# Probability and Statistics

## Topic 15 - The Normal Probability Distribution

Aamir Alaud Din, PhD

November 15, 2024

# TABLE OF CONTENTS

# TABLE OF CONTENTS

# RECAP

- Area under a normal curve is equal to 1.
- Area under a normal curve represents the probability of the continuous random variable $x$.
- If we have multiple (millions and billions) data sets, it is kind of impossible to develope the table of probabilities for all the data sets with their mean and standard deviation.
- Therefore, we convert the value of the continuous random variable $x$ into corresponding $z$-score and consult the table of area under normal curve with mean 0 and standard deviation 1 (the table constructed with mean 0 and standard deviation 1 *i.e.*, $z$-score data.
- The areas within first, second, and third standard deviations of the mean for any data set are 68%, 95%, and 99.7%, respectively.
- As areas under normal curve don't change for any data set, we make use of a single table of areas for $z$-score data to compute probabilities and also the value of the continuous random variable $x$ if probability is the input.

# TABLE OF CONTENTS

# OBJECTIVES

After learning this topic and studying, you should be able to:

1. Find and interpret the area under a normal curve
2. Find the value of a normal random variable

# TABLE OF CONTENTS

# THE WHY SECTION

- On the basis of our understanding of data, we know that the data is normally distributed.
- For a normal distribution, we can determine the area under the normal curve given the value of the continuous random variable.
- For 100 different data sets with 100 different (totally different) means and standard deviations, it is not feasible to draw 100 normal curves each for a data set and then find area for any data value in the data set.
- We make use of z-scores as the z-score obtained from 100 different data will have a 0 mean and 1 standard deviation.
- Using this normal curve we can find area under the normal curve for the data point in any data set using the corresponding z-score and using the curve of z-scores.
- How can we do it, is the main purpose of this topic.

# TABLE OF CONTENTS

# FINDING AND INTERPRETING AREA UNDER NORMAL CURVE

- If $X$ is a normally distributed random variable, the area under the normal curve represents the proportion of a population with a certain characteristic, or the probability that a randomly selected individual from the population has the characteristic.
- The question then is, "How do I find the area under the normal curve?"
- We have two options which are 1) by-hand calculations with the aid of a table or 2) technology.
- We use z-scores to help find the area under a normal curve by hand.
- Recall, the z-score allows us to transform a random variable $X$ with mean m and standard deviation s into a random variable $Z$ with mean 0 and standard deviation 1.

### Standardizing a Normal Random Variable

Suppose that the random variable $X$ is normally distributed with mean $\mu$ and standard deviation $\sigma$.Then the random variable

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

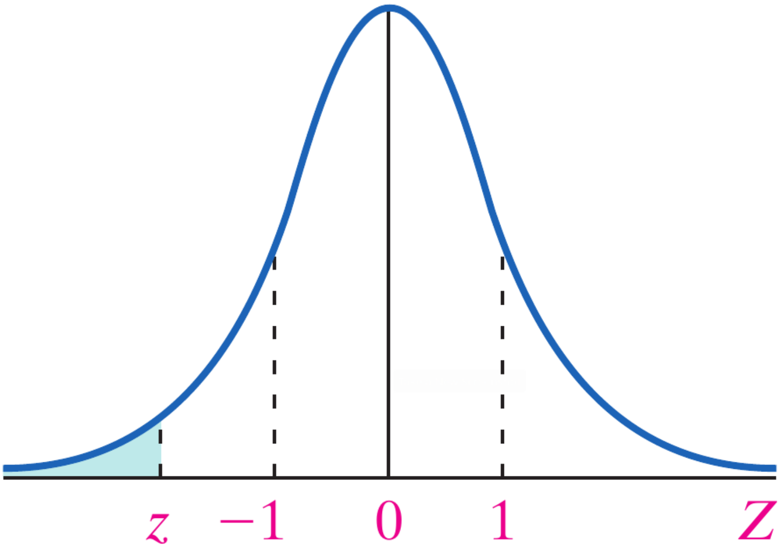
is normally distributed with mean $\mu = 0$ and standard deviation $\sigma = 1$. The random variable $Z$ is said to have the standard normal distribution.

