Normal Distribution

Normal Distribution: Lecture Notes & Practice

Introduction

The **Normal Distribution** is one of the most fundamental concepts in statistics. It is a bell-shaped curve that describes the probability distribution of a continuous random variable.

This lecture covers: - **Properties of Normal Distribution - Empirical Rule (68-95-99.7) - Standard Normal Distribution & Z-scores - Computing Probabilities with R - Applications to real-world scenarios**

Properties of the Normal Distribution

1. **Symmetry:** The curve is symmetric around the mean (μ).
2. **Empirical Rule:** Approximately:
   - **68%** of data falls within $\pm\mathbf{1}$ (one standard deviation).
   - **95%** within $\pm\mathbf{2}$.
   - **99.7%** within $\pm\mathbf{3}$.
3. **Standard Normal Distribution:** Standardizing with

$$Z = \frac{X - \mu}{\sigma}$$

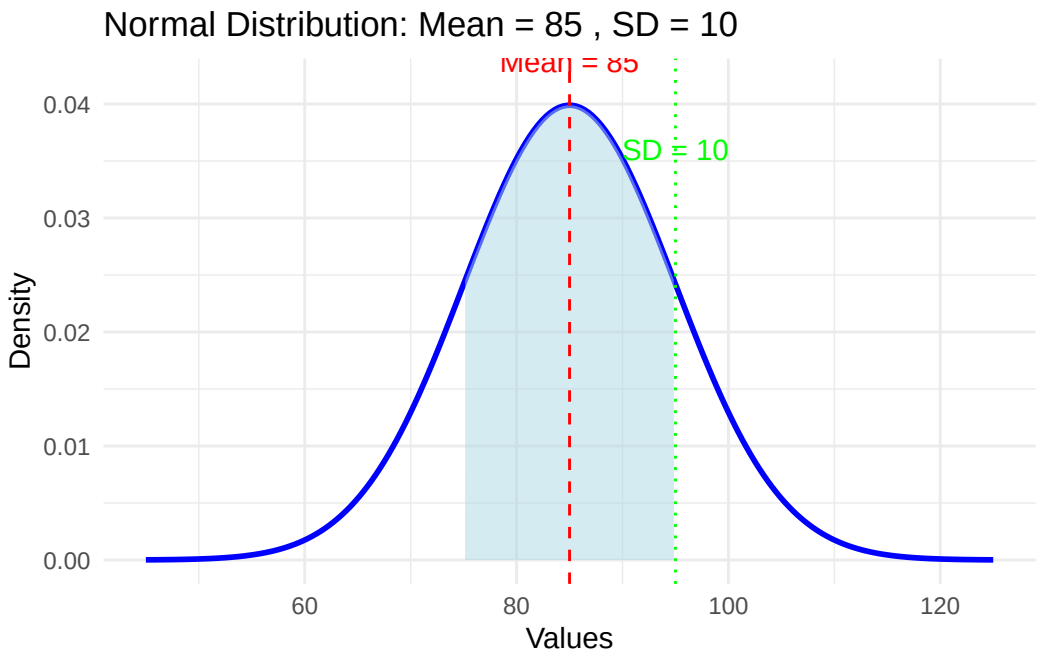
transforms the normal distribution into one with **mean = 0** and **standard deviation = 1**.

Tip

Why is this important?
The normal distribution is used in statistics to **approximate real-world data** such as test scores, IQ levels, and measurement errors.

Visualizing the Normal Curve

Let's **generate a normal distribution curve in R**.



Modify the mean and sd values to see how the curve changes! This helps visualize how different datasets may have different spreads.

The Empirical Rule

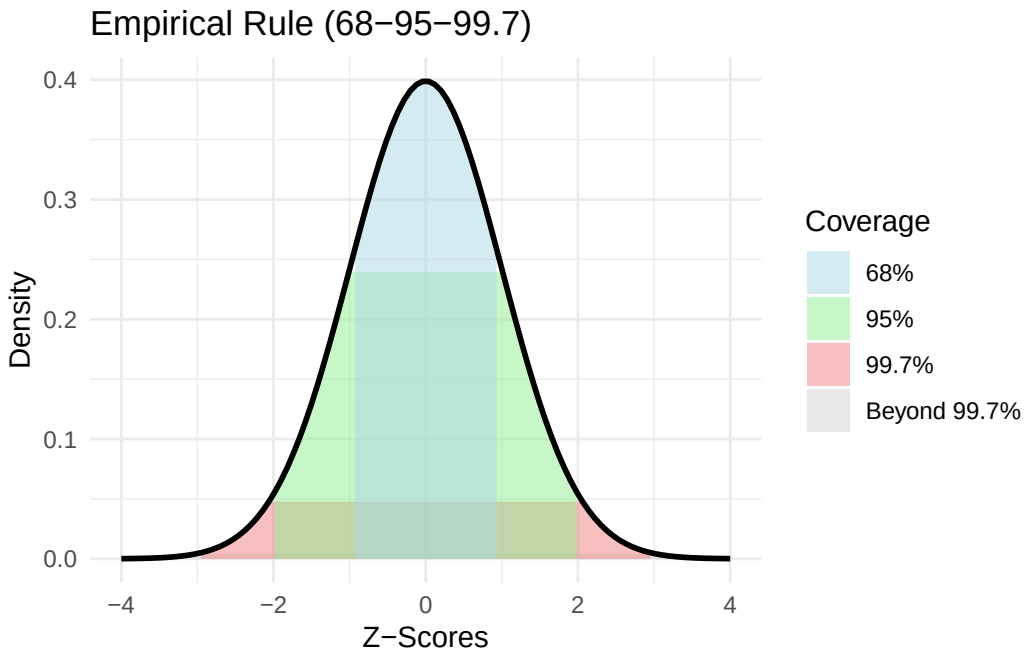
The **Empirical Rule** states that:

