Chapter 1

Sampling

Objectives

After studying this chapter, you should:

- Equip students with a comprehensive understanding of sampling and its critical role in statistical analysis.
- Foster analytical skills necessary for evaluating sampling techniques, avoiding common pitfalls, and applying statistical methods effectively.
- Prepare students to conduct independent research by providing the foundational knowledge required for proper sampling and data analysis.

Introduction

Sampling is a fundamental aspect of statistical analysis, serving as a bridge between theoretical concepts and practical applications. In many research scenarios, it is often impractical or impossible to collect data from an entire population due to constraints such as time, cost, and accessibility. Consequently, researchers rely on samples—subsets of the population—to draw conclusions and make inferences about the larger group.

This chapter provides a comprehensive overview of sampling techniques and their significance in statistical methodology. We will explore various sampling methods, including random, stratified, and systematic sampling, each with its advantages and challenges. Understanding these methods is crucial for minimizing bias and ensuring that samples accurately represent the population.

Furthermore, we will discuss common pitfalls that researchers may encounter when sampling and strategies to avoid them, enhancing the integrity of the data collected. Key statistical concepts related to samples, such as empirical mean and variance, will also be examined to equip students with the necessary tools to analyze and interpret sample data effectively.

By the end of this chapter, students will gain a solid understanding of sampling principles, enabling them to conduct independent research with confidence and precision. The knowledge acquired will not only enhance their analytical skills but also prepare them to navigate the complexities of data collection and analysis in various fields of study.

1.1 Notions of samples

Sampling is a fundamental concept in statistics, and it's essential to understand how it works, why it's used, and how to implement it. Here are more details about sampling, along with practical examples:

Why Sampling is Used: Sampling is used in statistics for several reasons:

1. **Cost-Efficiency:** Collecting data from an entire population can be time-consuming and expensive. Sampling allows researchers to gather a smaller, more manageable set of data points.

2. **Practicality:** In some cases, it's impossible to collect data from an entire population, especially if the population is large or widely dispersed.

3. **Destruction of Items:** When the data involves destructive testing or examination, such as in medical research, it's not feasible to collect data from the entire population.

4. **Accuracy:** When done correctly, sampling can provide accurate and representative information about the entire population.

1.1.1 Sampling Methods

There are different methods of sampling, depending on the research objectives and available resources. Some common sampling methods include:

1. Simple Random Sampling (SRS)

Simple Random Sampling (SRS) is a fundamental sampling technique used in statistics to select a subset of individuals from a larger population, ensuring that each individual has an equal chance of being chosen. This method helps obtain unbiased and representative data, which is crucial for making accurate inferences about the entire population.

Key Characteristics of SRS

- **Equal Probability**: Each member of the population has an identical chance of being included in the sample.
- **Random Selection**: Individuals are chosen randomly, often using random number generators or other unbiased selection methods.
- **Independence**: The selection of one individual does not affect the selection of another, ensuring true randomness.

Steps for Implementing SRS

1. **Define the Population**: Clearly outline the group from which you will draw your sample.
2. **Determine Sample Size**: Decide the number of individuals you need to represent the population accurately.
3. **Select Individuals Randomly**: Use methods like random number tables or software tools to ensure unbiased selection.

Example: If you have a population of 1,000 students and want a sample of 100, each student is assigned a number. A random number generator selects 100 numbers, ensuring each student has an equal likelihood of being chosen.

Solution: In this scenario, simple random sampling is employed to select a sample of students from a population of **1,000 students**. Each student is assigned a unique number from 1 to 1,000, and a random number generator is used to select 100 students. This method ensures that every student has an equal chance of being chosen. Here's a step-by-step breakdown of the process:

1. Define the Population

The population consists of **1,000 students**, each assigned a unique number from 1 to 1,000.

Assign Numbers

Each student is assigned a number as follows:

- Student 1: Number 1
- Student 2: Number 2
- Student 3: Number 3
- $\vdots$
- Student 1000: Number 1000

Randomly Select Sample

A random number generator selects **100 unique numbers** from the range of 1 to 1,000. For this example, assume that the selected numbers are:

- 5, 12, 37, 45, 78, 89, 123, 234, 345, 456, 567, 678, 789, 890, 901, 934, 987, 1000, etc.

