# APS606 - Lecture Note (Week 2)

Norman Lo, M.S. in Quantitative Economics

9/30/2020

## Introduction to Probability Distribution

I. Probability of Normal Distribution
II. Standardized Score (z-score)
III. Confidence Interval
IV. Probability Questions with Normal Distribution

### I. Probability and Normal Distribution

In the previous section, we demonstrate how to calculate the probability for a given event, a sequence of events, or events by conditions. In this section, we would like to discover how we can apply the probability concepts to different types of data and visualize it on a distribution.

A probability distribution is a function that describes the likelihood of obtaining the possible values that a random variable can assume. In other words, the values of the variable vary based on the underlying probability distribution.

Suppose you draw a random sample and measure the heights of the subjects. As you measure heights, you can create a distribution of heights. This type of distribution is useful when you need to know which outcomes are most likely, the spread of potential values, and the likelihood of different results.

#### Continuous Probability Distribution

A continuous random variable can assume any value in an interval on the real line or in a collection of intervals. It is not possible to talk about the probability of the random variable assuming a particular value. Instead, we talk about the probability of the random variable assuming a value within a given interval.

The probability of the random variable assuming a value within some given interval from $x_1$ to $x_2$ is defined to be the area under the graph of the **probability density function** between $x_1$ and $x_2$.

#### Normal Probability Distribution

The **normal probability distribution** is the most important distribution for describing a continuous random variable. It is widely used in statistical inference. It has been used in a wide variety of applications including: height of peole, test scores, amounts of rainfall, scientific measurements,etc. Abraham de Moivre, a French mathematician, published The Doctrine of Chances in 1733, derived the normal distribution.

**Normal Probability Density Function:**

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

where:

$\mu$ = mean

$\sigma$ = standard deviation

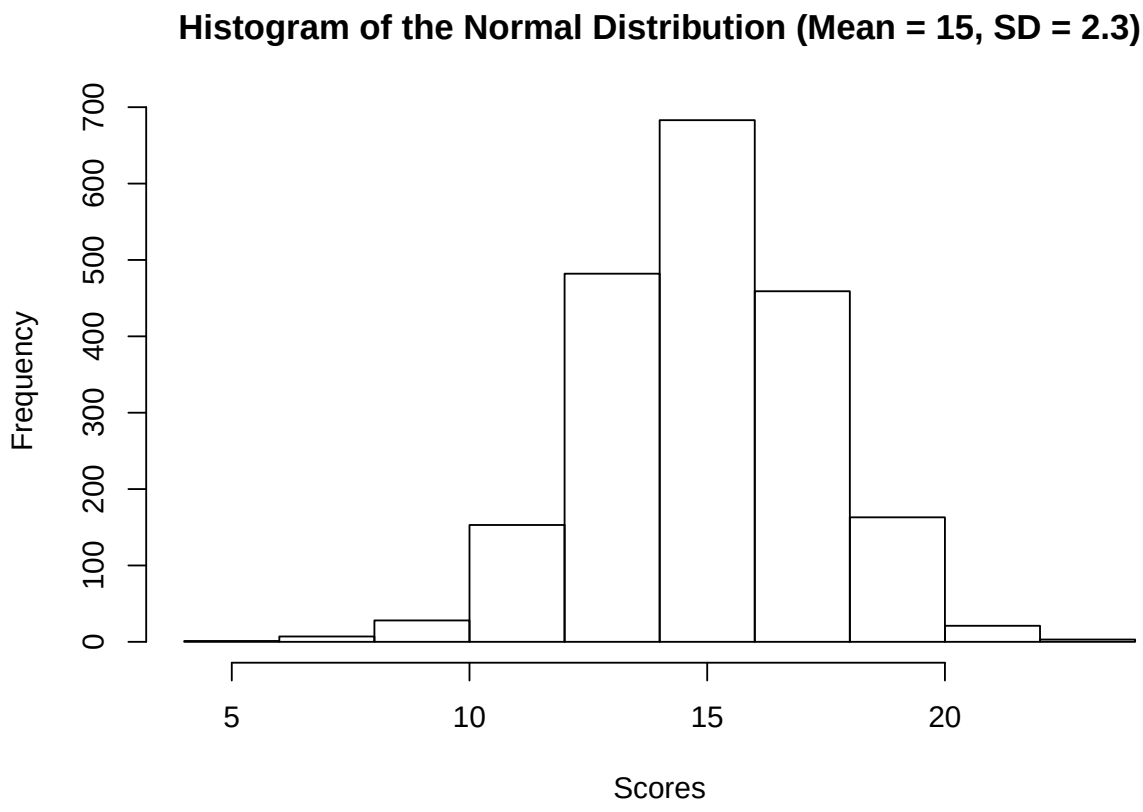
$\pi = 3.14159\ldots$

e = 2.71828...

**Characteristics of Normal Probability Distribution**

1. The distribution is **symmetric**; its skewness measure is zero.
2. The entire family of normal probability distributions is defined by its mean $\mu$ and its standard deviation $\sigma$.
3. The highest point on the normal curve is at the mean, which is also the median and mode.
4. The mean can be any numerical value: negative, zero, or positive.
5. The standard deviation determines the width of the curve: larger values result in wider, flatter curves.
6. Probabilities for the normal random variable are given by areas under the curve. The total area under the curve is 1 (0.5 to the left of the mean and 0.5 to the right).
7. A random variable having a normal distribution with a mean of 0 and a standard deviation of 1 is said to have a **standard normal probability distribution**.

Here is a simulated dataset that demonstrate a normal distribution with mean = 15 and standard deviation = 2.3.



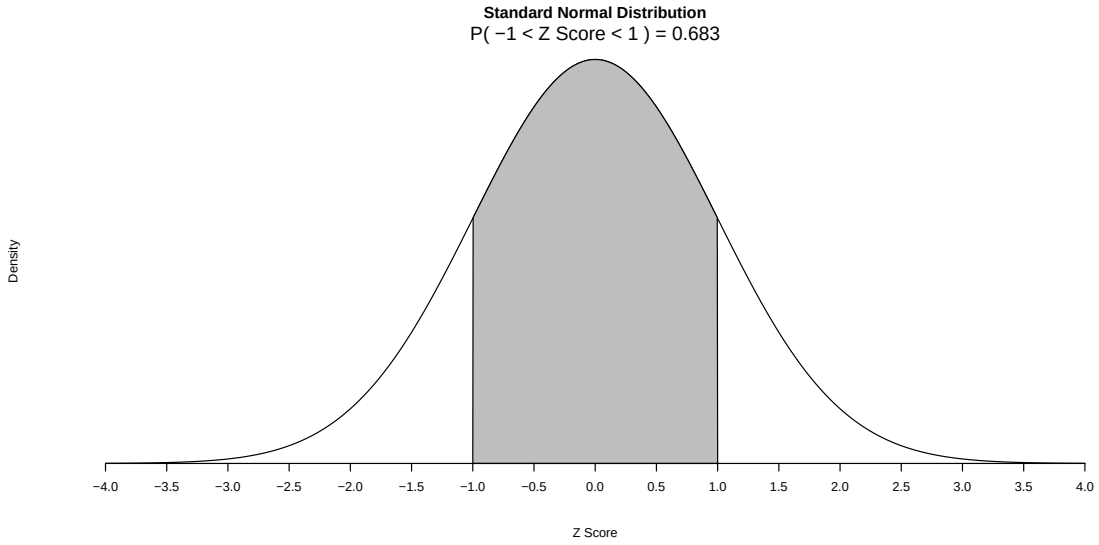
## II. Standardized Score (z score)