- This result is powerful!
- If a normal random variable X has mean different from 0 or a standard deviation different from 1, we can transform X into a standard normal random variable Z whose mean is 0 and standard deviation is 1.

# FINDING AND INTERPRETING AREA UNDER NORMAL CURVE

- Then we can use Table V (found on the inside back cover of the text and in Appendix A) to find the area to the left of a specified z-score, z, as shown in figure below, which is also the area to the left of the value of $x$ in the distribution of $X$.
- The graph in figure below is called the standard normal curve.
- For example, IQ scores can be modeled by a normal distribution with $\mu = 100$ and $\sigma = 15$.
- An individual whose IQ is 120, is $z = \frac{x-\mu}{\sigma} = \frac{120-100}{15} = 1.33$ standard deviations above the mean (recall, we round z-scores to two decimal places).
- We look in Table V and find the area under the standard normal curve to the left of $z = 1.33$ is 0.9082.
- See figure to see how to consult table with shaded area of 0.9082.

# FINDING AND INTERPRETING AREA UNDER NORMAL CURVE



# FINDING AND INTERPRETING AREA UNDER NORMAL CURVE

**Standard Normal Distribution**

| $z$ | .00 | .01 | .02 | .03 | .04 | .05 | .06 |
|---|---|---|---|---|---|---|---|
| **0.0** | 0.5000 | 0.5040 | 0.5080 | 0.5120 | 0.5160 | 0.5199 | 0.5239 |
| **0.1** | 0.5398 | 0.5438 | 0.5478 | 0.5517 | 0.5557 | 0.5596 | 0.5636 |
| **0.2** | 0.5793 | 0.5832 | 0.5871 | 0.5910 | 0.5948 | 0.5987 | 0.6026 |
| **0.3** | 0.6179 | 0.6217 | 0.6255 | 0.6293 | 0.6331 | 0.6368 | 0.6406 |
| **0.4** | 0.6554 | 0.6591 | 0.6628 | 0.6664 | 0.6700 | 0.6736 | 0.6772 |
| **0.5** | 0.6915 | 0.6950 | 0.6985 | 0.7019 | 0.7054 | 0.7088 | 0.7123 |
| **0.6** | 0.7257 | 0.7291 | 0.7324 | 0.7357 | 0.7389 | 0.7422 | 0.7454 |
| **0.7** | 0.7580 | 0.7611 | 0.7642 | 0.7673 | 0.7704 | 0.7734 | 0.7764 |
| **0.8** | 0.7881 | 0.7910 | 0.7939 | 0.7967 | 0.7995 | 0.8023 | 0.8051 |
| **0.9** | 0.8159 | 0.8186 | 0.8212 | 0.8238 | 0.8264 | 0.8289 | 0.8315 |
| **1.0** | 0.8413 | 0.8438 | 0.8461 | 0.8485 | 0.8508 | 0.8531 | 0.8554 |
| **1.1** | 0.8643 | 0.8665 | 0.8686 | 0.8708 | 0.8729 | 0.8749 | 0.8770 |
| **1.2** | 0.8849 | 0.8869 | 0.8888 | 0.8907 | 0.8925 | 0.8944 | 0.8962 |
| **1.3** | 0.9032 | 0.9049 | 0.9066 | 0.9082 | 0.9099 | 0.9115 | 0.9131 |
| **1.4** | 0.9192 | 0.9207 | 0.9222 | 0.9236 | 0.9251 | 0.9265 | [illegible] |
| **1.5** | [illegible] | [illegible] | [illegible] | [illegible] | 0.9382 | 0.9394 | [illegible] |

# FINDING AND INTERPRETING AREA UNDER NORMAL CURVE

- Therefore, the area under the normal curve to the left of $x = 120$ is 0.9082 as shown in figure below showing normal curve with the shaded area of 0.9082.