Survey Selected Students

The researcher surveys the students corresponding to the selected numbers. Each of the 100 selected numbers represents a student that will be included in the sample.

Summary of Simple Random Sampling

Total Students	1,000
Sample Size	100
Selected Student Numbers	Randomly Generated (e.g., 5, 12, 37, ...)

Table 1.1: Simple Random Sampling Summary

This simple random sampling approach allows the researcher to obtain a sample that is representative of the entire student population, minimizing selection bias.

91	**92**	**93**	**94**	**95**	**96**	**97**	**98**	**99**	**100**
81	**82**	**83**	**84**	**85**	**86**	**87**	**88**	**89**	**90**
71	**72**	**73**	**74**	**75**	**76**	**77**	**78**	**79**	**80**
61	**62**	**63**		**65**	**66**	**67**	**68**	**69**	**70**
51	**52**	**53**	**54**	**55**	**56**	**57**	**58**	**59**	**60**
	42	**43**	**44**		**46**	**47**	**48**	**49**	**50**
31	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**
21	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
11		**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**

Highlighted Students are Selected by Random Number Generator

Advantages and Limitations:
Advantages: SRS is easy to implement and provides unbiased samples, especially with large populations.
Limitations: It may not be feasible for very large populations without proper tools, and it may require a complete list of the population.

Conclusion: Simple random sampling is widely used in fields like survey research, market analysis, and social sciences for generating reliable, generalizable insights about a population.

Stratified Sampling

Stratified Sampling is a statistical sampling technique used to obtain a sample that represents various subgroups within a population. By dividing the population into distinct strata based on specific characteristics, this method ensures that each subgroup is adequately represented in the final sample, leading to more accurate and reliable results.

Key Characteristics of Stratified Sampling

- **Division into Strata**: The population is divided into distinct subgroups

based on shared characteristics such as age, gender, income, or education level.

- **Random Sampling within Strata**: Random samples are taken from each stratum, ensuring each subgroup is represented.
- **Proportional or Equal Allocation**: Samples can be drawn in proportion to the size of each stratum in the population or with equal numbers from each stratum.

Steps for Implementing Stratified Sampling

1. **Define the Population**: Clearly outline the entire population from which samples will be drawn.
2. **Identify Strata**: Determine relevant characteristics for stratification and categorize the population accordingly.
3. **Determine Sample Size**: Decide the total number of individuals needed in the sample.
4. **Select Samples from Each Stratum**: Use random sampling methods to select samples from each identified stratum.

Example: If a university has 1500 students and you want to study their study habits, you could divide the students into three strata: freshmen, sophomores, juniors, and seniors. If you want a sample of 150 students.

Solution:In this scenario, stratified sampling involves dividing the university's **1,500 students** into strata based on their academic levels—**Licence, Master, and Doctorat**. With a desired sample size of **150**, you would draw approximately **100 students from Licence**, **40 from Master**, and **10 from Doctorat** to achieve a balanced sample. Here's a step-by-step breakdown of the process:

1. Define the Population

The population consists of **1,500 students** at a university.

Create Strata

Divide the population into three strata based on academic level Licence, Master and Doctorat

Determine the Sample Size

The goal is to select a sample of **150** students. Allocate approximately 100 students from the Licence stratum, **40** students from the Master stratum and **10** students from the Doctorat stratum.

This allocation is based on the proportion of each stratum in the overall population. Assuming the distribution is as follows: 1,000 students in **Licence**, 400 students in **Master** and 100 students in **Doctorat**.

The calculations for sampling would be:

$$\begin{aligned} \text{Licence:} \quad & \frac{1000}{1500} \times 150 = 100 \\ \text{Master:} \quad & \frac{400}{1500} \times 150 = 40 \\ \text{Doctorat:} \quad & \frac{100}{1500} \times 150 = 10 \end{aligned}$$

Random Sampling within Each Stratum

For each academic level, select the designated number of students randomly to ensure that the sample is representative across all strata. This helps avoid potential bias related to differences in study habits by academic level.