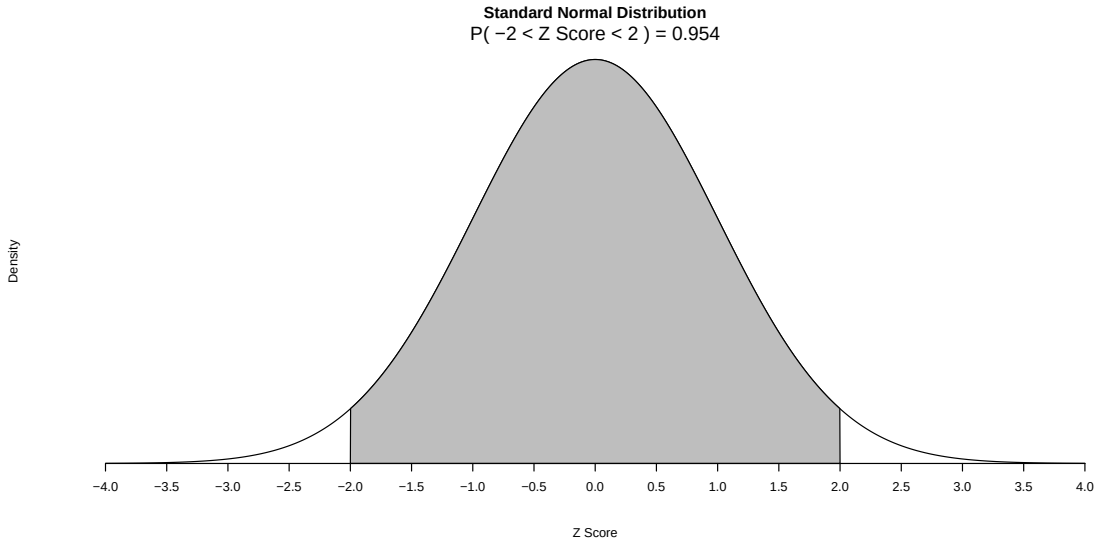
A **z-score** is a numerical measurement that describes a value's relationship to the mean of a group of values. Z-score is measured in terms of standard deviations from the mean. If a Z-score is 0, it indicates that the data point's score is identical to the mean score. A Z-score of 1.0 would indicate a value that is one standard deviation from the mean. Z-scores may be positive or negative, with a positive value indicating the score is above the mean and a negative score indicating it is below the mean.

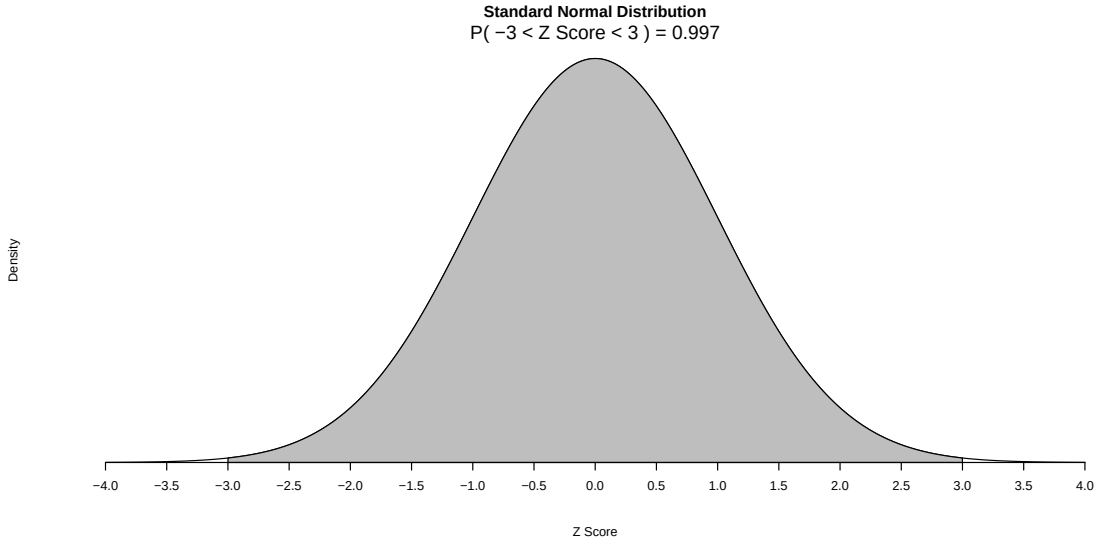
We can think of z as a measure of the number of standard deviations x is from $\mu$. Using the following formula, we can convert a numerical measurement to the standardized z score:

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

**Empirical Rule of Normal Distribution:**


Standard Normal Distribution
P( −1 < Z Score < 1 ) = 0.683
Density
−4.0 −3.5 −3.0 −2.5 −2.0 −1.5 −1.0 −0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0
Z Score