- To find the area to the right of the value of a random variable, use the Complement Rule and determine one minus the area to the left.

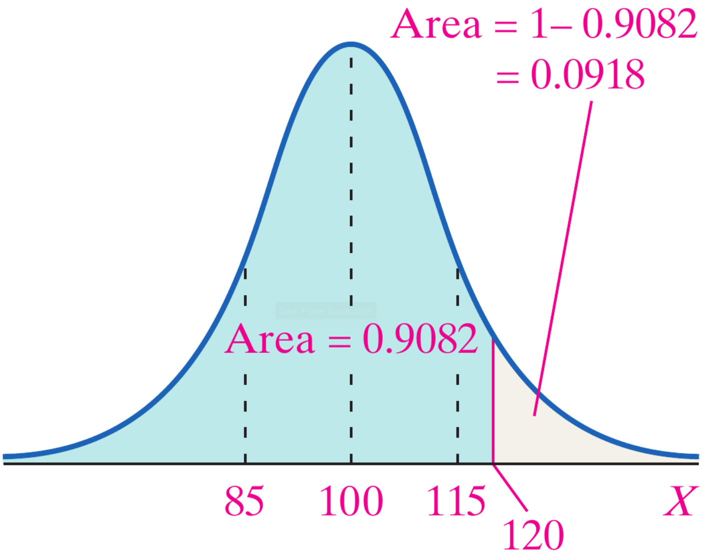
- For example, to find the area under the normal curve with mean $\mu = 100$ and standard deviation $\sigma = 15$ to the right of $x = 120$, compute

$$\text{Area} = 1 - 0.9082 = 0.0918$$

  as shown in figure below.

- Now, it's time to use the table in Example 1 below.

# FINDING AND INTERPRETING AREA UNDER NORMAL CURVE



# FINDING AND INTERPRETING AREA UNDER NORMAL CURVE

## EXAMPLE 1

A pediatrician obtains the heights of her three-year-old female patients. The heights are approximately normally distributed, with mean 38.72 inches and standard deviation 3.17 inches. Use the normal model to determine the proportion of the three-year-old females that have a height less than 35 inches.

- According to the results of Example 1, the proportion of three-year-old females who are shorter than 35 inches is approximately 0.12.
- If the normal curve is a good model for determining proportions (or probabilities), then about 12% of the three-year-olds in Table 1 (from Topic 14) should be shorter than 35 inches.

- For convenience, part of Table 1 is repeated in Table 2.
- The relative frequency distribution in Table 2 shows that $0.005 + 0.005 + 0.005 + 0.025 + 0.02 + 0.055 = 0.115 = 11.5\%$ of the three-year-old females are less than 35 inches tall.
- The results based on the normal curve are close to the actual results.
- The normal curve accurately models the heights.
- If we wanted to know the proportion of three-year-old females whose height is greater than 35 inches, use the Complement Rule and find the proportion is 1 - $0.1210 = 0.879$ (using the "by-hand" computation).
- Because the area under the normal curve represents a proportion, we can also use the area to find percentile ranks of scores.

- Recall that the kth percentile divides the lower $k\%$ of a data set from the upper $100-k\%$.
- In Example 1, 12% of the females have a height less than 35 inches, and 88% of the females have a height greather than 35 inches, so a child whose height is 35 inches is at the 12$^{\text{th}}$ percentile.

### EXAMPLE 2

For the pediatrician presented in Example 1, find the probability that a randomly selected three-year-old girl is between 35 and 40 inches tall, inclusive. That is, find $P(35 \leq X \leq 40)$.

- According to the relative frequency distribution in Table 2, the proportion of three-year-old females with heights between 35 and 40 inches is $0.075+0.09+0.115+0.15+0.12=0.55$.

# FINDING AND INTERPRETING AREA UNDER NORMAL CURVE

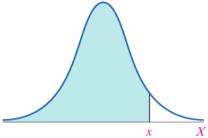
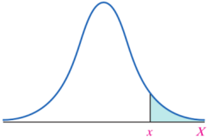
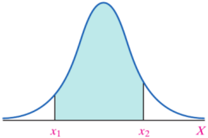
- This is very close to the probability found in Example 2.
- We summarize the methods for obtaining the area under a normal curve in Table 3.