Summary Table

Stratum	**Total Students in Stratum**	**Sampled Students**
Licence	1,000	100
Master	400	40
Doctorat	100	10
Total	**1,500**	**150**

Table 1.2: Stratified Sampling Summary

This stratified sampling approach ensures balanced representation across all academic levels, allowing conclusions drawn about study habits to reflect each subgroup proportionally.

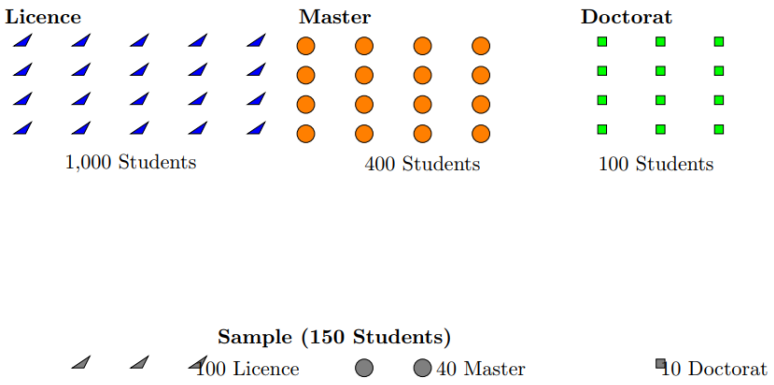


Figure 1.1: Probability Density Function (PDF) of a Normal Distribution

Advantages and Limitations:
Advantages: Stratified sampling increases precision and ensures that all relevant subgroups are represented, reducing sampling bias.
Limitations: This method can be more complex to implement than simple random sampling and requires detailed population information to define strata.

Conclusion: Stratified sampling is an effective sampling technique widely used in various fields, such as social sciences, market research, and health studies, to obtain reliable insights by ensuring all significant subgroups are included.

Systematic Sampling

Systematic Sampling is a statistical sampling technique used to select a sample from a larger population by choosing individuals at regular intervals. This method is straightforward and can be more efficient than random sampling methods, particularly when a complete list of the population is available.

Key Characteristics of Systematic Sampling

- **Fixed Interval Selection**: Individuals are selected at regular intervals from a randomly chosen starting point.
- **Simple to Implement**: The process is often easier to carry out than other sampling methods, especially for large populations.
- **Requires a Complete List**: A complete and ordered list of the population is necessary for systematic sampling to be effective.

Steps for Implementing Systematic Sampling

1. **Define the Population**: Clearly outline the group from which you will draw your sample.

2. **Determine Sample Size**: Decide how many individuals you need in the sample.

3. **Calculate the Sampling Interval**: Determine the sampling interval (k) by dividing the population size (N) by the desired sample size (n) using the formula $k = \frac{N}{n}$.

4. **Select a Random Starting Point**: Randomly select a starting point between 1 and k.

5. **Select Individuals at Regular Intervals**: Starting from the chosen point, select every k-th individual until the sample size is reached.

Example: Suppose you have a population of 1,000 employees in a company and you want a sample of 100 employees. If you calculate a sampling interval of $k = 10$ (i.e., $1000 \div 100 = 10$), you would randomly select a starting point between 1 and 10. If you choose 3, you would then select the 3rd, 13th, 23rd, and so on, until you reach the desired sample size.

Solution: In this scenario, systematic sampling is used to select a sample of **100 employees** from a company population of **1,000 employees**. The sampling interval is calculated as $k = 10$, meaning every 10th employee will be selected after a random starting point. Here's a step-by-step breakdown of the process:

1. Define the Population

The population consists of **1,000 employees** in a company.

Determine the Sampling Interval

To achieve the sample size, calculate the sampling interval as follows:

$$k = \frac{1000}{100} = 10$$

This means every 10th employee will be selected.

Select a Random Starting Point

Choose a random starting point between 1 and 10. For this example, assume we select **3** as the starting point.

Select Employees Using the Interval

Starting from employee **3**, select every 10th employee: **3, 13, 23, 33,** and so on, until 100 employees are chosen.