Standard Normal Distribution
P( −2 < Z Score < 2 ) = 0.954
Density
−4.0 −3.5 −3.0 −2.5 −2.0 −1.5 −1.0 −0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0
Z Score


Standard Normal Distribution
P( −3 < Z Score < 3 ) = 0.997
Density
−4.0 −3.5 −3.0 −2.5 −2.0 −1.5 −1.0 −0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0
Z Score

## III. Confidence Interval

Confidence intervals are used to indicate how accurate a calculated statistic is likely to be. Confidence intervals can be calculated for a variety of statistics, such as the mean, median, or slope of a linear regression. In this section, we focus on confidences intervals for means.

Most of the statistics we use assume we are analyzing a sample which we are using to represent a larger population. If extension educators want to know about the caloric intake of $7^{th}$ graders, they would be hard-pressed to get the resources to have every $7^{th}$ grader in the U.S. keep a food diary. Instead they might collect data from one or two classrooms, and then treat the data sample as if it represents a larger population of students.

The mean caloric intake could be calculated for this sample, but this mean will not be exactly the same as the mean for the larger population. If we collect a large sample and the values aren't too variable, then the sample mean should be close to the population mean. But if we have few observations, or the values are highly variable, we are less confident our sample mean is close to the population mean.

We will use confidence intervals to give a sense of this confidence.

Our sample mean is a point estimate for the population parameter. A point estimate is a useful approximation for the parameter, but considering the confidence interval for the estimate gives us more information.

As a definition of confidence intervals, if we were to sample the same population many times and calculated a sample mean and a 95% confidence interval each time, then 95% of those intervals would contain the actual population mean.

**Interval Estimate of a Population Mean:** ($\sigma$ is known)

$$\bar{x} \pm z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

where: $\bar{x}$ is the sample mean

1 - $\alpha$ is the confidence coefficient

$z_{\alpha/2}$ is the z value providing an area of $\alpha/2$ in the upper tail of the standard normal probability distribution

$\sigma$ is the population standard deviation

$n$ is the sample size

**Example:**

National Discount, Inc. has general merchandise retail outlets in 260 locations throughout the U.S. In considering possible new retail outlet locations, National evaluates each potential new location on several factors, one of which is the mean annual-income of the individuals in the marketing area serviced by the new outlet.

To estimate the population mean annual income m, National decided to use a simple random sample of size n = 64. In this sample, $\bar{x}$ = \$21,000 and s = \$5,600.

Based on similar annual income surveys, the standard deviation of annual incomes in the entire population is considered known with $\sigma$ = \$5,000.

Develop a 95% confidence interval estimate.

Solution:

Equation: $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

```r
z <- qnorm(0.025, lower.tail = FALSE)
margin_of_error <- z * 5000 / sqrt(64)
Upper <- 21000 + margin_of_error
Lower <- 21000 - margin_of_error
print(paste("The 95% confidence interval of the average income is $", round(Lower, digits = 0),"and $",
```

```
## [1] "The 95% confidence interval of the average income is $ 19775 and $ 22225 ."
```

**Most Commonly Used Confidence Levels:**

| Confidence Level | $\alpha$ | $\alpha$ | Table Look-up Area | $z_{\alpha/2}$ |
|---|---|---|---|---|
| 90% | 0.10 | 0.05 | 0.9500 | 1.645 |
| 95% | 0.05 | 0.025 | 0.9750 | 1.960 |
| 99% | 0.01 | 0.005 | 0.9950 | 2.576 |

Because 90% of all the intervals constructed using $\bar{x} \pm 1.645_{\sigma_{\bar{x}}}$ will contain the population mean, we say we are 90% confident that the interval $\bar{x} \pm 1.645_{\sigma_{\bar{x}}}$ includes the population mean $\mu$.

We say that this interval has been established at the 90% confidence level. The value .90 is referred to as the confidence coefficient.

