STAT 350 Worksheet #9

We have previously discussed the general form of probability density functions and cumulative distribution functions, as well as two fundamental examples of named continuous distributions the **Uniform** and **Exponential** distributions. Many real-world measurements, such as heights and measurement errors, naturally cluster around an average in a roughly symmetrical, bell-shaped pattern. The **Gaussian distribution,** also called the **Normal distribution,** is the most widely used model for describing this shape. Sums or averages of many independent non-Normal random variables often converge to a **Normal distribution**, which makes it a cornerstone of statistical inference and practical applications. In this worksheet, we explore the key properties of the **Normal distribution**, show how the standard Normal serves as a reference distribution, and practice essential probability calculations for finding areas and percentiles under the bell-shaped curve.

A continuous Normal random variable X has support the entire real line $\mathbb{R} = (-\infty, +\infty)$, and is defined by the following probability density function:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

The distribution has the following properties: it is **symmetrical**, **unimodal**, and **bell shaped** (concave down at the center and then concave up starting about one standard deviation on each side). It is completely determined by two parameters, the mean μ and **standard deviation σ.** However, computing probabilities from this distribution directly is **non-trivial** since the cumulative distribution function has no closed-form solution. Instead, probabilities must be obtained through numerical methods, such as integration algorithms, statistical tables, or software.

We will begin with the traditional approach of using statistical tables, and in the next section when we discuss the Central Limit Theorem, we will incorporate R for these calculations. However, since there are infinitely many Normal distributions, one for each choice of mean μ and **standard deviation σ**, we first need a reference distribution, the standard Normal distribution, which allows us to standardize any Normal variable and use a common probability table for calculations.

A continuous Normal random variable Z with mean $\mu = 0$ and standard deviation $\sigma = 1$ is known as the **standard Normal distribution** and is defined by the following density function:

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}.$$

The standard Normal distribution will serve as a reference distribution for our probability calculations for Normal random variables.

However, let's start by using the properties of the Normal distribution curve to answer some basic questions. One of the most useful properties of the Normal distribution is the **Empirical Rule**, also known as the **68-95-99.7 rule**. This rule provides a quick way to estimate probabilities based on how data is distributed around the mean. Specifically, for any Normal distribution:

- **Approximately 68%** of values fall within **one standard deviation** of the mean $(\mu \pm \sigma)$:
$$P(\mu - \sigma < X < \mu + \sigma) = 0.68.$$
- **Approximately 95%** of values fall within **two standard deviations** of the mean $(\mu \pm 2\sigma)$:
$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.95.$$
- **Approximately 99.7%** of values fall within **three standard deviations** of the mean $(\mu \pm 3\sigma)$:
$$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.997.$$
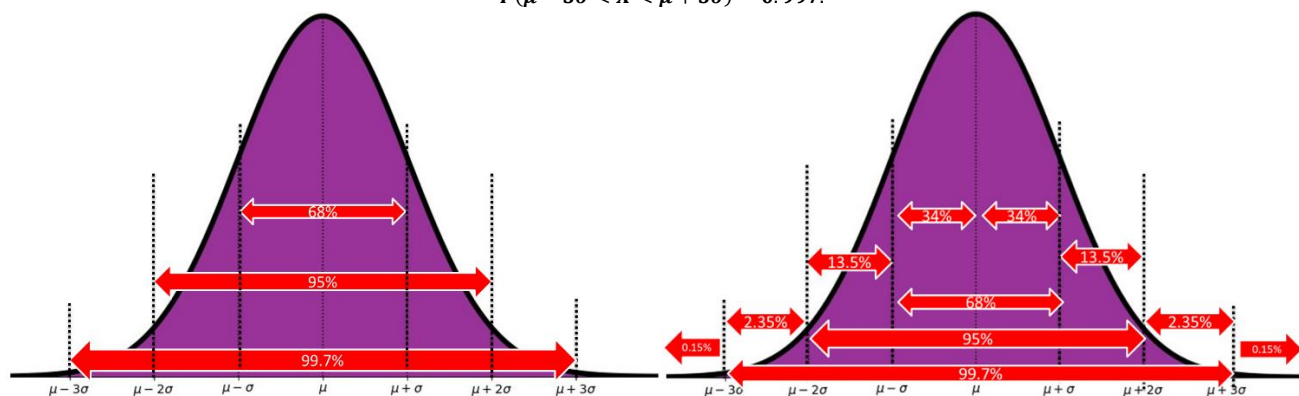
As shown in the above Figure, the empirical rue allows us to quickly assess the likelihood of an observation occurring within a certain range without the need for complex calculations. Before using statistical tables for more precise probability computations, we will first apply this rule to answer some fundamental probability questions about the Normal distribution.

1. A manufacturing company produces metal rods that follow a **Normal distribution** with a mean length of **50 cm** and a standard deviation of **2 cm**.

   a) Using the **Empirical Rule**, determine the probability that a randomly selected rod has a length within the following ranges. Clearly write out the probability statement before computing each value.

      i. Greater than 52 cm.

      ii. Less than 48 cm.

      iii. Between 44 cm and 54 cm.

      iv. Greater than 54 cm or less than 44 cm.

   b) Suppose you randomly select a rod from a pile labeled "greater than 52 cm". Using the **Empirical Rule**, determine the probability that the rod is also **greater than 54 cm**.

   c) Using the **Empirical Rule determine** the **84th percentile** of rod lengths.

   d) A quality control engineer decides that rods must be **between 46 cm and 54 cm** to be considered within acceptable tolerance limits. A batch contains **10,000 rods**, and rods are considered **defective** if their lengths fall **outside the range 46 cm to 54 cm**. How many rods would be expected to fail the quality check?