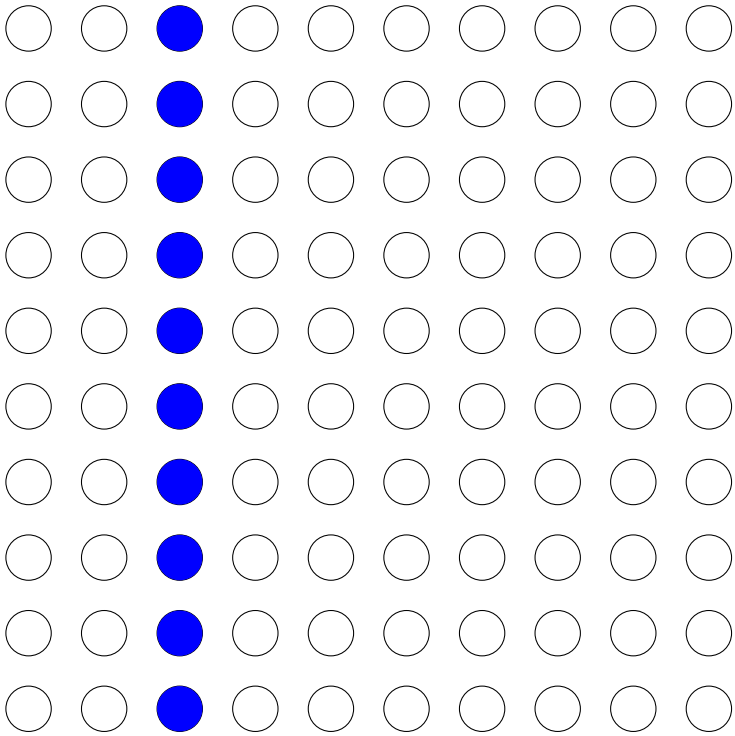
Sampling Interval	10
Starting Point	3
Employees Selected	3, 13, 23, 33, . . . , 993

Table 1.3: Systematic Sampling Summary

Summary of Systematic Sampling

This systematic sampling approach ensures that every part of the population has an equal chance of being included, based on the interval.

Population of 1,000 Employees



Sample of 100 Employees Selected (in Blue)

Advantages and Limitations:
Advantages: Systematic sampling is easy to implement and can be more efficient than random sampling. It can also be less time-consuming.
Limitations: It may introduce bias if there is a hidden pattern in the population that corresponds to the sampling interval, and it requires a complete list of the population.

Conclusion: Systematic sampling is a practical and efficient technique widely used in various fields such as quality control, survey research, and social sciences, where a straightforward method for sampling is required.

Cluster Sampling

Cluster Sampling is a statistical sampling technique used to select a sample from a larger population by dividing the population into groups or clusters and then randomly selecting entire clusters. This method is particularly useful when the population is large and widely dispersed, making it impractical to conduct a complete sampling.

Key Characteristics of Cluster Sampling

- **Division into Clusters**: The population is divided into non-overlapping groups or clusters based on specific characteristics, such as geographic location or organizational units.
- **Random Selection of Clusters**: Entire clusters are randomly selected, and all or a sample of individuals within those clusters are surveyed.
- **Cost-Effective**: This method can reduce costs and time when surveying widely dispersed populations.

Steps for Implementing Cluster Sampling

1. **Define the Population**: Clearly outline the entire population from which samples will be drawn.
2. **Divide into Clusters**: Identify the clusters within the population based on relevant characteristics.
3. **Randomly Select Clusters**: Randomly select a few clusters to be included in the sample.
4. **Collect Data from Selected Clusters**: Survey all individuals within the selected clusters or a random sample from those clusters.

Example: In a study of the dietary habits of school children in a large city, a researcher might divide the city into different school districts (clusters). If there are 9 districts, the researcher could randomly select 3 districts and then survey all the children in those districts.

Solution: In this scenario, cluster sampling is used to study dietary habits among school children in a large city. Here, **9 school districts** represent clusters, and the researcher will **randomly select 3 of these districts** to survey all the children within those clusters. Here's a step-by-step breakdown of the process:

1. Define the Population

The population consists of all school children across **9 school districts** in a large city.

Define Clusters

Each of the **9 districts** represents a cluster of school children.