## HOMEWORK **(EXTENDIBLE)**

1. Enter the data shown on page A-11 and A-12 of the book in either (1) a spreadsheet *e.g.*, Microsoft Excel or (2) a text file.
2. Write a python function which takes (1) mean, (2) standard deviation, and (3) the value of the continuous random variable $x$ as input arguments.
3. The program must calculate the probability of the continuous random variable using the data table developed in step 1.
4. The program must produce the graph showing the shaded area.

# FINDING AND INTERPRETING AREA UNDER NORMAL CURVE

**Table 3**

| Problem | Approach | Solution |
|---|---|---|
| Find the area to the left of $x$. | Shade the area to the left of $x$.   | • Convert the value of $x$ to a $z$-score. Use Table V to find the row and column that correspond to $z$. The area to the left of $x$ is the value where the row and column intersect.<br>or<br>• Use technology to find the area. |
| Find the area to the right of $x$. | Shade the area to the right of $x$.   | • Convert the value of $x$ to a $z$-score. Use Table V to find the area to the left of $z$ (which is also the area to the left of $x$). The area to the right of $z$ (also $x$) is 1 minus the area to the left of $z$.<br>or<br>• Use technology to find the area. |
| Find the area between $x_1$ and $x_2$. | Shade the area between $x_1$ and $x_2$.   | • Convert the values of $x$ to $z$-scores. Use Table V to find the area to the left of $z_1$ and the area to the left of $z_2$. The area between $z_1$ and $z_2$ is (area to the left of $z_2$) − (area to the left of $z_1$).<br>or<br>• Use technology to find the area. |

# TABLE OF CONTENTS

# VALUE OF NORMAL RANDOM VARIABLE

- Often, we do not want to find the proportion, probability, or percentile given a value of a normal random variable.
- Rather, we want to find the value of a normal random variable that corresponds to a certain proportion, probability, or percentile.
- For example, we might want to know the height of a three-year-old girl who is at the 20th percentile.
- Or we might want to know the scores on a standardized exam that separate the middle 90% of scores from the bottom and top 5%.
- Time to solve a problem to see how to apply things and getting into a coming hell.

# VALUE OF NORMAL RANDOM VARIABLE

- Often, we do not want to find the proportion, probability, or percentile given a value of a normal random variable.
- Rather, we want to find the value of a normal random variable that corresponds to a certain proportion, probability, or percentile.
- For example, we might want to know the height of a three-year-old girl who is at the 20th percentile.
- Or we might want to know the scores on a standardized exam that separate the middle 90% of scores from the bottom and top 5%.
- Time to solve a problem to see how to apply things and getting into a coming hell.
- BOOOO!!! The hell is coming.

# VALUE OF NORMAL RANDOM VARIABLE

## EXAMPLE 3

The heights of a pediatrician's three-year-old females are approximately normally distributed, with mean 38.72 inches and standard deviation 3.17 inches. Find the height of a three-year-old female at the $20^{\text{th}}$ percentile.

## EXAMPLE 4

The scores earned on the mathematics portion of the SAT, a college entrance exam, are approximately normally distributed with mean 516 and standard deviation 116. What scores separate the middle 90% of test takers from the bottom and top 5%? In other words, find the $5^{\text{th}}$ and $95^{\text{th}}$ percentiles. Source: The College Board