While the **Empirical Rule** provides a useful approximation, most real probability calculations do not fall neatly into the **68-95-99.7%** intervals, requiring a more precise approach. The **standard Normal table** provides cumulative probabilities for the **standard Normal distribution ($\mu = 0$, $\sigma = 1$)**, serving as a universal reference for probability calculations. Since the cumulative distribution function has no closed-form solution, the values in the table were obtained using numerical approximation methods. To use this table, we transform any Normal variable into standard Normal form through a process called the **z-score transformation**, which is analogous to u-**substitution** in integration, shifting and scaling the variable so that probabilities can be determined from a single distribution.

Even though we cannot obtain a closed-form solution for the cumulative distribution function (CDF), the **standard Normal distribution** is often referenced using the notation $\Phi(z)$ to represent its CDF. That is,

$$\Phi(z) = P(Z \leq z)$$

where Z follows a standard Normal distribution, denoted as:

$$Z \sim N(\mu = 0, \sigma = 1).$$

The function $\Phi(z)$ gives the probability that a standard Normal variable takes on a value less than or equal to z and is computed using numerical approximation methods. The standard Normal table provides precomputed values of $\Phi(z)$, allowing us to determine probabilities without directly integrating the probability density function.

$$P(Z \leq z) = \Phi(z)$$
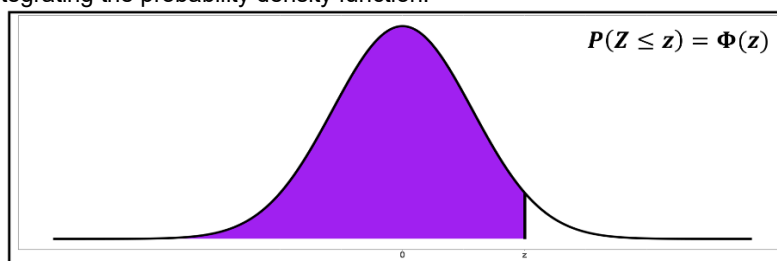
2. Using the standard normal table found on Brightspace under "Extra Documents" answer the following questions. If you get stuck try drawing the normal curve and shading in the region you need to find before using the table.

a) Compute the following probabilities:

    i. $\Phi(2.34) =$

    ii. $P(Z > -0.12) =$

    iii. $P(-1.64 < Z < 1.64) =$

b) Determine the following percentiles.

    i. $P(Z < z_p) = 0.9192$

    ii. $P(Z < z_p) = 0.95$

c) Determine the follow upper percentiles.

    i. $P(Z > z_p) = 0.017$

    ii. $P(Z > z_p) = 0.9990$

d) Determine two values **a** and **b** symmetric about zero such that $P(a \leq Z \leq b) = 0.95$.

Not all Normal distributions have a mean of 0 and a standard deviation of 1. In many cases, we deal with **non-standard Normal distributions**, which have different means (μ) and standard deviations (σ). Since the **standard Normal table** only provides probabilities for the **standard Normal distribution**, we must first transform any Normal variable into standard form. This is done using the **z-score transformation**, given by

$$z = \frac{x - \mu}{\sigma}.$$

This transformation expresses how many standard deviations a given value x from any **Normal distribution** $N(\mu, \sigma)$ is from its mean. By converting to the **standard Normal scale**, we can use a single probability table for all Normal distributions.

There are two main types of problems involving the Normal distribution: **forward problems** and **backward problems**.

- **Forward problems** involve computing probabilities for a given Normal value by first converting it to a **z-score** and then using the standard Normal table to determine the probability. Recommended steps to follow:
    1. **Draw** the probability region of interest and identify what you need to calculate.
    2. **Standardize:** Convert to **z-score**.
    3. **If necessary, change the probability statement** so you can use the table.
    4. **Round z-score** to 2 decimal places and look it up on the table.
    5. **Write** your conclusion in the context of the problem.

- **Backward problems** involve finding a specific value x when given a probability or percentile. In these cases, we look up the probability in the standard Normal table to find the corresponding **z-score** and then use the **z-score** transformation in reverse to solve for x. Recommended steps to follow:
    1. **Draw** the probability region of interest and determine the location you need to find.
    2. **Set up and modify your probability statements as needed**
    3. **Find the z-score** by looking up the probability in the main body of the normal table.
    4. **Convert** the **z-score** to x using

      $z = \dfrac{x - \mu}{\sigma}$

      solve for x

      $x = \mu + \sigma z$

    5. **Write** your conclusion in the context of the problem.

3. Let X be a **Normal random variable** with mean $\mu = 10$ and variance $\sigma^2 = 25$. Answer the following questions.

   a) Using the forward process determine the following probabilities.

       i. $P(X < 19.8) =$

       ii. $P(X > -10) \approx$

       iii. $P(0.2 < X < 19.8) =$

   b) Using the backward process determine the following percentiles in terms of the distribution of X, in other words find x_p.

       i. $P(X < x_p) = 0.7486$

       ii. $P(X < x_p) = 0.8$

   c) Using the backward process determine the following upper percentile $P(X > x_p) = 0.99$.

   d) Determine inter quartile range of the distribution. In other words find two values x_1 and x_2 centered around the mean such that $P(x_1 \leq X \leq x_2) = 0.5$.

4. An insurance company is analyzing policyholders who have all survived to **age 60**. Historical data suggests that their eventual **ages at death**, can be **approximated** by a **Normal distribution** with **mean** of **85 years** and a **standard deviation** of **4 years**.
   a) A newly turned 60-year-old policyholder asks, "What is the probability I will live past **90**?" Compute the probability.

   b) The insurance company wants to define an **age cutoff** above which only **5%** of these policyholders will live. Find the value $x_{0.05}$ the age threshold above which only **top 5%** of the 60+ insured population are expected to live.