Randomly Select Clusters

The researcher randomly selects **3 districts** from the 9 available districts. For this example, assume that Districts **2, 5, and 8** are selected.

Survey All Children in Selected Clusters

Once the districts are selected, the researcher surveys all school children within Districts **2, 5, and 8**.

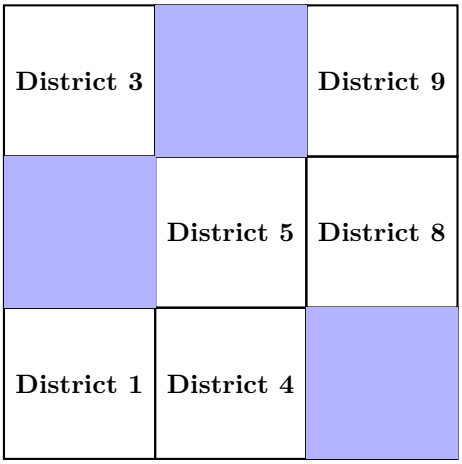
Summary of Cluster Sampling

Total Districts (Clusters)	9
Districts Selected	3 (Districts 2, 5, 8)
Sampled Children	All children in selected districts

Table 1.4: Cluster Sampling Summary

This cluster sampling approach allows the researcher to gather data from specific geographic areas within the city, making it more efficient than sampling children across all districts.

Population of 9 School Districts



Selected Districts (2, 5, and 8) are Highlighted in Blue

Advantages and Limitations:
Advantages: Cluster sampling is cost-effective and practical for large populations, especially when it is difficult to create a complete list of individuals.
Limitations: It can introduce higher sampling error compared to other methods, as individuals within clusters may be more similar to each other than to the broader population.

Conclusion: Cluster sampling is extensively used in fields like epidemiology, education, and social sciences for studying large and geographically dispersed populations.

1.1.2 Common Pitfalls to Avoid:

When conducting sampling, it's crucial to avoid common pitfalls, such as:

1. **Sampling Bias:** This occurs when the sampling method systematically excludes or over-represents certain groups in the population.
2. **Non-Response Bias:** If a significant portion of those selected for the sample does not participate, it can lead to a non-response bias.
3. **Sampling Error:** This is the natural variation that occurs when working with samples instead of the entire population. It can be minimized by increasing the sample size.
4. **Confounding Variables:** Failure to control for confounding variables (variables that are related to both the independent and dependent variables) can affect the results of the study.

In summary, sampling is a critical technique in statistics used to gather data efficiently and effectively from a subset of a larger population. The choice of sampling method depends on research goals and available resources. Proper sampling techniques are essential to ensure the validity and reliability of statistical analyses and research findings.

1.2 Statistics of samples: empirical mean, empirical variance

Data statistics often involve calculating various descriptive statistics to summarize and understand the characteristics of a dataset. Two essential statistics are the empirical mean (also known as the sample mean) and the empirical variance (sample variance).

1.2.1 Empirical Mean (Sample Mean):

The empirical mean, often denoted as $\bar{x}$, represents the average or central tendency of a dataset. It is calculated as the sum of all data points divided by the number of data points in the dataset.

The formula for the empirical mean is:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

where:

$\bar{x}$: Empirical mean
n : Number of data points in the dataset
x_i : Individual data points

Theorem 1.2.1. *(Distribution of $\bar{X}$): Let $(X_1, \ldots, X_n)$ be a simple random sample of size n of a random variable $N(\mu, \sigma^2)$. Then, the sample mean $\bar{X} := \frac{1}{n} \sum_{i=1}^{n} X_i$ satisfies*

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

.

Proof. To show that $\bar{X}$ has the stated distribution, we compute its expectation and variance, and then verify its distribution.

1. Expectation of $\bar{X}$:

Using the linearity of expectation, we have:

$$E[\bar{X}] = E\left[\frac{1}{n} \sum_{i=1}^{n} X_i\right] = \frac{1}{n} \sum_{i=1}^{n} E[X_i] = \frac{1}{n} \cdot n\mu = \mu.$$

2. Variance of $\bar{X}$:

Since the X_i are independent and identically distributed, we have:

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}.$$

3. Distribution of $\bar{X}$:

Since each X_i follows a normal distribution, and $\bar{X}$ is a linear combination of independent normal random variables, $\bar{X}$ itself follows a normal distribution.