---

## IV. Probability Questions with Normal Distribution

A social scientist collected a sample data set from several Asia cities about personal weekly spending on food and beverage. The sample mean spending is \$15 and a standard deviation of \$6. The scientist is wondering what is the probability selecting a random person from these cities would spend more than \$20 or more per week on food and beverage.

$$P(x > 20) = ?$$

**Step 1:** Convert x to the standard normal distribution

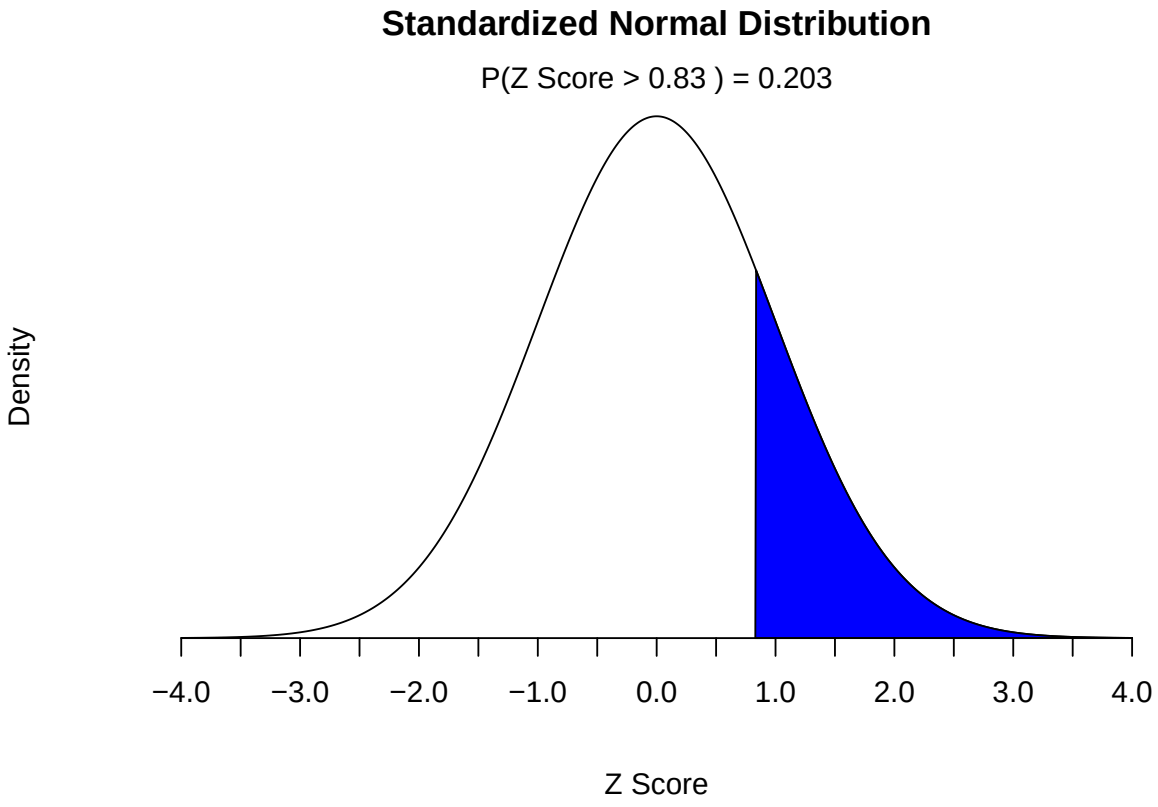
$$z = \frac{20 - 15}{6} = 0.83$$

**Step 2:** Find the area under the standard normal curve to the right of z = 0.83

Use the Probability Table for the Standard Normal Distribution

$$P(z > 0.83) = 0.2033 \; or \; 20.33\%$$

**Graphical Solution:**




Suppose the scientist is interested to find the 95% confidence interval of the personal weekly spending on food and beverages in the observed Asian cities? (Hint: Given a probability, we can use the standard normal table in an inverse fashion to find the corresponding z value.)

$$x = \mu \pm z_{0.025}\sigma$$
$$x = 15 \pm 1.645 * 6$$
$$(5.13 \leq x \leq 24.87)$$

The 95% confidence interval for the personal weekly spending in the observed Asian cities are between \$5.13 and \$24.87.