p \times U_{\text{Heaven}} + (1-p) \times U_{\text{Worship}}. \tag{4}$$

2. *Not Believing in God:* Similarly, the expected utility for not believing in God, $EU_{\text{Not Believe}}$, is the combination of the utility of Hell and the gain of time saved. This is expressed as:

$$EU_{\text{Not Believe}} = p \times U_{\text{Hell}} + (1-p) \times U_{\text{Time Saved}}. \tag{5}$$

Assigning Values to Utilities

To proceed with the calculations, we need to assign values to the utilities. Given the nature of Pascal's Wager, we assign:

- $U_{\text{Heaven}} = +\infty$ (representing an infinite reward),
- $U_{\text{Hell}} = -\infty$ (representing an infinite loss),
- U_{Worship} and $U_{\text{Time Saved}}$ as finite values, where $U_{\text{Worship}} < U_{\text{Time Saved}}$ since the loss due to worship is considered less significant than the gain of time saved.

Evaluating the Decision

Given these values, the expected utilities can be evaluated:

- For EU_{Believe}, any term multiplied by $+\infty$ (representing the infinite utility of Heaven) will dominate the equation. Therefore, EU_{Believe} effectively becomes $+\infty$.
- Similarly, for $EU_{\text{Not Believe}}$, the term involving $-\infty$ (the infinite loss of Hell) dominates, making $EU_{\text{Not Believe}} = -\infty$.

In Pascal's Wager, from a purely utilitarian perspective, the decision to believe in God offers a higher expected utility due to the infinite reward of Heaven. Much like hypothesis testing, the cost of making an incorrect decision (Type I or Type II errors) must be weighed. In hypothesis testing, the consequences of incorrectly rejecting a true null hypothesis (Type I error) versus failing to reject a false null hypothesis (Type II error) are considered.

2.2 Practical Decision Applications

2.2.1 Climate Change

In contemporary discussions, Pascal's Wager finds a parallel in the debate over climate change. The dilemma of whether to acknowledge and address climate change or to dismiss it can be examined through a similar analytical lens.

	Climate Change is Real	**Climate Change Isn't Real**
Act to Combat It	Avert catastrophic outcomes	Unnecessary expenditure of resources
Do Nothing	Risk grave consequences	Conservation of resources

This decision matrix contrasts the outcomes of action versus inaction on climate change, considering global impacts, economic ramifications, and the well-being of future generations.
In this scenario, the utilities are defined as follows:

- U_{Action} symbolizes the utility derived from combating climate change, potentially preventing dire consequences.
- U_{Inaction} represents the utility of not engaging in climate action, which might include short-term economic gains or convenience.
- U_{Adverse} stands for the negative utility associated with the severe impacts of unchecked climate change.
- $U_{\text{Redundant}}$ denotes the utility lost through futile actions if the severity of climate change is overestimated.

Considering the decision to act against climate change, the expected utility, EU_{Action}, is calculated. It balances the benefit of positive action, U_{Action}, against the cost of potentially unnecessary efforts, $U_{\text{Redundant}}$. This is mathematically expressed as:

$$EU_{\text{Action}} = p \cdot U_{\text{Action}} + (1 - p) \cdot U_{\text{Redundant}}, \tag{6}$$

where p is the probability of the adverse effects of climate change materializing.
Alternatively, the decision to refrain from climate action involves calculating the expected utility of inaction, EU_{Inaction}. This utility combines the immediate benefits of inaction, U_{Inaction}, with the long-term risks of adverse effects, U_{Adverse}. The formula for this utility is:

$$EU_{\text{Inaction}} = p \cdot U_{\text{Adverse}} + (1 - p) \cdot U_{\text{Inaction}}. \tag{7}$$

These formulations provide a structured approach to evaluating the complex decision-making process related to climate change, accounting for both immediate and future impacts.

2.2.2 Medical Decision-Making

In medical decision-making, doctors must often decide on treatment plans based on diagnostic tests and patient data. These decisions can be modeled using decision theory to optimize patient outcomes while minimizing risks.

Decision	**State (θ_1): Patient is sick**	**State (θ_2): Patient is not sick**	**Loss Function**
D_1: Treat	Low loss	Moderate loss	$L(\theta, D_1)$
D_2: Do not treat	High loss	No loss	$L(\theta, D_2)$

Table 1: Decision Table for Medical Treatment Scenario

In this table, the decisions D_1 and D_2 represent treating and not treating the patient, respectively. The loss associated with each decision depends on the actual health state of the patient (θ).
Consider two decision rules:

1. Rule 1 (R1): Treat if the test is positive, otherwise don't.
2. Rule 2 (R2): Always treat, regardless of the test.

In terms of effectiveness:

- R1 could be effective if the test is reliable. It balances treating the sick and avoiding unnecessary treatments.

- R2 is less effective if the test is reliable, as it leads to unnecessary treatments. R1 is a better rule in this case.

An effective rule is one where no other rule is always better. R1 could be effective with a good diagnostic test, while R2 is less so due to its inefficiency.

2.3 Decision Making Under Uncertainty

In statistical decision theory, we explore how to make the best choices under conditions of uncertainty. This involves understanding decisions d based on data x and considering unknown parameters θ. The key is to assess the effectiveness of these decisions, which is done using a loss function $L(\theta, d)$. This function measures the cost or penalty of making decision d when the true situation is θ.

Definition 3. *The* risk function *in decision theory, denoted as $R(\theta, d)$, is defined as the expected loss associated with a decision d when the true state of nature is θ. Mathematically, it is expressed as:*

$$R(\theta, d) = E_\theta[L(\theta, d)], \tag{8}$$

where E_θ represents the expectation under the probability distribution of the observed data x, given the parameter θ.

Our goal is to find a decision rule that reduces this risk function as much as possible, regardless of the actual value of θ.

To properly set up a statistical decision problem, we need to consider:

- The probability model P_θ, describing data distribution.
- The decision space D, which includes all possible decisions.
- The loss function $L(\theta, d)$, representing the cost of each decision.

A decision rule is effective if there's no other rule that always results in a lower risk across all possible θ. This means the rule is rational and efficient under the given assumptions of the model.

Now consider a medical treatment model, where we consider a matrix **X** of explanatory variables, and a binary 'action' variable indicating treatment decisions (1 for treatment, 0 for no treatment). This action is separate from the variables in **X**.
The matrix **X** and the action variable are arranged as follows:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}, \tag{9}$$

$$\text{action} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}, \tag{10}$$

where each a_i is either 0 or 1, indicating the treatment decision for each case.
We define a new observation vector $\mathbf{x_{new}}$ from **X** and an action value a_{new} for this new case:

$$\mathbf{x_{new}} = \begin{bmatrix} x_{\text{new}1} & x_{\text{new}2} & \cdots & x_{\text{new}n} \end{bmatrix}, a_{\text{new}} = \text{binary value (0 or 1).} \tag{11}$$

In our regression model, we aim to predict the decision variable Y by considering both the new observations $\mathbf{x_{new}}$ and the impact of the treatment decision a_{new}. The model is expressed as:

$$\hat{y} = \mathbf{x_{new}}\hat{\beta} + a_{\text{new}}\hat{\beta}_{\text{action}}, \tag{12}$$

where $\hat{\beta}$ are the estimated coefficients for the explanatory variables **X**, and $\hat{\beta}_{\text{action}}$ is the estimated coefficient reflecting the effect of the treatment decision. This model provides insights into how the combination of observed data and treatment decision influences the decision variable Y.