68% of the data is within **1 standard deviation** of the mean.

95% is within **2 standard deviations**.

99.7% is within **3 standard deviations**.

We can visualize this using R:



Computing Probabilities Using R

Instead of using Z-tables, we can use R commands to find probabilities.

Example 1: Exam Scores

A statistics exam has scores normally distributed with:

- Mean = 72
- Standard Deviation = 15.2.

What percentage of students scored below 84?

```
pnorm(84, mean=72, sd=15.2)
```

```
[1] 0.7850824
```

What percentage of students scored above 84?

```
pnorm(84, mean=72, sd=15.2, lower.tail = FALSE)
```

```
[1] 0.2149176
```

or we can use:

```
1-pnorm(84, mean=72, sd=15.2)
```

```
[1] 0.2149176
```

Example 2: Percentile Calculations

The SAT scores are normally distributed with:

- Mean = 1000
- Standard Deviation = 100.

Find the score corresponding to the 90th percentile.

```
qnorm(0.90, mean=1000, sd=100)
```

```
[1] 1128.155
```

Interpretation: This tells us the SAT score above which only 10% of students score.

Student Challenge!

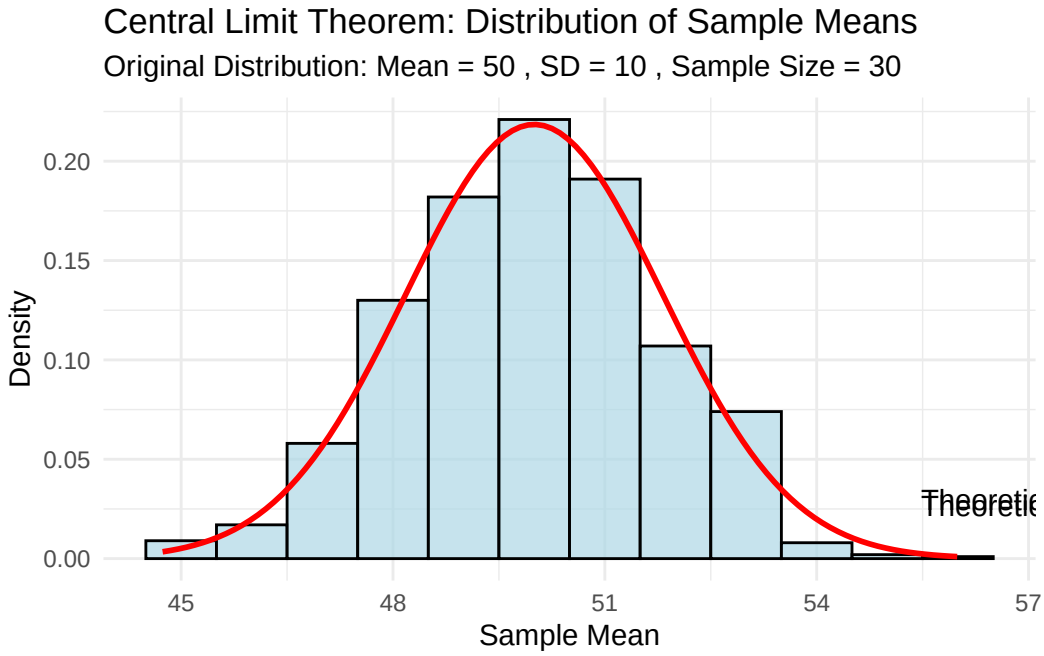
Try to answer these using pnorm() and qnorm():

1 What is the probability that a student scores below 65 on a test with • $\mu = 75$, $\sigma = 10$?

2 Find the score that corresponds to the 75th percentile for a test with • $\mu = 80$, $\sigma = 5$

Central Limit Theorem (CLT)

The Central Limit Theorem states that as sample size increases, the distribution of sample means becomes approximately normal, regardless of the population's shape.



This **simulates taking many samples** and computing the mean of each. Notice how the histogram forms a normal curve, even if the original data wasn't normally distributed.