Therefore, we conclude that:

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

□

1.2.2 Empirical Variance (Sample Variance):

The empirical variance, often denoted as S^2, measures the spread or variability of data points in the dataset. It quantifies how much individual data points

deviate from the mean. The larger the variance, the more dispersed the data points are. The sample variance is given by

$$S^2 := \frac{1}{n}\sum_{i=1}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1} X_i^2 - \bar{X}^2.$$

The sample quasi-variance will also play a relevant role in inference. It is defined by simply replacing n with $n-1$ in the factor of S^2:

$$S'^2 := \frac{1}{n-1}\sum_{i=1}(X_i - \bar{X})^2 = \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{i=1} X_i^2 - \frac{n}{n-1}\bar{X}^2.$$

Before establishing the sampling distributions of S^2 and S'^2, we obtain in the first place their expectations. For that aim, we start by decomposing the variability of the sample with respect to its expectation μ in the following way:

$$\sum_{i=1}(X_i - \mu)^2 = \sum_{i=1}(X_i - \bar{X})^2 + n(\bar{X} - \mu)^2.$$

Taking expectations, we have

$$n\sigma^2 = nE[S^2] + \frac{n\sigma^2}{n},$$

and then, solving for the expectation,

$$E[S^2] = (n-1)\sigma^2.$$

Therefore,

$$E[S'^2] = \frac{n}{n-1}E[S^2] = \sigma^2.$$

Recall that this computation does not employ the assumption of sample normality, hence it is a general fact for S^2 and S'^2 irrespective of the underlying distribution. It also shows that S^2 is not "pointing" towards σ^2 but to a slightly smaller quantity, whereas S'^2 is "pointing" directly to σ^2. This observation is related to the bias of an estimator and will be treated in detail in Section 3.1.

In order to compute the sampling distributions of S^2 and S'^2, it is required to obtain the sampling distribution of the statistic $\sum_{i=1} X_i^2$ when the sample is generated from a $N(0,1)$, which will follow a chi-square distribution.

Theorem 1.2.2. *2.2 (Fisher's Theorem) If $(X_1, \ldots, X_n)$ is a simple random sample of a $N(\mu, \sigma^2)$ random variable, then S^2 and $\bar{X}$ are independent, and*

$$\frac{nS^2}{\sigma^2} = (n-1)S'^2 \sim \chi^2_{n-1}.$$

Proof. To demonstrate that the ratio $\frac{nS^2}{\sigma^2}$ follows a chi-squared distribution with $n-1$ degrees of freedom, we can use the concept of chi-squared distribution and the properties of the sample variance.

Consider the random variable $\frac{nS^2}{\sigma^2}$:

$$\frac{nS^2}{\sigma^2} = \frac{n}{\sigma^2} \cdot \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

The expectation of $\frac{nS^2}{\sigma^2}$:

$$E\left(\frac{nS^2}{\sigma^2}\right) = \frac{n}{\sigma^2} \cdot E(S^2) = \frac{n}{\sigma^2} \cdot \sigma^2 = n$$

The expectation of $\frac{nS^2}{\sigma^2}$ is n, which is the degrees of freedom of a chi-squared distribution. Since the degrees of freedom of a chi-squared distribution can be expressed as $k-1$ (where k is the number of degrees of freedom), we have:

$$\frac{nS^2}{\sigma^2} \sim \chi^2(n-1)$$

So, the ratio $\frac{nS^2}{\sigma^2}$ follows a chi-squared distribution with $n-1$ degrees of freedom. □

1.3 The Density Function

The **probability density function (PDF)**, denoted as $f(x)$, describes the relative likelihood of a continuous random variable X taking on a specific value. Unlike discrete probability distributions, where probabilities can be assigned to specific outcomes, the PDF provides a way to describe the distribution of probabilities across a continuum of possible values.

Properties of the PDF

- **Non-negativity**: The value of the PDF is always non-negative, meaning $f(x) \geq 0$ for all x. This is because probabilities cannot be negative.
- **Normalization**: The total area under the curve of the PDF across all possible values of x is equal to 1:

  $$\int_{-\infty}^{\infty} f(x)\, dx = 1.$$

  This ensures that the total probability of the random variable falling within its possible range is 1.