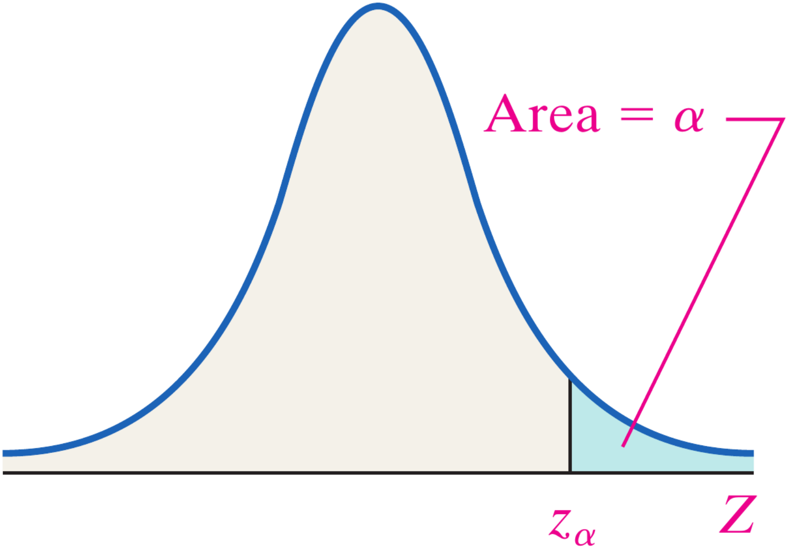
# VALUE OF NORMAL RANDOM VARIABLE

- We could also obtain the by-hand solution to Example 4 using symmetry.
- Because the normal curve is symmetric about its mean, the z-score that corresponds to an area of 0.05 to the left will be the additive inverse (i.e., the opposite) of the z-score that corresponds to an area of 0.05 to the right.
- Since the area to the left of $z = -1.645$ is 0.05, the area to the right of $z = 1.645$ is 0.05.

**Important Notation for the Future**

In upcoming chapters, we will need to find the z-score that has a specified area to the right. We have special notation to represent this situation. The notation $z_\alpha$ (pronounced "z sub alpha") is the z-score such that the area under the standard normal curve to the right of $z_\alpha$ is $\alpha$. Figure below illustrates the notation.

# VALUE OF NORMAL RANDOM VARIABLE



# VALUE OF NORMAL RANDOM VARIABLE

## HOMEWORK (FURTHER EXTENDIBLE)

1. Extend the python program developed to incorporate probabilities as input argument and return the corresponding values of the continuous random variable $x$.
2. If a value doesn't exist in the table developed, use interpolation to get the value stated in the problem.
3. Solve example 5 using the program developed.
4. Solve all examples of topic 15 by using the python program developed and validate your program.
5. In case of errors, debug the program and complete the point 4 stated above.

# TABLE OF CONTENTS

# SUMMARY

- Area under the normal curve represents the probability of occurrence of a continuous random variable $x$ given the mean and standard deviation.
- In order to consult Table V, we must convert the value of the continuous random variable $x$ into the corresponding $z$-score.
- Given the probability $P(x)$of the continuous random variable $x$, the value of the $x$ can be obtained given the mean and standard deviation using formula

$$x = \mu + z\sigma$$

- Different software packages compute the probability (area under the curve) using different methods/models/simulations *e.g.*, Monte Carlo Simulation.

# SUMMARY

- Area under the normal curve represents the probability of occurrence of a continuous random variable $x$ given the mean and standard deviation.
- In order to consult Table V, we must convert the value of the continuous random variable $x$ into the corresponding $z$-score.
- Given the probability $P(x)$of the continuous random variable $x$, the value of the $x$ can be obtained given the mean and standard deviation using formula

$$x = \mu + z\sigma$$

- Different software packages compute the probability (area under the curve) using different methods/models/simulations *e.g.*, Monte Carlo Simulation.
- BOOOO!!!???

# SUMMARY

- Area under the normal curve represents the probability of occurrence of a continuous random variable $x$ given the mean and standard deviation.
- In order to consult Table V, we must convert the value of the continuous random variable $x$ into the corresponding $z$-score.
- Given the probability $P(x)$of the continuous random variable $x$, the value of the $x$ can be obtained given the mean and standard deviation using formula

$$x = \mu + z\sigma$$

- Different software packages compute the probability (area under the curve) using different methods/models/simulations *e.g.*, Monte Carlo Simulation.
- BOOOO!!!??? OK!! ENJOY!!


Thank You!