Sampling Distribution of the Sample Mean

Sampling Error

- **Sampling error** is the error resulting from using a sample to estimate a population characteristic.

Sampling Distribution of the Sample Mean

- For a variable x and a given sample size, the distribution of the variable $\overline{x}$ is called the **sampling distribution of the sample mean**.
- The sampling distribution of the sample mean is a distribution of the sample means of all possible samples of the same size from a population.
- The sampling distribution of the sample mean is different from the population distribution. The population distribution is the distribution of the variable x, while the sampling distribution of the sample mean is the distribution of the sample mean $\overline{x}$.

- The sampling distribution of the sample mean is also different from the distribution of the sample. The distribution of the sample is the distribution of the data in a particular sample. The sampling distribution of the sample mean is the distribution of the sample means of all possible samples of the same size from a population.
- The sampling distribution of the sample mean is used to make inferences about the population mean. For example, we can use the sampling distribution of the sample mean to calculate the probability that the sample mean is within a certain distance of the population mean.

Sample Size and Sampling Error

- The larger the sample size, the smaller the sampling error tends to be in estimating a population mean, μ, by a sample mean, $\overline{x}$.
- The sampling error is the difference between the sample mean and the population mean.
- The sampling error is due to the fact that the sample is not a perfect representation of the population.
- The larger the sample size, the more closely the sample mean will approximate the population mean.
- As the sample size increases, the sampling distribution of the sample mean becomes narrower and taller. This means that the sample mean is more likely to be close to the population mean when the sample size is large.
- The sampling distribution of the sample mean is a normal distribution when the sample size is large enough. This is true even if the population distribution is not normal.
- The standard deviation of the sampling distribution of the sample mean is called the **standard error of the mean**.
- The standard error of the mean is a measure of the variability of the sample mean.
- The standard error of the mean decreases as the sample size increases. This means that the sample mean is more likely to be close to the population mean when the sample size is large.

Key Formulas

- **Sample Mean:**
  - $\overline{x} = \frac{\sum x_i}{n}$
  - where x_i are the individual observations and n is the sample size.
- **Standard Error of the Mean:**

  - $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$
  - where σ is the population standard deviation and n is the sample size.

- **If the population standard deviation (σ)** is unknown, we use the sample standard deviation (s) to estimate the standard error:
  - $s_{\overline{x}} = \frac{s}{\sqrt{n}}$

Central Limit Theorem

- For a relatively large sample size, the variable $\overline{x}$ is approximately normally distributed, regardless of the distribution of the variable under consideration. The approximation becomes increasingly better as the sample size increases.
- For samples of size n from a normally distributed population, the variable $\overline{x}$ is exactly normally distributed, no matter what the sample size is.
- For most populations, a sample size of 30 will ensure that the variable $\overline{x}$ is approximately normally distributed.

Visualizations

- **Figure 7.4:** Shows how the sampling distribution of the sample mean becomes narrower and taller as the sample size increases.
- (a) Normal distribution for IQs; (b) sampling distribution of the sample mean for n = 4; (c) sampling distribution of the sample mean for n = 16
- **Figure 7.6:** Shows the sampling distributions of the sample mean for normal, reverse-J-shaped, and uniform variables.
- Sampling distributions of the sample mean for (a) normal, (b) reverse-J-shaped, and (c) uniform variables
- **Figure 7.5:** Relative-frequency histogram for household size

Normal Distribution Practice Problems

1. Most graduate schools of business require applicants for admission to take the Graduate Management Admission Council's GMAT examination. Scores on the GMAT are roughly normally distributed with a mean of 527 and a standard deviation of 112. What is the probability of an individual scoring above 500 on the GMAT?
2. How high must an individual score on the GMAT in order to score in the highest 5%?

3. The length of human pregnancies from conception to birth approximates a normal distribution with a mean of 266 days and a standard deviation of 16 days. What proportion of all pregnancies will last between 240 and 270 days (roughly between 8 and 9 months)?
4. What length of time marks the shortest 70% of all pregnancies?
5. The average number of acres burned by forest and range fires in a large New Mexico county is 4,300 acres per year, with a standard deviation of 750 acres. The distribution of the number of acres burned is normal. What is the probability that between 2,500 and 4,200 acres will be burned in any given year?
6. What number of burnt acres corresponds to the 38th percentile?
7. The Edwards's Theater chain has studied its movie customers to determine how much money they spend on concessions. The study revealed that the spending distribution is approximately normally distributed with a mean of \$4.11 and a standard deviation of \$1.37. What percentage of customers will spend less than \$3.00 on concessions?
8. What spending amount corresponds to the 87th percentile?

Summary & Key Takeaways

The Normal Distribution follows a bell-shaped curve.

The Empirical Rule (68-95-99.7) helps approximate probabilities.

Z-scores standardize normal distributions for easier comparison.

R functions like pnorm() and qnorm() help compute probabilities.

Using R for homework & tests: Students are encouraged to compile all formulas & R code in RStudio and use them for tests.