Interpretation of the PDF

The height of the PDF at any point x, denoted $f(x)$, indicates the **relative likelihood** of X being close to x. Higher values of $f(x)$ suggest that values around x are more likely compared to those around points where $f(x)$ is lower.

While $f(x)$ provides a density, it does not represent the probability of X taking on a specific value. Instead, probabilities are computed over intervals. For example, the probability that X falls between two values a and b is given by the area under the curve between these two points:

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx.$$

Graphical Representation

The following graphical representation illustrates the concept of the PDF using a normal distribution.

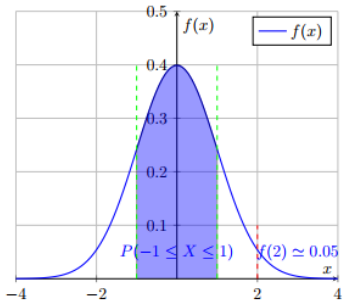


Figure 1.2: Probability Density Function (PDF) of a Normal Distribution

Explanation of the Graph

- **Blue Curve**: The blue line represents the PDF $f(x)$ of a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$.
- **Shaded Area**: The shaded area under the curve between the points $a = -1$ and $b = 1$ represents the probability $P(-1 \leq X \leq 1)$. This area indicates the likelihood of the random variable X falling within this interval.
- **Vertical Dashed Lines**: The dashed vertical lines at a and b help to visually identify the interval for which the probability is being calculated.

The probability density function (PDF) is a crucial concept in statistics and probability, allowing for the modeling and understanding of continuous random variables. By visualizing the PDF, we gain insights into how probabilities are distributed and how to compute probabilities over specific intervals.

1.4 Gaussian Samples

A Gaussian sample, also known as a normal sample, represents a set of data points that follow a Gaussian (normal) distribution. The Gaussian distribution is characterized by its mean (μ) and standard deviation (σ).

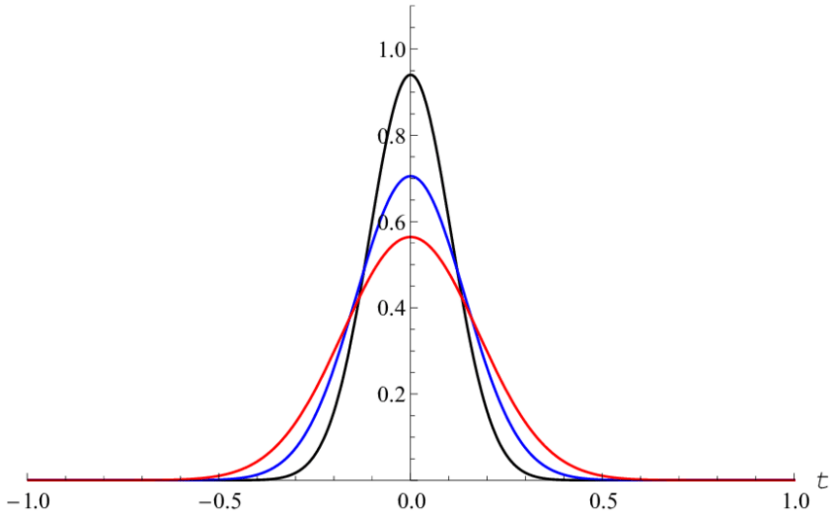


Figure 1.3: Gaussian sample

For example, let's consider a Gaussian sample with the following properties:

μ : Mean
σ : Standard Deviation

Suppose we have a Gaussian sample of 100 data points:

$$X = \{x_1, x_2, \ldots, x_{100}\}$$

where x_i represents an individual data point.
The Gaussian distribution for this sample can be expressed as:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

To provide concrete values, let's say:

$$\mu = 10$$
$$\sigma = 2$$

Then, a few data points from this Gaussian sample might be:

$$X = \{11.23, 9.75, 10.50, 9.98, \ldots\}$$

These data points are drawn from the Gaussian distribution with the specified mean and standard deviation.